Can Biotechnology Abolish Suffering?

by David Pearce

Edited by Magnus Vinding

Published by The Neuroethics Foundation, North Carolina

Table of Contents

Part IV: Consciousness

Part V: The Sentience Explosion

Preface

By Magnus Vinding

The first time I heard of David Pearce was in late 2013. It was through the Youtube video "PostHuman: An Introduction to Transhumanism", which features some of Pearce's main ideas. My immediate reaction was skeptical. "That stuff sounds simplistic!… that's not wholly accurate… who does this guy think he is?"

Nonetheless, the video made me curious and compelled me to seek out more information about the ideas of David Pearce. The deeper I probed into these ideas, the more intrigued I was, and the more important I considered them to be. Let me try to explain why.

First of all, David Pearce focuses on what matters. Particularly, like so many thinkers before him, from Buddha and Mahavira to Schopenhauer and Popper, he focuses on the alleviation of suffering, and views this as humanity's main imperative. As he writes in "The Abolitionist Project" (echoing Karl Popper): "There isn't a moral urgency to maximizing superhappiness in the same way as there is to abolishing suffering." In more technical terms, Pearce is a self-identified negative utilitarian.

But more than just focusing on what is important, Pearce is also uniquely ambitious about it. For Pearce is not content with a mere reduction of suffering. He wants to abolish it completely throughout the living world – what he refers to as **The Abolitionist Project**. This is Pearce's *raison d'être*.

Yet Pearce's high level of ambition does not end here. Beyond the abolition of suffering, he argues that we should make life even better still, by making sentient beings animated by gradients of (ever greater) bliss. However, the abolition of suffering remains the overriding goal. Everything beyond that is frosting on the cake.

In addition to focusing on the alleviation of suffering, David Pearce is also a unique and important thinker due to his well-considered view of the nature of consciousness, including suffering. In a nutshell: our states of consciousness are, according to Pearce, concrete physical states in our heads. This is hardly an original view. What *is* original to Pearce, in my opinion, is his level of appreciation of this insight. Sure, most well-informed people would call themselves physicalists and say that they believe in some kind of identity theory as the solution to the mind-body problem. Yet the ability to express such a thin string of words is many light years away from truly appreciating that the entire conscious experience one is having, including all that is "out there", indeed *is* a concrete physical state residing in one's head. A world-simulation, as Pearce calls it.

I have come to view this world-simulation model of perception as the master key to understanding the philosophy of David Pearce (who himself, it seems to me, considers it so obvious that he often misses its uniqueness and significance). It is the key to understanding his ontology, epistemology, and view of consciousness. The best explanation Pearce has provided of this world-simulation model is arguably found in the section of *The Hedonistic Imperative* titled "Alone Amongst the Zombies" and the essay "Terminological Note for Philosophers" found in this volume (yet perhaps also see his review of David Chalmers' *The Conscious Mind*).

This view of the world is also what makes Pearce believe in the possibility of abolishing suffering in all sentient beings. The root of suffering is not something "out there" in the mind-independent world, as our direct realist intuitions might have us believe. Rather, all

forms of suffering are mediated by concrete structures in our heads. Hence, Pearce contends, abolishing suffering is ultimately an engineering problem all about changing certain physical structures. And in the case of existing sentient life forms, these structures have a genetic basis. Thus, with the right genetic tweaking, Pearce argues, this class of physical structures, i.e. suffering, can be abolished. (This is not to say that Pearce focuses *only* on the biological causes of suffering as opposed to cultural ones, such as discrimination; rather, he stresses the importance of a twin-track approach that addresses both.)

These are some of the main reasons I consider the ideas of David Pearce to be important, although this brief exposition obviously does no real justice to the importance and uniqueness of these ideas. To get a better sense of that, the reader will have to consult the essays in this volume. These essays are arranged in five parts.

Part I outlines and defends Pearce's Abolitionist Project, arguing both for its technological feasibility and moral importance on a wide variety of value systems, as well as outlining how it might happen (see "The Reproductive Revolution") and responding to many of the objections against it.

Part II, Bioethics, mostly addresses fundamental issues in moral philosophy, particularly differences between classical and negative utilitarianism.

Part III is about our obligations toward the vast majority of sentient beings on the planet: non-human animals. Those obligations being, in short, that we should stop harming them and start helping them. Beyond outlining a technology-catalyzed road to global veganism (or "its ethical invitrotarian equivalent"), Pearce also describes how

technology might enable us to help free-living non-human animals, from insects (see "Compassionate Biology") to elephants ("A Welfare State for Elephants").

In Part IV, the focus is on consciousness. Pearce presents and defends an original – and in his own words "bizarre" – conjecture about the nature of consciousness. A surprising implication of this conjecture is that (non-trivial) digital sentience is impossible. As Pearce readily admits, this is highly speculative stuff.

Finally, in Part V, we find two essays on "the technological singularity" and the future of intelligence. Here, Pearce provides a critique of the notion of an intelligence explosion in I.J Good, Eliezer Yudkowsky, and Nick Bostrom's sense, along with a criticism of what Pearce characterizes as "our narrow conception of intelligence". In contrast, Pearce presents the idea of "full-spectrum superintelligence", and predicts a future bio-intelligence explosion in which humans recursively improve themselves by editing their biology beyond Darwinian recognition.

The essays in this volume by no means comprise the complete works of David Pearce, but merely what I consider his most important essays to date (and the fault for bad choices in terms of the content of this volume is therefore entirely mine). All of these essays have been published elsewhere, in the HedWeb ecosystem, yet they have not been published in book form. Not until now. It is my hope that with this publication, the work of David Pearce will reach more readers and thereby, most importantly, help create a better future. A future with less suffering.

Deep thanks go to Cynthia Stewart for proofing this volume with impressive speed and enthusiasm, to James Evans for the cover design, and to Tom Richards and Katie Willis

for their feedback that greatly helped improve the content of this volume. Lastly, I wish to thank David Pearce for writing these essays, and for his dedication to reducing the suffering of all sentient beings.

Magnus Vinding

Copenhagen

August 2017

Introduction

Think of your best "peak experience". Everyday future life could be better. In 1995, I wrote an online manifesto, *The Hedonistic Imperative*, which advocates using biotechnology to abolish pain and suffering throughout the living world in favour of gradients of intelligent bliss. The proposal to reprogram the biosphere sounds like science fiction. But in 1998, Swedish-born philosopher Nick Bostrom and I helped set up the World Transhumanist Association (H+) to promote a transhuman agenda. Transhumanists urge the use of technology to overcome our biological limitations. Post-Darwinian life can be founded on the "three supers" of superintelligence, superlongevity and superhappiness. Transhumanists "advocate the well-being of all sentience, including humans, non-human animals, and any future artificial intellects, modified life forms, or other intelligences to which technological and scientific advance may give rise." (The Transhumanist Declaration, 1998, 2009)

Over the past two decades, I've run websites aimed at raising awareness of potential biological-genetic solutions to the problem of suffering. Suffering and malaise have been a defining feature of Darwinian life over the past 540 million years. Natural selection didn't design biological organisms to be happy. CRISPR genome-editing can repair the deficit. Imminent mastery of our reward circuitry promises a future of genetically preprogrammed hedonic enrichment, together with a *recalibration* of the hedonic

treadmill. The hedonic treadmill is the vicious but adaptive set of negative feedback mechanisms that prevents lifelong well-being for all but a few genetic outliers today.

On a personal note, I am what is known, uninvitingly, as a negative utilitarian. Advocates of suffering-focused ethics believe that intelligent moral agents have an overriding obligation to prevent suffering in human and nonhuman animals alike. Preventing pain should always trump creating pleasure. Critically, however, you don't need to be a utilitarian of *any* kind to support phasing out the biology of suffering and replacing it with gradients of superhuman bliss. Most people still find the prospect of a living world completely without pain and misery hard to imagine. My imagination often fails too. Darwinian life can be grim. Yet the accelerating biotech revolution means we are living in the final century when experience below "hedonic zero" need be anything other than optional. Perhaps a few centuries from now, the world's last experience below "hedonic zero" will mark a major evolutionary transition: the dawn of the first civilisation worthy of the name. Transhuman life will be wonderful, and perhaps sublime.

Many thanks to editor Magnus Vinding for putting together this selection of essays.

Part I: The Abolitionist Project

THE ABOLITIONIST PROJECT

INTRODUCTION

This essay is about suffering and how to get rid of it.

The abolitionist project outlines how biotechnology will abolish suffering throughout the living world.

Our descendants will be animated by gradients of genetically preprogrammed well-being that are orders of magnitude richer than today's peak experiences.

First, I'm going to outline why it's *technically* feasible to abolish the biological substrates of any kind of unpleasant experience - psychological pain as well as physical pain.

Second, I'm going to argue for the overriding *moral* urgency of the abolitionist project, whether or not one is any kind of ethical utilitarian.

Third, I'm going to argue why a revolution in biotechnology means it's going to happen, albeit not nearly as fast as it should.

1: WHY IT IS TECHNICALLY FEASIBLE

Sadly, what *won't* abolish suffering, or at least not on its own, is socio-economic reform, or exponential economic growth, or technological progress in the usual sense, or any of the traditional panaceas for solving the world's ills. Improving the external environment

is admirable and important; but such improvement can't recalibrate our hedonic treadmill above a genetically constrained ceiling. Twin studies confirm there is a [partially] heritable set-point of well-being - or ill-being - around which we all tend to fluctuate over the course of a lifetime. This set-point varies between individuals. [It's possible to *lower* an individual's hedonic set-point by inflicting prolonged uncontrolled stress, but even this re-set is not as easy as it sounds: suicide rates typically go down in wartime; and six months after a quadriplegia-inducing accident, studies[1] suggest that we are typically neither more nor less unhappy than we were before the catastrophic event.] Unfortunately, attempts to build an ideal society can't overcome this biological ceiling, whether utopias of the left or right, free-market or socialist, religious or secular, futuristic high-tech or simply cultivating one's garden. Even if *everything* that traditional futurists have asked for is delivered - eternal youth, unlimited material wealth, morphological freedom, superintelligence, immersive VR, nanotechnology, etc - there is no evidence that our subjective quality of life would, on average, significantly surpass the quality of life of our hunter-gatherer ancestors - or a New Guinea tribesman today - in the absence of reward pathway enrichment. This claim is difficult to prove in the absence of sophisticated neuroscanning; but objective indices of psychological distress, e.g. suicide rates, bear it out. *Un*enhanced humans will still be prey to the spectrum of Darwinian emotions, ranging from terrible suffering to petty disappointments and frustrations - sadness, anxiety, jealousy, existential angst. Their biology is part of "what it means to be human". Subjectively unpleasant states of consciousness exist because they were genetically adaptive. Each of our core emotions had a distinct signalling role in our evolutionary past: they tended to promote behaviours that enhanced the inclusive fitness of our genes in the ancestral environment.

So if manipulating our external environment alone can never abolish suffering and malaise, what *does* technically work?

Here are three scenarios in ascending order of sociological plausibility:

1) wireheading

2) utopian designer drugs

3) genetic engineering and - what I want to focus on - the impending **reproductive revolution** of designer babies

1) Recall **wireheading** is direct stimulation of the pleasure centres[2] of the brain via implanted electrodes. Intracranial self-stimulation shows no physiological or subjective tolerance, i.e. it's just as rewarding after two days as it is after two minutes. Wireheading doesn't harm others; it has a small ecological footprint; it banishes psychological and physical pain; and arguably it's a lot less offensive to human dignity than having sex. Admittedly, lifelong wireheading sounds an appealing prospect only to a handful of severe depressives. But what are the *technical* arguments against its adoption?

Well, wireheading is not an evolutionarily stable solution: there would be selection pressure against its widespread adoption. Wireheading doesn't promote nurturing behaviour: wireheads, whether human or non-human, don't want to raise baby wireheads. *Uniform*, indiscriminate bliss in the guise of wireheading or its equivalents would effectively bring the human experiment to an end, at least if it were adopted globally. Direct neurostimulation of the reward centres destroys informational sensitivity to environmental stimuli. So assuming we want to be smart - and become smarter - we have a choice. Intelligent agents can have a motivational structure based on gradients of ill-being, characteristic of some lifelong depressives today. Or intelligent agents can have

our current typical mixture of pleasures and pains. Or alternatively, we could have an informational economy of mind based entirely on [adaptive] gradients of cerebral bliss - which I'm going to argue for.

Actually, this dismissal of wireheading may be too quick. In the far future, one can't rule out offloading *everything* unpleasant or mundane onto inorganic supercomputers, prostheses and robots while we enjoy uniform orgasmic bliss. Or maybe not orgasmic bliss, possibly some other family of ideal states that simply couldn't be improved upon. But that's speculative. Whatever our ultimate destination, it would be more prudent, I think, to aim for both superhappiness and superintelligence - at least until we understand the full implications of what we are doing. There isn't a moral urgency to maximizing superhappiness in the same way as there is to abolishing suffering.

[It's worth noting that the offloading option assumes that inorganic computers, prostheses and robots don't - or at least needn't - experience subjective phenomenal pain even if their functional architecture allows them to avoid and respond to noxious stimuli. This absence of inorganic suffering is relatively uncontroversial with existing computers - switching off one's PC doesn't have ethical implications, and a silicon robot can be programmed to avoid corrosive acids without experiencing agony if it's damaged. It's debatable whether any computational system with a classical von Neumann architecture will ever be interestingly conscious. I'm sceptical; but either way, it doesn't affect the offloading option, unless one argues that the subjective texture of suffering is functionally essential to *any* system capable of avoiding harmful stimuli.]

2) The second technical option for eradicating suffering is futuristic **designer drugs**. In an era of mature post-genomic medicine, will it be possible to rationally design truly ideal pleasure-drugs that deliver lifelong, high-functioning well-being without unacceptable

side-effects? "Ideal pleasure drugs" here is just a piece of shorthand. Such drugs can in principle embrace cerebral, empathetic, aesthetic and perhaps spiritual well-being - and not just hedonistic pleasure in the usual one-dimensional and amoral sense.

We're *not* talking here about recreational euphoriants, which simply activate the negative feedback mechanisms of the brain; nor the shallow, opiated contentment of a Brave New World; nor drugs that induce euphoric mania, with its uncontrolled excitement, loss of critical insight, grandiosity and flight of ideas. Can we develop true wonderdrugs that deliver sublime well-being on a sustainable basis, recalibrating the hedonic treadmill to ensure a high quality of life for everyone?

A lot of people recoil from the word "drugs" - which is understandable given today's noxious street drugs and their uninspiring medical counterparts. Yet even academics and intellectuals in our society typically take the prototypical dumb drug, ethyl alcohol. If it's socially acceptable to take a drug that makes you temporarily happy and stupid, then why not rationally design drugs to make people perpetually happier and smarter? Presumably, in order to limit abuse potential, one would want any ideal pleasure drug to be akin - in one limited but important sense - to nicotine, where the smoker's brain finely calibrates its optimal level: there is no uncontrolled dose escalation.

There are, of course, all kinds of pitfalls to drug-based solutions. Technically, I think these pitfalls can be overcome, though I won't try to show this here. But there is a deeper issue. If there weren't something fundamentally wrong - or at least fundamentally inadequate - with our existing natural state of consciousness bequeathed by evolution, then we wouldn't be so keen to change it. Even when it's not unpleasant, everyday consciousness is *mediocre* compared to what we call “peak experiences”. Ordinary everyday consciousness was presumably adaptive in the sense that it helped our genes

leave more copies of themselves on the African savannah; but why keep it as our default-state indefinitely? Why not change human nature by literally repairing our genetic code?

Again, this dismissal of pharmacological solutions may be too quick. Arguably, utopian designer drugs may always be useful for the *fine-grained* and readily reversible control of consciousness; and I think designer drugs will be an indispensable tool to explore the disparate varieties of conscious mind. But wouldn't it be better if we were all *born* with a genetic predisposition to psychological superhealth rather than needing chronic self-medication? Does even the most ardent abolitionist propose to give cocktails of drugs to all children from birth; and then to take such drug cocktails for the rest of our lives?

3) So thirdly, there are **genetic** solutions, embracing both somatic and germline therapy.

By way of context, today there is a minority of people who are always depressed or dysthymic, albeit to varying degrees. Studies with mono- and dizygotic twins confirm a high degree of genetic loading for depression. Conversely, there are some people who are temperamentally optimistic. Beyond the optimists, there is a very small minority of people who are what psychiatrists call “hyperthymic”. Hyperthymic people aren't manic or bipolar; but by contemporary standards, they are always exceedingly happy, though not uniformly so. Hyperthymic people respond appropriately and adaptively to their environment. Indeed, they are characteristically energetic, productive and creative. Even when they are blissful, they aren't "blissed out".

Now what if, as a whole civilisation, we were to opt to become genetically hyperthymic - to adopt a motivational system driven entirely by adaptive gradients of well-being? More radically, as the genetic basis of hedonic tone is

understood, might we opt to add multiple copies of hyperthymia-promoting genes/allelic combinations and their regulatory promoters - not abolishing homeostasis and the hedonic treadmill, but shifting our hedonic set-point to a vastly higher level?

Three points here:

First, this genetic recalibration might seem to be endorsing another kind of uniformity; but it's worth recalling that happier people - and especially hyperdopaminergic people - are typically responsive to a broader range of potentially rewarding stimuli than depressives: they engage in more exploratory behaviour. This makes getting stuck in a sub-optimal rut less likely, both for the enhanced individual and posthuman society as a whole.

Second, universal hyperthymia might sound like a gigantic experiment; and in a sense, of course, it is. But *all* sexual reproduction is an experiment. We play genetic roulette, shuffling our genes and then throwing the genetic dice. Most of us flinch at the word "eugenics"; but that's what we're effectively practising, crudely and incompetently, when we choose our prospective mates. The difference is that within the next few decades, prospective parents will be able to act progressively more rationally and responsibly in their reproductive decisions. Pre-implantation genetic screening is going to become routine; artificial wombs will release us from the constraints of the human birth-canal; and a revolution in reproductive medicine will begin to replace the old Darwinian lottery. The question is not whether a reproductive revolution is coming, but rather what kinds of being - and what kinds of consciousness - do we want to create?

Third, isn't this reproductive revolution going to be the prerogative of rich elites in the West? Probably not for long. Compare the brief lag between the introduction of, say,

mobile phones and their world-wide adoption with the 50 year time-lag between the introduction and world-wide adoption of radio; and the 20 year lag between the introduction and world-wide penetration of television. The time-lag between the initial introduction and global acceptance of new technologies is shrinking rapidly. And so is the price.

Anyway, one of the advantages of genetically recalibrating the hedonic treadmill rather than abolishing it altogether, at least for the foreseeable future, is that the *functional* analogues of pain, anxiety, guilt and even depression can be preserved without their nasty raw feels as we understand them today. We can retain the functional analogues of discontent - arguably the motor of progress - and retain the discernment and critical insight lacking in the euphorically manic. Even if hedonic tone is massively enhanced, and even if our reward centres are physically and functionally amplified, it's still possible *in principle* to conserve much of our existing preference architecture. If you prefer Mozart to Beethoven, or philosophy to pushpin, then you can still retain this preference ranking even if your hedonic tone is hugely enriched.

Now personally, I think it would be better if our preference architecture were radically changed, and we pursued [please pardon the jargon] a "re-encephalisation of emotion". Evolution via natural selection has left us strongly predisposed to form all manner of dysfunctional preferences that harm both ourselves and others for the benefit of our genes. Recall Genghis Khan: “The greatest happiness is to scatter your enemy, to drive him before you, to see his cities reduced to ashes, to see those who love him shrouded in tears, and to gather into your bosom his wives and daughters.”

Now I'm told academia isn't quite that bad, but even university life has its forms of urbane savagery - its competitive status-seeking and alpha-male dominance rituals: a

zero-sum game with many losers. Too many of our preferences reflect nasty behaviours and states of mind that were genetically adaptive in the ancestral environment. Instead, wouldn't it be better if we rewrote our own corrupt code? I've focused here on genetically enhancing hedonic tone. Yet mastery of the biology of emotion means that we'll be able, for instance, to enlarge our capacity for *empathy*, functionally amplifying mirror neurons and engineering a sustained increase in oxytocin-release to promote trust and sociability. Likewise, we can identify the molecular signatures of, say, spirituality, our aesthetic sense, or our sense of humour - and modulate and "over-express" their psychological machinery too. From an information-theoretic perspective, what is critical to an adaptive, flexible, intelligent response to the world is not our absolute point on a hedonic scale, but that we are informationally sensitive to differences. Indeed information theorists sometimes simply *define* information as a "difference that makes a difference".

However, to stress again, this re-encephalisation of emotion is optional. It's technically feasible to engineer the well-being of all sentience *and* retain most but not all of our existing preference architecture. The three technical options for abolishing suffering presented here - wireheading, designer drugs and genetic engineering - aren't mutually exclusive. Are they exhaustive? I don't know of any other viable options. Some transhumanists believe we could one day all be scanned, digitized and uploaded into inorganic computers and reprogrammed. Well, perhaps. I'm sceptical, but in any case, this proposal doesn't solve the suffering of existing organic life unless we embrace so-called destructive uploading - a holocaust option I'm not even going to consider here.

2: WHY IT SHOULD HAPPEN

Assume that within the next few centuries we will acquire these Godlike powers over our emotions. Assume, too, that the *signalling* function of unpleasant experience can be

replaced - either through the recalibration argued for here, or through the offloading of everything unpleasant or routine to inorganic prostheses, bionic implants or inorganic computers - or perhaps through outright elimination in the case of something like jealousy. Why should we all be abolitionists?

If one is a **classical utilitarian**, then the abolitionist project follows: it's Bentham plus biotechnology. One doesn't have to be a classical utilitarian to endorse the abolition of suffering; but all classical utilitarians should embrace the abolitionist project. Bentham championed social and legislative reform, which is great as far as it goes; but he was working before the era of biotechnology and genetic medicine.

If one is a scientifically enlightened **Buddhist**, then the abolitionist project follows too. Buddhists, uniquely among the world's religions, focus on the primacy of suffering in the living world. Buddhists may think that the Noble Eightfold Path offers a surer route to Nirvana than genetic engineering; but it's hard for a Buddhist to argue in principle against biotech if it works. Buddhists focus on relieving suffering via the extinction of desire; yet it's worth noting this extinction is technically optional, and might arguably lead to a stagnant society. Instead it's possible both to abolish suffering *and* continue to have all manner of desires.

Persuading followers of **Islam** and the **Judeo-Christian** tradition is more of a challenge. But believers claim - despite anomalies in the empirical evidence - that Allah/God is infinitely compassionate and merciful. So if mere mortals can envisage the well-being of all sentience, it would seem blasphemous to claim that God is more limited in the scope of His benevolence.

Most contemporary philosophers aren't classical utilitarians or Buddhists or theists. Why should, say, an **ethical pluralist** take the abolitionist project seriously?

Here I want to take as my text Shakespeare's

"For there was never yet philosopher

That could endure the toothache patiently"

[*Much Ado About Nothing*, Scene Five, Act One (Leonato speaking)]

When one is gripped by excruciating physical pain, one is always shocked at just how frightful it can be.

It's tempting to suppose that purely "psychological" pain - loneliness, rejection, existential angst, grief, anxiety, depression - can't be as atrocious as extreme physical pain; yet the reason over 800,000 people in the world take their own lives every year is mainly psychological distress. It's not that other things - great art, friendship, social justice, a sense of humour, cultivating excellence of character, academic scholarship, etc - aren't valuable; but rather when intense physical or psychological distress intrudes - either in one's own life or that of a loved one - we recognize that this intense pain has immediate *priority* and *urgency*. If you are in agony after catching your hand in the door, then you'd give short shrift to someone who urged you to remember the finer things in life. If you're distraught after an unhappy love affair, then you don't want to be tactlessly reminded it's a beautiful day outside.

OK, while it lasts, extreme pain or psychological distress has an urgency and priority that overrides the rest of one's life projects; but so what? When the misery passes, why not just get on with one's life as before?

Well, natural science aspires to "a view from nowhere", a notional God's-eye view. Physics tells us that no here-and-now is privileged over any other; all are equally real.

Science and technology are shortly going to give us Godlike powers over the entire living world to match this Godlike perspective. I argue that so long as there is any sentient being who is undergoing suffering similar to our distress, that suffering should be tackled with the same priority and urgency as if it were one's own pain or the pain of a loved one. With power comes complicity. Godlike powers carry Godlike responsibilities. Thus the existence of suffering 200 years ago, for instance, may indeed have been terrible; but it's not clear that such suffering can sensibly be called "immoral" - because there wasn't much that could be done about it. But thanks to biotechnology, now there is - or shortly will be. Over the next few centuries, suffering of any kind is going to become optional.

If you're *not* a classical ethical utilitarian, the advantage of recalibrating the hedonic treadmill rather than simply seeking to maximise superhappiness is that you are retaining at least a recognizable descendant of our existing preference architecture. Recalibration of the hedonic treadmill can be made consistent with your existing value scheme. Hence even the ill-named "**preference utilitarian**" can be accommodated. Indeed, control over your emotions means that you can pursue your existing life projects more effectively.

And what about the alleged character-building function of suffering? "That which does not crush me makes me stronger", said Nietzsche. This worry seems misplaced. Other things being equal, enhancing hedonic tone strengthens motivation - it makes us psychologically more robust. By contrast, prolonged low mood leads to a syndrome of learned helplessness and behavioural despair.

I haven't explicitly addressed the value nihilist - the **subjectivist** or ethical sceptic who says all values are simply matters of opinion, and that one can't logically derive an "ought" from an "is".

Well, let's say I find myself in *agony* because my hand is on a hot stove. That agony is intrinsically motivating, even if my conviction that I ought to withdraw my hand doesn't follow the formal canons of logical inference.

If one takes the scientific world-picture seriously, then there is nothing ontologically special or privileged about *here-and-now* or *me* - the egocentric illusion is a trick of perspective engineered by selfish DNA.

If it's wrong for me to be in agony, then it is wrong for anyone, anywhere.

3: WHY IT WILL HAPPEN

OK, it's technically feasible. A world without suffering would be wonderful; and full-blown paradise-engineering even better. But again, so what? It's technically feasible to build a thousand-metre cube of tofu. Why is a pain-free world going to happen? Perhaps it's just wishful thinking. Perhaps we'll opt to retain the biology of suffering indefinitely[3].

The counterargument here is that whether or not one is sympathetic to the abolitionist project, we are heading for a **reproductive revolution** of designer babies. Prospective parents are soon going to be choosing the characteristics of their future children. We're on the eve of the Post-Darwinian Transition, not in the sense that selection pressure will be any less severe, but evolution will no longer be "blind" and "random": there will no longer be natural selection, but unnatural selection. We will be choosing the genetic makeup of our future offspring, selecting and designing alleles and allelic combinations *in*

anticipation of their consequences. There will be selection pressure against nastier alleles and allelic combinations that were adaptive in the ancestral environment.

Unfortunately, this isn't a rigorous argument, but imagine you are choosing the genetic dial-settings for mood - the hedonic set-point - of your future children. What settings would you choose? You might not want gradients of lifelong superhappiness, but the overwhelming bulk of parents will surely want to choose happy children. For a start, they are more fun to raise. Most parents across most cultures say, I think sincerely, that they want their children to be happy. One may be sceptical of parents who say happiness is the *only* thing they care about for their kids - many parents are highly ambitious. But other things being equal, happiness signals success - possibly the ultimate evolutionary origin of why we value the happiness of our children as well as our own.

Of course, the parental choice argument isn't decisive. Not least, it's unclear how many more generations of free reproductive choices lie ahead before radical anti-aging technologies force a progressively tighter collective control over our reproductive decisions - since a swelling population of ageless quasi-immortals can't multiply indefinitely in finite physical space. But even if centralised control of reproductive decisions becomes the norm, and procreation itself becomes rare, the selection pressure against primitive Darwinian genotypes will presumably be intense. Thus it's hard to envisage what future social formations would allow the *premeditated* creation of any predisposition to depressive or anxiety disorders - or even the "normal" pathologies of unenhanced consciousness.

Non-Human Animals

So far I've focused on suffering in just one species. This restriction of the abolitionist project is parochial; but our anthropocentric bias is deeply rooted. Hunting, killing, and exploiting members of other species enhanced the inclusive fitness of our genes in the ancestral environment. [Here we are more akin to chimpanzees than bonobos.] So unlike, say, the incest taboo, we don't have an innate predisposition to find, say, hunting and exploiting non-human animals wrong. We read that Irene Pepperberg's parrot, with whom we last shared a common ancestor several hundred million years ago, had the mental age of a three-year-old child. But it's still legal for so-called sportsmen to shoot birds for fun. If sportsmen shot babies and toddlers of our own species for fun, they'd be judged criminal sociopaths and locked up.

So there is a contrast: the lead story in the news media is often a terrible case of human child abuse and neglect, an abducted toddler, or abandoned Romanian orphans. Our greatest hate-figures are child abusers and child murderers. Yet we routinely pay for the industrialized mass killing of other sentient beings so we can eat them. We eat meat even though there's a wealth of evidence that functionally, emotionally, intellectually - and critically, in their capacity to suffer - the non-human animals we factory-farm and kill are equivalent to human babies and toddlers.

From a notional God's-eye perspective, I'd argue that morally we should care just as much about the abuse of functionally equivalent non-human animals as we do about members of our own species - about the abuse and killing of a pig as we do about the abuse or killing of a human toddler. This violates our human moral intuitions; but our moral intuitions simply can't be trusted. They reflect our anthropocentric bias - not just a moral limitation, but an intellectual and perceptual limitation too. It's not that there are

no differences between human and non-human animals, any more than there are no differences between black people and white people, freeborn citizens and slaves, men and women, Jews and gentiles, gays or heterosexuals. The question is rather: are they *morally* relevant differences? This matters because morally catastrophic consequences can ensue when we latch on to a real but morally irrelevant difference between sentient beings. [Recall how Aristotle, for instance, defended slavery. How could he be so *blind*?] Our moral intuitions are poisoned by genetic self-interest - they weren't designed to take an impartial God's-eye view. But greater intelligence brings a greater cognitive capacity for empathy - and *potentially* an extended circle of compassion. Maybe our superintelligent/superempathetic descendants will view non-human animal abuse as no less abhorrent than we view child abuse: a terrible perversion.

True or not, surely we aren't going to give up eating each other? Our self-interested bias is too strong. We like the taste of meat too much. Isn't the notion of global veganism just utopian dreaming?

Perhaps so. Yet within a few decades, the advent of genetically-engineered vatfood means that we can enjoy eating "meat" tastier than anything available today - without any killing or cruelty. As a foretaste of what's in store, the In Vitro Meat Consortium was initiated at a workshop held at the Norwegian University of Life Sciences in June 2007. Critically, growing meat from single stem cells is likely to be scalable indefinitely: its global mass consumption is potentially cheaper than using intact non-human animals. Therefore - assuming that for the foreseeable future we retain the cash nexus and market economics - cheap, delicious vatfood is likely to displace the factory-farming and mass-killing of our fellow creatures.

One might wonder sceptically: are most people really going to eat gourmet vatfood, even if it's cheaper and more palatable than flesh from butchered non-human animals?

If we may assume that vatfood is marketed properly, yes. For if we discover that we prefer the taste of vat-grown meat to the taste of carcasses of dead animals, then the moral arguments for a cruelty-free diet will probably seem much more compelling than they do at present.

Yet even if we have global veganism, surely there will still be terrible cruelty in Nature? Wildlife documentaries give us a very Bambified view of the living world: it doesn't make good TV spending half an hour showing a non-human animal dying of thirst or hunger, or slowly being asphyxiated and eaten alive by a predator. And surely there has to be a food chain? Nature is cruel; but predators will always be essential on pain of a population explosion and Malthusian catastrophe.

Not so. *If* we want to, intelligent agents can use cross-species depot-contraception[4], redesign the global ecosystem, and rewrite the vertebrate genome to get rid of suffering in the rest of the natural world too. For non-human animals don't need liberating; they need *looking after*. We have a duty of care, just as we do to human babies and toddlers, to the old, and the mentally handicapped. This prospect might sound remote; but habitat destruction means that effectively all that will be left of Nature later this century is our wildlife parks. Just as we don't feed terrified live rodents to snakes in zoos - we recognise that's barbaric - will we really continue to permit cruelties in our terrestrial wildlife parks because they are "natural"?

The last frontier on Planet Earth is the ocean. Intuitively, running compassionate ecosystems might seem too complicated. But the exponential growth of computer power and nanorobotic technologies means that we can, in theory, comprehensively re-engineer

marine ecosystems too. Currently such re-engineering is still impossible; but in a few decades, the task will be computationally feasible but challenging. Eventually, it will be technically trivial. So the question is: will we actually do it? *Should* we do it - or alternatively, should we conserve the Darwinian status quo? Here we are clearly in the realm of speculation. Yet one may appeal to what might be called "The Principle Of Weak Benevolence". Unlike the controversial claim that superintelligence entails superempathy, The Principle Of Weak Benevolence *doesn't* assume that our technologically and cognitively advanced descendants will be any more morally advanced than we are now.

Let's give a concrete example of how the principle applies. If presented today with the choice of buying either free-range or factory-farmed eggs, most non-vegan consumers will pick the free-range eggs. If battery-farmed eggs are one penny cheaper, most people will still pick the "cruelty-free" option. No, one shouldn't underestimate human malice, spite and bloody-mindedness; but most of us have at least a *weak* bias towards benevolence. If any non-negligible element of self-sacrifice is involved, for example if free-range eggs cost even 20 pence more, then sales fall off sharply. My point is that if - and it's a big if - the sacrifice involved for the morally apathetic could be made non-existent or trivial, then the abolitionist project can be carried to the furthest reaches of the living world.

The Reproductive Revolution

Selection Pressure in a Post-Darwinian World

Here are three predictions about life one thousand years from now:

1) Suffering of any kind will be biologically impossible. Our descendants will lead lives of genetically pre-programmed bliss whose worst lows surpass today's peak experiences. A thousand years hence, the heritable hedonic set-point of ordinary waking life will have been ratcheted upwards so that everyday existence feels sublime.

2) Our genetically enhanced successors won't grow old and die, but will be effectively immortal, barring accidents, which mean certain brains have to be restored from digital backup.

3) Posthumans will be innately smarter than us, not just in the narrow autistic sense of intelligence measured by contemporary IQ tests, but also in the sense that they will have a more empathetic intelligence. To use a non-scientific term, our descendants will be "wiser" than contemporary humans.

These are bold claims. They could, of course, be completely mistaken: futurology doesn't have a brilliant track record. However, I'm going to argue why these three seemingly unrelated developments - superhappiness, superlongevity and superintelligence - are intimately linked. We are on the brink of a revolution in reproductive medicine - the coming era of designer babies, a fundamental transition in the evolution of life in the

universe. Evolution will shortly cease to be "blind" and "random", as it has been for the past four billion years. Instead, intelligent agents are going to choose and design genotypes *in anticipation of* their likely behavioural and psychological effects. Specifically, prospective parents will increasingly choose the genetic makeup of their future children rather than playing genetic roulette. Natural selection is going to be replaced by "unnatural" selection.

But first, let us outline a very different, bioconservative vision, perhaps best represented today by the distinguished geneticist at University College London, Professor Steve Jones.

Two Contrasting Views of Future Human Evolution

1) BIOCONSERVATIVISM: ["*The End of Evolution*"?] "If you want to know what Utopia is like, just look around - this is it", says Professor Jones in a Royal Society debate in Edinburgh. In a talk[1] entitled "Is Evolution Over?" Prof. Jones says: "Things have simply stopped getting better, or worse, for our species." Professor Jones explains how there were three components to human evolution – natural selection, mutation and random change. "Quite unexpectedly, we have dropped the human mutation rate because of a change in reproductive patterns."

"In ancient times half our children would have died by the age of 20. Now, in the Western world, 98 per cent of them are surviving to 21", says Professor Jones in a recent interview[2] with *The Times*. The mutation rate is also slowing down. Although chemicals and radioactive pollution could cause genetic changes, one of the most important mutation triggers was advanced age in men. "Perhaps surprisingly, the age of reproduction has gone down - the mean age of male reproduction means that most

conceive no children after the age of 35. Fewer older fathers means that if anything, mutation is going down."

It's worth adding that some scientists and right-wing commentators go further than Steve Jones. They argue that because nominally more intelligent people have fewer children than nominally less intelligent people, the intelligence of the human species as a whole is actually going to decline. This prediction isn't borne out by the long-term increase in IQ scores over the last century, the "Flynn Effect". However, believers in the so-called dysgenic fertility hypothesis counter that it is possible for genotypic IQ to decline even while phenotypic IQ rises throughout the population, at least in the short run. They explain this paradox by environmental effects such as better schooling, improved nutrition, and even television viewing.

By contrast to the bioconservative perspective:

2) BIOREVOLUTION: Human evolution is about to accelerate. Selection pressure isn't going to slacken. On the contrary, we're on the eve an era of *un*natural or artificial selection - a different kind of selection pressure, but a selection pressure that will be extraordinarily intense, favouring a very different set of adaptations than traits that were genetically adaptive in the ancestral environment on the African savannah.

Let's quickly review some background. The Human Genome Project (HGP) was the international scientific research project that aimed to determine the sequence of chemical base pairs of our DNA: the genetic make-up of our species. Researchers identified, physically and functionally, the 25,000 or so genes of the human genome. The project was formally declared complete in 2003, though in reality there are a lot of loose ends to be tied up. The full implications of our deciphered code have scarcely been glimpsed. They may take centuries to unravel.

Currently [2009], if you want your whole genome of three billion odd base pairs sequenced, the price is several thousand dollars. This figure is prohibitively expensive for most people. [In 2015, the price had fallen to around one thousand dollars.] But in a decade or so, the cost on some estimates could be as little as ten dollars. Whatever the exact price or timing, the cost of access to one's own source code is poised to collapse. Routine access to one's personal genome will usher in an era of personalised medicine - individual drugs, dosages and gene therapies targeted at the individual rather than the scatter-gun approach we see in clinical pharmacology (and recreational drug use) today.

Yet we're not just heading for an era of personalized medicine - we're on the eve of an era of personalized reproductive medicine: "designer babies", to use the popular term. The phrase suggests something frivolous, akin to designer clothes. But choosing the genetic make-up of your child may soon become the badge of responsible parenthood - as distinct from throwing the genetic dice and hoping they roll the right way, as now. A reluctance to pass on harmful code to our children won't just apply to obvious autosomal dominant conditions like the neurological disorder Huntington's disease. What prospective parent, if offered the choice, is deliberately going to pass on genes for haemophilia, sickle-cell anaemia or muscular dystrophy? It has been estimated that on average we each carry four lethal recessive genes. In a future of post-genomic reproductive medicine, the selection pressure against, say, the cystic fibrosis allele, the cause of the most common life-limiting autosomal recessive disease among people of European heritage, is going to become intense, as indeed is selection pressure against a whole range of genes that cause or contribute to physical disease. Currently, we're used to Googling prospective partners on the Net to find out more about them. Looking ahead, what responsible prospective parent will neglect to check their partner's DNA - and their own - before having children? This *doesn't* mean that anyone who wants a child will

reject an asymptomatic partner who carries a recessive copy of a "nasty" gene. Instead, responsible parents can use preimplantation genetic diagnosis and germline gene therapy to ensure that potentially harmful genes like the recessive cystic fibrosis allele aren't passed on to their children.

Genetic Roulette Versus Designer Babies

Yet how about heritable psychological traits - "personality genes" that contribute to psychological pain? Not merely is there no consensus on whether some of their less pleasant variants should be classed as pathological, here too things are much more complex technically than for monogenic disorders like cystic fibrosis. This is because there is no such thing as a single gene "for" depression or anxiety disorders or jealousy or obsessive compulsive disorder (OCD) and so forth. But there *are* alleles and genotypes that predispose for depression or anxiety disorders or jealousy or obsessive compulsive disorder - and other polygenic, multifactorial psychological conditions. So if there is a particular allele - a variant gene - that makes it, say, 5% more likely that a particular trait such as low mood or chronic anxiety will be expressed, or an allele that makes its bearer 5% more or less anxious or more or less depressive, then what percentage of prospective parents will purposely choose the less pleasant variant for their children? Yes, there are numerous complications, for instance pleiotropy, where a single gene influences multiple phenotypic traits; alternative splicing, whereby a single gene may produce different proteins in different settings; genomic imprinting, a parent-dependent form of gene expression; non-Mendelian inheritance in the form of transgenerational epigenetic effects; and so forth. More generally, critics of the new genetic medicine worry about creating "designer personalities". Other things being equal, however, most informed parents will presumably choose the more compassionate option for their child.

Indeed, one Oxford Professor of Ethics goes further. Julian Savulescu argues that we are are morally *obligated* to select genetic blueprints for children with the greatest chance of leading the best life: what Prof. Savulescu dubs the Principle of Procreative Beneficence.

This conjecture isn't premature. For example, people who inherit two copies of a short version of the chromosome 17 serotonin transporter gene, 5-HTTLPR, have an 80 per cent chance of becoming clinically depressed if they experience three or more negative life-events in five years. By contrast, genetically resilient people who inherit the long version have only a 30 per cent chance of developing mental illness in similar circumstances. If offered the choice via preimplantation diagnosis (PGD), would you opt for the short or the long serotonin transporter gene variant for your future child? Or would you decline to choose, putting your faith in a God or Mother Nature?

Right now, of course, this kind of scenario still sounds far-fetched. Later this century and beyond, are prospective parents really going to enroll in courses in behavioural genetics and molecular biopsychiatry before having kids? For sure, certain genetic decisions are in principle straightforward, for example gender selection, or whether to pass on a cystic fibrosis allele. Such decisions are taken by some prospective parents in a few countries already. But other genetic decisions will be much more complicated, not least for "mood genes" that help determine a person's average level of well-being or ill-being over a lifetime.

For what it's worth, I personally think that taking advanced courses in behavioural genetics, or at least seeking genetic counselling, will be morally incumbent on *anyone* before assuming the immense responsibility of having a child. Yet this kind of education is unlikely to be widespread in the foreseeable future. The argument presented here doesn't depend on it. Instead, in an era of mature reproductive medicine, we may

forecast an abundance of user-friendly software tools to enable prospective parents to take responsible genetic decisions - as distinct from blindly taking their chances in the genetic lottery of Darwinian life. For the exponential growth in computing power can be harnessed to a new growth industry of sophisticated baby-authoring software. So the average parent will no more be required to understand molecular genetics than the average contemporary Windows PC user is required to understand machine code. And the parallel goes further. If it's ethically acceptable to spend hours redesigning your Windows PC desktop the way you like it, then why not at least take a few hours to make sure that your future child is psychologically and physically healthy too?

Of course, such authoring tools open up an ethical and regulatory minefield of gargantuan proportions. Yet so does sexual reproduction: playing the genetic equivalent of Russian roulette with a child's life.

Recalibrating the Hedonic Treadmill

OK, maybe prospective parents will choose to avoid alleles and allelic combinations associated with depression or anxiety disorders or schizophrenia when they prepare to have children. But what grounds are there for thinking that the *average* hedonic set-point of humankind as a whole will be ratcheted ever upwards? Recall that we all have a kind of inbuilt hedonic treadmill that prevents most of us from remaining extremely happy or extremely miserable for very long - though of course extreme misery can seem like an eternity while it lasts. Our hedonic treadmill tends to have an approximate hedonic set-point around which we fluctuate over time. This hedonic set-point crudely determines the average level of subjective well-being or ill-being that most people experience throughout a lifetime. Of course we're all buffeted by external events, both pleasant and unpleasant, that affect us acutely for good or ill; but over time, we mostly

revert to a [partly] heritable individual mean. In some people, the hedonic set-point tends to be fixed below the Darwinian average: such people have a gloomy temperament - what the ancients would have called an excess of black bile. In other people, the hedonic set-point is fixed above average: they are temperamentally optimistic. Some people's mood oscillates sharply, other people are more equable. But the current range of hedonic diversity aside, why may we predict that the typical default state of well-being of the human population is going to increase indefinitely - even after genes predisposing to anxiety disorders and clinical depression have been weeded out of the gene-pool?

The plain answer is that we can't know for sure. So this is speculation. Yet here is a thought experiment. Imagine that you have the option of choosing the genetic dial-settings of the hedonic set-point of your future child: the degree to which your child is temperamentally depressive or happy - or superhappy. To keep things simple, I won't yet consider the richer forms of emotional well-being, just normal hedonic tone, which we know is partly heritable. What average level of hedonic tone would you choose for your future child on a 10-point scale? [Here again I am being deliberately simplistic.] On the unscientific basis of a few straw polls conducted over the years, I'd estimate that most people, if pressed, would opt for a hedonic 8 or 9. Yet a high number of respondents say "10": they would like their children to be as temperamentally happy as possible.

Realistically, perhaps only a minority of prospective parents will initially want to have children disposed to be naturally *super* happy by contemporary norms. But most parents will want happy children, as distinct from depressive, moody, anxiety-ridden children. Not least, happy children are more fun to raise. Happy, resilient, self-confident children are also more likely to be "successful" over-achievers in the traditional Darwinian sense.

We needn't suppose that prospective parents care only about the happiness of their future kids: many parents-to-be are, of course, highly ambitious for their offspring. Anyhow, on this argument, the average, genetically constrained set-point of emotional well-being of our species is destined to rise over time as a reflection of these individual parental choices, as tomorrow's enhancement technologies shift social norms of well-being and become the next generation's remedial therapies. The depressive realism of one century may become the affective psychosis of the next. Over time, an analogous selection pressure may be exerted in favour of alleles and allelic combinations predisposing to high intelligence - and perhaps even genius and supergenius - although here any contribution to enhanced quality of life will be indirect. In any event, over a whole spectrum of physical and psychological traits, we may predict that germline enhancement will become germline remediation as the average level of biological well-being improves across human society. As biophysicist Gregory Stock notes in *Redesigning Humans* (2002), "The arrival of safe, reliable germline technology will [...] transform the evolutionary process by drawing reproduction into a highly selective social process that is far more rapid and effective at spreading successful genes than traditional sexual competition and mate selection." Thus the tempo of worldwide mood enrichment may accelerate.

Critically, the genetic mood enrichment conjecture *doesn't* hypothesise the future existence of any mega-project to make a happier world. The possibility of such a pan-global project can't be excluded - grandiose and fanciful as the idea of some kind of Hedonistic Imperative (1995) now sounds. Currently, only the tiny Himalayan Kingdom of Bhutan officially exalts Gross National Happiness (GNH) over Gross National Product (GNP). If hedonic enrichment *were* internationalized and pursued with scientific rigour, then the selection pressure against nastier Darwinian genotypes would be even more

severe than anticipated here. Now personally, I advocate a world-wide Abolitionist Project laid down as official United Nations policy. Not least, only a global mega-project can ever extend the abolition of suffering to the rest of the living world. Ecosystem redesign, cross-species depot-contraception, and eventually rewriting the whole vertebrate genome can't be achieved via private initiative. However, such a mega-project isn't imminent. Less extravagantly, global mood enrichment may be the collective outcome of billions of personal reproductive decisions made by individual parents-to-be during the next century and beyond.

Phrased in the language of designer babies, the prospect of species-wide hedonic enrichment evokes sinister images - even though it promises to make the world a much happier place. Do we really want parents controlling the destiny of their future children? But we have to be careful about how we frame the issue here. Just as good physical health is empowering, and doesn't determine what you *do* with your life, likewise being temperamentally happy and psychologically robust doesn't determine what you actually do with your life either. Like physical health, mental health tends to empower rather than constrain. Genetically hardwired mental *super* health is potentially even more empowering. It makes you psychologically indestructible. It stops you from ever becoming depressed or anxiety-ridden - and from suffering the crippling loss of life-opportunities that such conditions entail. Moreover, in the future anybody who *isn't* satisfied with aspects of their core personality, and who *doesn't* want to use consciousness-altering drugs to change it, can practise somatic gene therapy. We won't always be at the mercy of a scrambled mix of our parent's genes as now, whether those genes have been passed on by accident or design.

Future Nociception: The End of Physical Pain?

So far I've talked about the abolition of suffering, and how psychological pain can be genetically eliminated over time. But what about the terrible scourge of raw physical pain? Surely, the sceptic might wonder, genes that promote pain-sensitivity in response to tissue damage will be as adaptive one thousand years from now as they are today - and as they were in the ancestral environment. So the prediction that one thousand years hence, the worst experiences that anyone undergoes will be richer than today's peak experiences sounds like a pipe-dream. How is this even technically possible, let alone sociologically realistic?

Well, there is a short-to-medium term answer and a longer-term answer. Let's consider the short-to-medium term options first.

The Cyborg Solution versus Radical Recalibration.

At present there are different "natural" genetic variants that promote varying degrees of pain sensitivity, e.g. variant alleles of the gene SCN9A coding for the a-subunit of the voltage-gated sodium channel Nav1.7 in nociceptive neurons; the *mu*-opioid receptor gene; and the gene encoding catecholamine-O-methyltransferase (COMT). Few prospective parents are going to want kids who are *hyper*sensitive to physical pain. Most parents, if given the choice, will presumably seek no more than mild-to-modest pain-sensitivity for their offspring. Thus, if genetically planned parenthood ever becomes the norm, then our pain thermostats (or "algostats", as one might call them) are likely to be genetically re-set over time too.

But this recalibration doesn't actually abolish suffering, it just diminishes its prevalence and intensity when physical pain occurs. Moreover, as attested by rare cases of

congenital anaesthesia, children born without *any* capacity to suffer pain are currently liable to undergo all manner of life-threatening medical complications. So does this mean we are stuck with pain in some guise or other for ever?

No, though there are formidable technical challenges to overcome. If we are to abolish physical pain altogether, I think there are two long-term options. These two options are not mutually exclusive, but I will consider them separately. Recall how silicon robots with the right functional architecture can get by fine without the nasty "raw feels" of phenomenal pain; they can be programmed to avoid and respond flexibly and adaptively to noxious stimuli. Clearly, there is a distinction between the physiological function of nociception and the subjective experience of phenomenal pain; they are dissociable even in organic robots like us, not just in our inorganic counterparts. So likewise, *in theory* future humans could computationally offload everything nasty or routine onto prosthetic devices, nanobots and the like, preserving only the life-enriching forms of sentience and discarding the ugly Darwinian junk. This is what we may call the Cyborg Solution. The main advantage of the Cyborg Solution in the long run is that it permits maximum lifelong bliss for all sentient life. Thus, its ultimate adoption would seem mandatory on a classical utilitarian ethic. But assuming that we don't go down the cyborg route, there is another option. In principle, we can radically reset the scale of the pleasure-pain axis in the mind/brain. All that is needed for an organism to respond adaptively to a changing and potentially hostile environment is informational-sensitivity to fitness-relevant changes - including the binary "wonderful" versus "not-quite-as-wonderful" - regardless of the tidal range of our emotions on an absolute hedonic scale. A narrow compass of pleasure gradients can, in theory, play a role analogous to pain gradients in some victims of chronic pain syndrome today.

This hypothesis is counterintuitive. One might imagine that if people always feel more-or-less *super well* - both physically and psychologically - then they won't be motivated to act circumspectly; and therefore they will tend to hurt themselves, whether physically or emotionally or both. Who could respond adaptively to the world if consumed by a perpetual whole-body orgasm? Yet this doesn't follow. As we know today, the happiest people, the keenest life-lovers, tend to be the *most* motivated people. It's depressives who tend to be unmotivated. Yes, there *are* forms of happiness associated with indolence, for example opiated bliss. But there are also forms of happiness associated with intense motivation, forward planning and goal-directed behaviour - so-called hyperdopaminergic states. Either way, our descendants, and possibly our elderly selves, will have a choice of what *kinds* of physical and emotional well-being they want to enjoy, and a choice of what kinds of genetic predisposition to pass on to the next generation. If you don't want to bring any more suffering into the world, then your only option right now is to not have children. In the future, however, we'll be able to have cruelty-free children with a clear conscience - on that score at least.

Gradients of Bliss?

What's true of physical pain and depression is true of other negative states of mind. Thus the prediction that life a thousand years hence will feel orders of magnitude better than now *isn't* a claim that posthumans will all be *uniformly* happy, or that future life will be perfect, whatever that might mean. Indeed one can argue that discontent is the motor of progress, and that the *functional* analogues of discontent are likely to endure one thousand years from now, just as the raw feels of discontent exist at present. Admittedly, it's hard to know whether fourth millennium (post)humans will be endowed with anything even functionally resembling the same core emotions that define our lives

today. The molecular signature of some kinds of emotion, for example disgust, panic or jealousy, might be abolished altogether, both phenomenally and functionally, whereas genes and regulatory code for novel life-enriching emotions may be customised and spliced into the genome. Our perceptual and cognitive architecture is likely to be genetically reshaped too - probably in ways beyond the contemporary human imagination. But such innovation isn't essential for an improved quality of life. The *functional* analogues of anxiety and depression could still persist, and yet life could always be subjectively wonderful - since it's technically possible to decouple functional role from the subjective texture of unpleasant experience as we feel it now.

Critically, I'm not arguing that our descendants will enjoy uniform bliss, and certainly not that they will be manic or "blissed out", simply that their genetically constrained floor of *comparative* ill-being will be higher than our absolute ceiling of well-being. Continual germline-enhancement across the generations will create a novel motivational system. Its mechanisms of emotional homeostasis will transcend the Darwinian pleasure-pain axis. Thanks to the unfolding Reproductive Revolution, there will be continual selection pressure in favour of the biology of a subjectively improved quality of life. Equating net value and net happiness in the manner of classical utilitarian ethics may or may not be simplistic; but acknowledgement of the connection between enhanced value and enhanced emotional well-being is common to a whole range of ethical systems, both religious and secular. Few ethical systems give *no* weight to emotional well-being. Thus, if a piece of music sounds a thousand times more enchanting than its predecessor, or if a work of art looks a thousand times more beautiful than anything physiologically possible at present, then I think the default assumption must be that such overpowering beauty is indeed a good thing - in the absence of cogent arguments to the contrary. The new

germinal choice technologies allow the creation of subjectively valuable experience on a truly prodigious scale. So other things being equal, we should embrace their use.

Spiritual Well-Being?

The approach I've sketched so far probably sounds crudely reductionist. But one needn't interpret superhappiness in just a narrow one-dimensional sense. Take, for example, spirituality and spiritual well-being. In future, if you are very spiritual and want to have hyperspiritual children, then you can opt to over- or under-express the relevant genes or allelic combinations promoting a spiritual temperament; and perhaps ultimately design angelic "spiritual" genomes for your children. Indeed, if you want to be naturally superspiritual yourself and don't want to take entheogenic drugs, you could use autosomal gene enhancement and add extra copies or over-express variants of alleles and allelic combinations associated with spirituality. Secular rationalists, on the other hand, may prefer to lay the genetic foundations of a more worldly well-being.

To take another example of multi-dimensional well-being, prospective parents may be able to choose genes and genotypes associated, not just with intelligence in the simple-minded conventional sense, but with an increased capacity for empathy, involving functionally amplified mirror neurons and enhanced social cognition. Prospective parents will have the opportunity to endow their kids with an enriched oxytocin system, leading to greater trust, generosity of spirit, and pro-social behaviour, potentially with immense benefits for society as a whole. Such scenarios are, of course, speculative.

A Reproductive Elite?

An obvious question arises: Won't these new reproductive technologies be solely for the rich, or at least mainly for members of the prosperous developed nations that can buy the best genes, undercutting the argument from selection pressure advanced here?

Initially, surely yes. But not for long, even assuming [implausibly] that the world's poorest nations will remain poor indefinitely. Consider how rapidly web-enabled cell phones have spread through even impoverished sub-Saharan Africa. If personal genome sequencing always costs anything like the $200,000 it does now [December 2008; year 2013 = c.$10,000], then only an elite of affluent people could benefit from such breakthroughs. If personal genome sequencing cost ten dollars or less, then effectively everyone can have it. The nature of information and information technology entails that IT-based services don't involve the consumption of scarce natural resources in the way material goods do, where one person's gain is frequently another person's loss. Only a handful of people in the world can ever own a Rolls Royce or a Maserati, and even fewer can own an original Picasso or an Old Master; but an unlimited number of people can listen to the world's entire catalogue of music, enjoy access to all its electronic games, its computer software, its movies, or indeed the whole Library of Congress. Information is effectively free, or at least it will be soon. Later this century, reproductive technologies like preimplantation genetic screening (PGS) and diagnosis (PGD) - techniques used to identify genetic defects in embryos created through *in vitro* fertilization before pregnancy - are going to become dirt-cheap too. Already, crude personal genotyping services are available for a few hundred dollars.

Of course, it's easy to sing a happy tune with the word "soon". I'm glossing over a host of problems in the transitional era between old-fashioned sexual reproduction and true

planned parenthood. "Soon" in this context may mean decades, and perhaps centuries. But even on the most conservative timescales, we're on the brink of a major discontinuity in the four-billion-year odyssey of the evolution of life on Earth.

Some Unknowns

Human Cloning

One big unknown affecting any conjectures about future selection pressure is the role of human cloning. Whether human reproductive cloning takes another five years or fifty years, it is going to happen. What's less clear is the cost and expertise involved when the technology matures, and what its global implications for selection pressure will be. If human cloning will always take a large team of research professionals, complex medical equipment, many failed attempts and a great deal of money, then it will presumably always be rare. But if it can ever be done cheaply and safely at home, perhaps via DIY cloning kits available for purchase over the Net, then human cloning could become a common way to make babies, regardless of official laws and regulations.

For the sake of argument, let's suppose that human cloning *does* eventually become a common mode of reproduction. It's not clear this is a bad development *per se*, any more than identical twins or triplets are intrinsically bad. Either way, this possibility might seem to throw a big spanner into the argument from selection pressure I'm making here, since genetically identical babies are likely to suffer from the same problems as their father or mother if exposed to a similar environment.

Yet it seems a reasonable assumption that most future human cloners won't seek to create *exact* genetic duplicates of themselves, but will instead aspire to have offspring free of defects or unwanted characteristics possessed by their parent. To use a trivial

example, a human cloner with thinning hair wouldn't necessarily want to have a cloned child with a predisposition to grow bald. Granted, most Asian people who want a clone will want to have children who are Asian-looking, and most blue-eyed people will arguably want blue-eyed clones, but presumably carriers of the cystic fibrosis allele won't seek to pass the defective gene on to their cloned offspring. Likewise, for the most part, depressive people who might like to clone themselves *aren't* likely to want depressive children. Cases of "negative enhancement", akin to the existing use of preimplantation genetic diagnosis to select an embryo for the presence of a particular disability such as deafness shared by the parent(s), will presumably be uncommon. So yes, if human cloning becomes widespread, and certainly if human cloning becomes cheap and ubiquitous, then its spread makes the argument from selection pressure defended here more complex; but the practice wouldn't fundamentally undercut its conclusion.

Autosomal Gene Therapy and Enhancement

Another unknown that adds to the complexity of the selection pressure argument is the future extent of autosomal gene therapy. I've been focusing on reproduction and germ-line gene therapy and genetic enhancement; but somatic gene therapy is sure to become available and probably extensively used too. After all, if offered the choice of either taking a drug to remedy some physical or psychological defect for the rest of your life, or curing that deficit with a one-off course of gene therapy, which would you choose? The same is true of future enhancement technologies - though remediation versus enhancement is a naïve dichotomy.

Potential Pitfalls

The Spectre of Coercive Eugenics

Anyone uncritically enthusiastic about the Reproductive Revolution in prospect would do well to reflect on the history of the twentieth century. In the words of bioethicist Nicholas Agar, "Those who do not learn from the history of human enhancement may be doomed to repeat it". One recalls the forced segregation, sterilization, racial hygiene, the euthanasia program and ultimately the genocide practised in the pseudo-scientific name of eugenics. Might the impending Reproductive Revolution lead to similar horrors? After all, there are still plenty of people in the world convinced that some races are intellectually or morally superior to other races. Might history repeat itself?

The short answer is yes, though I think such scenarios are unlikely. For a start, the totalitarian dictatorships of the twentieth century, not least the Third Reich, all depended on censorship and a state-monopoly of information. The Internet makes the creation of totalitarian dictatorships much harder; as has been well said, the Internet interprets censorship as damage and re-routes. However, this is obviously a huge topic. All I'll say here is that there is a fundamental difference between a regulatory system where eugenics [under whatever name] is practised for the well-being of the *individual* - whether human or non-human - and an authoritarian society where eugenics is practised for the notional benefit of a class, race or nation.

Even so, there are clearly lots of problems with so-called liberal eugenics. For instance, there are pitfalls with prospective parents choosing enhancements that offer a merely *positional advantage* to their children. To give a concrete example, if parents pick genes likely to allow their child to grow taller than the current average, then there is no net benefit to either the child or society if most other parents do the same. Indeed, if human stature were to become significantly higher than today, then we would all be prone to multiple health difficulties under Earth's gravitational regime. Even enhancements such

as genes that may contribute to superior intelligence - over-expressing or adding extra copies of the NRP2 or ASPM or microcephalin gene to use a contentious example - that sound as though they could confer intrinsic benefit might arguably amount to positional goods like height. Thus, women tend to find intelligence sexy in prospective mates; but presumably what's advantageous to the brainy male bearer in terms of enhanced sex-appeal is relative - and not absolute - intelligence. A counter to this argument might be that there are inherent benefits to high male intelligence aside from attracting women.

In contrast with interventions that confer positional advantage, genetic enhancements that enrich subjective well-being - crudely, whether you are temperamentally happy or superhappy - would be *intrinsically* beneficial; they can potentially benefit *everyone*, regardless of where one falls on any comparative scale of well-being. Indeed, technologies that biologically enrich emotional well-being are arguably the *only* enhancements that are *intrinsically* good as distinct from positionally or instrumentally good. This claim is obviously controversial; it would be contested by many bioethicists who aren't classical utilitarians.

Other pitfalls?

Although designer genomes can, in principle, lead to vastly greater diversity, might designer genomes lead, in practice, to greater genetic uniformity? Would most parents strive to have similar kinds of "ideal" children, the supernormal reflections of preferences adaptive in our Darwinian past? Admittedly, some kinds of genetic uniformity are presumably desirable. Thus, by common consent, it would be a blessing if there were no gene for Huntington's disease (HD). But twentieth century eugenicists didn't take account of phenomena such as heterozygote advantage - normally defined as cases where the heterozygote genotype has a higher relative fitness than either the homozygote

dominant or homozygote recessive genotype. Heterozygote advantage explains why some kinds of genetic variability persist, most famously the gene for sickle-cell anaemia. Analogous heterozygote advantage may exist for psychological traits too, though this is unproven.

Whatever their evolutionary origin, here are three examples where the issues are complicated.

The Future of Homosexuality: Even if you have absolutely no prejudices at all about homosexuality, would you choose so-called gay genes for your child - variant alleles that predispose your child to be gay? Now of course it's possible that in 50 or 150 years time, homophobia will have been relegated to the dustbin of history where it belongs; but I wouldn't count on it. In the meantime, what percentage of prospective parents, whether straight or gay or bisexual, will deliberately choose to have a gay child knowing the greater social problems that child would likely encounter in life due to social prejudice? If this is the case, and if there is indeed a Reproductive Revolution as outlined here, then it is quite likely that genes predisposing to homosexuality and possibly even bisexuality will be strongly selected against. They may even die out. If one looks in human history from classical antiquity to the present at the contribution made by people whom we would probably classify as gay or bisexual, and likewise at the contribution of their close genetic relatives, then this is not an outcome to be contemplated lightly. On the other hand, it's also possible that many gay couples will use the new reproductive technologies to have gay children, rendering the gay extinction scenario moot.

The Future of Bipolar Disorder: Chronic unipolar depression may be an unmitigated evil; but what about Bipolar Disorder, formerly known as manic depression? Bipolar Disorder can undoubtedly cause terrible suffering both to its victims and their families.

Yet many creative high achievers in art, science and politics have at the very least been soft bipolars. Is there a danger that something valuable will be lost if in future prospective parents weed out of the gene-pool alleles associated with bipolarity? Again, this is a huge topic.

The Future of Autism Spectrum Disorders: Classical autism is characterized by varying degrees of "mindblindness" and deficits in social interaction; deficits in language, communication, and the capacity for social play; and multiple stereotypies of behaviour. The three most common forms of autism spectrum disorders (ASD) are classical autism; pervasive developmental disorder not otherwise specified (PDD-NOS); and Asperger's syndrome. Whereas children with, say, trisomy 21 (Down syndrome) or Williams syndrome can be abnormally sociable - and therefore rewarding to raise - by contrast autistic children with an absent or underdeveloped theory of mind commonly cause great distress to their caregivers. It is hard to bond with someone who always treats you as an object. Thus, any genetic disposition to autism might seem a prime candidate for elimination from the gene-pool as the Reproductive Revolution gathers pace. However, some of the greatest scientists who ever lived, notably Newton, Einstein and Dirac, fulfilled many or all of the diagnostic criteria for Asperger's syndrome. To what extent was their scientific acumen separable from their pathologies of mind?

Calculating Risk-Reward Ratios

If there are likely to be so many possible adverse and/or unintended consequences of the new reproductive medicine - and perhaps dystopian outcomes no one has even considered - then why forge ahead? Why not outlaw the new reproductive technologies altogether, or at least drastically restrict their use to simple Mendelian genetic diseases of the body rather than complex disorders of the mind/brain? After all, there is no way

we can computationally model all the ramifications of even modest rewrites of the human genome.

Here the question comes down to an analysis of risk-reward ratios - and our basic ethical values, themselves shaped by our evolutionary past. Lest extension of the new reproductive medicine seem too rashly experimental even to contemplate, it's worth recalling that each act of old-fashioned sexual reproduction is itself an untested genetic experiment, the outcome of random mutations and meiotic shuffling of the genetic deck. So just who are we to accuse of reckless gambling? As it stands, all of us are genetically predestined to grow old and die; and in the course of a lifetime, the great majority of humans will experience periods of intense psychological distress, for instance loneliness and heartache after an ended love affair. Our social primate biology ensures that most of us sometimes experience, to a greater or lesser degree, all manner of nasty states that were genetically adaptive in the ancestral environment, e.g. jealousy, resentment, anger, and so forth. Hundreds of millions of people in the world today suffer bouts of depression; others live with chronic anxiety. One might say these phenotypes are part of what it means to be human. Worse, we pass a heritable predisposition to these horrible states on to our children.

Bioconservatives, religious traditionalists, and social reformers alike would contest this bleak analysis. If you believe that human life today is fundamentally good, and viciously unpleasant states of mind are an aberration that can be mostly remedied by improving society, then you will need compelling reasons before wanting to change the regime of ordinary sexual reproduction as it exists now. Most likely, you will be loathe to support anything like the Reproductive Revolution predicted here; and focus entirely on its potential dangers. The spectre of "Brave New World" will probably loom large in any

discussion. If, on the other hand, you think that Darwinian life is cruel and tragic by its very nature, then you are more likely to be willing to contemplate radical alternatives to the genetic *status quo*, despite the possible risks.

My own view of the risks and uncertainties is that there is a critical distinction between trying to abolish suffering exclusively via social reform, and abolishing suffering directly via biotechnology. As we know, utopian social experiments typically go wrong, sometimes hideously wrong, and end up causing a lot of suffering instead. The abolitionist project of eradicating the biological substrates of suffering sounds like just another utopian scheme, whether it's touted as a grandiose species-project, or simply as a byproduct of the Reproductive Revolution explored here. Although the abolition of psychological pain is arguably no more utopian in principle than pain-free surgery, it could presumably go wrong in unanticipated ways. Perhaps we'll unwittingly create a fool's paradise. But if and when we ever abolish the molecular underpinning of unpleasant experience, and it becomes *physiologically* impossible for any sentient being to suffer, we thereby change the very meaning of what it means for anything to "go wrong". Unwelcome surprises where sentient beings do not get hurt are very different from unwelcome surprises where they do. For what it's worth, I think the abolition of involuntary suffering is the precondition of any civilised posthuman society; and therefore a risk worth taking.

The End of Sexual Reproduction?

OK, I've outlined grounds for believing that our nastier Darwinian emotions will be selected against in future. Yet there is a fundamental objection to the argument from selection pressure that I've sketched so far. Surely most people, not least teenagers, will carry on producing babies by having sex together, regardless of any so-called

Reproductive Revolution of laboratory-mediated conception. Unplanned pregnancies are extremely common even in an age where contraceptives are widely available. Yes, maybe responsible, forward-looking parents will seek to ensure that they have children who are free of genetic handicaps, who are joyful, ultra-intelligent, super-empathetic and psychologically robust; and maybe in future such responsible parents-to-be will practise preimplantation genetic diagnosis, use germline gene therapy and pursue some of the futuristic interventions described here. But that won't stop feckless teenagers having unplanned babies. In addition, *billions* of people may be reluctant to embrace the new reproductive technologies for traditional moral or religious reasons, or simply out of custom and habit. It stretches the imagination to envisage genetically planned parenthood ever becoming as prevalent as, say, anaesthetics to guarantee pain-free surgery. If most women continue to bear genetically *un*enriched babies by the conventional route, then surely our inbuilt genetic tendency to all forms of Darwinian suffering is going to express itself indefinitely?

Maybe so. It's a powerful argument. Yet there are strong grounds for thinking that traditional-style sexual reproduction can't continue for more than a few generations. The reason is bound up with the coming revolution in anti-aging medicine.

Throughout most of human history, radical life-extension, let alone the prospect of eternal youth, has been the province of quacks and charlatans. To some extent it still is; swallowing a bunch of vitamin pills each day isn't going to let you live for ever. But over the next few centuries, and possibly before, aging and the genes that promote or allow senescence are going be phased out. This is, of course, a bold claim that I won't even attempt to defend in detail here. If you are sceptical and haven't read the book already, I'd recommend Aubrey de Grey's *Ending Aging: The Rejuvenation Breakthroughs That*

Could Reverse Human Aging in Our Lifetime (2007). Now I am more pessimistic than Aubrey de Grey about timescales. Yet the genetic and pharmacological interventions that we are already trying in nonhuman animals will eventually be tried in the human animal too. One hesitates to embrace what sounds like a facile technological determinism; but I think we can say, quite dogmatically, that if and when radical anti-aging technologies become available, the overwhelming majority of people will use them - regardless of any rationalizations of death and aging we express now. Moreover, most people will also want such treatments for their family pets; the Anti-aging Revolution won't be confined to one species.

Let's assume for the sake of argument that this is the case, i.e. there will be *both* a Reproductive Revolution *and* an anti-aging Revolution. If post-genomic medicine dramatically extends our lifespan, and fewer and fewer people die of the traditional diseases of old age, then our planet will soon reach its carrying capacity. Looking centuries ahead, a rapidly expanding population of eternally youthful quasi-immortals means that human reproduction of any kind will have to become rare, and eventually a momentous event, and tightly controlled in every respect. It's here that I foresee both the greatest ethical dilemmas arising from the Reproductive Revolution and also the intimate link between superhappiness, superintelligence and superlongevity.

Selection Pressure in an Age of Quasi-Immortality

When the Earth reaches its carrying capacity - the maximum packing density of sentient beings consistent with sustainable life - there will have to be immensely greater centralized control of the human reproductive system on pain of complete Malthusian catastrophe. This does indeed sound like a truly sinister prediction. Perhaps one can imagine the existence of a mandatory regime of depot-contraception from an early age.

Yet could depot-contraception really be made fail-safe? How would such fertility control be enforced? Moreover, the problem isn't just preventing reproductive accidents. The urge to have one's "own" children can be extraordinarily strong, as attested by the anguish caused by involuntary childlessness today; and for many childless couples, this yearning could eclipse any general worries about the carrying capacity of the planet. A majority of people will want both to stay forever young *and* to have children. If radical anti-aging technologies are indeed widely adopted, then a central and unavoidably intrusive control of human reproduction may be inevitable, though one may trust such powers will be accountable to democratic control. In an era of mass superlongevity, every intellectually competent citizen will presumably recognize, in the abstract, that unlimited reproduction is physically impossible. On the other hand, some people will presumably try to have unregulated, unsanctioned children, just as they do in the People's Republic of China (PRC) today, albeit without the promise of eternal youth. This is not an attractive parallel. Of course there are other social perils associated with mass superlongevity: in an era of genetically pre-programmed eternal youth, the ruling power elites may prove almost immovable in the absence of adequate democratic safeguards. But the potential loss of bodily autonomy and procreative liberty is especially troubling to the liberal conscience - and to any libertarian life-extensionist.

A counterargument here is that the urge to bear children is under genetic control; and that urge will itself be amenable to biological intervention. Manipulation of our first-order desires is likely to prove biologically easier than defeating aging. Yet if most of one's enhanced fellow citizens do act responsibly and forgo or postpone reproduction, then any predisposition to "cheat" and have children might be highly (genetically) adaptive, at least in the short-run. Such an outcome would be disastrous in an already overpopulated global megalopolis. Plausible group selectionist scenarios aren't easy to construct, even

for the far future. Hence, the price of posthuman superlongevity is the likelihood of ever greater state intervention in the (hitherto) private realm - although such intrusiveness need not be subjectively *distressing* in any sense we would recognise today, since the functional analogue of distress might suffice. Long before any era of post-genomic medicine, Plato believed that human reproduction should be monitored and controlled by the state, a portent of totalitarian societies to come; but once we transcend the biology of human mortality, some sort of collective control of reproductive decision-making may prove inescapable even in a liberal democracy. The only alternative to such control would be draconian, state-enforced rationing of anti-aging therapies: a scarcely credible re-enactment of *Logan's Run*. It's important to note that this argument doesn't turn on whether it transpires that the ultimate carrying capacity of our planet is 15 billion, or 150 billion, or conceivably even higher packing densities. Yes, we can colonise the solar system. In theory, too, in some era of the distant future, the authorities on Earth could tell anyone who wants to have a child that they must do so on one of the extrasolar planetary systems that we colonise. But for the next few centuries at least, and possibly millennia, the prospect of some kind of Galactic adaptive radiation is pure science fiction. For it is hard to overstate the technical obstacles to mass interstellar travel. Quite possibly posthumans *will* go to the stars, and perhaps even colonise our local galactic supercluster in a few million years or so. Realistically, this doesn't solve the near-term demographic challenge of a massively overcrowded Earth.

Admittedly I am making a number of contestable assumptions here. I will note just three. First, intelligent life won't wipe itself out altogether in the next few decades. [Doomsday scenarios are conceivable; but they are much harder to construct once self-sustaining colonies are established on other planets later this century.] Second, there is a

unique past and a unique future. [This simplifying assumption is inconsistent with quantum cosmology and most likely false. However, consideration of the "branch density" measure of alternative, classically inequivalent histories in post-Everett quantum mechanics would take us too far afield in this talk.] Third, unlike futurists who believe in "uploading", I am assuming that our (post)human descendants will retain an organic substrate - maybe augmented by web-enabled neurochips, nanobots, bionic implants and the like - and hence that humans won't scan, digitize and "upload" themselves to dwell in another computational medium where the constraints of the Earth's ecosystem don't apply. [There is no evidence that your PC is any more conscious than an abacus, despite its greater processing power; and if a souped-up version of your PC contained a digitized representation of you, this would doubtless facilitate restoration from backups, but there are no grounds for thinking such lines of code would be conscious either - let alone "you". Yes, artificial intelligence will hasten the Reproductive Revolution; and perhaps one day we will all become web-enabled cyborgs. And who knows what kinds of exotic postbiological artificial life can be evolved if and when our descendants run mature quantum computers. Yet there is simply no evidence that inorganic systems with a classical von Neumann architecture support "raw feels", or that they intrinsically *matter*: the notion that our species might destructively upload ourselves from basement Reality into digital nirvana is unworkable.] So here at least I am being tamely bioconservative in assuming that the Earth 1000 years hence will support a densely populated primordial "meatworld" of our flesh-and-blood post-human descendants.

Anyhow, to summarise, assume that the creation of new quasi-immortal beings will indeed become exceedingly rare later this millennium. The Earth will be (almost) literally full. I'd argue that on such historic occasions as the creation of a new posthuman-being, it is unlikely that superhappy, superintelligent agents will create the genetic malware for

unpleasant, stupid, senile substrates of consciousness, i.e. archaic *Homo sapiens*. Our posthuman descendants are more likely to create fellow "smart angels" instead. The triumph of the Reproductive Revolution will have reshaped the post-Darwinian fitness landscape beyond all recognition. Hence my (tentative) prediction that the biology of suffering and senescence is destined to pass into evolutionary history.

HIGH-TECH JAINISM

Introduction

"May all that have life be delivered from suffering", said Gautama Buddha. The vision of a happy biosphere isn't new. Jains, for instance, aim never to hurt another sentient being by word or deed. But all projects of secular and religious utopianism have foundered on the rock of human nature. Evolution didn't design us to be happy.

Yet the living world is poised for a major evolutionary transition. Natural selection has thrown up a species able to self-edit its own genetic source code; phase out experience below "hedonic zero"; and engineer the well-being of all sentience in our forward light-cone. Intelligent agents will shortly be able to pre-select their own hedonic range: its upper and lower bounds, and hedonic set-points. Posthuman life can be animated by gradients of intelligent bliss - a default hedonic tone orders of magnitude richer than today's peak experiences.

Why Does Suffering Exist?

No one knows why suffering exists at all. To the best of our knowledge, unpleasant experience doesn't play any irreplaceable or computationally unique role in intelligent agents. Inorganic robots can be programmed or trained up to avoid and respond to noxious stimuli without undergoing subjective distress. Likewise, nonbiological machines can functionally replicate the role of our nastier core emotions without their "raw feels" - the ugly implementation detail that blights so many lives today.

Fortunately, solving the problem of suffering doesn't depend on our first solving the Hard Problem of consciousness. Neuroscanning and the tools of molecular biology are deciphering the "neural correlates of consciousness". If we use biotechnology to eradicate the molecular signature of experience below "hedonic zero", then on some fairly modest assumptions, phenomenal suffering becomes physically impossible.

So a practical question arises. Which existing psychological *functions* should we enrich, replicate or scrap? What kinds of function are best offloaded onto smart prostheses rather than biologically tweaked? Ideally, adaptations such as a predisposition to jealous behaviour might be abolished along with their nasty subjective textures. Such Darwinian traits have few defenders, even among bioconservatives. Other roles, notably nociception, will presumably be *functionally* essential for sentient beings to flourish for the foreseeable future - and perhaps indefinitely. Initially, preimplantation genetic screening of prospective children can ensure tomorrow's humans are endowed with benign, "low-pain" alleles of e.g. the SCN9A(1) gene to modulate pain-sensitivity. People blessed with high pain tolerance aren't vulnerable to the life-threatening information-processing deficits of congenital analgesia. Eventually, the avoidance of noxious stimuli

can be offloaded onto smart inorganic prostheses, allowing life based entirely on information-sensitive gradients of bliss.

The Reproductive Revolution

Natural selection hasn't favoured a motivational architecture of gradients of bliss: it's "blind". Genetic mutations are effectively random; sexual reproduction is a crapshoot. As long as reproduction endures, some form of selection pressure is inevitable. But the nature of selection pressure is transformed when rational agents pre-select and customise the genomes of their future children *in anticipation of* the likely behavioural and psychological effects of their choices. Far-seeing "artificial" selection changes the rules of the game - in human and nonhuman animals alike. Selection pressure against traits scripted by our nastier code will intensify as the reproductive revolution gathers pace.

Clearly, "life events" matter hugely to each of us: genetic determinism is facile and simplistic. Genes and culture have co-evolved. Epigenetics, the heritable changes in gene activity not caused by changes in DNA sequence, and the purely conditional activation of different genes and allelic combinations "for" particular psychological traits, complicate the simple-minded storyline told here. Yet twin studies confirm that hedonic set-points - crudely, whether we are temperamentally happy or gloomy - have a high degree of genetic loading. More specifically, studies of e.g. the role of variant alleles of the 5-HTT serotonin transporter(2) ("the depression gene"); the COMT gene(3) (high versus low reward); and deletion variant of ADA2b(4) ("the pessimism gene") corroborate twin studies(5) of our heritable hedonic range. In an era of routine preimplantation genetic screening, prospective parents will presumably select code predisposing to emotional, intellectual and physical superhealth for their children in preference to disease and frailty.

It's unclear when selection pressure for a predisposition toward greater subjective well-being will plateau. Why settle for the mediocre when we can enjoy the sublime?

Credible modelling of the long-term trajectory of selection pressure in the post-Darwinian era is a formidable challenge. Yet it's not complete guesswork. Here a single example must suffice. Imagine you can use preimplantation genetic screening to choose, approximately, the hedonic set-point of your future child. What default hedonic tone would you pick? Oversimplifying, let's designate "minus 10" a predisposition to chronic severe depression; "0" to be hedonically neutral; and "plus 10" a predisposition to lifelong gradients of bliss. Informal straw-polling suggests a mean preference for "plus 7"s or "plus 8"s - with a perhaps surprising number of "plus 10"s. Today, depressives with sub-zero hedonic baselines form a significant minority of the population. A majority of people in the course of a lifetime cluster quite tightly to either side of hedonic zero - with varying degrees of equability or emotional volatility. Either way, we needn't assume for the purposes of this thought-experiment that most parents care primarily about their children's happiness *per se*. For evolutionary reasons, many parents are intensely ambitious for their offspring. Other things being equal, psychologically resilient children tend to be "winners" - and more fun to raise too. So let's assume such anecdotal and impressionistic evidence is borne out by well-controlled studies. What hedonic dial-settings will these (super)happy children choose in turn when, as parents-to-be, they decide to have children of their own? In consequence of such individual parental choices - and perhaps top-down medical paternalism - default levels of subjective well-being will presumably be ratcheted upwards world-wide as the reproductive revolution unfolds later this century and beyond. Further genetically engineered reward pathway enhancements open up the prospect of an immensely richer hedonic ceiling and an elevated hedonic floor: true genomic rewrites rather than simple preimplantation screening. The negative

feedback mechanisms of the hedonic treadmill can still play out; but on an exalted plane. The pitfalls are legion. So are the potential psychological rewards. Living in Heaven is fun.

When will the reproductive revolution take off?

Might traditional sexual reproduction predominate indefinitely?

Early in the twenty-first century, the use of genomic medicine to phase out terrible genetic disorders like cystic fibrosis or Tay–Sachs disease commands widespread but not universal assent. More controversial among bioethicists and prospective parents alike will be phasing out genes and allelic combinations predisposing to anxiety disorders and depression. In the West, if not China(6), the spectre of coercive eugenics hangs over the debate. Critics of "designer babies" claim that misery and malaise are "part of what it means to be human". No doubt the critics are right. But low mood is at least as devastating to the quality of life of depressives as genetic disorders like cystic fibrosis. Depressive disorder causes almost a million people in the world to take their own lives each year.

Where should genetic remediation - or enhancement - stop?

The World Health Organisation (WHO) defines health as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.”

Why not take the WHO definition literally?

As so defined, complete health can be secured only via post-genomic medicine.

Untreated humans would be prone to malaise in the Garden of Eden.

Rapid Genome Self-Editing

The reproductive revolution is potentially good news for our children and grandchildren. But what about existing humans? Like the prediction that medical science will deliver a cure for ageing shortly after your death, the possibility that our descendants may enjoy lifelong gradients of bliss can elicit mixed feelings. However, the CRISPR$_{(Z)}$ gene-editing revolution heralds an era when rapid self-editing of your own genome can become the norm. Rapid genome self-editing promises radical enhancement options not just for our descendants, but our future selves. And just as today using computers no longer entails writing low-level machine code, likewise powerful suites of user-friendly editing tools can revolutionise the user experience of genomic self-modification. The same is true of cybernetic self-enhancement. The smarter our artificially intelligent machines, the more effectively organic robots can edit our own biological wetware in a recursive cycle of self-improvement.

Why Recalibration Matters

Why not *maximise* happiness?

A venerable tradition in philosophical ethics, namely classical or "hedonistic" utilitarianism, bids us to maximise the happiness of the greatest number of sentient beings. For reasons we don't understand, the pain-pleasure axis seemingly discloses the world's inbuilt metric of (dis)value. Naturally, we don't know what intensities of bliss our distant successors may choose, or its guises; conceivably, posthuman superintelligence may opt for some kind of utilitronium shockwave propagating across the cosmos. In the meantime, we have abundant grounds for caution.

Two advantages of hedonic recalibration are worth noting here.

First, combining hedonic enrichment *and* recalibration ensures that critical insight, reinforcement learning, and social responsibility can potentially be retained while simultaneously massively enriching subjective quality of life. Recalibration undercuts the dilemma of hedonic enhancement as standardly posed. Should we choose gritty reality or an escapist fantasy world of self-delusion? The messy real world or Nozick's "Experience Machine"? Authenticity versus drug-addled life on soma? The Red Pill or the Blue Pill? And so forth.

Yet it's a false dilemma. We needn't choose between the Red Pill and the Blue Pill. Genetically enlightened agents can take the Purple Pill, so to speak, combining the benefits of realism and recalibration. For sure, mood-congruent cognitive *biases* are potentially a risk anywhere on the hedonic scale; without them, low mood might never have evolved in the first instance(8). But rose-tinted spectacles, so to speak, can be corrected with the tools of AI and decision-theoretic rationality. Genetic case-studies can also be conducted on contemporary high-functioning hedonic outliers, i.e. fortunate souls who are temperamentally "hyperthymic" but not manic. There's no evidence that traits such as empathy, intellectual prowess and virtues of character are less common in hyperthymics than depressives(9). What's clear is that other things being equal, a life animated by happy experiences is appreciated as more valuable than a life of mediocre experience, just as mediocre experiences are more valuable than nastiness. Compare our appreciation of art, music or literature. If it's not rewarding, it's no good. Other things being equal, superhappy life is supervaluable too - subjectively at any rate: philosophers can endlessly debate its transcendental (in)significance. More informally, take care of happiness, and the meaning of life takes care of itself.

A second reason for embracing recalibration rather than happiness-maximisation is human cultural diversity. We're social primates; most of our preferences implicate

others. Across the world, diverse people have diverse religious and secular value systems; and trillions of inconsistent and frequently irreconcilable preferences, trivial and profound. Theists and atheists, deontologists and utilitarians, virtue theorists and contractualists, pluralists and theory-scorning pragmatists dispute how best to live. Logically, let alone practically, there is no way to satisfy even the "idealised" preferences of liberals and conservatives, Christians and Muslims, jealous rivals in love - or fanatical Manchester United and Manchester City supporters. By contrast, reward pathway enhancements can radically enrich everyone's quality of life without forcing choices between "winners" and "losers": the zero-sum games endemic to Darwinian life. Recalibrating your hedonic set-point does not entail reducing neurodiversity, or buying into other people's utopias or their conception of the good life - unless of course you're opposed to the principle of recalibration itself. Hedonic enhancement can preserve whatever values, preferences and human relationships you hold most dear, while discarding states of mind you would gladly lose.

In practice, of course, the pleasures and preferences of posthumans may be humanly inconceivable.

Who Benefits?

Historically, the blessings of consumer capitalism have been enjoyed mainly by a privileged elite, slowly and erratically percolating socially downwards. Information-based technologies are different. The price of genome sequencing is collapsing. Preimplantation genetic screening is already more common in India and China than the West. The nature of information-based services is such that their price trends effectively to zero. Genomic rewrites will be cost-effective both ethically and financially. Thus, the burden of depression, both clinical and subclinical, currently costs hundreds of billions of dollars

each year to the global economy - quite aside from the misery of its victims. Critically, information-based services don't need to be rationed. For example, "counterfeit" genome-editing tools will be as inferior as "counterfeit" copies of Microsoft Word - an affront to intellectual property lawyers, no doubt, but not the bane of end-users.

The Rise of Full-Spectrum Superintelligence

Naively, a happy world is an intellectually stagnant world. Huxley's dystopian classic *Brave New World* shapes our preconceptions about universal bliss. In reality, the enterprise of knowledge has scarcely begun. Natural science can mathematically describe the formal, structural properties of the physical world with astonishing fidelity and predictive power. Yet first-person states of mind are an enigma. Drug-induced psychedelia(10) hints at the existence of immense state-spaces of consciousness as different as dreaming is from waking: a tantalising mental *terra incognita* beyond the bounds of normal human experience. Alas, at present, exploration of psychedelia is unsafe to all but the most mentally robust psychonauts. The risk of nightmarish "bad trips" lurks within our dysfunctional reward circuitry. Hence the controlled status of psychedelic drugs in contemporary society - and lame *a priori* philosophising about consciousness in academia. A foundation of invincible well-being can inaugurate a future post-Galilean science of mind - a knowledge explosion to complement the happiness explosion.

The Plight of the Cognitively Humble

Human civilisation is based on an animal holocaust. Billions of incarcerated nonhuman animals suffer and die in factory farms and slaughterhouses each year. Their sentience and sapience is comparable to human infants and toddlers. Over the next few decades, mass-manufactured *in vitro* meat promises to replace the barbarities of factory farming.

The technology of cultured meat products can amplify mankind's minimal and uneven benevolence to our fellow creatures.

More controversially, technology can accelerate the transition from harming to helping free-living sentient beings: mankind's fitfully expanding "circle of compassion". The civilising process needn't be species-specific, but instead extend to free-living dwellers in tomorrow's wildlife parks. Every cubic metre of the biosphere will soon be computationally accessible to surveillance, micro-management and control. Fertility regulation via immunocontraception can replace Darwinian ecosystems governed by starvation and predation. Any species of obligate carnivore we choose to preserve can be genetically and behaviourally tweaked into harmlessness. Asphyxiation, disembowelling, and agonies of being eaten alive can pass into the dustbin of history.

Critics warn darkly of hubris. Yet *Homo sapiens* already "plays God": humans massively interfere with the rest of the living world. What's in question is whether we will act as callous or benevolent deities. Power breeds deepening complicity. It's hard to predict whether recognisable approximations of human or nonhuman Darwinian life will be preserved by posthumans. Perhaps we'll transform ourselves into post-Darwinian superbeings and consign primordial life to oblivion. In the meantime, an ethic of compassionate conservatism offers a compromise between the cruelties of orthodox conservation biology and the outright extinction of Darwinian life-forms. Free-living human or nonhuman animals do not lose some mysterious species-essence when they cease to be "wild"; on this score if no other, conservationists should sleep easy.

Suffering and Existential Risk

The problem of suffering and the problems of global and existential catastrophic risk(11) might seem tangential. In reality, maintenance of the biological status quo is hazardous

to the prospects of civilisation and perhaps life itself. Evolution "designed" male humans to be hunters and warriors. The existence of suffering in a world of weapons of mass destruction presents profound global and existential catastrophic risks. Angst-ridden depressives, misanthropes, doomsday cultists and anti-natalists are more likely to believe sentience is a mistake and act accordingly. How many suicidal depressives would take the world down with them if the apocalyptic technology were at hand? A world of ubiquitous life-lovers is safer than a world full of smart but tormented Darwinians. Other things being equal, the more that intelligent beings love life, the more motivated we are to preserve it.

Paradise Engineering?

Other risks are more subtle. Imagine we stumble across an advanced civilisation that has abolished ageing, disease and unpleasant experience of any kind: a "Triple S" civilisation of superintelligence, superlongevity and superhappiness. The inhabitants enjoy lives animated by gradients of lifelong bliss. What arguments might human critics use to persuade its members to reintroduce involuntary suffering, predation, parasitism, ageing and the miseries of the ancestral past? The extraterrestrials would regard us as crazy: primitives in the grip of some kind of depressive psychosis.

Yet contrary to appearances, the advanced civilisation *is* guilty of one ethically catastrophic mistake. Its inhabitants have embraced the hedonistic imperative(12) too avidly and turned inwards too soon. They could have launched cosmic rescue missions for pain-ridden Darwinian life on Earth; and assumed responsible stewardship of the physical universe within their cosmological horizon.

In the real world, maybe we're alone. The skies look empty. Cynics might point to the mess on Earth and echo C.S. Lewis: "Let's pray that the human race never escapes from

Earth to spread its iniquity elsewhere." Yet our ethical responsibility is to discover whether other suffering sentients exist within our cosmological horizon; establish the theoretical upper bounds of rational agency; and assume responsible stewardship of our Hubble volume. Cosmic responsibility entails full-spectrum superintelligence: to be blissful, but not "blissed out" - high-tech Jainism on a cosmological scale. We don't yet know whether the story of life has a happy ending.

* * *

REFERENCES

1. Reimann. F. *et al*. "Pain perception is altered by a nucleotide polymorphism in SCN9A." *Proc Natl Acad Sci* U S A. 2010 Mar 16;107(11):5148-53. doi: 10.1073/pnas.0913181107.

2. Lykken, D., Tellegen, A. "Happiness Is a Stochastic Phenomenon." *Psychological Science* Vol.7, No. 3, May 1996.

3. Karg, K., Burmeister, M., Shedden, K., Sen, S. "The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation." *Arch Gen Psychiatry*. 2011 May;68(5):444-54. doi: 10.1001/archgenpsychiatry.2010.189.

4. Wichers, M. *et al*. "The catechol-O-methyl transferase Val158Met polymorphism and experience of reward in the flow of daily life." *Neuropsychopharmacology*. 2008 Dec;33(13):3030-6.

5. Todd R.M., *et al.* "Genes for emotion-enhanced remembering are linked to enhanced perceiving". *Psychol Sci*. 2013 Nov 1;24(11):2244-53. doi: 10.1177/0956797613492423.

6. Specter, M. (2014). "The Gene Factory: A Chinese firm’s bid to crack hunger, illness, evolution—and the genetics of human intelligence". *The New Yorker*, January 6, 2014.

7. Sander J.D., Joung J.K. (2014). "CRISPR-Cas systems for editing, regulating and targeting genomes". *Nature Biotechnology*. doi:10.1038/nbt.2842. PMID 24584096.

8. Sloman, L. (Ed.), Gilbert, P. (Ed.) (2000). "Subordination and Defeat: An Evolutionary Approach To Mood Disorders and Their Therapy". Routledge.

9. But see Jamison, R.J. (1996). "Touched with Fire: Manic-Depressive Illness and the Artistic Temperament". Free Press.

10. Shulgin, A. (1995). "PiHKAL: A Chemical Love Story". Berkeley: Transform Press.

11. Bostrom, N. (Ed), Cirkovic, M.M. (Ed). (2008). "Global Catastrophic Risks". OUP Oxford.

12. Pearce, D.C. (1995, 2014). "The Hedonistic Imperative". https://www.hedweb.com.

BRAVE NEW WORLD?

A Defence of Paradise-Engineering

Brave New World (1932) is one of the most bewitching and insidious works of literature ever written.

An exaggeration?

Tragically, no. Brave New World has come to serve as the false symbol for *any* regime of universal happiness.

For sure, Huxley was writing a satirical piece of fiction, not scientific prophecy. Hence to treat his masterpiece as ill-conceived futurology rather than a work of great literature might seem to miss the point. Yet the knee-jerk response of "It's Brave New World!" to *any* blueprint for chemically-driven happiness has delayed research into paradise-engineering for all sentient life.

So how does Huxley turn a future where we're all notionally happy into the archetypal *dys*topia? If it's technically feasible, what's wrong with using biotechnology to get rid of mental pain altogether?

Brave New World is an unsettling, loveless and even sinister place. This is because Huxley endows his "ideal" society with features calculated to alienate his audience. Typically, reading BNW elicits the very same disturbing feelings in the reader which the

society it depicts has notionally vanquished - not a sense of joyful anticipation. In *Brave New World Revisited* (1958), Huxley describes BNW as a "nightmare".

Thus BNW *doesn't*, and isn't intended by its author to, evoke just how wonderful our lives could be if the human genome were intelligently rewritten. In the era of post-genomic medicine, our DNA is likely to be spliced and edited so we can all enjoy life-long bliss, awesome peak experiences, and a *spectrum* of outrageously good designer-drugs. Nor does Huxley's comparatively sympathetic account of the life of the Savage on the Reservation convey just how nasty the old regime of pain, disease and unhappiness can be. If you think it does, then you enjoy an enviably sheltered life and an enviably cosy imagination. For it's all sugar-coated pseudo-realism.

In *Brave New World*, Huxley contrives to exploit the anxieties of his bourgeois audience about *both* Soviet Communism and Fordist American capitalism. He taps into, and then feeds, our revulsion at Pavlovian-style behavioural conditioning *and* eugenics. Worse, it is suggested that the price of universal happiness will be the sacrifice of the most hallowed shibboleths of our culture: "motherhood", "home", "family", "freedom", even "love". The exchange yields an insipid happiness that's unworthy of the name. Its evocation arouses our unease and distaste.

In BNW, happiness derives from consuming mass-produced goods, sports such as Obstacle Golf and Centrifugal Bumble-puppy, promiscuous sex, "the feelies", and most famously of all, a supposedly perfect pleasure-drug, soma.

As perfect pleasure-drugs go, soma underwhelms. It's not really a utopian wonderdrug at all. It does make you high. Yet it's more akin to a hangoverless tranquilliser or an opiate - or a psychic anaesthetising SSRI like Prozac - than a truly life-transforming elixir.

Third-millennium neuropharmacology, by contrast, will deliver a vastly richer product-range of designer-drugs to order.

For a start, soma is a very one-dimensional euphoriant. It gives rise to only a shallow, unempathetic and intellectually uninteresting well-being. Apparently, taking soma doesn't give Bernard Marx, the disaffected sleep-learning specialist, more than a cheap thrill. Nor does it make him happy with his station in life. John the Savage commits suicide soon after taking soma [guilt and despair born of serotonin depletion!?]. The drug is said to be better than (promiscuous) sex - the only sex brave new worlders practise. But a regimen of soma doesn't deliver anything sublime or life-enriching. It doesn't catalyse any mystical epiphanies, intellectual breakthroughs or life-defining insights. It doesn't in any way promote personal growth. Instead, soma provides a mindless, inauthentic "imbecile happiness" - a vacuous escapism which makes people comfortable with their lack of freedom. The drug heightens suggestibility, leaving its users vulnerable to government propaganda. Soma is a narcotic that raises "a quite impenetrable wall between the actual universe and their minds".

If Huxley had wished to tantalise, rather than repel, emotional primitives like us with the biological nirvana soon in prospect, he could have envisaged utopian wonderdrugs which *reinforced* or enriched our most cherished ideals. In our imaginations, perhaps we might have been allowed - via chemically-enriched brave new worlders - to turn ourselves into idealised versions of the sort of people we'd most like to be. In this scenario, behavioural conditioning, too, could have been used by the utopians to sustain, rather than undermine, a more sympathetic ethos of civilised society and a life well led. Likewise, biotechnology *could* have been exploited in BNW to encode life-long fulfilment and super-

intellects for everyone - instead of manufacturing a rigid hierarchy of genetically-preordained castes.

Huxley, however, has an altogether different agenda in mind. He is seeking to warn us against scientific utopianism. He succeeds all too well. Although we tend to see *other* people, not least the notional brave new worlders, as the hapless victims of propaganda and disinformation, we may find it is we ourselves who have been the manipulated dupes.

For Huxley does an effective hatchet-job on the very sort of "unnatural" hedonic engineering that most of us so urgently need. One practical consequence has been to heighten our already exaggerated fears of state-sanctioned mood-drugs. Hence millions of screwed-up minds, improvable even today by clinically-tested mood-boosters and anti-anxiety agents, just suffer in silence instead. In part this is because people worry they might become zombified addicts; and in part because they are unwilling to cast themselves as humble supplicants of the medical profession by taking state-rationed "antidepressants". Either way, the human cost in fruitless ill-being is immense.

Fortunately, the Net is opening up a vast trans-national free-market in psychotropics. It will eventually sweep away the restrictive practices of old medical drug cartels and their allies in the pharmaceutical industry. The liberatory potential of the Net as a global drug-delivery and information network has only just begun.

Of course, Huxley can't *personally* be blamed for prolonging the pain of the old Darwinian order of natural selection. Citing the ill-effects of *Brave New World* is not the same as impugning its author's motives. Aldous Huxley was a deeply humane person as well as a brilliant polymath. He himself suffered terribly after the death of his adored mother. But

death and suffering will be cured only by the application of bioscience. They won't be abolished by spirituality, prophetic sci-fi, or literary intellectualism.

So what form might this cure take?

In the future, it will be feasible *technically* - at the very least - for pharmacotherapy and genetic medicine to re-engineer us so that we can become - to take one example among billions - a cross between Jesus and Einstein. Potentially, transhumans will be endowed with a greater capacity for love, empathy and emotional depth than anything neurochemically accessible today. Our selfish-gene-driven ancestors - in common with the cartoonish brave new worlders - will strike posterity as functional psychopaths by comparison; and posterity will be right.

In contrast to Brave New World, however, the death of ageing won't be followed by our swift demise after a sixty-odd year life-span. We'll have to reconcile ourselves to the prospect of living happily ever after. Scare-mongering prophets of doom notwithstanding, a life of unremitting bliss isn't nearly as bad as it sounds.

The good news gets better. Drugs - not least the magical trinity of empathogens, entactogens and entheogens - and eventually genetic engineering will open up revolutionary new *state spaces* of thought and emotion. Such modes of consciousness are simply unimaginable to the drug-innocent psyche. Today, their metabolic pathways lie across forbidden gaps in the evolutionary fitness landscape. They have previously been hidden by the pressure of natural selection: for nature has no power of anticipation. Open such spaces up, however, and new modes of selfhood and introspection become accessible. The Dark Age of primordial Darwinian life is about to pass into history.

In later life, Huxley himself modified his antipathy to drug-assisted paradise. *Island* (1962), Huxley's conception of a *real* utopia, was modelled on his experiences of mescaline and LSD. But *until* we get the biological underpinnings of our emotional well-being securely encoded genetically, then psychedelia is mostly off-limits for the purposes of paradise-engineering. Certainly, its intellectual significance cannot be exaggerated; but unfortunately, neither can its ineffable weirdness and the unpredictability of its agents. Thus mescaline, and certainly LSD and its congeners, are not fail-safe euphoriants. The possibility of nightmarish bad trips and total emotional Armageddon is latent in the way our brains are constructed under a regime of selfish-DNA. Uncontrolled eruptions within the psyche must be replaced by the precision-engineering of emotional tone, if nothing else. If rational design is good enough for inorganic robots, then it's good enough for us.

In Brave New World, of course, there are no freak-outs on soma. One suspects that this is partly because BNW's emotionally stunted inhabitants don't have the imagination to have a bad trip. But mainly it's because the effects of soma are no more *intellectually* illuminating than getting a bit drunk. In BNW, our already limited repertoire of hunter-gatherer emotions has been constricted still further. Creative and destructive impulses alike have been purged. The capacity for spirituality has been extinguished. The utopians' "set-point" on the pleasure-pain axis has indeed been shifted. But it's flattened at both ends.

To cap it all, in Brave New World, life-long emotional well-being is *not* genetically pre-programmed as part of everyday mental health. It isn't even assured from birth by euphoriant drugs. For example, juvenile brave new worlders are traumatised with electric shocks as part of the behaviorist-inspired conditioning process in childhood. Toddlers

from the lower orders are terrorised with loud noises. This sort of aversion-therapy serves to condition them against liking books. We are told the inhabitants of Brave New World are happy. Yet they periodically experience unpleasant thoughts, feelings and emotions. They just banish them with soma: "One cubic centimetre cures ten gloomy sentiments".

Even then, none of the utopians of any caste come across as very happy. This seems all too credible: more-or-less chronic happiness *sounds* so uninteresting that it's easy to believe it must *feel* uninteresting too. For sure, the utopians are mostly docile and contented. Yet their emotions have been deliberately blunted and repressed. Life is nice - but somehow a bit flat. In the words of the Resident Controller of Western Europe: "No pains have been spared to make your lives emotionally easy - to preserve you, as far as that is possible, from having emotions at all."

A more ambitious target would be to make the world's last unpleasant experience a precisely dateable event; and from this minimum hedonic baseline, start aiming higher. "Every day, and in every way, I am getting better and better". Coué's mantra of therapeutic self-deception needn't depend on the cultivation of beautiful thoughts. If harnessed to the synthesis of smarter mood-enrichers and genetically-enhanced brains, it might even come true.

Of course, it's easy today to write (mood-congruent) tomes on how everything could go wrong. This review essay is an exploration of what it might be like if they go right. So it's worth contrasting the attributes of Brave New World with the sorts of biological paradise that *may* be enjoyed by our ecstatic descendants.

S t a s i s

Brave New World is a benevolent dictatorship: a static, efficient, totalitarian welfare-state. There is no war, poverty or crime. Society is stratified by genetically-predestined caste. Intellectually superior Alphas are the top-dogs. Servile, purposely brain-damaged Gammas, Deltas and Epsilons toil away at the bottom. The lower orders are necessary in BNW because Alphas - even soma-fuelled Alphas - could allegedly never be happy doing menial jobs. It is not explained why doing menial work is inconsistent - if you're an Alpha - with a life pharmacological hedonism - nor, for that matter, with genetically-precoded wetware of invincible bliss. In any case, our descendants are likely to automate menial drudgery out of existence; that's what robots are for.

Notionally, BNW is set in the year 632 AF (After Ford). Its biotechnology is highly advanced. Yet the society itself has no historical dynamic: "History is bunk". It is curious to find a utopia where knowledge of the past is banned by the Controllers to prevent invidious comparisons. One might imagine history lessons would be encouraged instead. They would uncover a blood-stained horror-story.

Perhaps the Controllers fear historical awareness would stir dissatisfaction with the "utopian" present. Yet this is itself revealing. For Brave New World is not an exciting place to live in. It is a sterile, productivist utopia geared to the consumption of mass-produced goods: "Ending is better than mending". Society is shaped by a single all-embracing political ideology. The motto of the world state is "Community, Identity, Stability."

In Brave New World, there is no depth of feeling, no ferment of ideas, and no artistic creativity. Individuality is suppressed. Intellectual excitement and discovery have been

abolished. Its inhabitants are laboratory-grown clones, bottled and standardised from the hatchery. They are conditioned and indoctrinated, and even brainwashed in their sleep. The utopians are never educated to prize thinking for themselves. In Brave New World, the twin goals of happiness and stability - both social and personal - are not just prized, but effectively *equated*.

This surprisingly common notion is ill-conceived. The impregnable well-being of our transhuman descendants is more likely to promote greater diversity, both personal and societal, not stagnation. This is because greater happiness, and in particular enhanced dopamine function, doesn't merely extend the depth of one's motivation to act: the hyper-dopaminergic sense of *things to be done*. It also broadens the *range* of stimuli an organism finds rewarding. By expanding the range of potential activities we enjoy, enhanced dopamine function will ensure we will be *less* likely to get stuck in a depressive rut. This rut leads to the kind of learned helplessness that says nothing will do any good, nature will take its revenge, and utopias will always go wrong.

In Brave New World, things do occasionally go wrong. But more to the point, we are led to feel the whole social enterprise that BNW represents is horribly misconceived from the outset. In BNW, nothing much really changes. It is an alien world, but scarcely a rich or inexhaustibly diverse one. Tellingly, the monotony of its pleasures mirrors the poverty of our own imaginations in conceiving of radically different ways to be happy. Today, we've barely even begun to conceptualise the range of things it's possible to be happy about. For our brains aren't blessed with the neurochemical substrates to do so. Time spent counting one's blessings is rarely good for one's genes.

BNW is often taken as a pessimistic warning of the dangers of runaway science and technology. Scientific progress, however, was apparently frozen with the advent of a

world state. Thus, ironically, it's not perverse to interpret BNW as a warning of what happens when scientific inquiry is suppressed. One of the reasons why many relatively robust optimists - including some dopamine-driven transhumanists - dislike Brave New World, and accordingly distrust the prospect of universal happiness it symbolises, is that their primary source of everyday aversive experience is boredom. BNW comes across as a stagnant civilisation. It's got immovably stuck in a severely sub-optimal state. Its inhabitants are too contented living in their rut to extricate themselves and progress to higher things. Superficially, yes, Brave New World is a technocratic society. Yet the free flow of ideas and criticism central to science is absent. Moreover, the humanities have withered too. Subversive works of literature are banned. Subtly but inexorably, BNW enforces conformity in innumerable different ways. Its conformism feeds the popular misconception that a lifetime of happiness will [somehow] be boring - even when the biochemical substrates of boredom have vanished.

Controller Mustapha Mond himself obliquely acknowledges the *dys*topian sterility of BNW when he reflects on Bernard's tearful plea not to be exiled to Iceland: "One would think he was going to have his throat cut. Whereas, if he had the smallest sense, he'd understand that his punishment is really a reward. He's being sent to an island. That's to say, he's being sent to a place where he'll meet the most interesting set of men and women to be found anywhere in the world. All the people who, for one reason or another, have got too self-consciously individual to fit into community life. All the people who aren't satisfied with orthodoxy, who've got independent ideas of their own. Everyone, in a word, who's anyone..."

Admittedly, Huxley's BNW enforces a much more benign conformism than Orwell's terrifying *1984*. There's no Room 101, no torture, and no war. Early child-rearing practices aside, it's not a study of *physically* violent totalitarianism. Its riot-police use

soma-vaporisers, not tear-gas and truncheons. Yet its society is as dominated by caste as any historical Eastern despotism. BNW recapitulates all Heaven's hierarchies (recall all those angels, archangels, seraphim, etc) and few of its promised pleasures. Its satirical grotesqueries and fundamental joylessness are far more memorably captured than its delights - with one pregnant exception, soma.

Unlike the residents of Heaven, BNW's inhabitants don't worship God. Instead, they are brainwashed into revering a scarcely less abstract and remote community. Formally, the community is presided over by the spirit of the apostle of mass-production, Henry Ford. He is worshipped as a god: Alphas and Betas attend soma-consecrated "solidarity services" which culminate in an orgy. But history has been abolished, salvation has already occurred, and the utopians aren't going anywhere.

By contrast, one factor of life spent with even mildly euphoric hypomanic people is pretty constant. The tempo of life, the flow of ideas, and the drama of events speed up. In a Post-Darwinian Era of universal life-long bliss, the possibility of stasis is remote; in fact, one can't rule out an ethos of permanent revolution. But however great the intellectual ferment of ecstatic existence, the nastiness of Darwinian life will have passed into oblivion with the molecular machinery that sustained it.

I m b e c i l i t y

Some drugs dull, stupefy and sedate. Others sharpen, animate and intensify.

After taking soma, one can apparently drift pleasantly off to sleep. Bernard Marx, for instance, takes four tablets of soma to pass away a long plane journey to the Reservation

in New Mexico. When they arrive at the Reservation, Bernard's companion, Lenina, swallows half a gramme of soma when she begins to tire of the Warden's lecture, "with the result that she could now sit, serenely not listening, thinking of nothing at all". Such a response suggests the user's sensibilities are numbed rather than heightened. In BNW, people resort to soma when they feel depressed, angry or have intrusive negative thoughts. They take it because their lives, like society itself, are empty of spirituality or higher meaning. Soma keeps the population comfortable with their lot.

Soma also shows physiological tolerance. Linda, the Savage's mother, takes too much: up to twenty grammes a day. Taken in excess, soma acts as a respiratory depressant. Linda eventually dies of an overdose. This again suggests that Huxley models soma more on opiates than the sort of clinically valuable mood-brightener which subverts the hedonic treadmill of negative feedback mechanisms in the CNS. The parallel to be drawn with opiates is admittedly far from exact. Unlike soma, good old-fashioned heroin is bad news for your sex life. But like soma, it won't sharpen your wits.

Even today, the idea that chemically-driven happiness must dull and pacify is demonstrably false. Mood-boosting psychostimulants are likely to heighten awareness. They increase self-assertiveness. On some indices, and in low doses, stimulants can improve intellectual performance. Combat-troops on both sides in World War Two, for instance, were regularly given amphetamines. This didn't make them nicer or gentler or dumber. Dopaminergic power-drugs tend to increase willpower, wakefulness and action. "Serenics", by contrast, *have* been researched by the military and the pharmaceutical industry. They may indeed exert a quiescent effect - ideally on the enemy. But variants could also be used on, or by, one's own troops to induce fearlessness.

A second and less warlike corrective to the dumb-and-docile stereotype is provided by so-called manic-depressives. One reason that many victims of bipolar disorder, notably those who experience the euphoric sub-type of (hypo-)mania, skip out on their lithium is that, when "euthymic", they can still partially recall just how wonderfully *intense* and euphoric life can be in its manic phase. Life on lithium is flatter. For it's the havoc wrought on the lives of others which makes the *uncontrolled* exuberance of frank euphoric mania so disastrous. Depressed or nominally euthymic people are easier for the authorities to control than exuberant life-lovers.

Thus one of the tasks facing a mature fusion of biological psychiatry and psychogenetic medicine will be to deliver enriched well-being and lucid intelligence to anyone who wants it *without* running the risk of triggering ungovernable mania. MDMA (Ecstasy) briefly offers a glimpse of what full-blooded mental health might be like. Like soma, it induces both happiness and serenity. Unlike soma, it is neurotoxic. But used sparingly, it can also be profound, empathetic and soulfully intense.

Drugs which commonly induce *dys*phoria, on the other hand, are truly sinister instruments of social control. They are far more likely to induce the "infantile decorum" demanded of BNW utopians than euphoriants. The major tranquillisers, including the archetypal "chemical cosh" chlorpromazine (Largactil), subdue their victims by acting as dopamine antagonists. At high dosages, willpower is blunted, affect is flattened, and mood is typically depressed. The subject becomes sedated. Intellectual acuity is dulled. They are a widely-used tool in some penal systems.

A m o r a l i t y

Soma doesn't merely stupefy. At face value, the happiness it offers is amoral; it's "hedonistic" in the baser sense. Soma-fuelled highs aren't a function of the well-being of others. A synthetic high doesn't force you to be happy for a *reason*: unlike people, a good drug will never let you down. True, soma-consumption doesn't actively promote anti-social behaviour. Yet the drug is all about instant gratification.

Drug-naïve John the Savage, by contrast, has a firm code of conduct. His happiness - and sorrows - don't derive from taking a soul-corrupting chemical. His emotional responses are apparently based on reasons - though these reasons themselves presumably have a neurochemical basis. Justified or unjustified, his happiness, like our own today, will always be vulnerable to disappointment. Huxley clearly feels that if a loved one dies, for instance, then one will not merely grieve: it is *appropriate* that one grieves, and there is good reason to do so. It would be *wrong* not to go into mourning. A friend who said he might be sad if you died, but he wouldn't let it spoil his whole day - for instance - might strike us as quite unfeeling, if rather droll: not much of a friend at all.

By our lights, the utopians show equally poor taste. They don't ever grieve or treat each others' existence as *special*. They are conditioned to treat death as natural and even pleasant. As children, they are given sweets to eat when they go to watch the process of dying in hospital. Their greatest kick comes from taking a drug. Life on soma, together with early behavioural conditioning, leaves them oblivious to the true welfare of others. The utopians are blind to the tragedy of death; and to its pathos.

Surely this is a powerful indictment of *all* synthetic pleasures? Shouldn't we echo the Savage's denunciation of soma to the Deltas: "Don't take that horrible stuff. It's poison, it's poison...Poison to the soul as well as the body...Throw it all away, that horrible poison". Don't all chemical euphoriants rob us of our *humanity*?

Not really; or only on the most malaise-sodden conception of what it means to be human. Media stereotypes of today's crude psychopharmacy are not a reliable guide to the next few million years. It is sometimes supposed that *all* psychoactive drug-taking must inherently be egotistical. This egotism is exemplified in the contemporary world by the effects of power-drugs such as cocaine and the amphetamines, or by the warm cocoon of emotional self-sufficiency afforded by opium and its more potent analogues and derivatives. Yet drugs - not least the empathogens such as Ecstasy - and genetic engineering can in principle be customised to let us be *nicer*; to *reinforce* our idealised codes of conduct. The complex role of the "civilising neurotransmitter" serotonin, and its multiple receptor sub-types, is hugely instructive - if still poorly understood. If we genetically re-regulate its receptors, we can make ourselves kinder as well as happier.

The crucial point is that, potentially, long-acting designer-drugs needn't supplant our moral codes, but chemically predispose us to act them out in the very way we would wish. Biotechnology allows us to conquer what classical antiquity called *akrasia* [literally, "bad mixture"]. This was a Greek term for the character flaw of weakness of the will where an agent is unable to perform an action that s/he knows to be right. Tomorrow's "personality pills" permit us to become the kind of people we'd most like to be - to fulfill our second-order desires. Such self-reinvention is an option that our genetic constitution today frequently precludes. Altruism and self-sacrifice for the benefit of anonymous strangers - including starving Third World orphans whom we acknowledge need

resources *desperately* more than we do - is extraordinarily hard to practise consistently. Sometimes it's impossible, even for the most *benevolent*-minded of the affluent planetary elite. Self-referential altruism is easier; but it's also different - narrow and small-scale. Unfortunately, the true altruists among our (non-)ancestors got eaten or outbred. Their genes perished with them.

More specifically; in chemical terms, very crudely, dopaminergics fortify one's will-power, mu-opioids enhance one's happiness, while certain serotonergics can deepen one's empathy and social conscience. Safe, long-lasting site-specific hybrids will do both. Richer designer cocktails spiced with added ingredients will be far better still. It is tempting to conceptualise such cocktails in terms of our current knowledge of, say, oxytocin, phenylethylamine, substance P antagonists, selective *mu*-opioid agonists and enkephalinase-inhibitors etc. But this is probably naïve. Post-synaptic receptor antagonists block their psychoactive effects, suggesting it's the post-synaptic intra-cellular cascades they trigger which form the heartlands of the soul. Our inner depths haven't yet been properly explored, let alone genetically re-regulated.

But our ignorance and inertia are receding fast. Molecular neuroscience and behavioural genetics are proceeding at dizzying pace. Better Living Through Chemistry doesn't have to be just a snappy slogan. Take it seriously, and we can bootstrap our way into becoming smart and happy while biologically deepening our social conscience too. Hopefully, the need for manifestos and ideological propaganda will pass. They must be replaced by an international biomedical research program of paradise-engineering. The fun hasn't even begun. The moral urgency is immense.

It's true that morality in the contemporary sense may no longer be *needed* when suffering has been cured. The distinction between value and happiness has distinctively

moral significance only in the Darwinian Era where the fissure originated. Here, in the short-run, good feelings and good conduct may conflict. Gratifying one's immediate impulses sometimes leads to heartache in the longer term, both to oneself and others. When suffering has been eliminated, however, specifically *moral* codes of conduct become redundant. On any utilitarian analysis, at least, acts of immorality become impossible. The values of our descendants will be predicated on immense emotional well-being, but they won't necessarily be focused on it; happiness may have become part of the innate texture of sentient existence.

In Brave New World, by contrast, unpleasantness *hasn't* been eradicated. That's one reason its citizens' behaviour is so shocking, and one reason they take soma. BNW's outright *im*morality is all too conceivable by the reader.

Typically, we are indignant when we see the callous way in which John the Savage is treated, or when we witness the revulsion provoked in the Director by the sight of John's ageing mother - the companion he had himself long ago abandoned for dead after an ill-fated trip to the Reservation. Above and beyond this, all sorts of sour undercurrents are endemic to the society as a whole. Bernard is chronically discontented, even "melancholic". The Alpha misfits in Iceland are condemned to a bleak exile. Feely-author Helmholtz Watson is frustrated by a sense that he is capable of greater things than authoring repetitive propaganda. The Director of Hatcheries is utterly *humiliated* by the understandably aggrieved Bernard. Boastful Bernard is himself reduced to tears of despair when the Savage refuses to be paraded in front of assorted dignitaries and the Arch-Community-Songster of Canterbury. Lesser problems and unpleasantnesses are commonplace. And appallingly, the utopians come to gawp at John in his hermit's exile and watch his suffering *for fun*.

Brave New World is a patently sub-standard utopia in need of some true moral imagination - and indignation - to sort it out.

False Happiness

Huxley implies that by abolishing nastiness and mental pain, the brave new worlders have gotten rid of the most profound and sublime experiences that life can offer as well. Most notably, they have sacrificed a mysterious deeper happiness which is implied, but not stated, to be pharmacologically inaccessible to the utopians. The metaphysical basis of this presumption is obscure.

There are hints, too, that some of the utopians may feel an ill-defined sense of dissatisfaction, an intermittent sense that their lives are meaningless. It is implied, further, that if we are to find true fulfilment and meaning in our own lives, then we must be able to contrast the good parts of life with the bad parts, to feel both joy and despair. As rationalisations go, it's a good one.

But it's still wrong-headed. If pressed, we must concede that the victims of chronic depression or pain today don't need interludes of happiness or anaesthesia to know they are suffering horribly. Moreover, if the mere relativity of pain and pleasure *were* true, then one might imagine that pseudo-memories in the form of neurochemical artefacts imbued with the texture of "pastness" would do the job of contrast just as well as raw nastiness. The neurochemical signatures of *deja vu* and *jamais vu* provide us with clues on how the re-engineering could be done. But this sort of stratagem isn't on Huxley's agenda. The clear implication of Brave New World is that *any* kind of drug-delivered

happiness is "false" or inauthentic. In similar fashion, *all* forms of human genetic engineering and overt behavioural conditioning are to be tarred with the same brush. Conversely, the natural happiness of the handsome, blond-haired, blue-eyed Savage on the Reservation is portrayed as more real and authentic, albeit transient and sometimes interspersed with sorrow.

The contrast between true and false happiness, however, is itself problematic. Even if the notion is both intelligible and potentially referential, it's not clear that "natural", selfish-DNA-sculpted minds offer a more authentic consciousness than precision-engineered euphoria. Highly selective and site-specific designer drugs [and, ultimately, genetic engineering] won't make things seem weird or alien. On the contrary, they can deliver a *greater* sense of realism, verisimilitude and emotional depth to raw states of biochemical bliss than today's parochial conception of Real Life. Future generations will "re-encephalise" emotion to serve *us*, sentient genetic vehicles, rather than selfish DNA. Our well-being will feel utterly natural; and in common with most things in the natural world, it will be so.

If desired, too, designer drugs can be used to trigger paroxysms of spiritual enlightenment - or at least the phenomenology thereof - transcending the ecstasies of the holiest mystic or the hyper-religiosity of a temporal-lobe epileptic. So future psychoactives needn't yield only the ersatz happiness of a brave new worlder, nor will euphoriant abuse be followed by the proverbial Dark Night Of The Soul. Just so long as neurotransmitter activation of the right sub-receptors triggers the right post-synaptic intra-cellular cascades regulated by the right alleles of the right genes in the right way indefinitely - and this is a *technical* problem with a technical solution - then we have paradise everlasting, at worst. If we want it, we can enjoy a liquid intensity of awareness

far more compelling than our mundane existence as contemporary sleepwalking *Homo sapiens*. It will be vastly more enjoyable to boot.

If sustained, such modes of consciousness can furnish a far more potent definition of reality than the psychiatric slumlands of the past. Subtly or otherwise, today's unenriched textures of consciousness express feelings of depersonalisation and derealisation. Such feelings are frequently nameless - though still all too real - because they are without proper contrast: anonymous angst-ridden modes of selfhood that, in time, will best be forgotten. "True" happiness, on the other hand, will feel totally "real". Authenticity should be a design-specification of conscious mind, not the fleeting and incidental by-product of the workings of selfish DNA.

Tomorrow's neuropharmacology, then, offers incalculably greater riches than souped-up soma. True, drugs can also deliver neurochemical wastelands of silliness and shallowness. A lot of the state-spaces currently beyond our mental horizons may be nasty or uninteresting or both. Statistically, most are probably just psychotic. But a lot aren't. Entactogens, say, [literally, to "touch within"] may eventually be as big an industry as diet pills; and what they offer by way of a capacity for self-love will be far more useful in boosting personal self-esteem.

"Entactogens", "empathogens", "entheogens" - these are fancy words. Until one is granted first-person experience of the states they open up, the phraseology invoked to get some kind of intellectual handle on Altered States may seem gobbledygook. What on earth does it all *mean*? But resort to such coinages isn't a retreat into obscurantism or mystery-mongering. It's a bid to bring some kind of order to unmapped exotica way beyond the drug-naïve imagination.

One can try to hint at the properties of even *seriously* altered states by syntactically shuffling around the lexical husks of the old order. But the kind of consciousness disclosed by these extraordinary agents provides the basis for new primitive terms in the language of a conceptual apparatus that hasn't yet been invented. Such forms of what-it's-likeness can't properly be defined or evoked within the state-specific resources of the old order. Ordinarily, they're not neurochemically accessible to us at all. Genetically, we're action-oriented hunter-gatherers, not introspective psychonauts.

So how well do we *understand* the sort of happiness Huxley indicts?

Even though we find the nature of BNW-issue "soma" as elusive as its Vedic ancestor, we think we can imagine, more or less, what taking "soma" might be like; and judge accordingly. Within limits, plain "uppers" and "downers" are intelligible to us in their effects, though even here our semantic competence is debatable - right now, it's hard to imagine what terms like "torture" and "ecstasy" really denote. When talking about drugs with (in one sense) more far-reaching effects, however, it's easy to lapse into gibbering nonsense. If one has never taken a particular drug, then one's conception of its distinctive nature derives from analogy with familiar agents, or from its behavioural effects on other people, not on the particular effects its use typically exerts on the texture of consciousness. One may be confident that other people are using the term in the same way only in virtue of their physiological similarity to oneself, not through any set of operationally defined criteria. Thus, until one has tried a drug, it's hard to understand what one is praising or condemning.

This doesn't normally restrain us. But are we *rationally* entitled to pass a judgement on *any* drug-based civilisation based on one fictional model?

No, surely not. Underground chemists and pharmaceutical companies alike are likely to synthesise all sorts of "soma" in future. Licitly or otherwise, we're going to explore what it's like; and we'll like it a lot. But to suppose that the happiness of our transhuman descendants will thereby be "false" or shallow is naïve. Post-humans are not going to get drunk and stoned. Their well-being will infuse ideas, modes of introspection, varieties of selfhood, structures of mentalese, and whole new sense modalities that haven't even been dreamt of today.

Brave New World-based soma-scenarios, by contrast, are highly conceivable. This is one reason they are so unrealistic.

T o t a l i t a r i a n

BNW is a benevolent dictatorship - or at least a benevolent oligarchy, for at its pinnacle there are ten world controllers. We get to meet its spokesman, the donnish Mustapha Mond, Resident Controller of Western Europe. Mond governs a society where all aspects of an individual's life, from conception and conveyor-belt reproduction onwards, are determined by the state. The individuality of BNW's two billion hatchlings is systematically stifled. A government bureau, the Predestinators, decides a prospective citizen's role in the hierarchy. Children are raised and conditioned by the state bureaucracy, not brought up by natural families. There are only ten thousand surnames. Value has been stripped away from the person as an individual human being; respect belongs only to society as a whole. Citizens must not fall in love, marry, or have their own kids. This would seduce their allegiance away from the community by providing a rival focus of affection. The individual's loyalty is owed to the state alone. By getting rid

of potential sources of tension and anxiety - and dispelling residual discontents with soma - the World State controls its populace no less than Big Brother.

Brave New World, then, is centred around control and manipulation. As ever, the fate of an individual depends on the interplay of nature and nurture, heredity and environment: but the utopian state apparatus controls both. Naturally, we find this control disquieting. One of our deepest fears about the prospect of tampering with our natural (i.e. selfish DNA-driven) biological endowment is that we will ourselves be controlled and manipulated by others. Huxley plays on these anxieties to devastating effect. He sows the fear that a future world state may rob us of the right to be unhappy.

It must be noted that this right is not immediately in jeopardy. Huxley, however, evidently feels that the threat of compulsory well-being is real. This is reflected in his choice of a quotation from Nicolas Berdiaeff as BNW's epigraph. "Utopias appear to be much easier to realize than one formerly believed. We currently face a question that would otherwise fill us with anguish: How to avoid their becoming definitively real?" Perhaps not all of the multiple ironies here are intended by BNW's author.

Huxley deftly coaxes us into siding with John the Savage as he defends the right to suffer illness, pain, and fear against the arguments of the indulgent Controller. The Savage claims the right to be unhappy. We sympathise. Intuitively but obscurely, he shouldn't have to *suffer* enforced bliss. We may claim, like the Savage, "the right to grow old and ugly and impotent; the right to have syphilis and cancer; the right to have too little to eat; the right to be lousy; the right to live in constant apprehension of what may happen tomorrow; the right to catch typhoid; the right to be tortured by unspeakable pains of every kind". Yet the argument against chemical enslavement cuts both ways. The point today - and at any other time, surely - is that we should have the right *not* to be

unhappy. And above all, when suffering becomes truly optional, we shouldn't force our toxic legacy wetware on others.

But what will be the price of all this happiness?

It's not what we might intuitively expect. Perhaps surprisingly, freedom and individuality can potentially be *enhanced* by chemically boosting personal well-being. Vulnerable and unhappy people are probably more susceptible to brainwashing - and the subtler sorts of mind-control - than active citizens who are happy and psychologically robust. Happiness is empowering. In real life, it is notable that mood- and resilience-enhancing drugs, such as the selective serotonin reuptake inhibitors, tend to reduce submissiveness and subordinate behaviour. Rats and monkeys on SSRIs climb the pecking order, or transcend it altogether. They don't seem to try to dominate their fellows - loosely speaking, they just stop letting themselves be messed around. If pharmacologically and genetically enriched, we may all aspire to act likewise.

Admittedly, this argument isn't decisive. It's a huge topic. Humans, a philosopher once observed, are not rats. Properly-controlled studies of altered serotonin function in humans are lacking. The intra-cellular consequences of fifteen-plus serotonin receptor sub-types defy facile explanation. But we do know that a dysfunctional serotonin system is correlated with low social-status. Enhancing serotonin function - other things being equal - is likely to leave an individual *less* likely to submit to authority, not docile and emasculated. *Brave New World* is exquisite satire, but the utopia it imagines is sociologically and biologically implausible. Its happy conformists are shallow cartoons.

Of course, *any* analysis of the state's role in future millennia is hugely speculative. Both minimalist "night-watchman" states and extreme totalitarian scenarios are conceivable.

In some respects, any future world government may indeed be far more intrusive than the typical nation-state today. If the ageing process and the inevitability of death is superseded, for instance, then decisions about reproduction - on Earth, at least - simply cannot be left to the discretion of individual couples alone. This is because we'd soon be left with standing room only. The imminence of widespread human cloning, too, makes increased regulation and accountability inevitable - quite disturbingly so. But challenges like population-control shouldn't overshadow the fact that members of a happy, confident, psychologically robust citizenry are far less likely to be the malleable pawns of a ruling elite than contented fatalists. A chemically-enslaved underclass of happy helots remains unlikely.

A n t h r o p o c e n t r i c

Brave New World is a utopia conceived on the basis of species-self-interest masquerading as a universal paradise. Most of the inhabitants of our planet don't get a look-in, any more than they do today.

Strong words? Not really. Statistically, most of the suffering in the contemporary world isn't undergone by human beings. It is sometimes supposed that intensity and degree of consciousness - between if not within species - is inseparably bound up with intelligence. Accordingly, humans are prone to credit themselves with a "higher" consciousness than members of other taxa, as well as - sometimes more justifiably - sharper intellects. Non-human animals aren't treated as morally and functionally akin to human infants and toddlers, i.e. in need of looking after. Instead, they are wantonly abused, exploited, and killed.

Yet it is a striking fact that our most primitive experiences - both phylogenetically and ontogenetically - are also the most vivid. For physical suffering probably has more to do with the number and synaptic density of pain cells than a hypertrophied neocortex. The extremes of pain and thirst, for example, are excruciatingly *intense*. By contrast, the kinds of experience most associated with the acme of human intellectual endeavour, namely thought-episodes in the pre-frontal region of the brain, are phenomenologically so anaemic that it is hard to introspect their properties at all.

Hardcore paradise-engineering - and not the brittle parody of paradise served up in BNW - will eradicate such nastiness from the living world altogether. None of Huxley's implicit criticism of the utopians can conceivably apply to the rest of the animal kingdom. For by no stretch of the imagination could the most ardent misery-monger claim animal suffering is essential for the production of great art and literature - a common rationale for its preservation and alleged redeeming value in humans. Nor would its loss lead to great spiritual emptiness. Animal suffering is just savage, empty and pointless. So we'll probably scrap it when it becomes easy enough to do so.

Whether pain takes the form of the eternal Treblinka of our Fordist factory farms and conveyor-belt killing factories, or whether it's manifested as the cruelties of a living world still governed by natural selection, the sheer viciousness of the Darwinian Era is likely to horrify our morally saner near-descendants. A few centuries hence - the chronological details are sketchy - hordes of self-replicating nanorobots armed with retroviral vectors and the power of on-board quantum supercomputers may hunt out the biomolecular signature of aversive experience all the way down the phylogenetic tree; and genetically eliminate it. Meanwhile, depot-contraception, not merciless predation, will control population in our wildlife parks. Carnivorous killing-machines - and that includes dear

misunderstood kitty, a beautiful sociopath - will be reprogrammed or phased out if the abolitionist project is to be complete. Down on the farm, tasty, genetically-engineered ambrosia will replace abused sentience. For paradise-engineering entails global veganism. Utopia cannot be built on top of an ecosystem of pain and fear. Unfortunately, this is an issue on which *Brave New World* is silent.

How is it possible to make such predictions with any confidence?

Properly speaking, one can't, or at least not without a heap of caveats. But as science progressively gives us the power to remould matter and energy to suit our desires - or whims - it would take an extraordinary degree of *malice* for us to sustain the painfulness of Darwinian life indefinitely. For as our power increases, so does our complicity in its persistence.

Even unregenerate humans don't tend to be sustainably ill-natured. So when genetically-engineered vat-food tastes as good as dead meat, we may muster enough moral courage to bring the animal holocaust to an end.

C a s t e - b o u n d

In BNW, genetic engineering *isn't* used straightforwardly to pre-code happiness. Instead, it underwrites the subordination and inferiority of the lower orders. In essence, Brave New World is a global caste society. Social stratification is institutionalised in a five-way genetic split. There is no social mobility. Alphas invariably rule, Epsilons invariably toil. Genetic differences are reinforced by systematic conditioning.

Historically, dominance and winning have been associated with good, even manically euphoric, mood; losing and submission are associated with subdued spirits and depression. Rank theory suggests that the far greater incidence of the internalised correlate of the yielding sub-routine, depression, reflects how low spirits were frequently more adaptive among group-living organisms than manic self-assertion. But in Brave New World, the correlation vanishes or is even inverted. The lower orders are *at least* as happy as the Alphas thanks to soma, childhood conditioning and their brain-damaged incapacity for original thought. Thus in sleep-lessons on class consciousness, for instance, juvenile Betas learn to love being Betas. They learn to respect Alphas who "work much harder than we do, because they're so frightfully clever." But they also learn to take pleasure in not being Gammas, Deltas, or the even more witless Epsilons. "Oh no," the hypnopedia tapes suggest, "I don't want to play with Delta children."

One might imagine that progress in automation technology would eliminate the menial, repetitive tasks so unsuitable for big-brained Alphas. But apparently this would leave the lower castes disaffected and without a role: allegedly a good reason for freezing scientific progress where it is. It might be imagined, too, that one solution here would be to stop producing oxygen-starved morons altogether. Why not stick to churning out Alphas? The Controller Mustapha Mond informs us that an all-Alpha society was once tried on an island. The result of the experiment was civil war. 19,000 of the 22,000 Alphas perished. Thus, the lower castes are needed indefinitely. The happiness that they derive from their routine-bound lives guarantees stability for society as a whole. "The optimum population", the Controller observes, "is modelled on the iceberg - eight-ninths below the waterline, one-ninth above".

There are evidently (strong!) counter-arguments and rebuttals that could be delivered against any specific variant of this scenario. But Huxley isn't interested in details. BNW is a deeply pessimistic blanket-warning against *all* forms of genetic engineering and eugenics. Shouldn't we keep the *status quo* and ban them altogether? Let's play safe. In the last analysis, Nature Knows Best.

As it stands, this argument is horribly facile. The ways in which the life sciences can be abused are certainly manifold. Bioethics deserves to become a mainstream academic discipline. But the idea that a living world organised on principles of blind genetic selfishness - the bedrock of the Darwinian Era - is inherently better than anything based on rational design is surely specious. Selfishness, whether in the technical or overlapping popular sense, is a *spectacularly* awful principle on which to base any civilisation. Sooner or later, simple means-ends-analysis, if nothing else, will dictate the use of genetic engineering to manufacture constitutionally happy mind/brains. Reams of philosophical sophistry and complication aside, that's what we're all after, obliquely and under another description or otherwise; and biotechnology is the only effective way to get it.

For despite how frequently irrational we may be in satisfying our desires, we're all slaves to the pleasure principle. No one ever leaves a well-functioning pleasure-machine because they get bored: unlike the derivative joys of food, drink and sex, the delightfulness of intra-cranial self-stimulation of the pleasure-centres shows no tolerance. Natural selection has "encephalised" emotion to disguise our dependence on the opioidergic and mesolimbic dopamine circuitry of reward. Since raw, unfocused emotion is blind and impotent, its axonal and dendritic processes have been recruited into innervating the neocortex. All our layers of cortical complexity conspire to help self-replicating DNA leave more copies of itself. Thus we fetishise all sorts of irrelevant

cerebral bric-a-brac ["intentional objects": loosely, what we're happy or upset "about"] that has come to be associated with adaptively nice and nasty experiences in our past. But the attributes of power, status and money, for instance, however *obviously* nice they seem today, aren't inherently pleasurable. They yield only a derivative kick that can be chemically edited out of existence. Their cortical representations have to be innervated by limbically-generated emotions in the right way - or the wrong way - for them to seem nice at all.

Rationally, then, if we want to modulate our happiness so that it's safe and socially sustainable, we must genetically code pre-programmed well-being in a way that shuts down the old dominance-and-submission circuits too. Such a shut-down is crudely feasible today on serotonergics, both recreational and clinical. But the shut-down can be comprehensive and permanent. Germ-line gene therapy is better than a lifetime on drugs.

Is this sort of major genetic re-write likely?

Yes, probably. A revolution in reproductive technologies is imminent. Universal pre-implantation diagnosis may eventually become the norm. But in the meantime, any unreconstructed power-trippers can get a far bigger kick in immersive VR than they can playing primate party-politics. If one wants to be Master Of The Universe, then so be it: *a chacun son gout*. The narrative software which supports such virtual worlds can even be pharmacologically enhanced in the user so that virtual world mastery is always better than The Real Thing - relegated one day, perhaps, to a fading antiquarian relic. The fusion of drugs and computer-generated worlds will yield greater verisimilitude than anything possible in recalcitrant old organic VR - the dynamic simulations which perceptual naïve realists call the world. For we live in a messy and frustrating regime

which passes itself off as The Real World, but is actually a species-specific construct coded by DNA.

OK. But can power-games really be confined exclusively to VR? Won't tomorrow's Alphas want to dominate both?

This question needs a book, not the *obiter dicta* of a literary essay. But if one can enjoy champagne, why drink meths, or even be tempted to try it in the first place? In common with non-human animals, we respond most powerfully to hot-button supernormal stimuli. Getting turned-on by the heightened verisimilitude of drugs-plus-VR from a very young age is likely to eclipse anything else on offer.

This isn't to deny that in any transitional era to a mature post-Darwinian paradise, there will have to be *huge* safeguards - no less elaborate than the multiple failsafe procedures surrounding the launch codes for today's nuclear weaponry. In the near future, for instance, prospective candidates for political leadership in The Real World will probably have their DNA profiles scrutinised no less exhaustively than their sexual peccadillos. For it will be imprudent to elect unenriched primitives endowed with potentially dangerous genotypes. If one is going to put oneself and one's children into, say, ecstasy-like states of loving empathy and trust, then one is potentially more vulnerable to genetic cavemen. But this is all the more reason to design beautifully enhanced analogues of ecstasy and coke which fuse the best features of both.

Even if a power-tripper's fantasy wish-fulfilment is confined to private universes, we are still likely to view it as an unnerving prospect. One of the reasons we find the very thought of being dominated and controlled and manipulated *à la* BNW so aversive is that we associate such images with frustration, nastiness and depression. For sure, the brave new worlders are typically happy rather than depressed. Yet they are all, bar perhaps the

Controllers, manipulated dupes. The worry that we ourselves might ever suffer a similar fate is unsettling and *depressing*. Brave New World gives happiness a bad name.

But it's misery that deserves to be stigmatised and stamped out. Brave New World dignifies unpleasantness in the guise of noble savagery just when it's poised to become biologically optional. And on occasion unpleasantness really can be *horrific* - too bad to describe in words. Some forms of extreme pain, for instance, are so terrible to experience that one would sacrifice the whole world to get rid of the agony. Pain just this bad is happening in the living world right now. It's misguided to ask whether such pain is *really* as bad as it seems to be - because the reality is the very appearance one is trying vainly to describe. The extremes of so-called "mental" pain can be no less dreadful. They may embody suicidal despair far beyond everyday ill-spirits. They are happening right now in the living world as well. Their existence reflects the way our mind/brains are built. Unless the vertebrate central nervous system is genetically recoded, there will be traumas and malaise in utopia - *any* utopia - too.

No behavioural account of even moderately severe depression, for instance, can do justice to its subjective awfulness. But a spectrum of depressive signs and symptoms will persist within even a latter-day Garden of Eden - in the absence of good drugs and better genes. We can understand why depressive states evolved among social animals in terms of the selective advantage of depressive *behaviour* in reinforcing adaptive patterns of dominance and subordination, avoiding damaging physical fights with superior rivals, or of inducing hypercholinergic frenzy of reflective thought when life goes badly wrong - for one's genes. Likewise, intense and unpleasant social anxiety was sometimes adaptive too. So was an involuntary capacity for the torments of sexual jealousy, fear, terror, hunger, thirst and disgust. Our notions of dominance and subordination are embedded

within this stew of emotions. They are clearly quite fundamental to our social relationships. They pervade our whole conceptual scheme. When we try to imagine the distant future, we may, of course, imagine hi-tech gee-whizzery. Yet emotionally, we also think in primitive terms of dominance and submission, of hierarchy and power structures, superiority and inferiority. Even when we imagine future computers and robots, we are liable to have simple-minded fantasies about being used, dominated, and overthrown. Bug-eyed extra-terrestrials from the Planet Zog, too, and their legion of hydra-headed sci-fi cousins, are implicitly assumed to have the motivational structure of our vertebrate ancestors. Superficially, they may be alien - all those tentacles - but really they're just like *us*. Surely they'll want to dominate us, control us, invade Earth, etc? Huxley's vision of control and manipulation is (somewhat) subtler; but it belongs to the same atavistic tradition.

For the foreseeable future, these concerns *aren't* idle. We may rightly worry that if some of us - perhaps most of us - are destined to get drugged-up, genetically-rewritten and plugged into designer worlds, then might not invisible puppet-masters be controlling us for their own ends, whatever their motives? Who'll be in charge of the basement infrastructure which sustains all the multiple layers of VR - and thus ultimately running the show? *Quis custodiet ipsos custodes?* as we say here in Brighton.

Admittedly, sophisticated and intellectually enriched post-humans are unlikely to be naïve realists about "perception"; so they'll recognise that what their ancestors called "real life" was no more privileged than what we might call, say, "the medieval world" - the virtual worlds instantiated by our medieval forebears. But any unenriched primitives still living in organic VR could still be potentially dangerous, because they could bring everything else tumbling down. In certain limited respects, their virtual worlds, like our

own, would causally co-vary with the mind-independent world in ways that blissed-up total-VR dwellers would typically lack. So can it *ever* be safe to be totally nice and totally happy?

These topics deserve a book - many books - too. The fixations they express are doubtless still of extreme interest to contemporary humans. Sado-masochistic images of domination-and-submission loom large in a lot of our fantasies too. The categories of experience they reflect were of potent significance on the African savannah, where they bore on the ability to get the "best" mates and leave most copies of one's genes. But they won't persist for ever. A tendency to such dominance-and-control syndromes is going to be written out of the genome - as soon we gain mastery of rewriting the script. For on the whole, we want our kids to be nice.

More generally, the whole "evolutionary environment of adaptation" is poised for a revolution. This is important. When any particular suite of alleles ceases to be the result of random mutation and blind natural selection, and is instead pre-selected by intelligent agents *in conscious anticipation* of their likely effects, then the criteria of genetic fitness will change too. The sociobiological and popular senses of "selfish" will progressively diverge rather than typically overlap. Allegedly "immutable" human nature will change as well when the genetic-rewrite gathers momentum and the Reproductive Revolution matures. The classical Darwinian Era is drawing to a close.

Unfortunately, its death agonies may be prolonged. Knee-jerk pessimism and outright cynicism abound among humanistic pundits in the press. They are common in literary academia. And, of course, any competent doom-monger can glibly extrapolate the trends of the past into the future. Yet anti-utopianism ignores even the *foreseeable* discontinuities that lie ahead of us as we mature into post-humans. Most notably, it

ignores the major evolutionary transition now imminent in the future of life. This is the era when we rewrite the genome in our own interest to make ourselves happy in the richest sense of the term. In the meantime, we just act out variations on dramas scripted by selfish DNA.

P h i l i s t i n e

Brave New World is a stupid society. For the most part, even the Alphas don't do anything more exalted than play Obstacle Golf. A handful of the Alphas are well-delineated: Bernard, Helmholtz, and Mustapha Mond. They are truly clever. Huxley was far too brilliant to write a novel with convincingly dim-witted lead characters. The Savage, in particular, is an implausibly articulate vehicle for Huxley's own sympathies. But in the main, brave new worlders are empty-headed mental invalids in the grip of terminal mind-rot - happy pigs rather than types of unhappy Socrates.

Since the utopians are (largely) contented with their lives, they don't produce Great Art. Happiness and Great Art are allegedly incompatible. Great Art and Great Literature are very dear to Huxley's heart. But is artistic genius really stifled without inner torment? Is paradise strictly for low-brows?

There is a great deal of ideological baggage that needs to be picked apart here; or preferably slashed like a Gordian knot. The existence of great art, unlike (controversially) great science, is not a state-neutral fact about the world. Not least, "great art" depends on the resonances it strikes in its audience. Today we're stuck with legacy wetware and genetically-driven malaise. It's frequently nasty and sometimes terrible. So we can

currently appreciate only too well "great" novels and plays about murder, violence, treachery, child abuse, suicidal despair etc. Such themes, especially when "well"-handled in classy prose, strike us as more "authentic" than happy pap. Thus a (decaying) Oxbridge literary intelligentsia can celebrate, say, the wonderful cathartic experience offered by Greek tragedies - with their everyday tales of bestiality, cannibalism, rape and murder among the Greek gods. It's good to have one's baser appetites dressed up so intelligently.

Yet after the ecstatic phase-change ahead in our affective states - the most *important* evolutionary transition in the future of life itself - the classical literary canon may fall into obscurity. Enriched minds with different emotions encephalised in different ways are unlikely to be edified by the cultural artefacts of a bygone era. Conversely, we might ourselves take a jaundiced view if we could inspect the artistic products of a civilisation of native-born ecstatics. This is because any future art which explores lives predicated on gradations of delight will seem pretty vapid from here. We find it hard enough to imagine even *one* flavour of sublimity, let alone a multitude.

The nagging question may persist: will posterity's Art and Literature [or art-forms expressing modes of experience we haven't even accessed yet] *really* be Great? To its creators, sure, their handiwork may seem brilliant and beautiful, moving and profound. But might not its blissed-out authors be simply conning themselves? Could they have lost true critical insight, even if they retain its shadowy functional analogues?

Such questions demand a treatise on the nature and objectivity of value judgements. Yet perhaps asking whether we would appreciate ecstatic art of 500 or 5000 years hence is futile in the first place. We simply can't know what we're talking about. For we are *un*happy pigs, and our own arts are mood-congruent perversions. The real philistinism to

worry about lies in the emotional illiteracy of the present. Our genetically-enriched posterity will have no need of our condescension.

Things Go Wrong

Even by its own criteria, BNW is *not* a society where everyone is happy. There are asylums in Iceland and the Falklands for Alpha-male misfits. Bernard Marx is disaffected and emotionally insecure; a mistake in the bottling-plant left him stunted. Lenina has lupus. If you run out of soma, a fate which befalls Lenina when visiting the Reservation, you feel sick: well-being is not truly genetically pre-programmed. Almost every page of the novel is steeped in negative vocabulary. Its idiom belongs to the era it has notionally superseded. On a global scale, the whole *society* of the world state is an abomination - science gone mad - in most people's eyes, at any rate. In Brave New World Revisited, Huxley clearly expects us to share his repugnance.

Surely *any* utopia can go terribly wrong? One thinks of Christianity; the Soviet experiment; The French Revolution; and Pol Pot. All ideas and ideals get horribly perverted by power and its pursuit. So what horrors might we be letting ourselves in for in a global species-project to abolish the biological substrates of malaise?

There is an important distinction to be drawn here. In a future civilisation where aversive experience is genetically impossible - forbidden not by social *diktat* but because its biochemical substrates are absent - then the notion of what it *means* for anything to go wrong will be different from today. If this innovative usage is to be adopted, then we're dealing with a separate and currently ill-defined - if not mystical - concept; and we run a

risk of conflating the two senses. For if we are incapable of aversive experience, then the notion of *things going wrong* with our lives - or anyone else's - doesn't apply in any but a Pickwickian sense. "Going wrong" and "being terrible" as we understand such concepts today are inseparable from the textures of nastiness in which they had their origin. Their simple transposition to the Post-Darwinian Era doesn't work.

Perhaps functional *analogues* of things going wrong will indeed apply - even in a secular biological heaven where the phenomenology of nastiness has been wiped out. So the idea isn't entirely fanciful. For the foreseeable future, functional analogues of phenomenal pain will be needed in early transhumans no less than in silicon robots to alert their bodies to noxious tissue damage, etc. Also, functional analogues of "things going wrong", at least in one sense, are needed to produce great science and technology, so that acuity of critical judgement is maintained; uncontrolled euphoric mania is not a recipe for scientific genius in even the most high-octane supermind. Yet directly or indirectly, the very notion of "going wrong" in the contemporary sense seems bound up with a distinctive and unpleasant phenomenology of consciousness: a deficiency of well-being, not a surfeit.

This doesn't stop us today from dreaming up scenarios of blissed-out utopias which strike us as distasteful - or even nightmarish - when contemplated through the lens of our own darkened minds. This is because chemically-unenriched consciousness is a medium which corrupts anything that it seeks to express. The medium is not the message; but it leaves its signature indelibly upon it. We may *imagine* future worlds in which there is no great art, no real spirituality, no true humanity, no personal growth through life-enriching traumas and tragedies, etc. We may conjure up notional future worlds, too, whose belief-systems rest on a false metaphysic: e.g. an ideal theocracy - is it a real utopia if it

transpires there's no God? But it's hard to escape the conclusion that "ill-effects" from which no one ever *suffers* are ontological flights of fancy. The spectre of happy *dys*topias may trouble some of us today rather than strike us as a contradiction on terms. But like Huxley's *Brave New World*, they are fantasies born of the very pathology that they to seek warn us against.

This is not to deny that the *transition* to the new Post-Darwinian Era will be stressful and conflict-ridden. We learn from the Controller that the same was true of Brave New World - civilisation as we know it today was destroyed in the Nine Years' War. One hopes, on rather limited evidence, that the birth-pangs of the new genetic order will be less traumatic. But the supposition that a society predicated on universal bliss engineered by science is *inherently* wrong - as Huxley wants us to believe - rests on obscure metaphysics as well as questionable ethics. Sin is a concept best left to medieval theologians.

C o n s u m e r i s t

Brave New World is a "Fordist" utopia based on production and consumption. It would seem, nonetheless, that there is no mandatory work-place drug-testing for soma; if there were, its detection would presumably be encouraged. In our own society, taking drugs may compromise a person's work-role. Procuring illicit drugs may divert the user from an orthodox consumer lifestyle. This is because the immediate rewards to be gained from even trashy recreational euphoriants are more intense than the buzz derived from acquiring more consumer fripperies. In BNW, however, the production and consumption of manufactured goods is (somehow) harmoniously integrated with a lifestyle of drugs

and sex. Its inhabitants are given no time for spiritual contemplation. Solitude is discouraged. The utopians are purposely kept occupied and focused on working for yet more consumption: "No leisure from pleasure".

Is this our destiny too?

Almost certainly not. Productivist visions of paradise are unrealistic if they don't incorporate an all-important genomic revolution in hedonic engineering. Beyond a bare subsistence minimum, there is no inherent positive long-term correlation between wealth and happiness. Windfalls and spending-sprees do typically bring short-term highs. Yet they don't subvert the hedonic treadmill of inhibitory feedback mechanisms in the brain. Each of us tends to have a hedonic set-point about which our "well"-being fluctuates. That set-point is hard to recalibrate over a lifetime without pharmacological or genetic intervention. Interlocking neurotransmitter systems in the CNS have been selected to embody both short- and long-term negative feedback loops. They are usually efficient. Unless they are chemically subverted, such mechanisms stop most of us from being contented - or clinically depressed - for very long. The endless cycle of ups and downs - our own private re-enactment of the myth of Sisyphus - is an "adaptation" that helps selfish genes leave more copies of themselves; in nature, alas, the restless malcontents genetically out-compete happy lotus-eaters. It's an adaptation that won't go away just by messing around with our external environment.

This is in no way to deny the distinct possibility that our descendants will be temperamentally ecstatic. They may well consume lots of material goods too - if they don't spend their whole lives in fantasy VR. Yet their well-being cannot derive from an unbridled orgy of personal consumption. Authentic mental health depends on dismantling

the hedonic treadmill itself; or more strictly, recalibrating its axis to endow its bearers with a motivational system based on gradients of immense well-being.

So what sort of scenario can we expect? If we opt for gradations of genetically pre-programmed bliss, just what, if anything, is our marvellous well-being likely to focus on?

First, in a mature IT society, the harnessing of psychopharmacology and biotechnology to ubiquitous virtual reality software gives scope for *unlimited* good experiences for everyone. Any sensory experience one wants, any experiential manifold one can imagine, any narrative structure one desires, can be far better realised in VR than in outmoded conceptions of Real Life.

At present, society is based on the assumption that goods and services - and the good experiences they can generate - are a finite scarce resource. But ubiquitous VR can generate (in effect) infinite abundance. An IT society supersedes the old zero-sum paradigm and Fordist mass-manufacture. It rewrites the orthodox laws of market economics. The ability of immersive multi-modal VR to make one - depending on the software title one opts for - Lord Of Creation, Casanova The Insatiable, etc, puts an entire universe at one's disposal. This can involve owning "trillions of dollars", heaps of "status-goods", and unlimited wealth and resources - in today's archaic terminology. In fact, one will be able to have all the material goods one wants, and any virtual world one wants - and it can all seem as "unvirtual" as one desires. A few centuries hence, we may rapidly take [im]material opulence for granted. And this virtual cornucopia won't be the prerogative of a tiny elite. Information isn't like that. Nor will it depend on masses of toiling workers. Information isn't like that either. If we want it, nanotechnology promises old-fashioned abundance all round, both inside and outside synthetic VR.

Nanotechnology is not magic. The self-replicating molecular robots it will spawn are probably more distant than their enthusiasts suppose, perhaps by several decades. We may have to wait a century or more before nanorobots can get to work remoulding the cosmos - to make it a home worth living in and calling our own. Details of how they'll be programmed, how they'll navigate, how they'll be powered, how they'll locate all the atoms they reconfigure, etc, are notoriously sketchy. But the fact remains: back in the boring old mind-independent world, applied nanoscience will deliver material superabundance beyond measure.

For the most part, admittedly, vast material opulence may not be needed thanks to VR. This is because we can all have the option of living in immersive designer-paradises of our own choosing. At first, our customised virtual worlds may merely ape and augment organic VR. But the classical prototype of an egocentric virtual world is parochial and horribly restrictive; the body-image it gives us to work with, for instance, is pretty shoddy and flawed by built-in obsolescence. Unprogrammed organic VR can be hatefully cruel as well - nature's genetic algorithms are nastily written and very badly coded indeed. Ultimately, artificial VR may effectively supersede its organic ancestor no less (in)completely than classical macroscopic worlds emerged from their quantum substrate. The transition is conceivable. Whether it will happen, and to what extent, we simply don't know.

Heady stuff. But is it sociologically plausible? Doesn't such prophecy just assume a naïve technological determinism? For it might be countered that synthetic drugs and VR experiences - whether interactive or solipsistic, deeply soulful or fantasy wish-fulfilment - will always be second-rate shadows of their organically-grown predecessors. Why will we want them? After a while, won't we get bored? For surely Real Life is better.

On the contrary, drugs plus VR can potentially yield a *heightened* sense of verisimilitude; and exhilarating excitement. Virtual worlds can potentially seem more real, more lifelike, more intense, and more *compelling* than the lame definitions of reality on offer today. The experience of *this-is-real* - like all our waking- or dreaming consciousness - comprises a series of neurochemical events in the CNS like any other. It can be amped-up or toned-down. Reality does not admit of degrees; but our sense of it certainly does. Tone, channel and volume controls will be at our disposal. But once we've chosen what we like, the authentic taste of paradise will indeed be addictive.

Thus, in an important sense, Brave New World is wrong. Our descendants may "consume" software, genetic enhancements and designer drugs. But the future lies in bits and bytes, not as workers engaged in factory mass-production or cast as victims of a consumer society. In some ways, BNW is prescient science fiction - uncannily prophetic of advances in genetic engineering and cloning. But in other ways, its depiction of life in centuries to come is backward-looking and quaint. Our attempts to envision distant eras always are. The future will be unrecognisably better.

L o v e l e s s

BNW is an essentially loveless society. Both romantic love and love of family are taboo. The family itself has been abolished throughout the civilised world. We learn, however, that the priggish Director of Hatcheries and Conditioning was guilty of an indiscretion with a Beta-minus when visiting the Reservation twenty years ago. When John the Savage falls on his knees and greets him as "my father", the director puts his hands over his ears. In vain, he tries to shut out the obscene word. He is *embarrassed*. Publicly

humiliated, he flees the room. Pantomime scenes like this - amusing but fanciful - contribute to our sense that a regime of universal well-being would entail our *losing* something precious. Utopian happiness, we are led to believe, is built on sacrifice: the *loss* of love, science, art and religion. Authentic paradise-engineering, by contrast, can enhance them all; not a bad payoff.

In BNW, romantic love is strongly discouraged as well. Brave new worlders are conditioned to be sexually promiscuous: "Everyone belongs to everyone else." Rather than touting the joys of sexual liberation, Huxley seeks to show how sexual promiscuity cheapens love; it doesn't express it. The Savage fancies lovely Lenina no less than she fancies him. But he *loves* her too. He feels having sex would dishonour her. So when the poor woman expresses her desire to have sex with him, she gets treated as though she were a prostitute.

Thus Huxley doesn't offer a sympathetic exploration of the possibility that prudery and sexual guilt has soured more lives than sex. In a true utopia, the counterparts of John and Lenina would enjoy fantastic love-making, undying mutual admiration, and live together happily ever after.

Fantastical? The misappliance of science? No. It's just one technically feasible biological option. In the light of what we do to those we love today, it would be a kinder option too. At any rate, we should be free to choose.

The utopians have no such choice. And they aren't merely personally unloved. They aren't individually respected either. Ageing has been abolished; but when the utopians die - quickly, not through a long process of senescence - their bodies are recycled as

useful sources of phosphorus. Thus, *Brave New World* is a grotesque parody of a utilitarian society in both a practical as well as a philosophical sense.

This is all good knockabout stuff. The problem is that some of it has been taken seriously.

Science is usually portrayed as dehumanising. *Brave New World* epitomises this fear. "The more we understand the world, the more it seems completely pointless" (Steven Weinberg). Certainly science can seem chilling when conceived in the abstract as a metaphysical world-picture. We may seem to find ourselves living in a universe with all the human meaning stripped out: participants in a soulless dance of molecules, or harmonics of pointlessly waggling superstrings and their braneworld cousins. Nature seems loveless and indifferent to our lives. What right have we to be happy?

Yet what right have we to sneeze? If suffering has been medically eradicated, does happiness have to be justified any more than the colour green or the taste of peppermint? Is there some deep metaphysical sense in which we *ought* to be weighed down by the momentous gravity of the human predicament?

Only if it will do anyone any good. The evidence is lacking. Paradise-engineering, by contrast, can deliver an enchanted pleasure-garden of otherworldly delights for everyone. Providentially, the appliance of biotechnology offers us the unprecedented prospect of *enhancing* our humanity - and the biological capacity for spiritual experience. When genetically-enriched, our pursuit of such delights won't be an escape from some inner sense of futility, a gnawing existential angst which disfigures so many lives at present. Quite the opposite: life will feel self-intimatingly *wonderful*. Wholesale genetic-rewrites tweaked by rational drug-design give us the chance to enhance willpower and

motivation. We'll be able to enjoy a hugely greater sense of *purpose* in our lives than our characteristically malfunctioning dopamine systems allow today. Moreover, this transformation of the living world, and eventually of the whole cosmos, into a heavenly meaning-steeped nirvana will in no way be "unnatural". It is simply a disguised consequence of the laws of physics playing themselves out.

And, conceivably, it will be a loving world. Until now, selection pressure has ensured we're cursed with a genome that leaves us mostly as callous brutes, albeit brutes with intermittently honourable intentions. We are selfish in the popular as well as the technical genetic sense. Love and affection are often strained even among friends and relatives. The quasi-psychopathic indifference we feel toward most other creatures on the planet is a by-product of selfish DNA. Sociobiology allied to evolutionary psychology shows how genetic dispositions to conflict are latent in every relationship that isn't between genetically identical clones. Such potential conflicts frequently erupt in overt form. The cost is immense suffering and sometimes suicidal anguish.

This isn't to deny that love is real. But its contemporary wellsprings have been poisoned from the outset. Only the sort of love that helps selfish DNA to leave more copies of itself - which enable it to "maximise its inclusive fitness" - can presently flourish. It is fleeting, inconstant, and shaped by cruelly arbitrary criteria of physical appearance which serve as badges of reproductive potential. If we value it, love should be rescued from the genes that have recruited and perverted the states which mediate its expression in blind pursuit of reproductive success. *Contra* Brave New World, love is not biologically inconsistent with lasting happiness.

This is because good genes and good drugs allow us, potentially, to love everyone more deeply, more empathetically and more sustainably than has ever been possible before.

Indeed, there is no fundamental biological reason why the human genome can't be rewritten to allow everyone to be "in" love with everyone else - if we should so choose. But simply loving each other will be miraculous enough; and will probably suffice. An empty religious piety can be transformed into a biological reality.

Love is versatile; so we needn't turn ourselves into celibate angels either. True love does not entail that we become disembodied souls communing with each other all day. "Promiscuous" sex doesn't have to be loveless. Bonobos ("pygmy chimps") are a case in point; they would appreciate a "Solidarity Service" rather better than we do. When sexual guilt and jealousy - a pervasive disorder of serotonin function - are cured, bed-hopping will no longer be as morally reckless as it is today. Better still, designer love-philtres and smarter sex-drugs can transform our concept of intimacy. Today's ill-educated fumblings will seem inept by comparison. Sensualists may opt for whole-body orgasms of a frequency, duration and variety that transcends the limp foreplay of their natural ancestors. Whether the sexual adventures of our descendants will be mainly auto-erotic, interpersonal, or take guises we can't currently imagine is a topic for another night.

Profound love of many forms - both of oneself and all others - is at least as feasible as the impersonal emotional wasteland occupied by Huxley's utopians.

Gene-Splicers Versus Glue-Sniffers

The molecular biology of paradise

The prospect of a lifetime of *genetically*-engineered sublimity strikes some contemporary Savages as no less appalling than getting high with drugs. The traditional conception of living happily-ever-after in Heaven probably hasn't thrilled them unduly either; but the unusual eminence of its Author has discouraged overt criticism. In any event, the consensus seems to be that God's PR representatives did a poor job in selling The Other Place to his acolytes. Today, many people find the idea of winning the national lottery far more appealing; and in fairness, it probably offers better odds. Possibly God's representatives on earth should have tried harder to make Heaven sound more appealing. One worries that an eternity spent worshipping Him might begin to pall.

But the Death Of God, or at least his discreet departure to a backstage role, shouldn't mean we're doomed to abandon any notion of heaven, and certainly not on Earth. Suffering, whether it's merely irksome or too terrible for words, doesn't have to be part of life at all.

Unfortunately, the proposal that aversive experience should be eliminated *in toto* via biotechnology tends to find itself assimilated to two stereotypes:

1. The image of an intra-cranially self-stimulating rat. Its degraded frenzy of lever-pressing is eventually followed by death from inanition and self-neglect.

2. Soma and visions of Brave New World.

And just as during much of the Twentieth Century, any plea for greater social justice could be successfully damned as Communist, likewise today, any strategy to eradicate suffering is likely to be condemned in similar reactionary terms: either wirehead hedonism or revamped Brave New World. This response is not just facile and simplistic. If it gains currency, the result is morally catastrophic.

Of course, the abolitionist issue rarely arises. Typically, universal bliss is still more-or-less unthinkingly dismissed as *technically* impossible. Insofar as the prospect is even contemplated - grudgingly - it is usually assumed that the new regime would be underwritten day-by-day with drugs or, more crudely, electrodes in the pleasure-centres.

These techniques have their uses. Yet in the medium-to-long-term, stopgaps won't be enough. *All* use of psychoactive drugs may be conceived as an attempt to correct something pathological with one's state of consciousness. There's something deeply wrong with our brains. If what we had now was OK, we wouldn't try to change it. But it isn't, so we do. Mature biological psychiatry will recognise inadequate innate bliss as a pandemic form of mental ill-health: good for selfish DNA in the ancestral environment where the adaptation arose, but bad for its throwaway vehicles, notably us. The whole gamut of behavioural conditioning, socio-economic reform, talk-therapies - and even euphoriant superdrugs - are just palliatives, not cures, for a festering global illness. Its existence demands a global eradication program, not idle philosophical manifestos and scientific *belles lettres*.

But one does one's best. The ideological obstacles to genetically pre-programmed mental super-health are actually more daunting than the technical challenges. To be cured, hypo-hedonia must be recognised as a primarily *genetic* deficiency-disorder. Designer mood-brighteners and anti-anxiety agents to alleviate it are sometimes branded

"lifestyle-drugs"; but this is to trivialise a serious medical condition which must be corrected at source. Happily, our hereditary neuropsychiatric disorder is likely to become extinct within a few generations as the Reproductive Revolution unfolds. Aversive experience, and the poisonous metabolic pathways that mediate its textures, will become physiologically impossible once the genes coding its neural substrates have been eliminated. We won't miss its corrupting effect when it's gone.

In the medium-term, the *functional* equivalent of aversive experience can help animate us instead. Late in the Third Millennium and beyond, its functional successors may be expressed as gradients of majestic well-being. On this scenario, our descendants will enjoy a civilisation based on information-bearing pleasure-gradients: whether steep or shallow, we simply don't know. Such a global species-project does not have the desperate *moral* urgency of eliminating the phenomenon of Darwinian pain - both "mental" and "physical", human and non-human alike. Abolishing raw nastiness - sometimes vile beyond belief - remains the over-riding ethical priority. One doesn't have to be an outright negative utilitarian to acknowledge that getting rid of agony takes moral precedence over maximising pleasure. But both genetic fundamentalists and gung-ho advocates of Better Living Through Chemistry today agree on one crucial issue. There is no sense in sustaining a legacy of mood-darkening metabolic pathways out of superstitious deference to our savage past.

* * *

When Bernard Marx tells the Savage he will try to secure permission for him and his mother to visit the Other Place, John is initially pleased and excited. Echoing Miranda in *The Tempest*, he exclaims: "O brave new world that has such people in it." Heavy irony. Like innocent Miranda, he is eager to embrace a way of life he neither knows nor

understands. And of course he comes unstuck. Yet if we swallow such fancy literary conceits, then ultimately the joke is on us. It is only funny in the sense there are "jokes" about Auschwitz. For it is Huxley who neither knows nor understands the glory of what lies ahead. A utopian society in which we are sublimely happy will be far better than we can presently imagine, not worse. And it is we, trapped in the emotional squalor of late-Darwinian antiquity, who neither know nor understand the lives of the God-like super-beings we are destined to become.

UTOPIAN SURGERY

Early arguments against anaesthesia in surgery, dentistry and childbirth

INTRODUCTION

Before the advent of anaesthesia, medical surgery was a terrifying prospect. Its victims could suffer indescribable agony. The utopian prospect of surgery without pain was a nameless fantasy - a notion as fanciful as the abolitionist project of life without suffering still seems today. The introduction of diethyl ether $CH_3CH_2OCH_2CH_3$ (1846) and chloroform $CHCl_3$ (1847) as general anaesthetics in surgery and delivery rooms from the mid-19th century offered patients hope of merciful relief. Surgeons were grateful as well: within a few decades, controllable anaesthesia would at last give them the chance to perform long, delicate operations. So it might be supposed that the adoption of painless surgery would have been uniformly welcomed too by theologians, moral philosophers and medical scientists alike. Yet this was not always the case. Advocates of the "healing power of pain" put up fierce if disorganised resistance.

The debate over whether to use anaesthetics in surgery, dentistry and obstetrics might now seem of merely historical interest. Yet it is worth briefly recalling some of the arguments used against the introduction of pain-free surgery raised by a minority of 19th century churchmen, laity and traditionally-minded physicians. For their objections parallel

the arguments put forward in the early 21st century against technologies for the alleviation or abolition of "emotional" pain - whether directed against the use of crude "psychic anaesthetisers" like today's SSRIs, or more paradoxically against the use of tomorrow's mood-elevating feeling-*intensifiers* i.e. so-called "empathogen-entactogens", hypothetical safe and long-acting analogues of MDMA.

It's worth recalling too that early critics of surgical and obstetric anaesthesia weren't (all) callous reactionaries or doctrinaire religious fundamentalists. Nor are all contemporary critics of the use of pharmacotherapy to treat psychological distress. The doubters, critics and advocates of caution were right to consider the potential diagnostic role of pain - and to emphasise that the risks, mechanisms and adverse side-effects of the new anaesthetic procedures were poorly understood. In Victorian Britain, around 1 in 2,500 people given chloroform anaesthesia died directly in consequence. Around 1 in 15,000 died as a direct result of being administered ether. This statistic pales beside the proportion that died from post-surgical infection; but it compares with the present-day mortality figure of 1 in around 250,000 people who die as a direct result of undergoing surgical anaesthesia in the UK. Safe and sustainable total anaesthesia that is 100% reliable - and reliably reversible - is as hard to achieve as safe and sustainable analgesia, euthymia, or euphoria. Yet the technical and ideological challenges ahead in banishing suffering from the world shouldn't detract from the moral case for its abolition.

* * *

HISTORICAL BACKGROUND

The effect of inhaling ether, chloroform, nitrous oxide and similar agents was christened by the physician-poet Oliver Wendell Holmes, Sr (1809-94). In a letter to etherization pioneer William Morton, who had solicited his opinion, Holmes coined the words "anæsthetic" and "anæsthesia" from the Greek *an* for "without" and *esthesia* for "sensibility". Holmes once remarked that if the whole pharmacopoeia of his era "were sunk to the bottom of the sea, it would be all the better for mankind, and all the worse for the fishes"; but he knew anaesthetics were a spectacular exception. Strictly speaking, the word for anaesthesia wasn't new - the Greeks themselves occasionally used the term, notably the herbalist physician and surgeon Dioscorides (c.40-c.90 AD). It had been revived on more than one occasion since: Bailey's *English Dictionary* (1724) defines anaesthesia as "a defect of sensation". But Holmes was the first to propose the term to designate the state of unconsciousness induced by gas-inhalation for painless surgery. Holmes apparently thought hard about his recommendation, and urged Morton to consult with other scholars too. For he recognised that as news of the revolution spread like wildfire across the globe, the term would be "repeated by the tongues of every civilized race of mankind."

The concept of pain-relief and even total insensibility for surgery wasn't original or unfamiliar. However, for thousands of years its reliable prospect had seemed impossibly utopian - as unrealistic as a future of lifelong bliss seems at present.

The single or combined use of stupefying agents such as ethyl alcohol, mandragora, cannabis and opium to deaden the sensibilities prior to surgery had been practised in classical antiquity. Herodotus (c.484-c.432 BC) relates how the Scythians induced stupor

by inhalation of the vapours of some kind of hemp, a remarkable if apocryphal feat in the low-THC era before superskunk.

Inhaling vapours to alter one's state of consciousness was practised too by the pythonesses of Delphi, sacred female oracles who breathed in vapours from a rock crevice in the course of their priestly duties. However, inhalation was performed for the purposes of divination rather than anaesthesia.

Egyptian surgeons apparently half-asphyxiated children undergoing circumcision by first almost strangling them. This practice sounds almost as barbarous as the operation itself.

Centuries later, Saint Hilary Bishop of Poitiers (315-367), exiled to the Orient in 356 A.D. by the Roman Emperor Constantius, described drugs that "lulled the soul to sleep". But if they were administered in adequate dosage, there was a risk that the soul would not wake up, in this world at least.

Apuleius, a 5th century compiler of medical literature, recommends that “if anyone is to have a member mutilated, burned or sawed let him drink half an ounce with wine, and let him sleep till the member is cut away without any pain or sensation.” Unfortunately, extreme pain tends to exert a sobering effect.

A few mediaevals were surprisingly resourceful. The 13th century occultist, alchemist and learned physician Arnold of Villanova (c.1238-c.1310) searched for an effective anaesthetic. In a book usually credited to him, a variety of medicines are named and different methods of administration are set out, designed to make the patient insensible to pain, so that "he may be cut and feel nothing, as though he were dead." For this purpose, a mixture of opium, mandragora, and henbane was to be used. This method was similar to inhaling the vapours of the soporific sponge mentioned around 1200 by Nicholas of Salerno, and sporadically in different sources from the 9th to 14th centuries.

Arnold's recipe was modified by the Dominican friar Theodoric of Lucca (1210-1298), who added the mildly narcotic juice of lettuce, ivy, mulberry, sorrel and hemlock to the opium-mandragora mixture. From this decoction, a new soporific sponge would be boiled and then dried; when needed again, it was dipped in hot water and applied to the nostrils of the afflicted. The typical effect still left much to be desired: general anaesthesia *avant la lettre* was more of an aspiration than an achievement. Yet if the outcome was often disappointing, so too are the response- and remission-rates for drugs licensed to treat emotional distress in the era of Big Pharma.

Other painkilling techniques for surgery had a longer pedigree. Blood-letting undoubtedly relieved pain, though it was carried out to dangerous and often fatal excess. Before the invention of the suture or stitch in the 16th century by the French military surgeon Ambroise Paré (c.1510-1590), patients undergoing surgery frequently died - either because of bleeding or as a result of the method used to close the wound. Wound-closure usually entailed cauterization by the application of hot oil or hot irons. To seal the exposed blood vessels after amputations, the stump of the bloodied limb might be dipped in boiling pitch. At a distance of several centuries, the use of boiling oil strikes us as brutal and primitive; but it's worth recalling that as late as the 20th century and beyond, the deficiencies of somatic and psychiatric medicine alike could still drive patients to suicidal despair.

Further options for surgical pain-relief were explored with limited success. Non-pharmacological methods besides blood-letting included the use of cold water, ice, distraction by counter-irritation with stinging nettles, carotid compression and nerve clamping. Concussion anaesthesia relied on the hammer stroke: the victim's head was first encased in a leather helmet, after which the surgeon delivered a solid blow to his

patient's skull with a wooden hammer. A less refined concussion method involved a knockout punch to the jaw. In the early 19th century, the most common technique was "Mesmerism", a pseudo-scientific hypnosis dressed up in the language of "animal magnetism". Its eponymous originator was Anton Mesmer (1734-1815). Mesmer believed that all living bodies contain a magnetic fluid. By manipulating this fluid into a state of balance within the body, physical health could allegedly be restored. It's hard to rescue such notions and their proponents from what E.P. Thompson called "the enormous condescension of posterity"; but equally, it's hard to know whether our own descendants will more probably feel condescension or compassion for today's lame theories of, say, consciousness or pain relief.

Across the Atlantic, the New World enjoyed the benefits of coca. Inca medicine men sucked coca leaf with vegetable ash and dripped saliva into the wounds of their patients. Thanks to Viennese ophthalmologist Karl Koller (1857-1944) the anaesthetic effects of the celebrated product of the coca plant were to prove a godsend for surgical operations on the eye. Cocaine relieves other forms of pain too, though these uses are now deprecated.

In the East, the Chinese developed a long tradition of acupuncture. Unlike total anaesthesia, which confers benefits on true believer and sceptic alike, acupuncture works well only with the highly suggestible, and far from reliably even then. But the endogenous opioids released by its application may be better for even the sceptical patient than no palliative relief at all. More obscurely, the Chinese physician Hua Tuo (c.110-c.207) reputedly used hemp boiled with wine to anaesthetize his patients. It is claimed Hua Tuo performed complex surgical operations on the abdominal organs, though only fragmentary details of his exploits with "foamy narcotic powder" are known.

During the Middle Ages in the West, the practice of using natural soporifics, sedatives and pain-relievers to ease the agonies of surgical intervention fell largely into disuse. This neglect was mainly due to the influence of the Christian Church, many of whose leading lights were more adept at causing pain than relieving it. Saving the soul from eternal damnation was conceived as more important than healing the mortal body - a reasonable inference given the assumptions on which it was based. Afflictions of the flesh were commonly understood as punishment for sin, original or otherwise. Pain was supposedly the result of Satanic influence, demonic possession or simply The Will of God rather than an evolved response to potentially noxious stimuli. Investigators who aspired to relieve mortal suffering and understand the workings of the body were not highly esteemed. In the words of Cistercian abbot Saint Bernard of Clairvaux (1090-1153): "...to consult physicians and take medicines befits no religion and is contrary to purity." Surgery and anatomical dissection were widely perceived as shameful activities, not least because they threatened the long-awaited Resurrection of the Flesh. In retrospect, it's clear the theological conception of disease retarded medical progress for generations - no less than the theological conception of mental disorder impedes progress toward a cruelty-free world to this day. Viewing our Darwinian pathologies of emotion as God-given rather than gene-driven obscures how biotechnology can abolish suffering of the flesh and spirit alike. Tomorrow's genetic medicine promises to turn heaven-on-earth from a pipedream into a policy option. Yet if pain is a punishment for original sin, then one would assume it is wicked as well as futile to try and escape it.

Realistic or otherwise, by the 19th century a new age of humanitarianism and scientific optimism about mankind's capacity for earthly self-improvement was (slowly) dawning. The synthesis of the atmospheric gases oxygen, carbon dioxide and nitrogen oxide by early scientific luminaries such as Black, Priestley, and Lavoisier gave birth to the ill-

conceived but seminal discipline of "pneumatic medicine". Its most famous champion was Thomas Beddoes (1760-1808), founder of the Pneumatic Medical Institution in Bristol. Beddoes hired the teenage Humphry Davy as its Research Director. Doctors and patients tried inhaling the newly discovered gases and vapours of volatile liquids to see if their inhalation cured any diseases.

The first gas recognised to have anaesthetic powers was nitrous oxide N_2O. Inert, colourless, odourless and tasteless, nitrous oxide was first isolated and identified in 1772 by the English chemist Joseph Priestley (1733-1804). Priestley was a remarkable polymath: a Unitarian clergyman, political theorist, natural philosopher and educator. Writing of his research on gases, he observed, "I cannot help flattering myself that, in time, very great medicinal use will be made of the application of these different kinds of airs..." [Priestley J., *Experiments and Observations on Different Kinds of Airs*. 6 vols. 1:228, 1774].

Nitrous oxide doesn't induce an anaesthesia nearly as deep or effective as ether: it's a strong analgesic in virtue of its tendency to promote opioid peptide release in the periaqueductal gray area of the midbrain; but unlike ether, it's only a weak anaesthetic. Nitrous oxide is not a muscle relaxant. Induction is rapid because of its low solubility. Its metabolism in the body is minimal, but it inhibits vitamin B-12 metabolism; chronic use of nitrous oxide can cause bone marrow damage. It also inactivates the enzyme methionine synthetase, critical to DNA synthesis and cell proliferation. Nitrous oxide is short-acting and generally regarded as safe to use. Even so, patients are in danger of hypoxia if it's employed at the very high concentrations needed when it's the sole anaesthetic agent.

The exhilarating effects of inhaling nitrous oxide were noted by English chemist Sir Humphry Davy (1778-1829). "Whenever I have breathed the gas," he wrote, "the delight has been often intense and sublime." "Sublime" may not be quite *le mot juste*: Davy found that inhaling the compound made him want to giggle uncontrollably until he passed out. So the illustrious scientist dubbed it "laughing gas". Regrettably, such a frivolous nickname probably discouraged the idea that the gas might serve a serious medical purpose. In like manner today, the racy street slang of short-acting recreational drugs belies the clues they offer to a post-genomic era of mental superhealth.

Sublime or otherwise, the nitrous oxide experience was so much fun Davy wanted to share it with his friends, notably the Romantic poets Samuel Taylor Coleridge (1772-1834) and Robert Southey (1774-1843). "I am sure the air in heaven must be this wonder working gas of delight", enthused Southey. Tantalisingly, Davy himself remarked on "the power of the immediate operation of the gas in removing intense physical pain"; in 1799 he inhaled nitrous oxide to banish the pain of an erupting molar tooth. Davy also discovered that taking the gas could induce “voluptuous sensations.” His early research at Beddoes' Pneumatic Medical Institute is well-documented, even though its implications were missed. In an 80,000-word book on nitrous oxide, *Researches, Chemical and Philosophical; Chiefly Concerning Nitrous Oxide, or Dephlogisticated Nitrous Air, and Its Respiration* (1800), Davy describes the different planes of anaesthesia [stage 1: analgesia; stage 2: delirium; stage 3: surgical anaesthesia; stage 4: respiratory paralysis], though without appreciating the significance of the third plateau suitable for surgical operations.

Most tantalisingly of all, Davy explicitly suggested the use of nitrous oxide as an analgesic during surgery, since it "...appears capable of destroying physical pain, it may

probably be used with advantage during surgical operations in which no great effusion of blood takes place". Unfortunately, this was an idea ahead of its time: several decades of continuing surgical mayhem were to pass before the worldwide anaesthetic revolution.

Davy's student, Michael Faraday (1791-1867), studied nitrous oxide too. He compared its pain-relieving effects with the action of sulphuric ether. In a brief, anonymous 1818 article in *The Quarterly Journal of Science and the Arts*, Faraday noted how:

"*When the vapour of ether mixed with common air is inhaled it produces effects very similar to those occasioned by nitrous oxide...a stimulating effect is at first perceived at the epiglottis, but soon becomes very much diminished...By the imprudent administration of ether, a gentleman was thrown into a very lethargic state, which continued with occasional periods of intermission for more than 30 hours.*"

In the years ahead, there were other missed opportunities, dashed hopes and false starts. In 1824, English country doctor Henry Hill Hickman (1800-30), a contemporary of Davy and Faraday, performed (allegedly) painless operations upon non-human animals using carbon dioxide-induced anaesthesia - thereby more-or-less asphyxiating the various mice, kittens, rabbits, puppies and an adult dog whose various body-parts he amputated. Hickman created in his victims a condition of what he called "suspended animation." This demonstration of inhalational anaesthesia didn't create the stir he anticipated, arousing the interest only of Napoleonic surgeon Baron Dominique-Jean Larrey (1766-1842). Hickman canvassed the possibility of pain-free surgery for humans, though the asphyxial narcosis induced by carbon dioxide makes this particular gas an unsuitable agent. In vain, he sent accounts of his work to the Royal Society of London. It seems Hickman's experiments reminded the Royal Society's President, the ageing Sir

Humphry Davy, of the undignified excesses of his youth. Nothing came of it beyond a footnote in the history books.

Instead, gases and vapours were used by medical instructors and their students for the purposes of hilarity and intoxication rather than the performance of pain-free surgery. Nitrous oxide in particular was exploited in stage shows. An advertisement for one such public entertainment promised that "the effect of the gas is to make those who inhale it either laugh, sing, dance, speak or fight, etc, according to the leading trait of their character. They seem to retain consciousness enough not to say or do that which they would have occasion to regret."

Admixed with oxygen, nitrous oxide remains in surgical use. But the first really effective and (relatively) safe general anaesthetic to gain acceptance was the now abandoned ether.

Ether is liquid at room temperature, but it vaporises very easily. It can therefore readily be either swallowed or inhaled. Unlike nitrous oxide, its vapour can induce anaesthesia without diluting the oxygen in room air to dangerously hypoxic levels. Ether itself had a long history before its use as a surgical anaesthetic. It was marketed under the brand name Anodyne by Halle medical professor Friedrich Hoffmann (1660-1742). Professor Hoffmann recommended Anodyne for pain due to earache, toothache, intestinal cramps, kidney stones, gallstones and menstrual distress. In England, *Materia Medica* (London, 1761) by W. Lewis describes ether as "one of the most perfect tonics, friendly to the nerves, cordial, and anodyne." Readers are advised that three to twelve drops should be taken on a lump of sugar, and swallowed down with water. In the 1790s, medical maverick James Graham (1745-1794), "a famous London quack, proprietor of the Temple of Hymen and owner of the Celestial Bed," habitually inhaled an ounce or two of

ether in public several times a day. He took ether "with manifest placidity and enjoyment". But no one who witnessed him seems to have thought of exploiting its effects for operations.

Ether was first discovered by Catalan philosopher chemist Raymundus Lullius (1232-1315). Lullius called it "sweet vitriol", its name until rechristened by German-born London chemist W.G. Frobenius in 1730. In Greek, "ether" means heavenly. Its synthesis described by German alchemist Valerius Cordus (1514-1554) in 1540. Soon afterwards, Philippus Aureolus Theophrastus Bombastus von Hohenheim (1490-1541), better known as Paracelsus, noted its tendency to promote sleep; and he observed how sweet vitriol/ether "...quiets all suffering without any harm and relieves all pain, and quenches all fevers, and prevents complications in all disease." Paracelsus observed how chickens take ether gladly, and they "...undergo prolonged sleep, awake unharmed". He had picked up much of his medical knowledge while working as a surgeon in several of the mercenary armies of the period; 16th century warfare was endemic, brutal and bloody. Paracelsus was not unduly afraid to challenge received medical wisdom or its proponents: "This is the cause of the world's misery, that your science is founded upon lies. You are not professors of the truth, but professors of falsehood", he informed his fellow doctors. Yet Paracelsus didn't make the intellectual leap needed to take advantage of the properties of ether for human surgical medicine. Had he done so, then given his undoubted brilliance as a publicist, centuries of untold suffering might have been averted.

The first use of general anaesthesia probably dates to early nineteenth century Japan. On 13th October 1804, Japanese doctor Seishu Hanaoka (1760-1835) surgically removed a breast tumour under general anaesthesia. His patient was a 60-year-old woman called Kan Aiya. Her sisters had all died of breast cancer; Kan sought Hanaoka's help. For the

anaesthetic, Hanaoka used "Tsusensan", an orally administered herbal preparation he had painstakingly developed over many years. Its main active ingredient seems to have been the plant Chosen-asagao. Many details of Hanaoka's early life and experiments are obscure. Scholars rely on "Mayaku-ko" (a collection of anaesthetics and analgesics), a pamphlet written by his close colleague Shutei Nakagawa's in 1796. As a young man, Hanaoka had arrived in Kyoto aged 23. He learned both traditional Japanese medicine and Dutch-inspired surgery. For centuries, Western presence in Japan was limited by law to a single island in Nagasaki Bay. The import of medical books was prohibited. But Japanese physicians were able to write down the orally transmitted medical lore of their Dutch counterparts. Critically, and allied to his surgical prowess, Hanaoka believed in "the duty to relieve pain". Apparently he performed numerous experiments on non-human animals in his search for a non-toxic anaesthetic. Hanaoka went on to perform scores of operations on human beings under anaesthesia; he even operated on his daughter and wife. Unfortunately, under the national seclusion policy of the Tokugawa Shogunate (1603-1868), Japan was essentially isolated. Western physicians and their patients knew nothing of Hanaoka's work and tradition.

The breakthroughs that heralded the modern era of anaesthesiology were to come in the New World. Once again the story is messy and involved, albeit better known. In *Artificial anesthesia and anesthetics* (New York, William Wood and Co., 1881), Henry M. Lyman records how in January 1842 the chemist and Berkshire Medical College student William E. Clarke (1818-78) administered ether on a towel to a Miss Hobbie, after which the dentist, Elijah Pope, extracted a painful tooth. It seems Clarke was inspired by his earlier experience of "ether frolics" in Rochester. Yet this was a one-off. Somehow Clarke and Pope failed to recognise the potentially momentous ramifications of what they had done. They neither wrote about nor repeated their feat, as far as we know. So conventionally,

the first clinical use of ether as a surgical general anaesthetic on humans is usually credited to rural Georgian pharmacist and physician Crawford Williamson Long (1815-78). On 30th March 1842, Dr Long removed a cyst from the neck of a Mr James Venable under ether anaesthesia; Mr Venable consented to be a test subject for the occasion on account of his "dread of pain". Dr Long had learned of its properties while ether-frolicking at medical school at the University of Pennsylvania. The use of such social intoxicants was as prevalent in the 1830s and 1840s as the MDMA-fuelled raves of a later era. Ether-filled balloons were liberally handed out for the enjoyment of the audience, a practice that might fruitfully enliven some academic lectures even today. In the era of ether frolics, medical students and budding chemists helped prepare gases for the festivities, a tradition of service that likewise continues more discreetly in the groves of academe even now. Historically, it seems likely that the medical connection may finally have helped several people, more-or-less independently, to draw the link between a form of stage-show entertainment and the opportunity to perform pain-free operations.

In any event, although Dr Long administered anaesthesia to his patients on various occasions, and extended its use to obstetrics, he didn't publicise his discovery beyond his local practice. Indeed until the publication in 1849 of his scholarly article for the *Southern Medical and Surgical Journal*, "An Account of the First Use of Sulfuric Ether by Inhalation as an Anesthetic in Surgical Operations", his work was mostly unknown to the wider world. Long's adoption of pain-free surgery was common knowledge in Jefferson, Georgia, at least: some local residents apparently suspected him of practising witchcraft, others thought merely that it was unnatural, and religious traditionalists objected that pain was God's way of cleansing the soul. Long's explanation of his early reticence was rational; it may even be true, though the full story is probably more complex. "The question will no doubt occur, why did I not publish the results of my experiments in

etherization soon after they were made? I was anxious, before making my publication, to try etherization in a sufficient number of cases to fully satisfy my mind that anaesthesia was produced by the ether, and was not the effect of the imagination, or owing to any peculiar insusceptibility to pain in the persons experimented on."

Whatever the reason, the anaesthetic revolution was now imminent. Connecticut dentist Horace Wells (1815-1848) attempted a public demonstration of surgical anaesthesia in January 1845. Wells had earlier been one of the stage-volunteers who tried inhaling nitrous oxide during a demonstration by P.T. Barnum's apt disciple "Professor" Gardner Quincy Colton (1814-98) at Union Hall in Hartford, Connecticut. One of the other volunteers, Samuel Cooley, a clerk at the local drugstore, injured his legs while agitated in the immediate aftermath of inhaling the gas. Wells afterwards asked the victim if his injury was painful. Cooley said he hadn't felt anything at all; he was surprised to find blood all over his leg.

As ever, chance proverbially favours the prepared mind: critically in this context, Wells was a tender-hearted dentist who hated to see his patients suffer. He had always sought ways to minimise their distress as best he could; dental pain was a notoriously terrible affliction, and so was its cure. Fatefully, Wells now conceived the notion of pain-relief/anaesthesia by gas inhalation. He asked Quincy Colton if he knew any reason why nitrous oxide couldn't be used for dental extractions. Colton said he didn't know any good reason. So the next day Wells submitted to the extraction of one of his own molars by fellow dentist Dr John Riggs. Colton administered the nitrous oxide. Almost insensible, Wells felt no more than a pinprick. Groggy at first, he soon recovered his senses. "A new era in tooth pulling!" he exclaimed; and also, "It is the greatest discovery ever made!" More conservatively, Stuart Hameroff nominates anaesthesia as the greatest invention of the past 2000 years.

Wells was overjoyed. Hugely encouraged at his success, Wells, together with his colleague Riggs, went on to extract teeth from their patients with the aid of nitrous oxide. Wells experimented energetically with ether and other agents too; but he preferred nitrous oxide because it was usually safer. He was now ready to spread news of his discovery as widely as possible. With the help of his former colleague Morton, Wells approached Dr John Collins Warren (1778-1856), founder of the *New England Journal of Medicine* and Massachusetts General Hospital, bearing an account of his marvellous innovation. Warren was sceptical; but with some reluctance he agreed to cooperate. If fate had been kinder, the name of Horace Wells might have echoed down the ages as one of the greatest benefactors of humankind.

Unfortunately, during the public demonstration at Massachusetts General Hospital that Wells staged to publicise his discovery, the patient stirred and cried out. He had been under-anaesthetised; the gasbag was withdrawn too soon. The reaction of Wells' audience, a class of irreverent medical students, was scornful. There was laughter and cries of "humbug". Wells was mortified. In the rest of his short life, it seems he never really recovered from the humiliation. Wells attempted to resume his normal practice back in Hartford. In the wake of the Massachusetts disaster, he tried using nitrous oxide anaesthesia once more the very next day. His determination not to under-medicate led him instead to administer too much gas; he almost killed his patient. Shortly thereafter Wells had some kind of nervous breakdown. For a time, he referred all his patients to his colleague Riggs. Nonetheless, Wells wrote a dissertation *A History of the Discovery of the Application of Nitrous Oxide Gas, Ether, and Other Vapours to Surgical Operations* (1847). He searched for an alternative to the nitrous oxide gas that had let him down in Massachusetts. Tragically, in the course of his experiments he became a chloroform addict. While intoxicated, he attacked a prostitute with sulfuric acid. Fearing he would

now be utterly discredited, and resentful that his unscrupulous protégé Morton was intent on stealing all the credit he deserved, Wells died shortly afterwards by his own hand, embittered and insane. *The Daily Hartford Courant* recorded:

"The Late Horace Wells. The death of this gentleman has caused profound and melancholy sensation in the community. He was an upright and estimable man, and had the esteem of all who knew him, of undoubted piety, and simplicity and generosity of character."

Historical curiosities aside, the era of surgical anaesthesia was inaugurated in a public demonstration inside the same surgical amphitheatre by Wells' former apprentice and colleague, William Morton (1819-1868). The date was 16 October 1846. Not wishing to risk the perceived fiasco of Wells' public demonstration, Morton sought a stronger anaesthetic agent. He was advised by the Boston physician and chemist Professor Charles Jackson (1805-1880) to use ether rather than nitrous oxide. Morton experimented secretly with ether vapour in his office. He also tried ether anaesthesia on a goldfish, his pet water spaniel, two assistants and himself. On 30 September 1846, Morton performed a dental extraction under ether on Eben Frost, a Boston merchant. Mr Frost said he "did not experience the slightest pain whatever". The event was reported over the next two days in the local Boston press, attracting the attention of Henry Bigelow (1818-1890), a smart, sensitive and compassionate young surgeon at Massachusetts General Hospital. Bigelow contacted Morton and Warren so they could liaise. Morton recognised that ether was suitable for full-blown hospital surgery as well as dentistry; he was now ready to enlighten the world. For the public performance, Morton's patient was a 20 year-old printer, Gilbert Abbott. Morton's surgeon was again Dr Warren, before whose audience Wells' disastrous demonstration had taken place less than two

years earlier. The spectators consisted of both medical students and surgeons. The operation consisted in the excision of a vascular tumour located under Mr Abbott's jaw.

Morton's audience was initially sceptical. The failure of Wells' demonstration was locally well known; Morton and Jackson had been present in the amphitheatre too, Morton because he had left Wells' practice and signed up as a medical student. This time, however, everyone who watched the spectacle was amazed. First, Dr Morton briefed his patient on what to do. Before an expectant gallery, Mr Abbot breathed for several minutes from the glass inhaler and its sulphuric ether-soaked sponge. Dr Warren then proceeded to perform the operation. It lasted about ten minutes. Mr Abbott appeared to sleep peacefully throughout, give or take the odd twitch. At no stage did he cry out, though his anaesthesia may not have been entirely complete: Abbott later recalled that he "...did not experience pain at the time, although aware that the operation was proceeding." When the operation was over, the suitably impressed Dr Warren said, "Gentlemen, this is no humbug". Astonished, the surgeons present rushed to try out the procedure themselves; and to spread the word to the rest of the continent - and across the Atlantic, where the innovation rapidly took hold. Bigelow published a report of Morton's triumph in the *Boston Medical Surgery Journal*.

The first use of general anaesthesia in Europe is generally credited to English surgeon Robert Liston (1794-1847). "This Yankee dodge, gentlemen, beats mesmerism hollow", Professor Liston observed after painlessly amputating a patient's leg. In principle, more ambitious surgical operations and investigations inside the abdomen, chest and skull were now feasible - though several decades were to pass before they became common. Operations no longer needed to be conducted at breakneck pace, though until Lister's

carbolic spray allowed antisepsis, vast numbers of patients still died of post-operative infection. The Boston surgical amphitheatre is now The Ether Dome.

Morton himself was eager to patent his procedure and get rich. From the outset, he had sought to disguise the nature of the agent he used: pure ether is a pungent, volatile, aromatic gas that was unpatentable owing to its long use for other purposes. Morton called his own secret ether-based concoction "The Letheon"; it contained various aromatic oils and opium as well as sulfuric ether. Unsurprisingly, the identity of its prime active ingredient soon leaked out. For the rest of his life, Morton would be engaged in rancorous disputes with rival claimants to priority. "In science the credit goes to the man who convinces the world, not to the man to whom the idea first occurs," wrote Francis Darwin in 1914. Morton's PR machine "won". More importantly, during the American Civil War (1861-65) Morton personally administered anaesthesia to thousands of Union and Confederate soldiers on the battlefield. His grave in Mount Auburn Cemetery near Boston bears the inscription:

WILLIAM T. G. MORTON

Inventor and Revealer of Anaesthetic Inhalation

Before Whom, in All Time, Surgery Was Agony

By Whom Pain in Surgery was Averted and Annulled

Since Whom Science Has Control of Pain

Suffering humanity the world over would find the last line cruelly ironic, and the priority assertion has been questioned; but the main claim of Morton's epitaph is in substance correct. The reasons for the persistence of suffering in the world are now more ideological than scientific. Pain - and pleasure - are controllable.

* * *

THE CASE FOR PAIN

Despite their obvious advantages, pain-free surgery, dentistry and (especially) pain-free childbirth were opposed by a conservative minority.

The City of Zurich initially outlawed anaesthesia altogether. "Pain is a natural and intended curse of the primal sin. Any attempt to do away with it must be wrong", averred the Zurich City Fathers (*Harpers* (1865); 31: 456-7). Their stance contrasts with the more enlightened Swiss attitude of the 1990s. Latter-day Zurich experimented with what became known as Needle Park. Addicts could openly buy narcotics and inject heroin without police intervention.

In Scotland, Sir James Young Simpson (1811-1870), eloquent advocate of chloroform anaesthesia and pioneer of painless delivery in childbirth, offended various Calvinist Scots by his presumption. For did not *Genesis* 3:16 declare: "Unto the woman he said, 'I will greatly multiply thy sorrow and thy conception; in sorrow thou shalt bring forth children'"? Religious traditionalists held that mothers ought to fulfil the "edict of bringing forth children in sorrow" as laid down in the Holy Bible. Simpson was accordingly denounced by a vocal minority of ministers and priests as a blaspheming heretic who uttered words put into his mouth by Satan. [see *Triumph over Pain* by René Fülöp-Miller, New York Library Guild, 1938]. One clergyman saw the new chloroform anaesthesia as "a decoy from Satan, apparently offering to bless woman; but, in the end, it will harden society and rob God of the deep earnest cries, which arise in time of trouble for help."

God's reaction to being robbed of the cries of women in labour is not on record; but there were mutterings that infants delivered painlessly should be denied the sacrament of baptism. This never came to pass: mid-Victorian religious opposition to anaesthesia was neither as widespread nor as organised as some historians were later to suggest. Yet a hostile reaction to human tampering with the God-given order of things hadn't always been empty rhetoric. In the text of his *A History of the Warfare of Science with Theology* (1896), A.D. White relates how "as far back as the year 1591, Eufame Macalyane, a lady of rank, being charged with seeking the aid of Agnes Sampson for the relief of pain at the time of the birth of her two sons, was burned alive on the Castle Hill of Edinburgh; and this old theological view persisted even to the middle of the nineteenth century."

Fortunately, Professor Simpson knew his Old Testament. He contended that the Biblical "sorrow" was better translated as toil, an allusion to the muscular effort a woman exerted against the anatomical forces of her pelvis in expelling her child at birth. Moreover he cited *Genesis* 2:21: "And the Lord God caused a deep sleep to fall upon Adam, and he slept: and he took one of his ribs, and closed up the flesh instead thereof". Casting God in the role of The Great Anaesthetist might seem at variance with the historical record; and not everyone was convinced. A Dr Ashwell (*The Lancet* (1848:1, p.291)) replied that "Dr Simpson surely forgets that the deep sleep of Adam took place before the introduction of pain into the world, during his state of innocence." Yet the suggestion that God Himself employed anaesthesia helped carry the day.

Simpson had risen from humble origins to become Professor of Obstetrics in Edinburgh. A strong-willed and opinionated controversialist, he was also a compassionate doctor who ministered to rich and poor alike. As a young man, he had almost abandoned his choice of a career in medicine after being shocked at witnessing the suffering that surgical practice then entailed. Patients undergoing the knife had first to be tightly strapped down

or held by several strong men so as to restrain their agonised writhings. Operating rooms had "hooks, rings and pulleys set into the wall to keep the patients in place during operations" (Julie M. Fenster, *Ether Day*, 2002); victims of surgery still underwent pain as excruciating as anything inflicted in a medieval torture chamber. In that respect, little had changed since the famous Roman physician Cornelius Celsus, writing in 30 A.D., claimed that the ideal surgeon should be "so far void of pity that while he wishes to cure his patient yet is not moved by his cries to go too fast or cut less than is necessary".

Eighteen hundred years later, surgery was still performed only as a desperate last resort. Operations were typically conducted against a backdrop of hideous screaming or groaning. "The escape from pain in surgical operations is a chimera...'Knife' and 'pain' in surgery are words which are always inseparable in the minds of patients", affirmed the great French surgeon Alfred-Armand-Louis-Marie Velpeau (1795-1867) in 1839. Surgery could be emotionally traumatic for surgeons as well as their patients. As President of Harvard, Edward Everett (1794-1865), noted with regret: "I do not wonder that a patient sometimes dies, but that the surgeon ever lives." Yet within little more than a decade, the anaesthetic revolution had spread across the globe, and its opponents vanquished.

Writing to a fellow doctor in 1836, Simpson had asked: "Cannot something be done to render the patient unconscious while under acute pain, without interfering with the free and healthy play of natural functions?" Simpson tried mesmerism; but it didn't work. He first learned of ether anaesthesia from his old tutor in London, Robert Liston. News had travelled to England by letter via the fastest possible route, transatlantic steamship. Simpson himself used ether in surgery three weeks later, publishing an account in *Edinburgh Monthly Journal of Medical Science* in March 1847. However, ether was disagreeably smelly, slow-acting and irritating to the bronchial tubes. Simpson sought a better agent, more suitable for women-in-labour. In October, his Liverpool manufacturing

chemist, David Waldie, sent him a chloroform sample. Simpson self-experimented. He then used chloroform successfully in his obstetric practice, publishing an enthusiastic account of the advantages of chloroform in the *Lancet* in November. Soon, he was insisting that "every operation without it is the most deliberate and cold-blooded cruelty". But Simpson went further. Among his patients, he favoured general anaesthesia in midwifery for *every* delivery. He quoted Galen: "Pain is useless to the pained". Simpson maintained: "All pain is *per se* and especially in excess, destructive and ultimately fatal in its nature and effects." Simpson's sentiments were admirable even if his medical science was sometimes flawed.

Professor Simpson didn't confine his use of anaesthesia to surgical practice. In his search for new and improved anaesthetics, he tried everything out on himself and his colleagues. Some accounts of his research on new anaesthetising agents read more like the exploits of a teenage glue-sniffer than a shining example of methodological rigour. Simpson was fond of using young women as test subjects. A larger-than-life figure, he was in the habit of administering chloroform to overawed dinner-party guests in drawing rooms across the country, and then kissing the young ladies who passed out under its influence - a form of experimentation now unlikely to pass muster with a medical ethics committee. "One of the young ladies, Miss Petrie, wishing to prove that she was as brave as a man, inhaled the chloroform, folded her arms across her breast, and fell asleep chirping 'I'm an angel! Oh, I'm an angel!'". René Fülöp-Miller describes one such scene:

"*On awakening, Simpson's first perception was mental. 'This is far stronger and better than ether,' said he to himself. His second was to note that he was prostrate on the floor. Hearing a noise, he turned round and saw Dr. Duncan beneath a chair – his jaw dropped, his eyes staring, his head bent under him; quite unconscious, and snoring in the most*

determined manner. Dr. Keith was waving feet and legs in an attempt to overturn the supper table. The naval officer, Miss Petrie and Mrs. Simpson were lying about on the floor in the strangest attitudes, and a chorus of snores filled the air."

"They came to themselves one after another. When they were soberly seated round the table once more, they began to relate the dreams and visions they had had during the intoxication with chloroform. When at length Dr. Simpson's turn came, he blinked and said with profound gratification: 'This, my dear friends, will give my poor women at the hospital the alleviation they need. A rather larger dose will produce profound narcotic slumber.'"

Unfortunately, Simpson failed to realise that chloroform is a potentially dangerous agent for the patient - or the recreational user - even if employed under ideal conditions. It can cause ventricular fibrillation of the heart, a potentially lethal complication. Initially, Simpson thought that chloroform anaesthesia was absolutely safe; and he then blamed early fatalities and adverse reactions to the procedure not on its depression of cardiovascular and respiratory function, but the incompetence of English physicians.

He was mistaken; but myths and misconceptions about the new operating procedure ran rife among professionals and laypeople alike. One popular rumour supposed that anaesthetics provoked carnal fantasies, converting childbirth into a gigantic orgasm. Some physicians thought likewise. The *American Journal of Medical Surgery* (1849 18:182) cites a leading obstetrician who "insist[ed] on the impropriety of etherization...in consequence of the sexual orgasm under its use being substituted for the natural throes of parturition". In *A Lecture on the Utility and Safety of the Inhalation of Ether in Obstetric Practice* (1847, *Lancet* 1, 321-323), Dr Tyler Smith reported the case of a young Frenchwoman who gave birth under ether anaesthesia and afterwards confessed

to have been dreaming of sexual intercourse with her husband. "To a woman of this country the bare possibility of having feelings of such a kind excited and manifested in outward uncontrollable actions would be more shocking even to anticipate than the endurance of the last extremity of physical pain", Dr Smith observed. As recounted in Linda Stratmann's illuminating *Chloroform: the Quest for Oblivion* (2003), this incident was taken up by Simpson's opponent Dr George Thompson Gream of Queen Charlotte's Lying-in Hospital. In *Remarks on the Employment of Anaesthetic Agents in Midwifery* (London, John Churchill, 1848), Gream offered his readers the further salacious detail that the wanton Frenchwoman had also offered to kiss a male attendant. Gream was confident that as soon as women in general heard what anaesthetics might do to them "they would undergo even the most excruciating torture, or I believe suffer death itself, before they would subject themselves to the shadow of a chance of exhibitions such as have been recorded....the facts are unfit for publication in a pamphlet that may fall into the hands of persons not belonging to the medical profession."

Fortunately, Gream overestimated the stoicism and virtue of English women. His views were extreme even among strait-laced prudes; most doctors did not take them seriously even at the time. But worries about drug-induced sexual disinhibition were scarcely peculiar to the Victorians. Periodic moral panics over drug-fuelled sex have always tended to be relatively independent of the pharmacodynamic properties of the agent in question. Thus in the popular press, Chinese immigrants in the age of "Yellow Peril" were intent on luring young white women into their opium dens to become sex slaves; GHB periodically turns chaste damsels into nymphomaniacs; and in the era of "Reefer Madness", marijuana supposedly transformed healthy youngsters into sexual deviants prone to inter-racial sex. Other examples are legion. On a more realistic note, cocaine use can indeed promote promiscuous hypersexuality, though not if used as a local

anaesthetic in dentistry; and the sense of universal love and trust promoted by MDMA can lead to "inappropriate bonding" and unprotected sex.

In Victorian Britain, not all women who experienced troublesome post-surgical imagery were deluded. A minority of doctors made a habit of seducing insensible female patients and ascribing any confused recollections of impropriety on their part to a known side-effect of the anaesthetic. But the notion that anaesthesia might promote lewd thoughts and disinhibited behaviour did little to promote the acceptance of painless delivery in polite society. Men especially were prone to believe that reducing mothers to a helpless state of unconsciousness while they enacted their life-defining childbearing role was unnatural and immoral. "The very suffering which a woman undergoes in labor is one of the strongest elements in the love she bears for her offspring," said one clergyman. In *The Lancet* 2 (1849), 537, English doctor Robert Brown explained how God and Nature "walked hand in hand"; painless delivery was an invention of the Devil. In an era when most people still subscribed to the metaphysics of vitalism, Simpson's opponents were convinced that the experience of pain must serve some essential purpose. "Pain in surgical operations is in a majority of cases even desirable, and its prevention or annihilation is for the most part hazardous to the patient", alleged Simpson's adversary Dr James Pickford, though without adducing any compelling evidence why this might be so.

At the South London Medical Society, sentiment ran strongly against painless surgery. Addressing a meeting held shortly after Simpson's original chloroform paper, the well-respected Dr Samuel Gull declared that it was a "dangerous folly to try to abolish pain". Even if its abolition were morally desirable, Dr Gull averred, "ether was a well-known poison". [F. Stanley. *For Fear of Pain, British Surgery 1790-1850*, (2003)]. Ether and chloroform were described by Gull's colleague Dr Cole as "pernicious". A Dr Nunn "failed

to see how surgeons could get on without pain". The views of Francois Magendie ("*La Douleur Toujours!*") across the Channel were quoted with approval. Dr Radford, drawing the meeting to a close, concluded that "there was nothing but evil" in the new-fangled procedure [see T. Dormandy *The Worst of Evils, The Fight Against Pain*, (2006)]. The rationalisation of human suffering is widely shared among foes of the medical prevention or annihilation of *emotional* pain today; and defended on equally tenuous metaphysical grounds.

In England, at least, the practice of anaesthesia during childbirth won greater respectability following its widely-publicised use on Queen Victoria. The delivery in 1853 of Victoria's eighth child and youngest son, Prince Leopold, was successful: chloroform was administered by Dr John Snow (1813-1858) of Edinburgh, the world's first anaesthesiologist/anaesthetist. In 1847 Snow had published *On the Inhalation of Ether in Surgical Operations*, a scientific milestone. Dr Snow sought to put the principles of anaesthesia on a sound scientific basis. Critically, Snow introduced inhalers designed to deliver an accurate and controlled "dose" of anaesthetizing agent to each closely monitored patient. Prudently, in Queen Victoria's case the dosage of chloroform induced analgesia rather than complete anaesthesia. "Dr Snow gave that blessed chloroform and the effect was soothing, quieting and delightful beyond measure", Her Majesty reported. If the Queen had died in consequence, then the progress of anaesthesia might have been set back a generation; fortunately, she survived unscathed. Anaesthesia *à la reine* even became fashionable in high society.

Patients and many mothers-to-be were understandably thrilled. One mother was so delighted by a painless delivery that she named her child Anaesthesia. Yet controversy didn't abate altogether. *The Lancet* was scandalised at the use of anaesthesia on the

Queen. The distinguished journal even professed to doubt if the story were true, since chloroform "has unquestionably caused instantaneous death in a considerable number of cases" ["Administration of Chloroform to the Queen", *The Lancet* 1 (May 14, 1853): 453)]. As its commentary noted with alarm, "Royal examples are followed with extraordinary readiness by a certain class of society in this country." *The Lancet* wasn't persuaded of the need for general anaesthesia in dentistry either. After a death in an Epsom dentist's chair in 1858, its editor warned: "It is chiefly fashionable ladies who demand chloroform. This time it was a servant girl who was sacrificed; the next time it may be a Duchess."

Though snobbish and rhetorically overblown, *The Lancet*'s caution was far from amiss. One problem was the lack of controlled clinical trials comparing use of chloroform and ether. Chloroform is faster-acting and easier to use, but ether is generally safer.

Chloroform use also had a shorter history. A colourless, volatile liquid with a characteristic smell and a sweet taste, chloroform was discovered in July 1831 by American physician Samuel Guthrie (1782-1848); and independently a few months later by Eugène Soubeiran (1797-1859) in France and Justus von Liebig (1803-73) in Germany. Prophetically, Guthrie's eight-year-old granddaughter Cynthia once anaesthetised herself by accident after inhaling chloroform vapour; she was in the habit of dipping her finger into the liquid and tasting it. "Guthrie's sweet whiskey" became a popular local tipple; its consumption induced what Guthrie described as "a lively flow of animal spirits, and consequent loquacity." Chloroform soon found its way into patent medicines. The most famous of these concoctions was chlorodyne, a tincture of chloroform and morphine designed as a remedy for cholera by British army surgeon Dr. J. Collins Browne.

Unlike ether, chloroform isn't flammable, an important virtue in a candlelit era. Chloroform is also less of a chemical irritant to the respiratory passageways. However, it is a cardiovascular depressant. More insidiously, chloroform has toxic metabolites that can cause delayed-onset damage to the liver. Like most anaesthetics, it has a relatively low therapeutic window. This posed a particular risk when chloroform was administered in the preferred Edinburgh fashion by folded silk handkerchief. There was no set dose; when chloroform was used in quantities suitable for anaesthesia rather than inebriation, it was simply administered until the patient became insensible. “The notion that extensive experience is required for the administration of chloroform is quite erroneous, and does harm by weakening the confidence of the profession in this invaluable agent”, declared the father of antiseptic surgery Joseph Lister (1827–1912), Surgeon to the Royal Infirmary and Professor of Surgery in the University of Glasgow.

With hindsight, this opinion was ill-judged and dangerously naïve. The first known death under anaesthesia was reported as early as January 1848: the case of Hannah Greener, a 15-year-old girl who died under chloroform while undergoing toenail excision. In response to such early tragedies, Dr Joseph Clover (1825-1882) developed in 1862 the first apparatus to provide chloroform in controlled concentrations; and a "portable regulating ether-inhaler" in 1877. Yet serious risks remained even as technology to control the depth of anaesthesia improved. Anaesthesiology as practised in the modern era is recognised as a technically demanding medical specialism with a long and arduous apprenticeship. Even now, anaesthetics can occasionally cause serious complications: liver or kidney damage, strokes, heart attacks, seizures, pneumonia, low blood pressure and allergic reactions are all known risks. In Victorian England, there was none of our sophisticated cardiorespiratory monitoring equipment, endotracheal intubation, ventilators and extensive perioperative care for the patient. Moreover, the epoch-making

transition to pain-free surgery wasn't initially accompanied by an appreciation of the germ theory of disease and the importance of asepsis: this critical breakthrough would await the discoveries of Semmelweis, Pasteur, Koch and Lister. Tragically, variations on "The operation was a success but the patient died" remained a common refrain in the aftermath of surgery for several decades to come. Almost half of patients undergoing some kinds of invasive surgery in the 19th century still died soon thereafter, mainly through septicaemia. There are in truth few *obstetric* situations today where general anaesthesia is either medically or humanely essential: use of local or regional anaesthesia usually suffices for natural childbirth, though millions of mothers in labour throughout the world endure grossly inadequate pain-relief. But four years after the birth of Prince Leopold, Dr Snow again used chloroform for the birth of Victoria's youngest daughter, Princess Beatrice. Dr Snow also delivered a baby for the daughter of the Archbishop of Canterbury. In the end, royal and ecclesiastical, if not divine, sanction was enough to silence the critics.

Rhetorical passions nonetheless ran as high across the Atlantic as they did in Great Britain. In 1847, *The Philadelphia Presbyterian* thundered, "Let everyone who values free agency beware of the slavery of etherization". The American Temperance movement took an equally dim view of surgical anaesthesia. It regarded etherization as a form of intoxication that posed a threat to the virtue of female patients. Although surgeons and their patients mostly embraced pain-free operations with gratitude, a motley collection of conventionally-minded doctors, dentists and scientists voiced vehement opposition. Dr William Henry Atkinson, first president of the American Dental Association (ADA), protested, "I think anesthesia is of the devil, and I cannot give my sanction to any Satanic influence which deprives a man of the capacity to recognize the law! I wish there were no such thing as anesthesia. I do not think men should be prevented from passing

through what God intended them to endure." [see *Sacred Pain: Hurting the Body For the Sake of the Soul*, By Ariel Glucklich, Oxford University Press, 2001]. Atkinson apparently conceived pain as spiritually uplifting. Pain wasn't an expression of God's punishment of man, but His paternal affection.

Theologians in particular were prone to believe that agony bravely borne was spiritually uplifting. In Milan, Cardinal Berlusconi, distant relative of the later Italian premier, delivered a much-cited sermon condemning advocates of painless surgery for seeking to abolish "one of the Almighty's most merciful provisions" [*Unsere Schmerzen* (Vienna, 1868)]. In human society, and especially the Judaeo-Christian tradition, heroes and heroines who stoically endure the greatest suffering are usually awarded the greatest esteem. An unheroic tendency toward self-pity is despised. Thus in Canada, surgeons-general in the army initially refused to use anaesthetics for operations on the grounds that their manly soldiers could take such trifles in their stride. In the USA, regular army surgeon John B. Porter banned use of anaesthetics on stricken soldiers under his command, allegedly on grounds of safety but probably in part because "the easiest pain to bear is someone else's". Our Darwinian concepts of moral strength and nobility of character are bound up with the ability to withstand extremes of suffering, whether the pain is called "physical" or "emotional" or both. Alas, sensitive psychological weaklings are seldom respected by Society or Mother Nature. In the case of "physical" pain, early critics of anaesthesia held that the prospect of rendering patients insensible for the purposes of surgery was dehumanising. Pain-free existence supposedly robbed human beings of their essential humanity and dignity. Unfortunately the dignity of unbearable pain is frequently lost on its victims.

The obscurantist view did not go theologically unchallenged. A few religiously-minded physicians used theological arguments to *justify* the medical use of anaesthetics. In *On*

the Property of Anaesthetic Agents in Surgical Operations (1855), Dr Eliza Thomas describes anaesthesia as "a second dispensation": a gift from God. But clerical enthusiasm, as distinct from acquiescence, was uncommon. How would God be able to punish His children for unrighteousness if the main weapon at His disposal, namely pain, were taken away? The argument that doctors and surgeons should not "play God" is common today even among those who pay homage to Nature rather than the Almighty. Naturopaths, homeopaths and herbalists were as hostile to "unnatural" anaesthesia as they are to the interventions of contemporary scientific medicine.

Critics of anaesthesia could count on academic allies. Doctor Charles Delucena Meigs (1792-1869), Professor of obstetrics and diseases of women at Jefferson Medical College, was of the opinion that labour-pains were "a most desirable, salutary and conservative manifestation of the life force." This "life force" was weakened by etherization, which should thus be avoided. Dr Meigs thought chloroform was objectionable too; he regarded taking it as little different from getting drunk. His degree of empathy with women in labour is captured in his remark of a woman that she "has a head almost too small for intellect and just big enough for love". More reasonably, Meigs pointed out that the mechanism and long-term effects of anaesthesia on the brain were unknown.

Antipathy to painless surgery soon entered scholarly print. The *New York Journal of Medicine* [9 (1847) 1223-25] declared that pain was vital to surgical procedure, and that its removal was harmful to the patient. This notion now sounds quaint, perhaps as quaint as our own assumption that a capacity for emotional pain is indispensable to health - or at least an essential diagnostic guide to problems - may one day seem to our descendants. But the anxiety which the journal's reaction expressed was common. Francois Magendie (1783-1855), the famous French physiologist, neurologist and puppy vivisectionist, held that pain was essential to life. Magendie believed that anaesthesia

reduced the "patient to a corpse". The loss of "vital spirit" induced by anaesthesia would supposedly endanger the patient in the operating theatre - and delay or prevent recovery after an operation. Like supporters of the "heroic medicine" of Benjamin Rush (1745-1813), Magendie supposed that etherization sapped the life-force. After Darwin and the triumphs of organic chemistry, we are more likely to view each other as neurochemical robots devoid of vital spirit; but physical pain had previously been so intimately bound up with life that many 19th century philosophers and scientists assumed that suffering must be inseparable from the mysterious life-force itself. Sections of the medical profession even valued pain and its manifestations as an encouraging sign of a patient's vitality - and the effectiveness of a doctor's prescription. In *Calculus of Suffering: Pain, Professionalism, and Anesthesia in Nineteenth-Century America*. New York, NY: Columbia University Press; 1985), Martin Pernick quotes physician Felix Pascalis: "The greater the pain, the greater must be our confidence in the power and energy of life". By contrast, anaesthesia evoked death.

Contemporary media commentators are prone to express similar sentiments if not idiom when conjuring up the spectre of a soma-driven Brave New World. Within the lifetime of people now living, biotechnology threatens to abolish life's rich tapestry of psychological distress. Suffering in its many guises is assuredly terrible, its rationalisers acknowledge; but its loss would deprive us of our humanity, freedom and dignity - and perhaps an indefinable vital energy too, though the expression itself has fallen into disuse. Pain, its apologists suggest, is somehow more *authentic* than happiness. Certainly, for evolutionary reasons euphoric well-being has hitherto been impossible to sustain for most of us, irrespective of its propositional content. So there is a tendency to regard its episodes as "false", or alternatively as rare and necessarily elusive "peak experiences".

Perhaps its "reality" may seem greater if invincible bliss becomes part of the genetically coded fabric of conscious life rather than a drug-induced aberration.

Other objections to the anaesthetic revolution were harder to refute. A number of critics were worried that anaesthetics merely immobilised the body and induced amnesia but didn't extinguish pain. The patient might then be left paralysed under the surgeon's knife but fully conscious - trapped in incommunicable agony. Although chloroform and ether are (we believe) innocent of any such charge, a terrible medical error *was* committed almost a century later with a neuromuscular blocking agent, the South American Indian arrow poison curare. In its day, curare represented a significant surgical advance. Although its use necessitated intubation of the trachea and mechanical ventilation of the patient's lungs, its introduction led to a decline in anaesthetic mortality. This is because curare lacks the depressant effects of general anaesthetics on the heart. Unfortunately, some surgeons and anaesthetists initially assumed that curare was an anaesthetic as well as a muscle-relaxant. A few patients endured surgery under curare while paralysed and awake. But rather than forgetting their nightmare afterwards, the victims were traumatised. Curare does not induce amnesia. Although this particular mistake has not been repeated, in operations on humans at least, the use of neuromuscular blocking agents in conjunction with anaesthetics increases the risk of awareness during surgery.

* * *

THE CONQUEST OF SUFFERING

So how close are the parallels between arguments used against technologies to relieve emotional pain and somatic pain? Where, if at all, does the analogy break down?

There are of course disanalogies between, on the one hand, the use of anaesthetics and analgesics to prevent pain in clinical medicine and, on the other hand, the use of therapeutic agents to dispel the "natural" mental pain of everyday Darwinian life. For a start, whereas painkillers typically diminish the intensity of consciousness, and general anaesthesia suppresses it, post-genomic medicine promises to deepen, diversify and intensify the quality of our awareness. By contrast, too, strong analgesics tend to diminish the functional capacity of the user, and anaesthetics effectively abolish it, whereas mood-enriching designer drugs and gene therapies of the future are more likely to extend our intellectual, physical, sensual and aesthetic capacities - and possibly even our spiritual, introspective, empathetic and moral sensibilities as well. There are further disanalogies. Undergoing total anaesthesia for surgery involves surrendering control of one's body to others: one early argument against surgical anaesthetics was that they left a woman defenceless - unable to defend her virtue should her half-naked body inflame the lust of her (male) surgeons, and perhaps a prey to wanton and disinhibited behaviour herself. By contrast, electing to take long-acting mood-brighteners is typically empowering. Other things being equal, heightened mood at once increases one's capacity for autonomous action, promotes enhanced social status in primate dominance-hierarchies, and strengthens one's sense of agency - the obverse of the "learned helplessness" and submissive behaviour characteristic of depression.

None of the above should obscure the critical similarity between the two fundamental categories of suffering. Insofar as they can be distinguished, both somatic and emotional

pain are at once profoundly distressing and, potentially, functionally *redundant* in the era of post-genomic medicine. Their functional roles can be multiply realised in other ways that don't involve the texture ("what it feels like") of unpleasantness - insofar as their functional roles need to be realised at all. Both somatic and emotional pain share common substrates in the molecular machinery of the nerve cell. Intuitively, extreme "physical" pain is worse. Yet it is unbearable "emotional" pain that causes almost a million people in the world to kill themselves each year. Emotional pain causes millions of "para-suicides" and cases of self-injurious behaviour; and emotional pain induces tens of millions of depressive people periodically to wish they could die or didn't exist. In practice, the two kingdoms of pain are intimately linked. Untreated pain commonly leads to depression, and depression is frequently manifested in somatic symptoms.

There are of course (many) complications before the conquest of suffering can ever be complete. Sustaining lifelong bliss *and* a capacity for critical insight isn't straightforward, especially if such bliss is to be empathetic and socially responsible rather than egotistic. Any intelligent organism depends on a complex web of interlocking, genetically regulated feedback mechanisms to flourish. So something as central to primordial human existence as aversive experience can't be edited out of our lives without ensuring a rich network of functional analogues to take its place - short of wireheading. Fortunately, there is nothing functionally indispensable to intelligent mind about the raw phenomenal texture of pain, whether it's the searing agony provoked by acute tissue trauma or the aching despair of melancholic depression. For phenomenal pain is not the same as sensory nociception; nor should its "psychological" counterparts be equated with the functional role they play in the informational economy of Darwinian minds.

Our imminent capacity to edit and rewrite our genetic code means that other information-processing options can be explored too. At the most basic, we can ratchet up our normal mood-levels so we can all feel happy and emotionally fulfilled. Critically, dips in gradients of an exalted well-being that (stably) fluctuates around an elevated "hedonic set-point" can potentially signal "danger" or "error" (and motivate us to avoid them) at least as powerfully as do gradients of suffering. If pleasure and pain were merely relative, then such homeostatic dips in exalted awareness would actively hurt; as it is, they may in future play an error-correcting role merely (dimly) analogous to the bestial horrors of the past. Opioid neurotransmitter system redesign will play a role in the recalibration; but re-engineering the architecture of the mesolimbic dopamine system will be a vital step toward recalibrating our reward circuitry so we can all be *dynamically* superwell throughout our lives. For the meso(cortico-)limbic dopamine system mediates, not just pleasure, but appetitive behaviour and so-called incentive-motivation. Revealingly, dopamine-releasing drugs act as powerful analgesics as well as euphoriants; by contrast, some 50% of victims of the "dopamine deficiency disorder" better known as Parkinson's disease report symptoms of physical pain. More generally, a large minority of people in contemporary human society are driven mainly by gradients of misery, discomfort and discontent. A small minority are animated primarily by gradients of well-being, and many - but not all - of this small minority are labelled either (hypo)manic or bipolar. Most people fall somewhere in between in their daily mood spectrum. "Hyperthymic" people animated by gradients of lifelong happiness without mania are currently medically rare freaks of nature, though not unknown.

Within the next few decades, however, humanity will have the pharmaceutical and genetic opportunity to choose whatever range of the affective axis we wish to occupy as

the backdrop to our lives. To date, we have barely glimpsed the potential extremes of the pleasure-pain axis; in the case of the dark side of sentience, it may be hoped (and statistically expected) that we never will. More ambitiously, the new biotechnologies promise to extend our range of choices way beyond tools for crude, unidimensional mood-modulation. For we'll have the tools to re-design the neural basis of our personalities to repair the deficits of natural selection. Even better, ethically speaking, the application of germline hedonic engineering can ensure that parenthood won't entail bringing any more suffering into the world. Misery becomes physically impossible if the genetic code for its biological substrates is missing. Thus having children needn't, as now, entail causing more heartache as well as episodic happiness. Procreation becomes permissible even for the negative utilitarian who finds it impossible to practise ethical parenthood with a Darwinian genome.

Yet will *some* form of real "emotional" pain still be endemic to future civilisation, just as "physical" pain was endemic to the lives of our ancestors - and as it lingers among disease-stricken victims of opiophobia even today? Or is it possible our post-Darwinian descendants will enjoy lifelong mental superhealth that's orders of magnitude richer than our own (though use of "health" terminology to describe our own malaise-ridden lives may be something of a misnomer)? From an information-theoretic perspective, what matters functionally and computationally to any neurochemical robot is not our absolute "hedonic set point" on the pleasure-pain axis. What counts is that we are informationally sensitive to fitness-relevant *changes* in our internal and external environment. Our contemporary pains and pleasures reflect the genetic "fitness tokens" of the African savannah; in consequence, we're stuck, scrabbling around in a severely sub-optimal homeostatic rut. It would be surprising if the genetic fitness tokens of our hominid

ancestors were to remain adaptive in a post-Darwinian era of planned parenthood and paradise-engineering.

* * *

NOCICEPTION WITHOUT TEARS

Not everyone has the physiological capacity to suffer pain. A few people quite literally do not understand what the term means. Several syndromes of congenital insensitivity to pain (CIP) are known. Their affective counterparts, sporadic cases of lifelong unipolar euphoric (hypo)mania and extreme hyperthymia without mania, occur in different subtypes; they are rare too. The opposite syndromes of chronic pain and hyperalgesia, and chronic unipolar depression or dysthymia, are much more common, presumably reflecting the comparative selection pressures of our ancestral environment. In most cases today, a lack of pain-sensitivity can plausibly be presented as a deficit in signal-processing capacity rather than a harbinger of post-Darwinian superhealth.

This judgment may be premature. In retrospect, 19th century opponents of painless surgery were wrong to claim that pain was an essential diagnostic aid to surgical medicine, and wrong to claim that anaesthesia extinguished a person's "vital spirit". Yet might opponents of genetically enriched life rooted in gradients of bliss be right to claim that *emotional* pain will always remain an indispensable diagnostic aid to danger and error?

Perhaps so. But abolitionists who seek life-long *high functioning* well-being can point to two families of alternative:

1. the futuristic "cyborg" solution. We know that silicon robots can be constructed with spectroscopic (etc) sensors that can "see" and "hear" more sensitively than human beings - even though this greater discriminative capacity isn't matched by the felt textures of phenomenal colour or sound. Artificial silicon (etc) systems can also be designed or trained up so as to be more sensitive to noxious insults and structural damage as well. In future, modular implants can benefit rare victims of congenital anaesthesia who are prone to life-threatening injury through lack of efficient feedback-signalling mechanisms. But *if* the rest of us, too, ever want to *augment* ourselves with modules performing an analogous adaptive role, i.e. efficient sensory nociception and avoidance behaviour without the cruel textures of phenomenal pain, then enlisting all sorts of smart neurochips and prostheses is technically feasible - whether or not their widespread adoption is ever sociologically realistic. Analogously, the information-theoretic role of our nastier emotions (jealousy, spite, etc) can in principle be replicated without their current sinister textures as bequeathed by evolution - though it may be wondered whether the "functional role" of modules mediating some of our baser feelings can't be discarded altogether along with their vicious "raw feels". It's hard to see what jealousy is good for beyond its tendency to maximise the inclusive fitness of our genes in the ancestral environment of adaptation. Our descendants may make the judgment that neither its texture nor functional role have any redeeming value; and may therefore elect to discard both. For sure, most people who aren't transhumanists are unexcited at the prospect of updating the "fitness tokens" of our evolutionary past, let alone implanting neural prostheses that tamper with their intimate soul-stuff. But it's worth stressing that this bionic strategy isn't committed to turning us into hyperintelligent "zombies". This is because desirable

facets of our subjective consciousness can be exquisitely enriched and amplified even as the nastier phenomenology of Darwinian life is phased out. Thus our descendants may not just be super-smart but hyper-sentient too. If so, then "awakened" life is likely to be founded on gradations of *blissful* hypersentience that replaces gradations of Darwinian malaise. The nature of what we'll be happy "about" is currently hard to guess; but this uncertainty reflects our ignorance rather than the likelihood of some kind of collective cosmic orgasm.

2. the alternative organic "softwire" or "wetware" option. This family of scenarios for high-functioning well-being takes either a) pharmacological or b) genetic guises; or combines both. But each variant assumes that organic biochemistry and molecular genetics can transcend their terrible genesis in a Darwinian world red-in-tooth-and-claw without significant *intra*cranial assistance from silicon. Critically, the biotech revolution will allow us progressively to rewrite our own genome. Later this century, new designer chromosomes can potentially be added to complement the expurgated code of our old DNA. Our post-human descendants may eventually opt to enjoy life lived on godlike planes of well-being - rather than simply ringing the changes within a Darwinian state-space of mediocre contentment or malaise. In these organic post-Darwinian scenarios, the imminent environmental threat of, say, acute tissue damage - or its neuropsychological counterparts - can be signalled by gradients of diminished bliss i.e. the functional analogues of aversive experience as we understand it now. This bliss-driven regime contrasts with the Darwinian order where eons of natural selection have spawned innumerable organisms driven mainly by gradients of pain, fear or gnawing dissatisfaction. As the impending reproductive revolution of designer babies unfolds, we are likely to pre-select enhanced "nice" rather than "nasty" genotypes for our future offspring.

> Few if any prospective parents will deliberately opt to raise depressive, anxiety-ridden children. This is not to deny that the reproductive psychology of our more distant descendants is anything other than speculative. Any contemporary account of the kinds of selection pressure at play in the era of (genetically) planned parenthood must be riddled with all kinds of conjecture. But if given the freedom to choose, most people would prefer intelligent, good-natured and happy phenotypes for their kids. When such choices become routinely available in the future, prospective parents will presumably choose genotypes to match.

* * *

CROSSING THE THRESHOLD

Humanity may or may not ever launch a global abolitionist project to eradicate suffering. The ethical urgency of engineering a cruelty-free world is not felt keenly by everyone. Hundreds of millions of people who *do* care deeply about others postpone hope of salvation to a mythical afterlife. Hence any more ambitious secular project to rewrite the vertebrate genome, switch to a cruelty-free diet of ambrosial vatfood, and perhaps redesign the planetary ecosystem is liable to sound even more infeasibly utopian than eradicating suffering in our own species. At present, such talk is confined to a few flaky dreamers. But in the impending post-genomic era of rational reproductive medicine, the *incremental* reduction of today's toll of human misery via the individual genetic choices of prospective parents may achieve results similar to the implementation of a grand abolitionist design - just more slowly. This overlap stands in contrast with the conquest

of suffering in non-human animals. The lion can lie down with the lamb only if whole populations are genetically reprogrammed for the designer habitats of our wildlife parks. So completion of any wider cross-species abolitionist enterprise may wait centuries until the task itself becomes technically trivial and the effort on our part negligible. Progress depends on how far and how fast the master species expands the "circle of compassion" across the phylogenetic tree.

Cynics will echo Bentham: "Dream not that men will move their little finger to serve you, unless their advantage in so doing be obvious to them. Men never did so, and never will, while human nature is made of its present materials." But this verdict may be (slightly) too pessimistic: most Darwinians *are* sometimes prepared to lift their little finger, so to speak, though it can be rash to count on us doing much more. Early pioneers of etherization, notably Morton and Jackson, may indeed have been consumed in later life more by sterile priority disputes than any sense of joy at the incalculable suffering they had relieved; but their flawed genius *did* recognise that the agonies of surgery were futile and preventable - and defeated them. Fortunately, the organisation if not the "present materials" of human nature will shortly be genetically upgraded. A greater capacity for altruistic finger-lifting will be offset by a diminished necessity for self-sacrifice. Either way, a ghastly legacy from our Darwinian past is poised to pass into evolutionary history. The heartbreaking suffering of the old order is destined to disappear, even if kinder and gentler implementations of its functional analogues are prudently retained.

Timescales for such a (hypothetical) major discontinuity in the evolution of life on Earth are inevitably uncertain. On a pessimistic analysis, centuries or even millennia of extreme global nastiness still lie ahead before any post-Darwinian transition is complete. Naturally, sceptics would argue that such a transition will never happen and predict that

suffering will endure as long as life itself - just as hard-headed "realists" like Velpeau argued as late as the early 1840s that pain and surgery were inseparable. The historical record certainly attests that formidable ideological as well as biomedical obstacles to the abolitionist project must be overcome if we are ever to live in a globally pain-free society. Our very language reflects a false dualist metaphysic of two different ontologies of suffering - the "mental" and the "physical" - whereas they share a common molecular substrate, texture of nastiness, and method of cure.

Yet a more optimistic family of scenarios can be modelled instead. The accelerating development of paradise-engineering technologies that are safe, life-enriching and *sustainable* may prove so empowering that we fast-track the emancipation of the living world from the pain chemistry of the old order - just as our Victorian forebears decided to abolish one whole class of ills of the flesh upon discovering controllable anaesthesia. Later this century, implementing the abolitionist revolution might conceivably take little longer than the revolutionary 19th century switchover to pain-free surgery or, less optimistically, the adoption of potent synthetic painkillers. The dawn of nanotechnology, quantum supercomputing and mature biotechnology prefigures an exhilarating abundance of ways to reshape the natural world - and detoxify "immutable" human nature. In principle, our genetically enriched descendants will be able to live sublime lives on a truly sublime planet - and perhaps populate the rest of the galaxy and beyond. After we cross the threshold of civilisation to a pain-free cosmos, it's even possible that the same cultural amnesia that has overtaken 19th century arguments against anaesthesia and analgesia will eventually befall arguments used against technologies to abolish emotional anguish too. In the aftermath of Year Zero, both the existential pain of old Darwinian life and the rationalisations that sustained it may pass into belated

oblivion. Alas, such a mental health revolution is a remote fantasy to countless suffering beings alive today.

UTOPIAN NEUROSCIENCE

SUPERHAPPINESS: Ten Objections To Radical Mood-Enrichment

INTRODUCTION

Transhumanists are ambitious. We want unlimited lifespan, unlimited intelligence, unlimited computer power. But this doesn't mean that we're ambitious about everything, for example height. Perhaps we want to be a bit taller, and we want to ensure that e.g. midgets have the opportunity to reach "normal" stature. Yet even in *Second Life*, or in tomorrow's immersive virtual realities, we don't for the most part want to be a thousand metres tall - despite freedom from the constraints of gravity. Of course, there are some very exotic creatures in *Second Life*: they might say the rest of us have stunted imaginations. But intuitively, there is quite a narrow optimum for body height. Moreover, height may be regarded as what economists call a "positional good". It's socially advantageous to be slightly taller than average; but if *everyone* were to become taller, then no one would be better off.

What about happiness - which I'm here going to use as a lame piece of shorthand for emotional well-being in the very richest sense. Is happiness best regarded as an absolute good, or as a positional good, like height? Is there an optimal range of hedonic tone that we should all aspire to - both for ourselves and for other sentient beings - just as there is for human body-stature under Earth's gravitational regime? Perhaps the heritable "set-point" of our hedonic treadmill might be genetically raised a little, just as some of us may

wish to be slightly taller. By the same token, perhaps victims of chronic low mood or anxiety disorders may benefit from gene-therapies or designer drugs so they can reach an idealised version of today's "normal" mental health - just as growth hormone can help the "abnormally" short.

There is a much more radical conception of well-being. Is happiness more akin to intelligence or lifespan, something that transhumanists should strive to enhance without limit - with the almost unimaginable implications that such an indefinite increase entails? The Transhumanist Declaration calls for the "well-being of all sentience". But well-being extends all the way from the barest contentment to peak experiences orders of magnitude more marvellous than unenriched humans can comprehend. Just how ambitious should rational agents aim to be in the scope of our reward pathway enhancements - both for ourselves and for other life-forms? What is *technically* feasible? What are the potential pitfalls? Could anything go catastrophically wrong if we succeed? Should some state-spaces of sentience be placed perpetually off-limits as too wonderful even to explore?

This question won't be answered here. As it happens, I tentatively predict that superintelligent posthumans will be animated by gradients of bliss that are literally *billions* of times richer than anything biologically accessible today; but whether or not such blissful civilisations exist beyond extremely low density branches of the universal wave-function is pure conjecture. Instead, I want to raise ten objections to the indefinite amplification of well-being - and sketch out ten possible replies.

1) The ETHICAL objection

Even talking about posthuman psychological superhealth is morally frivolous. Debating levels of posthuman bliss is akin to mediaeval theologians discussing the different levels of the celestial hierarchy - all those angels, archangels, cherubim, seraphim, and the like. Back in the real world, there are billions of sentient beings, human and non-human, who suffer varying degrees of ill-being - sometimes extreme ill-being. There's no sense in dwelling obsessively on the unpleasant side of life; but even the healthiest and happiest among us are in mortal danger of ending our lives "sans teeth, sans eyes, sans taste, sans everything". Ensuring a minimum of well-being for all sentient creatures is an immense enough technical and ideological challenge as it is. On a more positive note, much can be accomplished via incremental progress. Thus, the impending reproductive revolution of designer babies should lead to "unnatural" selection pressure against some of our nastier genes - allowing us to become smarter, happier, longer lived and, more controversially, perhaps nicer too. Critically for the well-being of all sentience, it's imperative that we stop killing and eating each other. If this utopian-sounding vision is to be realized, then cheap, palatable vatfood will need to replace flesh from factory farmed non-human animals in our diets; perhaps biotechnology plus market economics will succeed where moral argument fails. But ultimately, ending the Darwinian holocaust and securing the well-being of *all* sentient life entails an engineering mega-project: genomic rewrites, nanorobotics, and ecosystem redesign penetrating the furthest recesses of the oceans. So why ask for more? If and when the abolitionist project is complete, runs this objection, then we will have discharged all our ethical obligations. Or at least, only *after* suffering has been abolished throughout the living world should we consider truly revolutionary interventions to enrich our emotional lives. Maybe the critic here is a neo-Buddhist, or a negative utilitarian, or perhaps an enlightened bioconservative who shares

the desire to get rid of cruelty and [involuntary] suffering, but doesn't see any need to strive beyond its abolition.

POSSIBLE RESPONSE

I have a lot of sympathy with this objection. The moral urgency of using biotech to eradicate suffering *should* be carefully distinguished from speculative flights of fantasy about "paradise engineering" and so forth. Unless one is a strict classical utilitarian, the relief of suffering carries greater moral weight than enhancing well-being. So in that sense, the topic of this talk is *comparatively* unimportant - and arguably even morally trivial. However, it's hard to believe that there is anything inherently morally wrong with long-term planning. It's worth stressing that *none* of the things that transhumanists so ardently desire - unlimited lifespan, superintelligence, morphological freedom, novel sensory modalities and modes of consciousness, molecular nanotechnology, etc - will leave us significantly happier in the long-run *unless* we also redesign and recalibrate our hedonic treadmill. If we opt to do so, then it seems arbitrary to "freeze" its genetic calibration on the absolute minimum settings necessary to abolish the substrates of suffering - or to "lock in" merely a modest increase in the upper range of hedonic tone beyond that bare minimum. Why such poverty of ambition?

Clearly, this isn't the place for a philosophical treatise on the nature of value. Yet one needn't be any kind of hedonist or classical utilitarian to recognize that there are intimate links between the creation of life-long emotional well-being and the creation of value. Provisionally, let's just make a weak but still fertile working assumption. *Other things being equal*, the most rewarding music, comedy, art, computer game, virtual reality software, personal relationship, etc, is more valuable than its less enjoyable counterpart. A world with ever more richly rewarding experiences is, *other things being equal*,

preferable to comparatively emotionally impoverished worlds. Of course, as the critic will rightly insist, very often things *aren't* equal. We can all cite multiple counterexamples. But intuitively, it's departures from the default assumption that need justifying, not the default assumption itself.

Perhaps this response is a bit abstract. So for illustrative purposes, try to recall for a moment the most wonderful "peak experience" of your life. Imagine that its neuronal substrates could be identified, genetically enhanced, and conditionally activated at will. Assume, more controversially, that utopian neuroscience will be able to identify the complex molecular signatures of *any* valuable human experience and amplify their biological substrates. Will post-human experiences that *seem* millions of times more valuable than today's peak experiences *really* be millions of times more valuable? Or instead, as the moral nihilist claims, are value-judgements by their very nature truth-valueless? In other words, is this debate all just idle opinion, since the fact-value distinction is logically unbridgeable? Here I'll leave the question open; but if, provisionally, we may assume that some of our experiences are more valuable than the best experiences of, say, an earthworm, then one may wonder whether mature posthuman modes of sentience might not proportionately be more valuable than ours. So *if* value can be naturalised and biologically enhanced, then why not plan how to create a sustainable abundance of its molecular substrates by the most computationally effective means? Or at least, before passing judgement on posthuman well-being, let's first discover what we're missing.

2) The TECHNICAL objection(s)

It's intelligible to speak of becoming a thousand times taller - though the biomechanics might pose a problem. But does it even make sense to speak of becoming a thousand times happier - except as a rhetorical device? Can happiness sensibly be treated as a biological category at all? Is emotional well-being really a natural phenomenon that can be objectively measured and quantified? Do happiness and other desirable states of mind really have well-defined neurological substrates that can be selectively amplified indefinitely? Is there even a unidimensional pleasure-pain scale?

POSSIBLE RESPONSE

"Happiness" is indeed a crude label, evoking everything from the noblest triumphs of the human spirit to a nice day at the seaside. Identifying the molecular correlates of our emotional states in terms of receptor-density and neurotransmitter occupancy ratios, alternate splice variants, phosphorylated proteins, gene expression profiles, etc, is a daunting challenge for computational neuroscience. In future, our conceptual scheme for the emotions will need to be enriched along with our emotional repertoire itself. Eventually, some of our nastier emotions may be abolished: their fitness-enhancing computational role on the African savannah is now redundant. Others may be recalibrated: the posthuman analogue of boredom, for instance, needn't feel unpleasant to retain an analogous functional role; subjectively, its posthuman analogues need feel only *comparatively* less interesting than spellbound fascination. More speculatively, genes for novel core emotions may be spliced into the limbic pathways: our emotional palette may be genetically expanded. Whether a unidimensional pleasure-pain scale exists is controversial. In rats, at least, the ultimate "hedonic hotspots" are a cubic millimetre of tissue in the ventral pallidum and another comprising medium spiny neurons in the rostrodorsal region of the medial shell of the nucleus accumbens. But even if it transpires there is nothing akin to the final common pathway of reward in the

human brain, such complexity wouldn't fundamentally change the technical feasibility of indefinite emotional growth. As brain-scanning technology becomes ever more sophisticated and finer-grained, we'll be able to identify the multiple neural correlates of well-being and selectively "over-express" them in ways that transcend old-fashioned environmental tinkering.

More concretely, brainy "Doogie mice" with an extra copy of the NR2B subtype of NMDA receptor suffer from a chronically increased sensitivity to pain. That's a nasty example. Conversely, neuroscientists can in principle genetically splice in multiple extra copies of other subtypes of receptor e.g. the *mu*-opioid receptor, implicated in hedonic tone. Gene therapy can already be used experimentally to multiply a thousandfold the number of opioid receptors expressed on the surfaces of nerve cells carrying pain signals back and forth between an arthritic joint and the spinal cord; the pain is banished. In future, nerve cell responsiveness to naturally occurring endogenous opioids can be increased via receptor enrichment in the brain too. In principle, we can modulate their lifelong "over-expression", intermittently heightened (or gently diminished) by whatever kinds of personal and environmental contingencies we judge fit. Both functionally and anatomically, our reward pathways can be made "bigger and better". But intelligent emotional self-mastery will involve re-engineering the mind-brain so we derive the most intense rewards from activities we deem most lastingly worthwhile: i.e. prioritising our higher-order desires over legacy first-order appetites. Natural selection has "encephalised" our emotions to benefit our genes. Rational agents can "re-encephalise" our emotions to benefit us.

Long-term, perhaps the big challenge technically will not be amplifying "reward" circuitry or genetically re-editing "punishment" circuitry *per se*. The real challenge ahead could be doing so in ways that are socially responsible, intellectually insightful, sustainably

empathetic, preserve nurturing behaviour, avoid triggering psychosis or mania, and don't provoke adverse side-effects - either for the enriched individual or for society as a whole. These are severe constraints. For example, a problem with existing so-called antidepressants is not just that they are often ineffective and "dirty"; they can also trigger mania in the genetically susceptible instead of high-functioning well-being. [see also "Touched with Fire: Manic-Depressive Illness and the Artistic Temperament" (1993) by Kay Redfield Jamison] Becoming truly "better than well" entails not just an extended lifetime of feeling on-top-of-the-world, but retaining insight, intellectual acuity and social intelligence. In mania, critical judgement is lost.

I'm making a controversial assumption here. The traditional way to produce, say, aesthetic beauty is to create a painting or a sculpture that stirs a rewarding aesthetic response in one's audience. Hence the decorative arts. The advanced way to create awe-inspiring beauty is to use brain-scanning technology, identify the neural signature of aesthetic experience, purify its biomolecular essence and then amplify its substrates. Transcendentally beautiful experiences on-demand can then be selectively triggered far more potently than today - perhaps managed from a user-friendly interface as intuitive as your iPad, perhaps thought-activated, or perhaps stimulus-driven as now. Hence the claim that posthumans may have the innate capacity for aesthetic experiences that are billions of times more beautiful than anything accessible at present - possibly more so after the imbecilic constraints of the human birth-canal are overcome: artificial wombs are no more "unnatural" than artificial clothes. It's said that mystics and poets can "see the world in a grain of sand". In the future, why can't the rest of us raise our aesthetic default-settings so that our set-point of beauty-recognition fluctuates around a vastly higher baseline? Posthuman aesthetic appreciation (almost) certainly won't be uniform -

an undiscriminating cosmic "wow". But on at least one family of scenarios, everyday posthuman life may consist entirely of gradients of the sublime.

Or to use another speculative example: the traditional route to spiritual experience is via meditational discipline and prayer. The futuristic route – *if* one thinks spirituality is a valuable dimension of experience - is to identify the neural substrates of spiritual experience, perhaps even the neural substrates of divine revelation and the experience of God, and then amplify them, stripping out the incidental junk and amplifying both their molecular essence and the metabolic pathways that regulate their expression. It should be *technically* feasible for our descendants to enjoy daily experiences of the divine billions of times more profound than anything physiologically possible today. This argument can also be used to rebut the charge that transhumanists are all soulless materialists oblivious of the richer dimensions of experience. Some of us do indeed inhabit a spiritual wasteland. But ironically, it's religious bioconservatives who prevent the godless from communing with the divine; and it's traditional mystics who prevent the rest of us from accessing the technologies of mystical experience.

Admittedly, this kind of neurological reductionism can easily smack of phrenology. A critic might mock that one might as well speak of the brain having a "humour centre" - and "enhancing its biological substrates" too. Well, funnily enough, the brain *does* have a humour centre, not just functionally, but anatomically. Crudely stimulating a region of the left basal temporal cortex induces an *undiscriminating* sense of everything being hilariously funny. But instead of the crude neurostimulation of undiscerning mirth, our descendants [or future selves?] may decide to recalibrate the default-setting of their native humour response. Today we describe some people as temperamentally humourless; other people are prone to see the funny side of life. Well, assuming that a

keen sense of humour is valuable, what if we could reset our own propensity to find things funny? Is there an optimum humour-range for a given environment - low and infrequently expressed for brutish Darwinian life, modestly higher for posthumans? Or should the range of our sense of humour be ratcheted up indefinitely when conditions permit? For if we can identify the neural substrates of humour, then we can biologically enrich these substrates indefinitely too. In theory, given a post-human world without suffering, our descendants could appreciate humour many times more richly hilarious than anything possible now. The traditional route to comic genius has been to crack funnier jokes or write a comic masterpiece. The sophisticated posthuman route to cultivating a fantastic sense of humour is not (just) to be wittier; it's to amplify and enrich the neural substrates of amusement. This increase might seem a recipe for inanity. On the other hand, recall Wittgenstein's remark that good philosophical work could be written consisting entirely of jokes. In a Darwinian world full of suffering, this prospect might seem obscene; tomorrow such a mind-set may be perfectly appropriate.

Okay, that's a whimsical example. Yet exactly the same reasoning holds for information-signalling gradients of bliss; and given even a weak version of the pleasure principle, the adoption of a motivational system based on gradients of bliss is more sociologically plausible than an enhanced propensity to find everything funny. Thus the archaic route to improving well-being has been through manipulating the external environment - tempered on occasion by incompetent alcohol abuse. Environmentalist utopias invariably run aground on human nature and the inhibitory feedback mechanisms of the hedonic treadmill. Their polar opposite is wireheading: direct stimulation of the reward centres. Wireheading is effective but indiscriminate. It's not an evolutionarily stable solution. The mature posthuman route to happiness will presumably continue to embrace environmental improvement; but an environment perceived or simulated through what

kind of affective filters? Perhaps posthuman sensory input will be processed via an innately blissful medium of thought. Of course it's far harder technically to amplify gradients of complex "thick" social emotions than it is to amplify raw orgasmic bliss, or even spiritual raptures. Yet such amplification can be accomplished if so desired as our neuroscanning technology and gene-therapies improve. Technologies of sustained cerebral bliss are feasible in principle. The challenge is to use them wisely on a planetary scale and beyond. And unfortunately "wisdom" here isn't well-defined.

3) The 'EXPERIENCE MACHINE' objection

According to this objection, the prospect of "artificially" ratcheting up our hedonic set-point via biotech interventions just amounts to a version of Harvard University philosopher Robert Nozick's hypothetical Experience Machine. Recall the short section of *Anarchy, State, and Utopia* (1974) where Nozick purportedly refutes ethical hedonism by asking us to imagine a utopian machine that can induce experiences of anything at all in its users at will. A full-blown Experience Machine will presumably provide superauthenticity too: its users might even congratulate themselves on having opted to remain plugged into the real world - having wisely rejected the blandishments of Experience Machine evangelists and their escapist fantasies. At any rate, given this hypothetical opportunity to witness all our dreams comes true, then most of us wouldn't take it. Our rejection shows that we value far more than mere experiences. Sure, runs this objection, millennial neuroscience may be able to create experiences millions of times more wonderful than anything open to Darwinian minds. But so what? It's mind-independent facts in the real world that matter - and matter in some sense to *us* - not false happiness.

POSSIBLE RESPONSE

This Objection isn't fanciful. In future, technologies akin to Experience Machines will probably be technically feasible, perhaps combining immersive VR, neural nanobots and a rewiring of the pleasure centres. Such technologies may conceivably become widely available or even ubiquitous - though whether their global use could ever be sociologically and evolutionarily stable for a whole population is problematic. [If you *do* think Experience Machines may become ubiquitous, then you might wonder (shades of the Simulation Argument) whether statistically you're most probably plugged into one already. This hypothesis is more compelling if you're a life-loving optimist who thinks you're living in the best of all possible worlds than if you're a depressive Darwinian convinced you're living in the unspeakably squalid basement.]

However, feasible or otherwise, Experience Machines *aren't* the kind of hedonic engineering technology we're discussing here. Genetically recalibrating our hedonic treadmill at progressively more exalted settings needn't promote the growth of escapist fantasy worlds. Measured, incremental increase in normal hedonic tone can allow (post)humans to engage with the world - and each other - no less intimately than before; and possibly more so. By contrast, it's contemporary social anxiety disorders and clinical depression that are associated with behavioural suppression and withdrawal. Other things being equal, a progressively happier population will also be more socially involved - with each other and with consensus reality. At present, it's notable that the happiest people tend to lead the fullest social lives; conversely, depressives tend to be lonely and socially isolated. Posthuman mental superhealth may indeed be inconceivably different from the world of the happiest beings alive today: meaning-saturated and vibrantly authentic to a degree we physiologically can't imagine. Yet this wonderful outcome won't

be - or at least it *needn't* be - explicable because our descendants are escapists plugged into Experience Machines, but instead because posthuman life is intrinsically wonderful.

Perhaps. The above response to the Experience Machine objection is simplistic. It oversimplifies the issues because for a whole range of phenomena, there is simply no mind-independent fact of the matter that could potentially justify Experience Machine-style objections - and deter the future use of Experience Machine-like technologies for fear of our losing touch with Reality. Compare, say, mathematical beauty with artistic beauty. If you are a mathematician, then you want not merely to experience the epiphany of solving an important equation or devising an elegant proof of a mathematical theorem. You also want that solution or proof to be really true in some deep platonic sense. But if you create, say, a sculpture or a painting, then its beauty (or conversely, its ugliness) is inescapably in the eye of the beholder; there is no mind-independent truth beyond the subjective response of one's audience. For an aesthete who longs to experience phenomenal beauty, there simply isn't any fact of the matter beyond the quality of experience itself. The beauty is no less real, and it certainly *seems* to be a fact of the world; but it is subjective. If so, then why not create the substrates of posthuman superbeauty rather than mere artistic prettiness?

There's also a sense in which our brains already *are* (dysfunctional) Experience Machines. Consider dreaming. Should one take drugs to suppress REM sleep because our dreams aren't true? Or when awake, should one's enjoyment of a beautiful sunset be dimmed by the knowledge that secondary properties like colour are mind-dependent? [Quantum theory suggests that classical macroscopic "primary" properties as normally conceived are mind-dependent too; but that's another story.] If you had been born a monochromat who sees the world only in different shades of grey, then as a hard-nosed scientific

rationalist, should you reject colour vision gene therapy on the grounds that phenomenal colours are fake - and grass isn't intrinsically green? No, by common consent, visual experience enriches us, even if, strictly speaking, we are creating reality rather than simulating and/or perceiving it. Or to give another example: what if neural enhancement technologies could controllably modify our aesthetic filters so we could see 80-year-old women as sexier than 20-year-old women? Is this perception false or inauthentic? Intuitively, perhaps so. But actually, the perception is no more or less authentic than seeing 20-year-old women of prime reproductive potential as sexier. Evolution has biased our existing perceptual filters in ways that maximised the inclusive fitness of our genes in the ancestral environment; but in future, we can optimise the well-being of their bodily vehicles (i.e. us). Gradients of well-being billions of times richer than anything humans experience are neither more nor less genuine than the greenness of grass (or the allure of Marilyn Monroe). Could such states become as common as grass? Again, I suspect so; but speculation is cheap.

4) The INAPPROPRIATE BEHAVIOUR objection

Some critics are concerned that promoting superhappiness may lead to what one may call, informally, "inappropriate" behavioural responses. The scare quotes are necessary here because our sense of appropriateness is systematically biased by our evolutionary past. *All* our intuitions are tainted. But to give a concrete example of inappropriateness as commonly understood: suppose that you were to fall under the proverbial bus. Even if the accident didn't cause you to suffer, would you really want your friends to stay happy on hearing the news, despite your misfortune? Less dramatically, even as life gets better, we will presumably still make mistakes. There will be setbacks and disagreements,

perhaps strenuous disagreements. Negative feedback is vital to preserving critical insight. Even if suffering as we understand it today is abolished, then something analogous to anxiety and discontent will surely be needed as the engine of progress?

POSSIBLE RESPONSE

A counterargument here is that even radically enriching hedonic tone can preserve a full range of negative feedback mechanisms. Optionally, our range of hedonic contrast can actually be increased - even if posthumanity's genetically-predetermined affective floor is set higher than today's affective ceiling. For most purposes, however, fine gradations and nuances of hedonic tone can presumably serve well enough. Enriched posthumans can still be informationally sensitive to good and bad stimuli, even if our baseline hedonic set-point is elevated orders of magnitude beyond the contemporary norm. We can still experience the *functional analogues* of some of today's negative feelings even as the textures of consciousness get ever better.

Optionally as well, the greater part of our existing preference architecture can be preserved. If you prefer Beethoven to Brahms, or philosophy to pushpin, then enriching hedonic tone can still leave your preference architecture more-or-less intact. Hedonic contrast-ratios can in principle be conserved even if the scale is itself is recalibrated. Now of course there are serious grounds for asking whether we really want to leave our existing preference architecture unchanged. After all, a lot of our core desires and preferences are quite unpleasant: they have been shaped by humanity's red-in-tooth-and-claw evolutionary history to allow selfish DNA to make more copies of itself. Perhaps a lot of our nastier preferences should be abolished, not just recalibrated. But "preference conservatism" is consistent with the hedonic enrichment technologies canvassed here - at least as a theoretical option. In practice, a mood-congruent

conceptual revolution would (presumably) accompany global hedonic enrichment. Its nature and scope we can now scarcely imagine.

And what of mourning? Should grief be abolished before we have conquered death - a far more formidable biotechnological challenge than enriching subjective well-being? Well, if I were to fall under the proverbial bus, then I would indeed want, selfishly for sure, that such an accident diminish my friends' well-being. Otherwise I'd find it hard to conceive of them as friends. But if one truly values one's friends, then surely one wouldn't - surely one *shouldn't* - want them ever to *suffer* on one's account. A conditionally-activated reduction of their well-being, I'd argue, is the most one can appropriately ask for. If we're talking of "inappropriate" responses, then a prime candidate instead might be the Darwinian desire for others to suffer, including on occasion those we nominally "love".

More prosaically, one may hope that transhumanists will be careful crossing roads.

5) The CHARACTER-SAPPING objection

A fifth worry is that gradients of extreme well-being may be bad for our character. One thinks of partygoers who pursue a "hedonistic" lifestyle in the popular sense of the term, or drug addicts, or feckless parents who neglect their children. A futuristic example of character-deterioration might be variants of wireheading - perhaps in the guise of a neurochip that delivers undifferentiated bliss. In general, episodes of "unnaturally" extreme well-being tend to promote selfishness, egotism, impaired judgement, risk-taking, manic behaviour - and a lack of consideration for others. Surely, runs this objection, the future of life in the universe isn't foreshadowed by analogues of wireheading, heroin and crack cocaine?

POSSIBLE RESPONSE

Indeed not. A counterargument here is that true hedonic engineering, as distinct from mindless hedonism or reckless personal experimentation, can be profoundly good for our character. Character-building technologies can benefit utilitarians and non-utilitarians alike. Potentially, we can use a convergence of biotech, nanorobotics and information technology to gain control over our emotions and become better (post)human beings, to cultivate the virtues, strength of character, decency, to become kinder, friendlier, more compassionate: to become the type of (post)human beings that we might aspire to be, but aren't, and biologically couldn't be, with the neural machinery of unenriched minds. Given our Darwinian biology, too many forms of admirable behaviour simply aren't rewarding enough for us to practise them consistently: our second-order desires to live better lives as better people are often feeble echoes of our baser passions. Too many forms of cerebral activity are less immediately rewarding, and require a greater capacity for delayed gratification, than their lowbrow counterparts. Likewise, many forms of altruistic behaviour - giving even a paltry 10% of one's income to Oxfam, for instance - are less rewarding than personal consumption. But in future it should be feasible to derive gradients of richly flavoured bliss from studying sixteen hours a day, or being angelically kind and "insanely" generous. Posthuman control of our emotions should allow us to amplify the character traits that we regard as admirable, overcoming the limitations of Darwinian minds in ways that environmental manipulation alone cannot match. In a superficial manner, *Second Life* allows us to assume the personae of the type of beings we'd ideally like to be; but future enrichment technologies can empower us to become ideal beings in our First Life incarnations too.

One worry about such a rosy scenario is worth noting. Will genetically-underwritten superhappiness rob us of the opportunity for personal growth, character-building struggles against adversity, and the chance to practise heroic self-sacrifice?

Well, it was said of the late Madame de Staël that she would throw all her friends into the water for the pleasure of fishing them out again. Certainly, a civilisation run on gradients of superbliss would have no need of heroism in the traditional sense. But lifelong mental superhealth needn't turn us into milksops. Quite the reverse: superenriched reward circuitry promises to make us stronger-minded and thereby *more* able to fulfil our life projects - and promote the well-being of others. It's the clinically depressed and other victims of "learned helplessness" who give up too easily: sadly, there's more than a grain of truth in the popular stereotype of depressives as "weak". By contrast, genetically predestined superhappiness promises tomorrow's children "larger-than-life" personalities, uncompromising integrity, and a willpower stronger than anything neurologically feasible today. Potentially, superhappiness will also enable *non*-utilitarians to realise their projects more effectively.

Obviously, it remains an entirely open question whether we will in fact use such technologies prudently - if we use them at all. But given the terrible emotional shipwrecks of Darwinian life, why shouldn't we (re)design our personalities to at least exacting specifications that we demand of, say, our cars? Why shouldn't post-Darwinian life be robust, exhilarating *and* crash-proof?

6) The 'STUCK-IN-A-RUT' objection

This is the worry that directly enhancing well-being by neurobiological interventions will lead to a civilisation becoming trapped in a suboptimal rut. This *isn't* the historically-based objection that pursuing utopian visions inevitably leads to nightmarish dystopias. Indeed, perhaps there's an important sense in which nothing can go wrong, in the *ordinary* unpleasant sense of "going wrong", if you replace the biological substrates of suffering and malaise with adaptive gradients of bliss. But that's the underlying point of this objection: reaching too avidly or prematurely for what is on offer may lock us permanently into a *local* optimum that prevents us from maximising our full potential - whatever that full potential might ultimately be. One might think here of long-acting analogues of soma, Aldous Huxley's supposedly ideal pleasure drug, or more refined and globally sustainable analogues of wireheading. No, this isn't the gulag; but surely transhumanists are entitled expect more?

POSSIBLE RESPONSE

Again, this scenario can't be excluded. But its very conceivability is one reason why humanity would do well to think ahead strategically rather than collectively "stumbling on happiness", to borrow Daniel Gilbert's hopeful phrase. The credence we assign to such global-rut scenarios depends on the kinds of biologically enhanced well-being, if any, our descendants decide to embrace. For example, perhaps genetically encoding the substrates of contemplative, mystical well-being sounds potentially attractive to people of a troubled cast of mind today, especially the temperamentally anxious and angst-ridden. Buddhists, of course, identify the extinction of desire with Nirvana. However, globally engineering this kind of lifelong bliss might indeed lead to behavioural stagnation - and a whole civilisation in perpetual stasis - even if it delivers unprecedented spiritual growth. Now in response, one might say: so what? But rather than opting to become

constitutionally serene, perhaps policy-makers persuaded by the stuck-in-a-rut objection should instead promote elements of what (very) crudely one may label dopaminergically-enhanced well-being - with its tendency to enhanced novelty-seeking, exploratory behaviour and intellectual curiosity. Unfortunately, this kind of well-being has multiple pitfalls of its own. So modes of biological well-being radically different from *any* contemporary human stereotype deserve to be comprehensively researched too. But at least in the medium-term, "outward-looking" futures are presumably more likely to unfold than introverted civilisations based on varieties of meditative bliss. For an ecological niche remains to be populated in the shape of our local galaxy. Vacant ecological niches tend to get filled. Unless we were *all* to become contemplatives, or *all* opt to dwell in immersive virtual reality etc, then our descendants will probably radiate out and colonise the accessible universe within our forward light-cone. What they'll do next is unclear.

7) The SOCIALLY DISRUPTIVE objection

Biologically enhanced well-being might exert catastrophically disruptive effects on the wider structure of society. This objection is the very opposite of the commonly expressed concern that "artificial" happiness will make us contented dupes more vulnerable to control by the ruling elites (*cf*. Huxley's soma). Instead, the argument here is that super-enhanced well-being would be disruptive of the social pecking-order - the dominance hierarchies on which all existing social primate societies are based. Low mood and submissive behaviour evolved in social mammals as an adaptation to group living - itself an adaptation against predators. To abolish the substrates of social anxiety/low

mood/subordinate behaviour might turn us all into potential "alphas". Rampant alpha-plus behaviour would make society ungovernable, even in the minimal libertarian sense.

POSSIBLE RESPONSE

The counterargument here is that such scenarios just illustrate the importance of far-sighted planning. Uncontrolled mass mood-elevation - as distinct from emotional enrichment - might indeed provoke socially disruptive hypercompetitive behaviour, thereby worsening global catastrophic risk. Competitive alpha-male dominance behaviour in an age of nuclear, biological and chemical weaponry is perhaps the gravest threat to life on Earth. So this objection is actually much more serious than it sounds. On the other hand, mood-elevation can also be empathetic and pro-social. "Mirror neurons", for instance, can be multiplied and functionally amplified as well as hedonic tone, thereby enhancing our propensity to cooperative behaviour. Likewise, long-acting designer "hug-drugs", safe and sustainable analogues of MDMA and its congeners, are feasible too - as are their genetic equivalents. Social cohesion may thereby be biologically enhanced. The possible ramifications of radical mood-enrichment for existing social hierarchies are poorly understood because such scenarios have never been systematically modelled. Yet this neglect is no reason permanently to "freeze" the greater part of humanity in the biology of subordinate timidity - the condition of many "low ranking" social primates in the world today.

8) The SELECTION PRESSURE objection

It may be technically feasible, in the short run, directly to amplify the substrates of well-being across the lifespan. It may even be technically feasible to elevate our normal

hedonic set-point through somatic or germline gene-therapy. But in the long run, there will be selection pressure against escalating gradients of superhappiness. So the scenarios discussed here aren't realistic.

POSSIBLE RESPONSE

In a post-ageing world centuries hence, reproduction will need to be exceptionally rare and centrally-controlled - regardless of whether or not our quasi-immortal descendants practise hedonic engineering. Otherwise the Earth (or in theory our galaxy or local galactic supercluster, etc) will exceed its physical carrying capacity. However, this kind of speculation involves very complex arguments on the nature of selection pressure in an era when traditional childbearing has more-or-less ceased.

In the meantime, there *will* be intense selection pressure, but there are powerful grounds for believing such selection pressure will work against any genotypes/allelic combinations predisposing to Darwinian unpleasantness in all its forms. This is because we are on the brink of a reproductive revolution of designer babies. Prospective parents will shortly be choosing the personalities/genetic make-up of their future children rather than playing genetic roulette. As responsible child-planning becomes common, and preimplantation genetic screening becomes routine, severe selection pressure will come into play against genes/genotypes predisposing to the darker modes of human experience. This isn't the place to attempt formal game-theoretic modelling or a treatise on posthuman population genetics. So for illustrative purposes just imagine: If you were a prospective parent choosing the genetic make-up of your future children, what genetic dial-settings would you opt for? You wouldn't want genotypes predisposing to anxiety disorders, depressive illness, schizoid tendencies, and other undisputed pathologies of mind; but how high (or in theory, how low?) would be the settings you'd prefer for your children's normal

hedonic tone? Cross-culturally, parents typically say they want their children to be happy, albeit "naturally" so; but how happy? Redheads may prefer to have red-headed children; but few depressives will want depressive children. All that's needed for selection pressure to get to work here is a partially heritable slight preference for children who are modestly more temperamentally happy [or less gloomy] than oneself. Selection pressure is fundamentally different when evolution is no longer "blind" and random with respect to what is favoured by natural selection - i.e. when genes/allelic combinations are chosen/designed *in anticipation of* their likely effects. Such selection pressure is already manifest in non-human domestic animals; it will shortly come into play in humans. Hence we are entitled to speak of an impending post-Darwinian era - not because selection pressure will be absent (on the contrary!) but because we are poised to switch from the era of "natural" to "unnatural" selection.

This momentous reproductive shift certainly doesn't exclude the likelihood of continuing selection pressure against some modes of subjective well-being e.g. undifferentiated bliss. Thus wireheads and their natural analogues, for instance, will presumably always be at a reproductive disadvantage. But a motivational system of high-functioning gradients of superhappiness may be extremely adaptive *if* that's the behavioural phenotype we want for our children. Children genetically predisposed to be abundantly happy and affectionate are more rewarding to raise than surly, depressive children. It should be stressed that this optimistic scenario *doesn't* mean that posthuman social life will resemble a communal hug-in or an MDMA-driven rave. There can be functional analogues of depressive realism even in paradise.

9) The RISKS OF HASTE objection

The priority should be superintelligence, not superhappiness. Only after we are intelligent enough to understand the implications of what we're doing should we explore radical mood-enrichment. The risks of acting prematurely and building a fool's paradise are too great.

POSSIBLE RESPONSE

As it stands, this objection may well be correct. Only superintelligence can maximise the utility function of the universe. But emotional enrichment - as distinct from crude pleasure-amplification - is itself presumably a critical ingredient of superintelligence. So we should take care to avoid constructing a false dichotomy: mature superintelligence will presumably entail an unimaginably enriched capacity for empathetic understanding - a "God's eye view". This point is relevant because - given some fairly modest assumptions and even the slightest sense of moral urgency - we *should* be prepared, if necessary, to take risks to eliminate a terrible scourge, to prevent suffering and cruelty to our fellow creatures, or to act when the risks of inaction are greater than action. What's important is assessing risk-reward ratios. One obvious parallel is ageing. Bluntly, we are all dying. If you regard ageing as a horrible disease, then you may be prepared to run risks to retard its progression. Thus one might take a daily cocktail of supplements (e.g. resveratrol, selegiline, etc) that increases lifespan and life expectancy in "animal models", but whose efficacy and long-term safety is unproven in controlled longitudinal studies in humans. Perhaps the minority of "healthy" [i.e. dying] humans who adopt such a regimen misjudge the risk-reward ratio involved; but if so, the error doesn't reside in a willingness to take calculated risks - merely in their miscalculation. There are perils in inertia no less than in initiative. Likewise, current victims of intractable pain or chronic depression, whose quality of life is meagre (or worse), may justifiably take more therapeutic risks, and explore more experimental treatments, to alleviate their distress

than the psychologically robust who already enjoy life to the full - by mediocre Darwinian standards, at any rate.

A complication of this analysis is that *all* enhancement technologies may be viewed as remedial therapies by the enlightened standards of our successors. Yet there is a fundamental difference between taking risks to alleviate serious disease, chronic pain syndromes or prolonged psychological distress and taking risks to enhance pre-existing well-being.

Sadly, there *aren't* any short-cuts. So in that sense the objection is unanswerable. Current recreational euphoriants, for instance, may give their users a faint, fleeting, shallow foretaste of posthuman bliss; but for the most part, they activate the hedonic treadmill - and produce nasty side-effects, insidious or otherwise. It's worth recalling that some very smart people have been seduced. Twenty-eight-year-old Viennese neurologist Dr Sigmund Freud wrote a paean of scholarly praise for the therapeutic benefits of cocaine, newly isolated from the coca plant. Bayer introduced Heroin as a non-addictive remedy for coughs. And in the words of one intravenous heroin user: "It's so good. Don't even try it once." Any potential wonderdrug or gene-therapy that promises a miraculous breakthrough to posthuman nirvana needs to be investigated with *both* extraordinary urgency and extraordinary scepticism.

10) The CARBON CHAUVINISM Objection

This talk has focused on enriching the "biological substrates" of emotion. Yet given some quite widely accepted functionalist arguments in contemporary philosophy of mind, why not scan, digitize, and "upload" ourselves into silicon or another medium - and then

reprogram ourselves? The exponential growth of computing power promises to endow uploads with the self-reprogramming ability to cure ageing, infirmity and disease; attain true superintelligence; enjoy total morphological freedom; and amplify our reward pathways too. If the exponential growth of [inorganic] computer power continues unchecked, then this transformation may be only decades away - not the millennia that a meatware transition to posthumanity would presumably entail.

POSSIBLE RESPONSE

The range of opinions among transhumanists on uploading runs all the way from those who think it's inevitable to those who view it as some kind of millennialist death cult. If your overriding ethical goal is "merely" to eradicate suffering, then uploading could almost certainly achieve its abolition - one way or the other. However, most people aren't negative utilitarians. If you want "your" upload to achieve supersentience as well as superintelligence, or to enjoy posthuman levels of well-being, to achieve quasi-immortality, or simply to conserve your identity as understood today, then the existential risk posed by uploading is immense - perhaps the biggest existential risk the human species has ever contemplated. So *before* embarking on anything so revolutionary, it's vital that we have a compelling theory of consciousness - and a mathematically exact description of its myriad textures - on pain of creating zombies. Maybe you feel 99% certain that the sceptics are wrong, e.g. neurophilosophers who believe that unitary consciousness depends on quantum coherence, and hence any aspiration to non-trivial digital sentience falls foul of the "von Neumann bottleneck". But either way, the postulation of sentience *in silico* is not a testable scientific hypothesis. So advocates of uploading are placing a lot of faith in a metaphysical theory. Of course, the conviction that *anyone* else is conscious is a metaphysical theory too, albeit less controversial.

By way of [false] analogy, consider the game of chess. Imagine a misguided philosopher who claims that what matters when playing chess is not just the sequence of moves, but also the particular textures of the individual chess pieces; and that chess games played with wooden or metal pieces, say, or games played online via computer, can be different in character even if the sequence of moves played is the same. Surely, we would say, this fellow is simply confused: he is missing the point of chess. The particular textures of the pieces, and even the complete absence of any such textures in computer chess matches, are unimportant, since the textures, coloration, and physical composition (etc) of the pieces are functionally irrelevant to the gameplay - a mere implementation detail. The same game of chess can be multiply realised in different physical substrates. Now consider uploading. Imagine again a naïve-sounding bioconservative who insists that what matters for successful uploading is not just the behaviour [and behavioural dispositions] of hypothetical uploads, but also the particular textures [aka qualia: "what it feels like"] of their mental-cum-perceptual states. Now in one sense, yes, the phenomenal textures [if any] and substrate composition of a hypothetical upload are mere implementation details - functionally irrelevant insofar as the upload has the right functional architecture to support input-output relations identical to its meatworld counterpart. ["If it walks like a duck, quacks like a duck...", etc.] Yes, *if* we were exhaustively defined by our (macro)behavioural patterns, then the spectre of inverted qualia, "Martian pain", absent qualia, and so forth, is of no consequence. But in another, critically important sense, the analogy with chess fails. "What it feels like" to be me is of the very essence of my personal identity: it's not a trivial implementation detail, but definitive of who one is - one's intrinsic nature. If we had the slightest idea how to scan, record and digitise qualia, then uploading might be feasible; but alas we don't. It is

scarcely possible to overstate our scientific ignorance of consciousness. For now, at least, uploading belongs to the realm of science fantasy rather than science fiction.

However, let's assume for the sake of argument that sentient uploading will in future be technically and societally feasible - perhaps using quantum computers with a non-classical architecture. Given a mass-upload scenario, the fate of meatware "left behind" is unclear. Unless traditional organic life is to be liquidated - i.e. "destructive" uploading, the final solution to the organic life problem - then primordial Darwinian organisms will still need to be "rescued" by their postorganic descendants. So here we come back to the biological substrates of consciousness with which we began.

CONCLUSION

Superintelligence, Superlongevity and Superhappiness?

Centuries of technological and socio-economic "progress" haven't left us discernibly happier in the course of a lifetime than our hunter-gatherer ancestors. There's no compelling scientific evidence that thousands of years of reshaping our environment has cheated the hedonic treadmill one iota. Will the future resemble the past? Almost certainly not. Tomorrow's neuroscience promises to revolutionise subjective well-being, both individually and for our species as a whole. More speculatively, we may overcome our anthropocentric biases and enrich the rest of sentient life too.

But by how much? Unlike computing power, an exponential growth of happiness is (presumably) impossible, short of technologies beyond human imagination. Yet securing even an approximate linear growth of its biomarkers would represent a stunning discontinuity in the history of life to date. Posthuman versions of the Goldilocks zone - "not too hot, not too cold" - could *potentially* exceed the hedonic range adaptive for our

hominid ancestors by several orders of magnitude, if not more. Will our posthuman descendants eventually decide, to echo Bill McKibben, "Enough!". Possibly; but if so, it's unclear how, when and why.

It's worth emphasising that the sorts of scenarios for posthuman mood-enrichment explored here *aren't*, for the most part, an alternative to other transhuman scenarios of our future, notably superintelligence and superlongevity. On the contrary, a fine-grained control of our emotions together with motivational enhancement should enable us, other things being equal, to realise these scenarios more effectively - and to savour their outcome all the more appreciatively. Nor is hedonic enrichment some kind of prescription for *how* to live posthuman life - any more than being cured of a chronic pain condition dictates how one should lead a pain-free existence. "The world of the happy is quite different from that of the unhappy" observes Wittgenstein in the *Tractatus*. Yes, and the world of the superhappy is quite different from the human world. Whether we'll ever investigate its properties, however, is an open question.

Part II: Bioethics

THE PINPRICK ARGUMENT

Negative Utilitarianism

A counterintuitive consequence of negative utilitarianism (NU) is that it would seem to entail destroying the world rather than permitting its miseries to continue. If the destruction could be accomplished painlessly, then a negative utilitarian is logically compelled to accept this consequence. No amount of happiness, the negative utilitarian may argue, can outweigh the horrors of Auschwitz, or the recurrent tragedies of personal life.

However, planning and implementing the extinction of all sentient life couldn't be undertaken painlessly. Even contemplating such an enterprise would provoke distress. Thus a negative utilitarian is not compelled to argue for the apocalyptic solution. S/he may still privately believe that it would have been better if the world had never existed. This is a separate issue.

A more serious challenge to the intellectual coherence of NU is the Pinprick Argument. Would it really be better that life had never arisen if the only unpleasant experience that would otherwise occur is a pinprick? Surely some pains are too *trivial* to matter significantly?

A negative utilitarian could respond that the pain from a pinprick is of a qualitatively different nature than the pain of, say, bone cancer, or bereavement, or torture, or the

mass cruelties of genocide. A pinprick or its equivalent doesn't involve *suffering* - with its terrible baggage of emotional distress.

Yet this response to the Pinprick Argument seems *ad hoc*. It undermines the purity of the NU ethic. For where is the supposed cut-off point? When does pain become real suffering? How much mild pain/suffering is morally permissible? Who should determine these limits? If the avoidance of pain or suffering is accounted more morally important than happiness, but happiness is not accounted wholly morally negligible, then how can their relative importance be quantified? How can well-being and suffering be made commensurable? What kind of metric should be used? Should the fate of the world rest on an arbitrary, or at least a conventional, cut-off point on the pleasure-pain axis?

The negative utilitarian might reply that this formulation of the problem is misleading. We do not live in a notional world where only a pinprick, minor pains, or even just "mild" suffering exists. In the real world, frightful horrors as well as humdrum malaise occur every day. The intensity of suffering is sometimes so dreadful that its victims are prepared to destroy themselves to bring their torment to an end. Each year, some 800,000 people across the planet kill themselves while in the grip of suicidal despair. Tens of millions of people are severely depressed or suffer chronic neuropathic pain. By way of contrast, the genteel conventions of an ethics seminar in academic philosophy, or the scholarly technicalities of a journal article, simply fail to come to terms with the enormity of what's at stake. To talk of a "pinprick" is to *trivialise* the NU ethical stance.

This accusation may be true. Nonetheless, it's unclear how the intellectual coherence of NU can be restored. Less austere versions of NU are all intellectually *messy*. Weakened variants of the principle may capture our intuition that getting rid of a certain amount of suffering has more moral urgency than adding a "corresponding" amount of happiness

without discounting the moral value of happiness altogether. This sounds more plausible. However, hybrid ethical systems that give weighted priority to the relief of suffering over the promotion of happiness no longer embody pure NU. In theory, the negative utilitarian could bite the bullet and claim that even a pinprick is too much. But here it is the negative utilitarian who risks trivialising the moral seriousness of the NU ethic. A professed willingness to sacrifice the world to avert a mere pinprick violates our deepest moral intuitions.

Admittedly, it is unclear why intuition should be any better guide in ethics than it has been in folk physics or folk psychology. Our moral intuitions have been systematically biased by natural selection in ways that tend to maximise the inclusive fitness of our genes. Thus our moral intuitions are "deep" in the sense of being strongly felt rather than well-grounded, insightful or profound. Yet (almost) everybody would treat a bullet-biting response to the Pinprick Argument as the *reductio ad absurdum* of strict NU. Indeed, most philosophers have reckoned that *any* amount or intensity of suffering [though not necessarily their own] is a price worth paying for the precious gift of life, calling into question the sanity of anyone who suggests otherwise. The problem here is that while (almost) all of us have experienced the negligible pain of a pinprick, our judgement that there could be no suffering so unbearable that it justifies bringing the world to an end is not a claim we would be prepared to explore empirically. Tragically, thousands of people each year who have greater experience show that they disagree. Some kinds of suffering are so atrocious they can quite literally compel assent to NU - regardless of one's prior ethical views.

The classic rebuttal of NU, a doctrine whose implications are baldly alleged to be "wicked" and "absurd", is R.N. Smart, `Negative utilitarianism', *Mind* LXVII, 1958,

pp.542-3; see also J.J.C. Smart and B. Williams, *Utilitarianism: For and Against*, Cambridge UP: 1973, pp.28-9.

The Pinprick Argument is commonly conceived as a problem purely for the negative utilitarian. But an analogous argument confronts the "positive" utilitarian too. Once again, imagine a God-like superbeing with the power either to save or to extinguish the world - but this time governed by a classical utilitarian ethic. Imagine summating the pleasures and pains of all sentient creatures and discovering they are finely balanced. By the same token, the addition of a *single* pinprick's worth of pain apparently mandates world destruction. Whatever our value-scheme, can a trivial pinprick really bear such apocalyptic significance?

Direct versus Indirect Negative Utilitarianism

The world is not going to end any time soon, painlessly or otherwise. Human beings - or our immediate post-human descendants - will shortly colonise the solar system and (possibly) beyond, rendering most extreme catastrophe scenarios moot. Asteroid impacts, global warming, viral pandemics, bioterrorism or thermonuclear war may cause immense suffering and loss of life; but they will not kill everyone, or even sterilise the home planet. Nor will sentient life be brought to an end by collective human design. This is because it is psychologically and sociologically unrealistic for negative utilitarians to expect to convince most people of naïve NU. If a policy recommendation is certain to fail, and if plotting it would merely cause further suffering, then the sophisticated negative utilitarian is ethically obliged to act and argue *against* it. Collective global suicide is impossible. No "doomsday device" will (probably) ever get built. So the misguided negative utilitarian who argues for a wholesale compassionate nihilism is at best wasting his or her time. S/he also misconstrues the practical policy implications of NU. Life-lovers

will always tend to out-reproduce negative utilitarians, if only because life-affirming alpha males are likely to accrue more power, influence and greater reproductive opportunities than angst-ridden and depressive negative utilitarians. Believers in direct NU can scarcely go forth and multiply, since reproduction entails creating more suffering. Thus negative utilitarianism remains among the world's rarer ethical belief systems. Some forms of "status quo bias" are ineradicable. One might even expect that most advocates of NU will get weeded out of the gene-pool. Perhaps no more than a few hundred - or at most a few thousand - persons scattered across the globe currently acknowledge the NU title. They are unlikely ever to be effectively organised or led.

However, as the biotechnology revolution unfolds, it is possible that negative utilitarianism will prevail, albeit under a different description. Three particular developments are worth noting here.

First, within the next few years it is likely that neuroscientists will elucidate the final common pathway of pleasure in the brain. Once its molecular signature is identified, then happiness may be modulated, enriched, controlled and amplified effectively without limit; pure pleasure shows no physiological tolerance. Therapies to eradicate the molecular substrates of unpleasantness will probably follow too, permitting lifetimes of unalloyed bliss, or at least gradients of adaptive well-being. Initially, such interventions may be used by biological psychiatrists to treat conditions such as refractory "antidepressant-resistant" depression. But ever-richer varieties of "super-soma" are bound to leak out from the pharmaceutical grey market to the wider world via the burgeoning scientific counterculture. Whatever its guise, super-soma or its equivalents will inevitably prove extraordinarily popular. The Internet will vastly expand its international appeal and channels of distribution. Today, an apt objection to the use of most illicit recreational

mood-boosters is that they are ineffective and self-defeating. Contemporary fast-acting euphoriants activate the hedonic treadmill, not subvert it: street drugs typically give rise to more suffering, not less. But the advent of safer, cleaner, sustainable mood-brighteners that "re-set" our emotional thermostats obviates this objection. Less obviously, the advent of safe, sustainable empathogens will defeat the argument that drug-taking is inherently "selfish". Rationally-designed empathogens and entactogens promise to enrich our conception of mental health, introspective self-knowledge and social intelligence. Admittedly, talk today of "safe and sustainable" pleasure-drugs is liable to ring hollow given the dirty street drugs and crude mood-brightening medications currently available. The historical performance of Big Pharma in psychiatric medicine has been chequered at best. By the same token, even the most enlightened underground chemists have opened up a Pandora's Box of surprises. Yet mental pain is destined to become medicalised, optional, and perhaps one day obsolete.

Second, we are on the brink of a reproductive revolution of "designer babies". Within the next few decades, prospective parents will routinely start to choose the genetic makeup and personalities of their future children. The nastier alternative alleles and allelic combinations bequeathed by natural selection will be progressively edited out of the gene-pool as evolution ceases to be effectively random and "blind". Our evolutionary trajectory as a species will be shaped instead by quasi-rational agents. In future, novel designer genes and allelic combinations will be chosen in deliberate anticipation of their probable behavioural effects. When tomorrow's parents opt *not* to have depressive or anxiety-ridden children, most of such parents-to-be may have no grandiose ethical system in mind, let alone universal NU. But as the reproductive revolution spreads across the globe, the collective outcome of such acts of individual parental choice may be similar

to the fruits of grand utopian design. The great majority of parents will aspire to have superintelligent, happy, beautiful, affectionate kids. In turn, their superintelligent, happy, beautiful, affectionate kids will presumably want enriched children of their own - and from a much higher baseline of mental health. Thus the natural "set point" of our emotional well-being will be genetically ratcheted upwards, both individually and, statistically, for the "unnaturally" evolving (post)human species as a whole. Older humans trapped with legacy wetware may opt for somatic gene therapy as personalised medicine matures. Meanwhile, selection pressure against some of the nastier traits adaptive in our Darwinian past will be intense. Subtle functional analogues of pain and anxiety in the guise of gradients of diminished well-being will (probably) be retained to preserve our informational sensitivity to noxious stimuli and sustain critical insight; but the textures of raw suffering as we understand them today may be banished to evolutionary history.

Third, developments in single-celled protein technologies will soon enable us to grow genetically-engineered "vat food" that's at least as tasty as flesh from intact non-human animals. If so, then the process will presumably be scalable without limit. Critically, such vat-food will be cheaper. Given market economics, then on this scenario the factory farming "industry" will undergo world-wide collapse - or at least convert to the more efficient model. In fact there's a fair chance we'll witness global veganism by the second half of the century. The moral arguments for a cruelty-free diet will seem more cogent when their acceptance no longer demands renouncing the accustomed taste of some of our favourite foods. Elsewhere, Mother Nature, red-in-tooth-and-claw, won't disappear so swiftly. Yet at current rates of habitat destruction, no large mammals will survive in the wild later this century. Vestiges of old order may remain elsewhere in the living world;

but the residual forms of suffering, if any, that will be permitted in our wildlife parks or the deep oceans are far from certain. If we conclude that unpleasant states of consciousness are morally unacceptable, then genetic engineering, quantum computing and nanorobotics can be harnessed to redesign the global ecosystem and rewrite the vertebrate genome. The exponential growth of computing power to run complex simulations may eventually make such ecosystem transformation *trivial*. A technologically and ethically advanced civilisation can eradicate suffering in all sentient life.

* * *

It need scarcely be stressed that the three scenarios sketched above are speculative. No less speculative is the bioconservative prediction that we will opt to sustain suffering indefinitely.

Whatever the future holds, NU ethics will presumably still fail to resonate with the overwhelming majority of the population - especially after our emotional well-being increases as the adoption of enhancement technologies gathers pace. So perhaps the most effective way for a negative utilitarian to promote his/her ethical values is not to proselytize under that label at all. Instead, the negative utilitarian may find it instrumentally rational to give weight overtly to the "positive" values of ordinary classical utilitarians, preference utilitarians/preference consequentialists, and the far wider community of (mostly) benevolent non-utilitarians who share an aversion to "unnecessary" suffering. The indirect approach to NU is likely to yield the greatest payoff. Only by our striving to promote "positive" goals as well, and campaigning for greater individual well-being, is the ethic of NU ever likely to be realized in practice.

If the abolitionist project succeeds, whatever its ultimate time-scale, then should the negative utilitarian be morally satisfied with such an outcome? In an important sense yes: s/he will have discharged all his or her moral responsibilities. If this epoch-making transition in the history of life on Earth comes to pass, then it will be a revolution far more momentous and profound than anything to date. Moreover, unlike positive utilitarianism or so-called preference utilitarianism - neither of which can ever be wholly fulfilled - NU seems achievable in full.

The contrast is instructive. According to the felicific calculus of the positive utilitarian, advanced biotechnology mandates the molecular manufacture of happiness/value on a prodigious scale no less than the eradication of suffering. Indeed the impending biotech revolution ethically commits the classical "hedonistic" utilitarian to creating hypertrophied pleasure centres that generate levels of emotional well-being *orders of magnitude* more intense than anything accessible today. It is hard to express this implication soberly and without taint of sensationalism. Such a revolutionary application of the classic utilitarian ethic is a consequence that its originators can never have anticipated. Bentham and his contemporaries assumed that the felicific calculus would be most fruitfully applied via socio-political and legislative reform.

Looking to the future, what is the theoretical individual maximum of well-being/happiness/pleasure? Pleasure scientists don't know. It is presumably hard for organic nervous systems to sustain successive "warm" quantum coherent states beyond a given size and fleeting duration before thermally-induced decoherence sets in, ruling out a phenomenology of Jupiter-sized pleasure centres. But in our current state of ignorance, quantum mechanical accounts of upper bounds to the unity of consciousness

are unavoidably speculative. Bolder conjectures on the theoretical maximum of pleasure/value in the cosmos won't be pursued here.

By contrast, negative utilitarianism doesn't enjoin a never-ending amplification of our reward circuitry. In practice, most negative utilitarians would probably find such discussions morally frivolous. Here at least NU is closer to common sense - and perhaps the ethics and metaphysics of the Stone Age. However, there is a sense in which any satisfaction on the part of the negative utilitarian who envisages completion of the abolitionist project is misplaced. Strictly, the notion that suffering can be abolished rests on a pre-scientific conception of time. On the "block universe" scenario of modern physics, the horrors of the world perpetually occupy the spatiotemporal coordinates they do. All here-and-nows [tenselessly] exist and are equally real. The suffering characteristic of primordial life on Earth is not going to disappear from space-time. The intrinsic negative value of such suffering is ineradicable. Suffering - perhaps extreme agony beyond our comprehension - may also be located inaccessibly in other lifeforms elsewhere in the Multiverse. Worse, if eternal chaotic inflation scenarios of cosmology are correct, then the exponential increase of googols of "pocket universes" must spawn the exponential growth of suffering too - and possibly all manner of evils that humans haven't even conceptualised. Optimists may cherish Michael Faraday's dictum, "Nothing is too wonderful to be true, if it be consistent with the laws of nature"; but conversely, nothing is too terrible to be true if it is consistent with the laws of nature either. So the negative utilitarian may still believe that it would have been better if nothing existed at all. Less bleakly, in the vast expanse of space-time we informally call "the future", it is quite possible that beyond the 22nd century, say, no suffering whatsoever exists in our little island universe, or if it does, then it exists only in a vanishingly low-density region

of the universal wave function. [*For a more pessimistic analysis, see Suffering in the Multiverse*.]

So what will become of NU? If our genetically enriched descendants are by their very nature blissfully happy, then it is unlikely that they will explicitly endorse a negative utilitarian ethic, even assuming that their conceptual scheme is commensurable with our own. Psychologically superwell minds may find it constitutionally impossible to take NU from the bygone era seriously. The very possibility of NU may be cognitively closed off to them. Mature post-Darwinian consciousness may feel self-intimatingly valuable beyond anything we can grasp today. Indeed, posterity may enjoy norms of lifelong, multi-dimensional mental health too wonderful for present-day concepts to describe or even name. But if suffering of any kind, and even the merest "pinprick" of discomfort, becomes neurochemically impossible - perhaps replaced by information-theoretic gradients of well-being - then negative utilitarianism itself will have become irrelevant: a redundant historical curiosity. If so, it's an irrelevance that contemporary utilitarians should welcome.

Utilitarian Bioethics

TERMINOLOGY

"Utilitarianism" is an uninspiring name for an inspired ethic. The word derives from the Latin *utilis*, useful. It utterly fails to evoke the relief of suffering - and the prospect of sublime bliss - that the scientific application of the felicific calculus entails. It doesn't stir to action. It conveys no sense of moral urgency. "*Utilitarianism Now!*" will never serve as a rallying cry for anyone, perhaps with the (very) improbable exception of a small community of ethicists in academia. Even within a university setting, students typically associate utilitarian ethics more with scholarly logic-chopping, essay-writing and stressful examination rituals than a breathtakingly beautiful vision of life to come.

Beyond the academic treadmill, "utilitarian" in normal usage connotes a concern for usefulness without regard for beauty or even pleasantness. Such idiom has little currency on the street, but among the educated lay public, "utilitarian", "utility" and "utilitarianism" are terms more likely to evoke Thomas Gradgrind from Dickens' *Hard Times* than the abolition of pain - let alone the replacement of suffering by a gradients of profound happiness.

There is another challenge for the utilitarian activist. In ethics, probably more than in any other discipline, doctrines tend to become almost inseparable in the imagination from their most prominent advocates. Marx notoriously described Jeremy Bentham as "a desiccated calculating machine". Bentham's most visible legacy is his mummified corpse at University College, London - a remarkable sight to behold, but probably not the ideal

icon for a new era of life on Earth. Nor is Bentham's design for the Panopticon an inspirational symbol of a better world. The problem is not that the early utilitarians led depraved lives. On the contrary, the classical "hedonistic" utilitarians would appear to have been uncommonly virtuous, even in the light of ethical systems radically different from their own. Bentham was celibate; and it has been well said that no orgy was ever graced by the body of John Stuart Mill. Incredibly, however, utilitarianism seems rather "dull" - a historical, nineteenth century English affair. "Humans do not strive for happiness; only the Englishman does that" ["*Der Mensch strebt nicht nach Glück; nur der Engländer thut das*" - Nietzsche]. Bentham himself was trained as a lawyer. His prose and sometimes its content are lawyerly. The practical legislative issues that Bentham dealt with are especially associated with early industrial England; and not all of them are still pressing. This is scarcely a reproach. The idiom and preoccupations of malaise-ridden utilitarians of the early 21st century may seem no less quaint to our descendants. But the unfriendly language of utilitarianism and its numerous sub-species *is* a problem that remains unsolved to this day.

According to the *Oxford English Dictionary*, Bentham first used the term "utilitarian" in 1781. He was drawn to the term by the usage of "utility" in Hume. In later life, sensing its pitfalls, Bentham writes in a note of July, 1822, *Principles of Morals and Legislation*, ed. 1879, p. 1 n.7: “The word utility does not so clearly point to the ideas of pleasure and pain as the words happiness and felicity do: nor does it lead us to the consideration of the number of the interests affected.” So Bentham in his later years preferred "the greatest happiness principle". However, John Stuart Mill revived the term "utilitarianism", crediting the first use of the word "utilitarian" to a novel by Galt. As Mill explains, "Utilitarian supplies a want in the language, and offers, in many cases, a convenient

mode of avoiding tiresome circumlocution" (JS Mill; *Utilitarianism*, 1863, 210n); and the usage has stuck.

Why does getting the terminology right matter? ["a rose by any other name...", etc.] It matters to the utilitarian bioethicist because making the world a better place is difficult if not impossible when one flies under an ill-chosen banner. Confusing the name of something with the thing itself is such an obvious mistake that it might seem scarcely worth noting; but psychologically, we are prone to do it all the time. This kind of confusion can sometimes help, and sometimes hinder, advocacy of the value-system at issue. In the case of utilitarianism, the confusion represents a huge obstacle to success. Sadly, the principle of utility hasn't been applied to its own title.

Thus more than two centuries after its formulation by Bentham, utilitarianism and its revolutionary implications haven't captured the popular imagination - even if quasi-utilitarian intuitions do inform a lot of our moral and legislative practice. Opponents of utilitarian ethics would respond that this limited progress has little to do with an infelicitous choice of name and much more to do with the weakness of utilitarianism as a moral theory. These alleged shortcomings will not be addressed here, other than to note that most criticisms of utilitarianism to date stem from the allegedly bad *consequences* that follow from adopting a utilitarian ethic - whereas to show that applying utilitarian ethics leads to unpleasant net consequences is to show that the alleged policy prescription in question isn't really utilitarian at all. Certainly, doing felicific calculus isn't remotely straightforward. Bernard Williams once even argued that if utilitarianism were true, then one should try, on utilitarian grounds, to discourage anyone from believing it.

UTILITARIANISM BIOLOGISED: BENTHAM PLUS BIOTECH?

So how can the practical implications of utilitarianism best be conveyed in the modern era? Now that the human genome has been decoded, the ramifications of a utilitarian ethic go far beyond socioeconomic and legislative reform. In era of post-genomic medicine, they extend to control of the pleasure-pain axis itself. By unravelling the molecular substrates of emotion, biotechnology allied to nanomedicine permits the quantity, quality, duration and distribution of happiness and misery in the world to be controlled - ultimately at will. More controversially, the dilemmas of traditional casuistry will lose their relevance. This is because our imminent mastery of the reward centres ensures that *everyone* can be heritably "better than well" - a utopian-sounding prediction that currently still strikes most of us as comically childlike in its *naïveté*. However, unlike perennially scarce "positional" goods and services in economics, personal happiness doesn't need to be *rationed*. Within the next few centuries, a triple alliance of biotech, infotech and nanotech can - potentially - make invincible bliss a presupposition of everyday mental health. From a purely *technical* perspective at least, global happiness can be increased by many orders of magnitude; the substrates of suffering and depression can be abolished outright; genetically pre-programmed superhealth can become the norm; and well-being in the richest sense of the term can become ubiquitous.

Over-excited technofantasy? Well, perhaps. But instead, these rosy predictions may prove hopelessly conservative. For the melding of biotech, nanorobotics and quantum computing is going to be extraordinarily fertile - far beyond anything imaginable today. On the utilitarian conception of value, sentient life will become vastly more *valuable* as well, since value - in the form of an abundance of subjectively wonderful experiences - will be correlatively increased by orders of magnitude too. What kind of narrative

structures this diversity of valuable experiences will be woven into can only be speculated: future-gazing into the lives of our descendants is an idle parlour game at best. Yet when harnessed to biotechnology, the "greatest happiness principle" dictates the mass-manufacture of the molecular substrates of value on a prodigious - and perhaps one day cosmic - scale. Critics will view this "hedonistic" implication of a classical utilitarian ethic in the age of biotech as its *reductio ad absurdum*. Such critics also charge, inevitably, that utilitarians want to reduce us all to "happy pigs" - or the functional counterparts of utility-maximising wirehead rats. Utilitarianism itself has long been dismissed as a doctrine "worthy only of swine". Mischievously perhaps, Bentham himself kept a "beautiful pig" as a pet which would "grunt contentedly as he scratched its back and ears". Certainly, it *is* simplistic to view sentient beings as mere Benthamite pleasure machines - and not just because Darwinian life is typically "nasty, brutish and short". But quite aside from the critics' needless disparagement of our abused fellow creatures, a future world of mindless bliss - or some kind of collective cosmic orgasm - is less sociologically plausible than a post-human era of superintelligent, supersentient well-being. Regrettably, this mouth-watering vista of delights isn't immediately obvious from the "utilitarian" label.

THE BRANDING PROBLEM

One possibility is that this futuristic vision of heaven-on-earth should be promoted by marketing professionals, image consultants and branding specialists. The proposal that utilitarian ethicists should resort to the techniques of Madison Avenue to educate the wider community is likely to be met with a fastidious shudder of distaste - or outright incredulity - by (most) professional philosophers. One will be told that better scholarship,

not inspired propaganda, is needed to win over the sceptics - and stir the morally apathetic into action. But to hope that the cool light of reason alone will illuminate the case for a cruelty-free world, let alone secure its practical implementation, is optimistic and perhaps naïve. To be effective, utilitarians will need to organise, agitate and actively campaign - rather than simply talk and write papers in academic journals. Yet successful organisation-building demands a compelling label and a potent brand image - and a different set of skills from mere scholarly acumen. Alas, philosophers by temperament are rarely men of action. An effective "utilitarian" political organisation sounds fanciful. "The Utilitarian Party" today would be stillborn. A "utilitarian" mass-movement under that description seems out of the question. So what can be done?

If the idiom of "utilitarianism" and "utility" can't be salvaged, then there is a need to find a soul-stirring alternative - consistent with preserving the core utilitarian ethic. As Bentham recognised, "the greatest happiness principle" *does* resonate more strongly with most people. But the principle doesn't lend itself to any single-word "ism" - beyond "hedonism". As it happens, selfless hedonism is an apt description of Benthamite utilitarianism and its refinements. However, any slogan incorporating the word "hedonism" or its derivatives is likely to evoke shallowness, emptiness and amoralism - or at best a very one-dimensional kind of well-being. It won't work in conveying the marvellously enriched conception of mental health which tomorrow's biotechnologists have in store.

Unfortunately, none of the proposed terminological alternatives are satisfactory either.

Academic philosophers may be drawn to a term that incorporates "eudaimonism" or "eudaimonistic" [from *eu*: "good" or "well being"; and *daimon*: a "spirit" or minor "deity"; literally meaning "having a good guardian spirit" - frequently translated as "human

flourishing"]. No one really knows quite what it means, but its etymology is respectably ancient. Amongst scholars, at least, eudaimonistic idiom conveys a richer conception of human well-being than talk of unidimensional happiness - or indeed mere "pleasure", that handy catch-all antonym of "pain" with its regrettable penumbra of debased connotations. On the Aristotelian conception of happiness, the happy subject is one who focuses on developing the excellence of his character. Unfortunately, excellence of character is a notion that's hard to naturalise. Aristotle didn't think *eudaimonia* was possible for non-human animals at all: "we call neither ox nor horse nor any of the other animals happy" (*Nicomachean Ethics*; p.1099). So to conflate *eudaimonia* in Aristotle's sense with happy experiences is untenable. Worse, there is no way "eudaimonistic consequentialism" - or even just "eudaimonism" - is going to fire the popular imagination.

Negative utilitarians - and ethicists who give greater moral weight in general to alleviating distress than magnifying pleasure - focus on the abolition of suffering. Thanks to the unfolding revolution in the biological sciences, this scenario is technically feasible - though it may take centuries to complete. So there is "abolitionism", "abolitionist" and even "the abolitionist project". These terms convey something of the moral grandeur and seriousness of purpose of utilitarian ethics; and also reflect its transcendence of narrow species self-interest to encompass all sentient life. Here again, Bentham was ahead of his time in recognising our complicity in the plight of non-human animals - though he could scarcely have anticipated the growth of single-cell protein technologies that may one day inaugurate global veganism. Unfortunately, abolitionist terminology doesn't directly specify what is being abolished, in common usage at least. Further, not all abolitionists are utilitarians; and it may be unwise to imply that commitment to the eradication of suffering is the exclusive prerogative of one contested ethical theory. Buddhists, for

instance, locate *dukkha* [suffering; ill; unsatisfactoriness; imperfection] and its relief at the very heart of existence. Also, strict negative utilitarianism has [allegedly] counterintuitive consequences that have hitherto disqualified it from serious consideration. Either way, abolitionist vocabulary is problematic - but at least worth bearing in mind.

Other suggestions are problematic too. "Positive" utilitarians searching for the elusive dream makeover may be better disposed to a term like "paradise engineering". It's an expression that evokes the wonders ahead without discounting the relief of suffering. The quasi-religious metaphor inherits the favourable associations of Christian (and Islamic, etc) paradise shorn of its untenable theological commitments. "Heaven" could in theory play a similar lexical role, though "Heaven engineering" doesn't have the same ring. However, neither of these terms imparts a sense of moral urgency; and they may not be short and snappy enough.

The same applies to "Post-Darwinian transition". The term alludes to the impending reproductive revolution of so-called designer babies. By rewriting its own genome, our species is destined to transcend age-old "human nature". Beyond this century, prospective parents are unlikely to choose genotypes predisposing to depression, anxiety and malaise in their future children. Over time, the "unnatural" selection of designer genomes should weed our predisposition to emotional nastiness from the gene-pool - even in the absence of any grand ethical/ideological project. Our natural "set point" of emotional well-being should become progressively higher over the millennia - a form of hedonic enrichment possibly amounting to some kind of phase change in the nature of consciousness itself. But the nature of any Post-Darwinian transition is controversial even among scientifically informed utilitarians. There is no guarantee that the outcome of post-human reproductive medicine will accord with a utilitarian ethic - though this may

broadly be the case even under other labels. And the expression "Post-Darwinian transition" is a bit of a mouthful too.

A related term is "transhumanism". This convenient one-word label embraces a diverse family of belief and values that predict (and advocate) the transcendence of our biological heritage. According to Article 7 of the Transhumanist Declaration of the World Transhumanist Association (WTA), "transhumanism advocates the well-being of all sentience (whether in artificial intellects, humans, posthumans, or non-human animals)". But conceptions of the post-human realm differ widely. Not all transhumanists advocate the outright abolition of suffering, let alone the maximisation of happiness. So neither this term nor its cognates will serve the frustrated utilitarian either.

QUANTUM COMPUTERS AND THE FELICIFIC CALCULUS

There is a further difficulty with any possible replacement terminology. A strength of classical utilitarianism and the felicific calculus is that it provides, in principle, an objective criterion of whether an action - or rule of action - is right or wrong. Practical ethics becomes, in theory, a rigorous, exact, and mathematically quantifiable discipline - though this aspiration remains a pipe-dream even as neuroscientists elucidate the molecular substrates of happiness, sadness and other "core" emotions (anger, fear, disgust, surprise, etc) in the brain. By contrast, none of the proposed terminological alternatives capture this calculational feature, or indeed any kind of decision procedure for action - short of some very drastic stipulative (re)definitions. So what's needed is a catchier, sexier synonym for "utilitarianism" that retains the all-important criterion but at the same time is more evocative of the sublime - and sounds more morally urgent than "utilitarian". Undoubtedly this is a tall order. Perhaps a new word altogether should be

invented and explicitly defined from scratch. But neologisms have their drawbacks as well.

Practical utilitarianism *does* involve systematic calculation and planning - as the dour figure of Thomas Gradgrind might suggest. Most humans find the process of formal calculation painful; and calculus is typically associated with the miseries of school-day mathematics. But ultimately (most of) the calculation demanded by a global application of the felicific calculus needn't be done by humans, post-humans or indeed any sentient computational system. We can offload it. The exponential growth of computing power promises to revolutionise the discipline of futurology - and potentially practical ethics too. Most dramatically, the ability of [currently hypothetical] quantum supercomputers to run complex alternative simulations many orders of magnitude more powerful than their classical predecessors may transform the felicific calculus from a philosopher's fantasy into a scientific tool - and a utilitarian ethicist's dream.

This prospect leaves most people unmoved. In common with "utilitarian", the term "calculus" sounds cold, clinical, technocratic and disturbing, even when it's prefaced by the word "felicific". What place is there for romantic and poetic diction in a utilitarian ethicist's toolkit? Intuitively, we believe that the realm of feeling belongs to spontaneity - not premeditation. The joys of love, beauty and friendship shouldn't be subjected to a utilitarian cost-benefit analysis: here at least, Gradgrindian "fact" should submit to non-utilitarian "fancy". But spontaneity and romanticism can be practised safely only when the biological foundations of a civilised society are genetically in place. In a Darwinian world, they often lead to suffering and heartache. Likewise, in a future post-Darwinian world of rational reproductive decisions, honesty may be less hazardous than today. For

personal integrity is frequently impossible for utilitarians and non-utilitarians alike with a Darwinian genome: evolution by natural selection has spawned Machiavellian apes.

Whether or not we can effectively rehabilitate the terms "utilitarian", "calculus" and their cousins, the new lexicon of genetic engineering, CRISPR-based gene drives, drugs, wireheading, eugenics, and the technologies of mind-control stir deep anxieties too. The historical record of their application is not encouraging; and our conception of a notional utilitarian future owes more to Huxley's *Brave New World* than starry-eyed utopianism. In fact, our entire conceptual scheme is steeped in negatively-charged language - a legacy of the Darwinian emotions adaptive in our evolutionary past. *None* of the vocabulary we use today is unpolluted by the (ab)uses to which it has been put. Yet at the heart of utilitarianism is the most wonderful - and the most *valuable* - ethic ever discovered. Properly understood, the very name should induce an almost overpowering sense of delight at what's in prospect. Sadly, the ghost of Thomas Gradgrind still haunts the utilitarian project; and it's unclear how the terminological difficulty can best be overcome.

On Classical Versus Negative Utilitarianism

A response to Toby Ord's essay

Why I Am Not A Negative Utilitarian

Toby, a few thoughts...

1) World destruction? You write, "...a thoroughgoing Negative Utilitarian would support the destruction of the world (even by violent means)". No, a thoroughgoing *classical* utilitarian is obliged to convert your matter and energy into pure utilitronium, erasing you, your memories and indeed human civilisation. By contrast, the negative utilitarian believes that all our ethical duties will have been discharged when we have phased out suffering. Thus a negative utilitarian can support creating a posthuman civilisation animated by gradients of intelligent bliss where all your dreams come true. By contrast, the classical utilitarian is obliged to erase such a rich posthuman civilisation with a utilitronium shockwave. In practice, I don't think it's ethically fruitful to contemplate destroying human civilisation, whether by thermonuclear Doomsday devices or utilitronium shockwaves. Until we understand the upper bounds of intelligent agency, the ultimate sphere of responsibility of posthuman superintelligence is unknown. Quite possibly, this ultimate sphere of responsibility will entail stewardship of our entire Hubble volume across multiple quasi-classical Everett branches, maybe extending even into what we naively call the past (*cf*. "The Two-State Vector Formalism of Quantum Mechanics: an

Updated Review": http://arxiv.org/pdf/quant-ph/0105101v2.pdf). In short, we need to create full-spectrum superintelligence.

2) Negative utilitarians can (and do!) argue for creating immense joy and happiness. Indeed, other things being equal, negative utilitarians are ethically bound to do so. For the thought of a painless but joyless world strikes most people as *depressing*. Negative utilitarians are committed to phasing out even the faintest hint of disappointment! The prospect of an insipid pain-free life without peak experiences - mere muzak and eating potatoes, so to speak - sounds bleak. If a thought or deed causes the slightest unease or distress, then other things being equal, that thought or deed is *not* expressive of negative utilitarianism.

3) You write, "Absolute NU is a devastatingly callous theory". No: NU is a unsurpassably compassionate theory. You are "callous" only if you are indifferent to someone's suffering, not if you don't act to amplify the well-being of the already happy or act to create happiness *de novo* - although in practice, negative utilitarians should promote intelligent superhappiness too. Inducing sadness or disappointment is not NU.

[I think the force of your example depends on an untenable metaphysics of personal identity. If instead we use a more empirically supportable ontology of here-and-nows strung together in different sequences thanks to natural selection, no one is *harmed* by waking up happy in the morning rather than superhappy like their namesake the night before. So this is really an issue of population ethics, normally reckoned a different topic.]

Let us compare the callousness/compassion of classical utilitarianism and NU.

Since we're doing thought-experiments, imagine if a magic genie offers me super-exponential growth in my bliss at the price of exponential growth in your agony and

despair. If I'm a classical utilitarian, then I am ethically bound to accept the genie's offer. Each year, your torment gets unspeakably worse as my bliss becomes ever more wonderful. Indeed, the thought I'm ethically doing the right thing increases my bliss even further! By generating so much net bliss, I'm the most saintly person who ever existed! If you knew how incredibly superhumanly *wonderful* I'm feeling, then you'd realise that my super-bliss easily offsets your tortured despair. Your tortured despair is a trivial *pinprick* in comparison to my super-exponentially growing bliss!

Of course, as a real-life negative utilitarian, I'd politely decline the genie's offer.

But if you win me over to classical utilitarianism, I'll accept.

Which is the callous choice?

Classical utilitarianism offers perhaps the best hope of cheating Hume's Guillotine and naturalising value. But does it maximise *moral* value? Or something else?

On Utilitronium Shockwaves Versus Gradients of Bliss

Why is the idea of life animated by gradients of intelligent bliss attractive, at least to some of us, whereas the prospect of utilitronium leaves almost everyone cold? One reason is the anticipated loss of self: if one's matter and energy were converted into utilitronium, then intuitively the intense undifferentiated bliss wouldn't be *me*. By contrast, even a radical recalibration of one's hedonic set-point intuitively preserves the greater part of one's values, memories and existing preference architecture: in short, personal identity. Whether such preservation of self would really be obtained if life were animated by gradients of bliss, and whether such notional continuity is ethically significant, and whether the notion of an enduring metaphysical ego is even intellectually coherent, is another matter. Regardless of our answers to such questions, there is a tension between our divergent response to the prospect of cosmos-wide utilitronium and intelligent bliss. People rarely complain that e.g. orgasmic sexual ecstasy lasts too long, and that regrettably they lose their sense of personal identity while orgasm lasts. On the contrary: behavioural evidence strongly suggests that most men in particular reckon sexual bliss is too short-lived and infrequent. Indeed if such sexual bliss were available indefinitely, and if it were characterised by an intensity orders of magnitude greater than the best human orgasms, then would anyone - should anyone - wish such ecstasy to stop? Subjectively, utilitronium presumably feels more sublime than sexual bliss, or even whole-body orgasm. Granted the feasibility of such heavenly bliss, is viewing the history of life on Earth to date as a mere stepping-stone to cosmic nirvana really so outrageous?

For the foreseeable future, however, even strict classical utilitarians must work for information-sensitive gradients of intelligent bliss to raw undifferentiated pleasure. Classical hedonistic utilitarianism was originally formulated as an ethic for legislators, not biologists or computer scientists. Conceived in this light, the felicific calculus has been treated as infeasible. Yet a disguised implication of a classical utilitarian ethic in an era of mature biotechnology may be that we should be seeking to convert the world into utilitronium, generally assumed to be relatively homogenous matter and energy optimised for raw bliss. The "shockwave" in utilitronium shockwave alludes to our hypothetical obligation to launch von Neumann probes propagating this hyper-valuable state of matter and energy at, or nearly at, the velocity of light across our Galaxy, then our Local Cluster, and then our Local Supercluster. And beyond? Well, politics is the art of the possible. The accelerating expansion of the universe would seem to make further utilitronium propagation infeasible even with utopian technologies. Such pessimism assumes our existing understanding of theoretical physics is correct; but theoretical cosmology is currently in a state of flux.

Naively, the theoretical feasibility of utilitronium shockwave is too remote to sorry about. This question might seem a mere philosophical curiosity. But not so. Complications of uncertain outcome aside, any rate of time discounting indistinguishable from zero is ethically unacceptable for the ethical utilitarian. So on the face of it, the technical feasibility of a utilitronium shockwave makes working for its adoption ethically mandatory even if the prospect is centuries or millennia distant.

Existential risk? Utilitarian ethics and speculative cosmology might seem far removed. But perhaps the only credible candidate for naturalising value has seemingly apocalyptic implications that have never (to my knowledge) been explored in the scholarly literature.

And can we seriously hope to be effective altruists in the absence of serviceable model of reality?

Should existential risk reduction be the primary goal of: a) negative utilitarians? b) classical hedonistic utilitarians? c) preference utilitarians? All, or none, of the above?

The answer is far from obvious. For example, one might naively suppose that a negative utilitarian would welcome human extinction. But only (trans)humans - or our potential superintelligent successors - are technically capable of phasing out the cruelties of the rest of the living world on Earth. And only (trans)humans - or rather our potential superintelligent successors - are technically capable of assuming stewardship of our entire Hubble volume. Conceptions of the meaning of the term "existential risk" differ. Compare David Benatar's "Better Never To Have Been" with Nick Bostrom's "Astronomical Waste". Here at least, we will use the life-affirming sense of the term. Does negative utilitarianism or classical utilitarianism represent the greater threat to intelligent life in the cosmos? Arguably, we have our *long-term* existential risk-assessment back-to-front. A negative utilitarian believes that once intelligent agents have phased out the biology of suffering, all our ethical duties have been discharged. But the classical utilitarian seems ethically committed to converting *all* accessible matter and energy - not least human and nonhuman animals - into relatively homogeneous matter optimised for maximum bliss: "utilitronium".

Ramifications? Severe curtailment of personal liberties in the name of existential risk reduction is certainly conceivable. Assume, for example, that the technical knowledge of how to create and deploy readily transmissible, 100% lethal, delayed-action weaponised pathogens leaks into the public domain. Only the most Orwellian measures - a perpetual global totalitarianism - could hope to prevent their use, whether by a misanthrope or an

idealist. Such measures would most likely fail. By contrast, constitutively happy people would be incapable of envisaging the development and use of such a doomsday agent. The biology of suffering in intelligent agents *is* a deep underlying source of existential risk - and one that can potentially be overcome.

A theoretically inelegant but pragmatically effective compromise solution might be to initiate a utilitronium shockwave that propagates outside the biosphere - or realm of posthuman civilisation. The world within our cosmological horizon could then be tiled with utilitronium with the exception of a negligible island (or archipelago) of minds animated "merely" by gradients of intelligent bliss. One advantage of this hybrid option is that most *refusniks* would (presumably) be indifferent to the fate of inert matter and energy outside their lifeworld. Ask someone today whether they'd mind if some anonymous rock on the far side of the moon were converted into utilitronium and they'd most likely shrug.

In future, gradients of intelligent bliss orders of magnitude richer than today's peak experiences could well be a design feature of the post-human mind. However, I don't think intracranial self-stimulation is consistent with intelligence or critical insight. This is because it is *uniformly* rewarding. Intelligence depends on informational sensitivity to positive and negative stimuli - even if "negative" posthuman hedonic dips are richer and higher than the human hedonic ceiling.

In contrast to life animated by gradients of bliss, the prospect of utilitronium cannot motivate. Or rather the prospect can motivate only a rare kind of hyper-systematiser drawn to its simplicity and elegance. The dips of intelligent bliss need not be deep. Everyday hedonic tone could be orders of magnitude richer than anything physiologically feasible now. But will such well-being be orgasmic? Orgasmic bliss lacks - in the jargon of academic philosophy - an "intentional object". So presumably there will be selection

pressure against any predisposition to enjoy 24/7 orgasms. By contrast, information-sensitive gradients of intelligent bliss can be adaptive - and hence sustainable indefinitely, allowing universe maintenance: responsible stewardship of Hubble volume.

At any rate, posthumans may regard even human "peak experiences" as indescribably dull by comparison.

LIFE IN THE FAR NORTH

An information-theoretic perspective on Heaven

"If someone offered you a pill that would make you permanently happy, you would be well advised to run fast and run far. Emotion is a compass that tells us what to do, and a compass that is perpetually stuck on north is worthless."

Professor Daniel Gilbert

Department of Psychology, Harvard University

Many millions of people in the contemporary world have a compass that is perpetually "stuck on South". They are always unhappy and discontented. They endure chronic pain and/or depression. Some victims of severe anhedonia can't even imagine what it's like to be happy. A minor blessing is that not all of their days are quite as terrible as others. So in one sense, their emotional compass can point North as well as South: a motivational system of sorts still functions. But the whole of their lives is spent in an Antarctic wasteland of misery and despair.

At the other extreme, a small minority of people are blessed with a compass that seems perpetually "stuck on North". In pathological cases, they may be manic. But sometimes they are in varying degrees just "hyperthymic", i.e. the hedonic set-point around which their lives oscillate is unusually high compared to the Darwinian norm. Hyperthymic well-being is chronic; yet it's not *uniform*. Thus some days of hyperthymic life are even more

wonderful than others; pursuing their favourite activities makes hyperthymics even happier than otherwise. So again, the hyperthymic emotional compass is bidirectional: its scale is different, but it works. The relevant contrast here lies in the way hyperthymics are animated by information-signalling gradients of well-being, whereas dysthymics, depressives and victims of chronic pain spend their lives struggling to minimise ill-being. Either way, affective *gradients* rule.

"Normal" or so-called "euthymic" people are inclined to judge that hyperthymics/"optimists" view the world through rose-tinted spectacles. Their central information-processing system is systematically biased. Conversely, hyperthymics see the rest of us as unreasonably pessimistic. Chronic depressives, on the other hand, may view euthymic and hyperthymic people alike as deluded. Indeed victims of melancholic depression may feel the world itself is hateful and meaningless. For evolutionary reasons (*cf*. rank theory), a genetic predisposition to hyperthymia and euphoric unipolar mania are rarer than dysthymia or unipolar depression. Most of us fall somewhere in between these temperamental extremes, though the distribution is skewed to the southern end of the axis. Genetics plays a key role in determining our hedonic set-point, as does the ceaseless interplay between our genes and environmental stressors. Inadequate diet, imprudent drug use, and severe, chronic, uncontrolled stress can all reset an emotional thermostat at a lower level than its previous norm - though that norm may be surprisingly robust. Unlike recreational euphoriants, delayed-onset antidepressants may restore a lowered set-point to its former norm, or even elevate it. Antidepressants may act to reverse stress-induced hypertrophy of the basolateral amygdala and contrasting stress-induced dendritic atrophy in the hippocampus. Yet no mood-brightener currently licensed for depression reliably induces permanent bliss, whether information-signalling

or constant, serene or manic. A genetically-determined ceiling stops our quality of life as a whole getting better.

Is the future of mood and motivation in the universe destined to be an endless replay of life's evolutionary past? Are the same affective filters that were genetically adaptive for our hominid ancestors likely to be retained by our transhuman successors? Will superintelligent life-forms really opt to preserve the architecture of the primordial hedonic treadmill indefinitely? In each case, probably not, though it's controversial whether designer drugs, neuroelectrodes or gene therapies will make the biggest impact on recalibrating the pleasure-pain axis. In the long-run, perhaps germline genetic engineering will deliver the greatest global enhancement of emotional well-being. For a reproductive revolution of designer babies is imminent. Thanks to genomic medicine, tomorrow’s parents will be able to choose the genetic make-up and personality of their offspring. Critically, parents-to-be will be able to select the emotional dial-settings of their progeny rather than play genetic roulette. In deciding what kind of children to create, tomorrow's parents will (presumably) rarely opt for dysfunctional, depressive and malaise-ridden kids. Quite aside from the ethical implications of using old corrupt code, children who are temperamentally happy, loving and affectionate are far more enjoyable to bring up.

The collective outcome of these individual parental genetic choices will be far-reaching. In the new era of advanced biotechnology and reproductive medicine, a combination of designer drugs, autosomal gene therapies and germline interventions may give rise to a civilisation inhabiting a state-space located further "north" emotionally than present-day humans can imagine or coherently describe. Gradients of heritable, lifelong bliss may eventually become ubiquitous. The worst post-human lows may be far richer than the

most sublime of today's peak experiences. Less intuitively, our superwell descendants may be constitutionally smarter as well as happier than unenriched humans. Aided by synthetic enhancement technologies, fine-textured gradients of intense emotional well-being can play an information-signalling role at least as versatile and sophisticated as gradients of emotional ill-being or pain-sensations today. Simplistically, it may be said that posterity will be "permanently happy". However, this expression can be a bit misleading. Post-humans are unlikely to be either "blissed out" wireheads or soma-addled junkies. Instead, we may navigate by the gradients of a multi-dimensional compass that's designed - unlike its bug-ridden Darwinian predecessor - by intelligent agents for their own ends.

In theory, there may ultimately be no need for any information-signalling dips in *subjective* well-being at all. This is because both the nasty and simply mediocre side of life could in principle be computationally offloaded onto our smart machines and neural prostheses. You need a compass only until you reach your destination; it is then redundant. However, the possible existence of a cosmic *Ultima Thule* whose attainment makes a compass eventually superfluous is mere conjecture. So too is the nature of life's interim motivational architecture over the next few billion years. Less ambitiously, for human beings to envisage the abolition of suffering, or even the advent of paradise-engineering, it's not necessary to assume that we'll become full-blown cyborgs, upload ourselves, or pursue any of the more exotic possibilities floated by transhumanists. Simply recalibrating the genetic dial-settings that regulate our basal hedonic tone will suffice.

Of course a hypothetical motivation system based entirely on adaptive gradients of bliss still amounts to a major transition in the history of life in the universe. Its achievement

would mark an ethical revolution without precedent. Much more technically challenging, and also more speculative, will be the genetic design of new emotions and the "re-encephalisation" of the old. Our natural information-signalling system evolved to serve the interests of selfish DNA. Hypothetical future information-processing systems may be dedicated to the interests of DNA's sentient vehicles instead.

Such a revolution may never come to pass. Bioconservatives of all stripes disagree on principle with attempts to redesign human nature. They regard the abolition of suffering and the prospect of radical hedonic enrichment as romantic utopianism, at best. Secular bioconservatives believe that we should retain the same biologically predestined core emotions as our ancestors since time immemorial. Human life should continue to function around the same hedonic/dolorous "set point" that was fitness-enhancing in the ancestral environment of adaptation on the African savannah. Christian bioconservatives believe that we should preserve human nature because Man was created in God's image. If this is so, then the best that can be said is that we do not yet reflect very highly on God.

Well-intentioned or otherwise, bioconservatism is a recipe for perpetuating a never-ending cycle of pain and unhappiness. If Mother Nature really cared about us, then there might be a case for leaving well alone, as bioconservatives so desire. But ultimately such anthropomorphism is morally frivolous: it reflects a lack of any sense of moral *urgency* at the terrible hereditary propensity to suffering endemic to organic life - and the moral imperative to cure it by the only means possible, namely advanced biotechnology. Our primitive repertoire of emotions, and the dismal calibration of our hedonic treadmill, persists today only because they helped our genes leave more copies of themselves ["maximised their inclusive fitness"] when human life really was red-in-tooth-and-claw in the pre-modern era. Luckily, selection pressure in the coming age of "unnatural"

selection will favour a suite of adaptations that's radically different from the nastier traits adaptive in our Darwinian past. For as we decommission natural selection, evolution will no longer be "blind" and "random". Tomorrow's prospective parents will be quasi-rational agents who can choose future genomes *in anticipation of* their likely behavioural consequences. A more benevolent but no less intense kind of selection pressure will be at work. This trend can only accelerate with the conquest of ageing over the next few centuries. As the Earth reaches its carrying capacity of quasi-immortals, reproduction will need to be meticulously planned.

Post-human, post-Darwinian mental health late in the third millennium is likely to be far richer than its impoverished 21st century precursor. Indeed by the standards of our enlightened successors, perhaps malaise-ridden emotional primitives like us will be reckoned in the grip of a toxic affective psychosis. So if, fancifully, your guardian angel offers you a pill that would make you permanently happy, then perhaps you'd be *crazy*, in some sense, to say no. Alas, the nature of affective psychosis precludes full insight into the condition. Thus one may hesitate to swallow the pill. Critics like Professor Gilbert are already on hand to warn that you'd be well advised to "run fast and run far" from the pushers of any permanent-happiness potion.

Ironically, if you do succumb to a pleasure-pill pusher's wares, then as an incidental bonus to the promised lifelong bliss, you'll actually be able to run faster and farther in the direction of whatever you really care about than before. For greater well-being typically enhances motivation, will-power and the capacity to anticipate reward. Boosting mesocorticolimbic dopamine function also enhances the range of stimuli an organism finds rewarding. Counterintuitively, *if* hedonic enrichment is done wisely, then an emotional compass works *better*. Hedonic enrichment reverses the learned helplessness

and monotonous behavioural despair blighting the lives of depressives - and their less severe, subclinical analogues in "normal" Darwinian humans. Enhanced well-being is empowering. It's potentially liberating from the biochemical shackles of the *ancien régime*. In principle, enhanced well-being can be profoundly compassionate and empathetic. Hedonic enrichment heightens a love of life and the urge to self-preservation: an outlook that contrasts with the nihilistic despair of major depression. In future, a heritably blissful mindset should prove genetically adaptive *if* the designer genes/allelic combinations that promote it prove attractive to prospective parents.

Yet can an information-processing system that runs on gradients of lifelong bliss really sustain critical insight? In principle, at least, yes. The functional analogues of depressive realism can be sustained without the nasty textures of low mood. One snappy formulation sometimes used to define "information" is "a difference that makes a difference". On this basis, what's important in the context of the information-theoretic paradigm is not our absolute position on the pleasure-pain axis, but our differential sensitivity to emotionally-tagged variations in fitness-relevant stimuli. For the foreseeable future, an emotional compass *will* be needed to guide the psychologically superwell in everyday post-Darwinian Heaven, no less than in contemporary Darwinian purgatory. The binary coding scheme of a pleasure-pain axis is supremely economical for this navigational purpose. The difference in store for us is that shortly we'll be in a position to tame its cruelties by truncating the axis at the nasty end and vastly extending its reach at the other. An improved motivation system is not just technically feasible. Its hypothetical contours will be kinder to the end-user. Our navigational capabilities may be vastly improved too.

Perhaps a fictional analogy isn't amiss here. To complement a notional permanent happiness pill, consider the offer of a pill that induces permanent hilarity. The latter pill isn't (quite) so fantastical as it sounds. In 1998, neuroscientists at the University of California medical school discovered what may be called a "humour centre" in the brain, despite unfortunate echoes of phrenology. The neural basis of humour apparently lies in a tiny region in the left supplementary motor area. If electrically stimulated, then the subject not merely laughs, but finds *everything* irresistibly funny - just as the wirehead from science fiction finds everything indiscriminately pleasurable. In principle, once the neurological signature of pure humour is identified, then its molecular substrates could be amplified beyond anything natural selection has engineered to date. The amplification might come via pills, neuroelectrodes or genetic modification. Fancifully, one could even imagine a re-engineered civilisation whose inhabitants, by their very nature, found everything hilariously funny. Now intuitively, people who find everything funny are incapable of critical discernment. They are promiscuously amused by slapstick farce and terrible puns no less than by sublime literary wit. So if one values one's sophisticated sense of humour, then one might not be tempted by the offer of a pill that would leave one amused *indiscriminately* - just as one might run fast and run far from the offer of a pill that left one indiscriminately happy. For one wants to act and respond to a changing environment *appropriately* - in some admittedly ill-defined sense of "appropriate". But *if* an advanced, humour-valuing society ever wanted to make life perennially amusing, and yet its members also sought to preserve critical discernment and the quest for ever richer sources of humour, then there is nothing to stop them retaining an information-signalling role of hilarity-*gradients* by simply recalibrating the neurological default-settings of their humour scale. Thus there might arise a post-human civilisation whose gravest concerns were subjectively more hilarious than our funniest moments of comedy.

Of course, the above example is grotesque. It's not going to happen, even if it's a neurologically feasible option for test subjects in an experimental laboratory. Certainly, to generalise the possibility of biological humour-enhancement to a whole society is pure science fantasy - to the best of our knowledge, at any rate. Compared to the urgency of getting rid of suffering, abolishing humourlessness isn't even on our moral radar.

Yet one shouldn't underestimate the versatility of a biologically well-designed compass. Today, sentient life on Earth runs an informational economy of mind driven by gradients of discontent. Tomorrow, we'll have the option of an informational economy of mind run on gradients of well-being.

So is it wrong to swallow the pill?

Population Ethics, Aggregate Welfare, and the Repugnant Conclusion

"*For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living.*"

Derek Parfit

(*Reasons and Persons*. 1984)

Philosopher Derek Parfit's "repugnant conclusion" is eminently plausible; but it is also false. Strictly speaking, the way to maximise aggregate *and* individual welfare is literally to fill up the Earth (and eventually the accessible universe) with sentient beings whose reward circuitry is radically enriched.

Naïvely, the most efficient method to maximise the happiness of the biosphere would be to develop forms of wireheading: direct stimulation of the reward centres of thousands of billions of mind/brains. Wireheading and its genetic and/or pharmacological analogues are energy-efficient and ecologically friendly. However, wireheading is also evolutionarily unstable and socially implausible. Such a scenario will not be explored here except to note how the wirehead option is an existence-proof that unlimited lifelong well-being is feasible in an arbitrarily confined space; pure pleasure shows no physiological tolerance.

Actually, there is a complication. What used to be called the "pleasure centres" of the brain might better be called the "desire centres". Mesolimbic dopaminergic "wanting" is neurologically and anatomically distinct from *mu*-opioidergic "liking". Intracranial self-stimulation studies demonstrate that desire, not pleasure, shows no physiological tolerance. But the term "wireheading" will here be used for an entire family of scenarios involving exclusively direct reward pathway stimulation: indiscriminate and undifferentiated pleasure without end.

There is an alternative to wireheading that is harder to dismiss. This alternative relies on the standard weak assumptions of population ethics harnessed to futuristic computational neuroscience. In theory, maximal aggregate *and* individual welfare - with no trade-off - can be achieved on the twin foundations of:

1) radical enrichment of the pleasure and desire centres of the CNS. Irrespective of population density, suffering can in principle be abolished in all sentient life; and mind/brains motivated entirely by gradients of cerebral bliss. Ultimately, superintelligent posthumans may be animated by gradients of well-being that are billions of times richer than the range of hedonic tone adaptive for *Homo sapiens* in the ancestral environment.

2) a regime of global virtual reality, most memorably evoked in "The Matrix". The exponential growth of computer power (*cf*. Moore's Law) offers the prospect of lifelong immersive VR; a Matrix scenario minus its whimsical "Machines" dependent on pod-grown people for their bioelectrical energy. Most recently, Second Life and its cousins foreshadow what's possible. Next century's multimodal VR will be unimaginably more compelling.

On this "Paradise Matrix" scenario of reward circuitry enrichment plus immersive VR, the Earth's pain-ridden ecosystems can be progressively dismantled [though virtual wildlife

safaris will be optional]. Each envatted mind/brain/virtual world can dine on genetically-engineered single-celled total nutrition mix, subjectively tasting (perhaps) like the ambrosial food of the gods. In mature vatworld Matrix models, the carrying capacity of the Earth runs to *thousands of billions* of interconnected (post)humans. Each of these thousands of billions can enjoy lifelong well-being orders of magnitude richer than anything possible today. To maximise aggregate welfare on a cosmic scale, vatworlds could eventually be dispatched to seed and superpopulate other planets in our Local Group of galaxies - and indeed anywhere habitable or more-or-less terra-formable within our light-cone, saturating the universe with positive value.

For sure, this prospect sounds surreal. Vatworld paradise conjures up images from pulp science fiction - and a reflex response of "that's just Brave New World". To philosophers, the story carries echoes of Cartesian demons, or more strictly, Cartesian angels. Misleadingly, too, vatworld VR also raises the spectre of Harvard University professor Robert Nozick's "Experience Machine" argument. Nozick's thought-experiment purportedly refutes mental state welfarist theories by showing that we value - or at least think that we value - more than "mere" pleasurable experiences. Thus if given the chance to plug ourselves into a device that allows us to experience our fondest dreams-come-true, most of us would allegedly spurn the offer. This is because we value mind-independent truth, in some sense yet to be elucidated. However, it should be stressed that global VR plus reward-pathway enhancement can permit an arbitrarily high degree of mutual realism in each computationally interconnected virtual world. For as sketched here, an immersive virtual reality regime can be interactive and consensual, not solipsistic. If you write a novel, other people can read it. If you compose music, other people can enjoy it. If you want to chat with your friends, you can do so - just like now. What's different is that instead of literally hurtling around the world in planet-fouling cars

and planes using the traditional musculature of extracranial bodies, our sensorimotor stimuli can be computer-generated instead via brain-computer interfaces.

Of course, in practice our hypothetical VR-living descendants may program and dwell in virtual worlds with different laws of physics from Darwinian primitives. Our successors may occupy different modes of consciousness. Their VR social structures will presumably be transformed to reflect post-scarcity economics. Post-humans may take advantage of their limitless morphological freedom to assume a protean array of different VR bodily guises, or none at all; and they may opt to live in exotic designer heavens of their own devising. But this diversity of virtual worlds is optional. An advocate of Nozick's Experience Machine argument *can't* rely on the prospect of such alternative world-building to defeat the superpopulation scenario set out here. For computer-maintained vatworlds aren't *intrinsically* any more or less escapist than the virtual worlds of conventionally enskulled brains. When combined with radical mood-enrichment, vatworlds allow immensely more populous *and* ultra-high quality life to flourish than the brutish ecological naturalism of evolutionary history. "Heavenly" virtual worlds are neither computationally more demanding nor neurologically more energy-hungry by nature than their "Hellish" or mediocre Darwinian counterparts.

Intuitively, one may still recoil from any such paradise vatworld proposal - even though aggregate and individual welfare will be maximised i.e. both the sum and distribution of well-being are optimal. One recoils because all manner of distasteful images are evoked, not heaven-on-earth. The conclusion drawn here may sound even more repugnant than Parfit's. However, computer-choreographed "vatworlds" are no more (or less) prison-cells than traditional vertebrate skulls. So we won't be any more "trapped" than we are now; and in practice, we may feel empowered. The term "virtual" is unfortunate because it suggests the construction of an inferior simulacrum of Reality as we understand the

mind-independent world at present. On the contrary, utopian computational neuroscience offers the prospect of overpowering verisimilitude, dynamism, and seemingly boundless *Lebensraum*. Tomorrow's VR universe need not feel "crowded", let alone claustrophobic, even as the packing density of its substrates is maximised to ensure the greatest welfare of the greatest number. Optionally, "the World" in VR can be rendered no less obstinately mind-independent than it appears today. Likewise, designer drugs and genetic engineering can optionally be exploited to *enhance* our sense of authenticity - the very opposite of the derealisation and depersonalisation endemic to urban mass society. Indeed with selective use of supernormal stimuli, everything desirable can feel "more real".

Vatworlds sound ethereal since they are "disembodied". But they can incorporate an arbitrarily high level of sensuality and archaic bodily functions, if so desired. As phantom limb and similar phenomena attest, extracranial bodies are dispensable; our somatosensory cortex can't directly access "its" extracranial body even as evolved "naturally" under a Darwinian regime of natural selection. The only bodies we ever know are "virtual" bodies, whether our own, encoded pre-eminently in the somatosensory cortex, or other organically generated simulations.

The ethical assumptions underlying a Paradise-Matrix are modest and relatively uncontroversial. In the jargon of economics, a superpopulated VR vatworld scenario can be "Pareto-efficient" [Pareto-efficiency, aka Pareto-optimality, is a measure of efficiency in multi-criteria and multi-party situations. The Pareto criterion in welfare economics is normally regarded as morally undemanding. Yet the principle insists anything that can be done that would make at least one individual better off without making anyone worse off - a "Pareto improvement" or "Pareto optimization" - *should* be done.] Nonetheless the

outcome of applying these modest assumptions violates our everyday intuitions - and the usual pieties of population ethics. So which ethical theories mandate this bizarre conclusion?

Certainly, a superpopulated Paradise-Matrix is entailed by rigorous application of "hedonistic" utilitarianism. Indeed the classical utilitarian is implicitly committed to a full-blown "Pleasure Matrix" of ecstatically happy beings. The negative utilitarian may be satisfied, too, since suffering is eliminated via rewriting the genome; the extra happiness yielded by the abundance of extra inhabitants of the universe is morally redundant but unobjectionable. The case of so-called preference utilitarianism is more complicated, since the term is something of an oxymoron [or at least a misnomer] given our existing multitude of ill-conceived preferences. But some kind of Paradise Matrix is mandated by most forms of preference utilitarianism too, since both the sum and distribution of satisfied preferences are potentially maximised.

A complication for the preference utilitarian is that if and when anything akin to this VR scenario is ever seriously proposed by policy-makers, then some agents may form an explicit preference that their actions should be implemented via a traditionally routed causal chain rather than via the Matrix. Yet this newly explicit preference is presumably of limited weight when set against the astronomically wonderful payoff, i.e. the superabundance of realised preferences of trillions of post-humans pursuing their life-projects in vatworlds. The causal chain in Paradise Matrix-based civilisations is non-standard, by our lights. But it is not a contrived or "deviant" causal chain, as in Gettier-like examples against knowledge-claims. Moreover a preference for traditional embodiment is arguably a selfish preference. If sustained, the status quo will extinguish or preclude life for myriad other sentient beings: classical bodily existence carries a lethal ecological footprint. Assuming traditional embodied lifestyles are retained, then the Earth

can support only a few tens of billions of people at most; and the planet will paradoxically seem horribly crowded. By contrast, superpopulated VR vatworld scenarios permit maximal welfare [or satisfaction of preferences, etc] of the maximum number of people/post-humans [and perhaps their non-human animal companions]. Standard population ethics is radically life-denying. We don't know the names of its victims, but they are legion.

"Welfare" here is left purposely ill-defined; it's intended to embrace subjective well-being in the very richest sense for all sentient life. However, the notion of welfare isn't here tied specifically to utilitarian theory, even though non-utilitarians might view the sort of civilisation discussed in this fable as a *reductio* of applied utilitarian ethics. The construction of a Paradise Matrix *isn't* mandated by nonconsequentialist ethical theories (e.g. virtue ethics) that lack the motivating assumption of aggregate welfare-maximisation. But if, for example, you think aggregate and individual beauty [or whatever] should be maximised, then a "beauty Matrix" allied to aesthetic neural enrichment would maximise aggregate and individual beauty. Hybrid scenarios are possible too. The common feature of all these superpopulation models is that individual agents don't act out the contents of their egocentric virtual worlds independently of the Matrix; and their hedonic tone is pharmacologically and/or genetically enriched.

Yet a question naturally arises. Is the life of these thousands of billions of sublime vatworlds *really* valuable - "objectively" valuable as well as subjectively valuable? After all, the inhabitants of a Paradise Matrix are "merely" brains in vats (etc). But today we are "merely" brains in skulls. Is the value of life itself supposed to turn on a contingent historical distinction? Or on the metaphysics of perception? Either way, this discussion is intended only as an exploration of the disguised implications of the premises of standard

welfarist population ethics. The Meaning of Life or a defence of ethical realism are topics best tackled elsewhere.

One may wonder whether any adventurous souls born into a hypothetical Paradise Matrix will ever wish to be unplugged. Perhaps an inquisitive philosopher wants better to understand the (post)human predicament: confinement to "mere" stacks of mind/brains. By analogy, today a rare patient bound for the neurosurgical operating table might request that his or her brain-surgery under local anaesthesia be recorded on camera, confirming that s/he is really "just" a mind/brain/virtual world in a skull: just one skull-encased microcosm among billions. But for the most part such cranial inspection is unilluminating.

The technical challenges to developing VR vatworld civilisations are formidable by current standards. The story told here is notably light on details of the Transition from where we are now. So this fable certainly shouldn't be treated as a prediction. Yet the exponential growth of computing power could in theory deliver the computational resources to generate a rudimentary Matrix within a century or two; and biotech can deliver the reward pathway enhancements. Generating realistic virtual worlds for puny-minded *Homo sapiens* isn't unduly challenging for a mature civilisation since our visual world, for instance, is constructed from a mere 130,000,000 or so polygons a second. The technology needed is complex but not impossibly utopian. A global Paradise Matrix does not rely on speculative metaphysics or a hypothetical ontological revolution (*cf.* scanning, digitizing and "uploading" ourselves into inorganic computers, a potential recipe for zombies). But the timescale of any such revolution on Earth is of course unknown. Perhaps a Matrix will never take root; and the contrasting "empty world" regime entailed by traditional embodiment will persist indefinitely, with or without reward pathway enhancement.

So is this fable just an idle philosophers' thought-experiment designed to challenge our pre-reflective intuitions rather than serious science prophecy? Yes, quite possibly. For just who (or what) are the janitors of such a Paradise Matrix? Who are the sysadmins? Could the Matrix be hacked? *Quis custodiet ipsos custodies?* ["Who shall watch the watchers themselves?"]

Yet one reason such VR vatworld scenarios can't be excluded outright is that within a few centuries, we are likely to have conquered death and ageing. Thereafter our quasi-immortal descendants cannot procreate unchecked, not because superdense populations necessarily impair individual quality of life even if aggregate welfare is increased (Parfit's repugnant conclusion, aka the "mere addition paradox"), but because there is a physical limit to the number of mind/brains/virtual worlds that can be housed in a finite area - whether envatted or enskulled. On a cosmic scale, the Bekenstein bound presumably sets the ultimate limits to aggregate and individual welfare, at least within a given multiverse. We are unlikely to run up against this constraint in the near future. The calculation of cosmic utility functions is a task for mature superintelligence, not us.

Quantifying the well-being/life-satisfaction of a superpopulated biosphere is hard even assuming utopian neuroscanning techniques. The dilemmas of population ethics aren't eliminated altogether, even assuming some variant of the Paradise Matrix scenario outlined here. First, in order to maximise aggregate welfare it's unclear whether matter and energy should be configured to stack human-sized mind/brains/virtual worlds or alternatively to stack *supersized* posthuman mind/brains/virtual worlds. Anthropocentric bias aside, a single flourishing human mind/brain/virtual world is generally accounted superior by value theorists to 100,000 individually minimally conscious worms, say, even if aggregate vermal sentience is notionally greater. But by parity of reasoning, is a single

post-human mega-mind/brain/virtual world individually much more valuable than a multitude of diminutive speckles of human sentience? If so, should transhuman population ethicists advocate conversion of the latter into the former? How? One complication is that it's unclear whether massively supersized brains can sustain unitary consciousness (unless the temporal depth of their here-and-nows vastly exceeds traditional human awareness). Despite this uncertainty, there is no reason to suppose that posthuman mind/brain/virtual worlds won't physically be hugely bigger than their human ancestors once we are liberated from the cognitively incapacitating constraints of the human birth-canal.

A further obstacle to the exact quantification of individual and aggregate welfare in Paradise Matrices lies in quantum mechanics. Assuming universal QM, scenarios akin to some version of the superpopulated vatworlds mooted here are presumably real and physically inevitable in some branches of the multiverse; only their density in the universal wavefunction is unknown. Intuitively, their density/comparative frequency is extremely low. But the comparative abundance of sentient minds supported by such worlds relative to their sparsely populated counterparts makes it hard to be sure that living in a Paradise Matrix is atypical.

Should this wildly counter-intuitive implication be embraced by mainstream population ethics? Or should we revise our values on pain of inconsistency, supplementing our premises [i.e. maximise aggregate welfare without compromising individual welfare] with an *ad hoc* ban on vatworld-building? Perhaps so. Yet *if* we think our values are worth retaining, then it's irrational not to embrace their implications. Rationally, after mastering the technologies of invincible well-being, we should make the world a better place by

creating additional happy lifeforms. Indeed unless one is a strict negative utilitarian, perhaps we have a moral obligation to do so.

Part III: Non-Human Animals

The Antispeciesist Revolution

Speciesism.

When is it ethically acceptable to harm another sentient being? On some fairly modest[1] assumptions, to harm or kill someone simply on the grounds that they belong to a different gender, sexual orientation or ethnic group is unjustified. Such distinctions are real, but ethically irrelevant. On the other hand, species membership *is* normally reckoned an ethically relevant criterion. Fundamental to our conceptual scheme is the pre-Darwinian distinction between "humans" and "animals". In law, nonhuman animals share with inanimate objects the status of property. As property, nonhuman animals can be bought, sold, killed or otherwise harmed as humans see fit. In consequence, humans treat nonhuman animals in ways that would earn a life-time prison sentence without parole if our victims were human. From an evolutionary perspective, this contrast in status isn't surprising. In our ancestral environment of adaptation, the capacity to hunt, kill and exploit sentient beings of other species was fitness-enhancing[2]. Our moral intuitions have been shaped accordingly. Yet can we ethically justify such behaviour today?

Naively, one reason for disregarding the interests of nonhumans is the dimmer-switch model of consciousness. Humans matter more than nonhuman animals because (most) humans are more intelligent. Intuitively, more intelligent beings are more conscious than less intelligent beings; consciousness is the touchstone of moral status.

The problem with the dimmer-switch model is that it's empirically unsupported among vertebrates with central nervous systems, and probably in cephalopods such as the

octopus as well. Microelectrode studies of the brains of awake human subjects suggest that the most intense forms of experience, for example agony, terror and orgasmic bliss, are mediated by the limbic system, not the prefrontal cortex. Our core emotions are evolutionarily ancient and strongly conserved. Humans share the anatomical and molecular substrates of our core emotions with the nonhuman animals whom we factory farm and kill. By contrast, distinctively human cognitive capacities such as generative syntax, or the ability to do higher mathematics, are either phenomenologically subtle or impenetrable to introspection. To be sure, genetic and epigenetic differences exist between, say, a pig and a human being that explain our adult behavioural differences, e.g. the allele of the FOXP2(3) gene implicated in the human capacity for recursive syntax. Such mutations have little to do with raw sentience(4).

Antispeciesism.

So what is the alternative to traditional anthropocentric ethics? Antispeciesism is *not* the claim that "All Animals Are Equal", or that all species are of equal value, or that a human or a pig is equivalent to a mosquito. Rather, the antispeciesist claims that, other things being equal, equally strong interests should count equally. Experiences that are subjectively negative or positive in hedonic tone to the same degree must count for the same. And conscious beings of equivalent sentience often have equally strong interests, which (other things being equal) we must care for and respect equally - though other animals who may be less sentient can also have important interests as well. A pig, for example, is of comparable sentience to a prelinguistic human toddler. As it happens, a pig is of comparable (or superior) intelligence to a toddler as well(5). However, such cognitive prowess is ethically incidental. *If* ethical status is a function of sentience, then

to factory farm and slaughter a pig is as ethically abhorrent as to factory farm and slaughter a human baby. To exploit one and nurture the other expresses an irrational but genetically adaptive prejudice.

On the face of it, this antispeciesist claim isn't just wrong-headed; it's *absurd*. Philosopher Jonathan Haidt speaks of "moral dumbfounding"(6), where we just know something is wrong but can't articulate precisely why. Haidt offers the example of consensual incest between an adult brother and sister who use birth control. For evolutionary reasons, we "just know" such an incestuous relationship is immoral. In the case of any comparisons of pigs with human infants and toddlers, we "just know" at some deep level that any alleged equivalence in status is unfounded. After all, if there were no ethically relevant distinction between a pig and a toddler, or between a battery-farmed chicken and a human infant, then the daily behaviour of ordinary meat-eating humans would be sociopathic - which sounds crazy. In fact, unless the psychiatrists' bible, *Diagnostic and Statistical Manual of Mental Disorders*, is modified explicitly to exclude behaviour towards nonhumans, most of us do risk satisfying its diagnostic criteria for the disorder. Even so, we humans often conceive of ourselves as animal lovers. Despite the horrors of factory farming, and in general of slaughterhouses where farmed animals perish, most consumers of meat and animal products are clearly not sociopaths in the normal usage of the term; most factory farm managers are not wantonly cruel; and the majority of slaughterhouse workers are not sadists who delight in suffering. Serial killers of nonhuman animals are just ordinary people doing a distasteful job - "obeying orders" - on pain of losing their livelihoods.

Should we expect anything different? Political theorist Hannah Arendt spoke famously of the "banality of evil"(7). *If* twenty-first century humans are collectively doing something posthuman superintelligence will reckon monstrous, a crime against sentience akin to the

[human] Holocaust or Atlantic slave trade, then it's easy to assume our moral intuitions would disclose this to us. Our intuitions don't disclose anything of the kind; so we sleep easy. But both natural selection and the historical record offer powerful reasons for doubting the trustworthiness of our naive moral intuitions. So the possibility that human civilisation *might* be founded upon some monstrous evil should be taken seriously - even if the possibility seems transparently absurd at the time.

One possible speciesist response is to raise the question of "potential". Even if a pig is as sentient as a human toddler, there is a fundamental distinction between human toddlers and pigs. Only a toddler has the potential to mature into a rational adult human being.

The problem with this response is that it contradicts our treatment of humans who lack "potential". Thus we recognise that a toddler with a progressive disorder who will never live to celebrate his third birthday deserves at least as much love, care and respect as his normally developing peers - not to be packed off to a factory farm on the grounds it's a shame to let good protein go to waste. We recognise a similar duty of care for mentally handicapped adult humans and cognitively frail old people. For sure, historical exceptions exist to this perceived duty of care for vulnerable humans, e.g. the Nazi "euthanasia" program, with its eugenicist conception of "life unworthy of life". But by common consent, we value young children and cognitively challenged adults for who they are, not simply for who they may - or may not - one day become. On occasion, there may controversially be *instrumental* reasons for allocating more care and resources to a potential genius or exceptionally gifted child than to a normal human. Yet disproportionate *intra*species resource allocation may be justified, not because high IQ humans are more sentient, but because of the anticipated benefits to society as a whole.

Practical Implications.

1. Invitrotarianism.

The greatest source of severe, chronic and readily avoidable suffering in the world today is man-made: animal agriculture, most notably factory farming. Humans currently slaughter over fifty billion sentient beings each year. One implication of an antispeciesist ethic is that factory farms should be shut and their surviving victims rehabilitated.

In common with most ethical revolutions in history, the prospect of humanity switching to a cruelty-free diet first strikes most practically-minded folk as utopian dreaming. "Realists" certainly have plenty of hard evidence to bolster their case. As English essayist William Hazlitt observed, "The least pain in our little finger gives us more concern and uneasiness than the destruction of millions of our fellow-beings." Without the aid of twenty-first century technology, the mass slaughter and abuse of our fellow animals might continue indefinitely. Yet tissue science technology promises to allow consumers to become moral agents without the slightest hint of personal inconvenience. Lab-grown *in vitro* meat produced in cell culture rather than a live animal has long been a staple of science fiction. But global veganism - or its ethical invitrotarian equivalent - is no longer a futuristic fantasy. Rapid advances in tissue engineering mean that *in vitro* meat will shortly be developed and commercialised. Today's experimental cultured mincemeat can be supplanted by mass-manufactured gourmet steaks for the consumer market. Perhaps critically for its rapid public acceptance, *in vitro* meat does not need to be genetically modified - thereby spiking the guns of techno-luddites who might otherwise worry about "FrankenBurgers". Indeed, cultured meat products will be more "natural" in some ways than their antibiotic-laced counterparts derived from farmed animals.

Momentum for commercialisation is growing. Non-profit research organisations like New Harvest(8), working to develop alternatives to conventionally produced meat, have been joined by hard-headed business executives. Visionary entrepreneur and Stanford academic Peter Thiel(9) has just funnelled $350,000 into Modern Meadow, a start-up that aims to combine 3D printing with *in vitro* meat cultivation. Within the next decade or so, gourmet steaks could be printed out from biological materials. In principle, the technology should be scalable. While work on *in vitro* meat continues, rapid advances are being made in the development of so-called plant meats. Beyond Meat(10), for example, has already brought to market the first plant-based meat with a texture almost identical to chicken flesh.

Tragically, billions of nonhuman animals will atrociously suffer and die this century at human hands before the dietary transition is complete. Humans are not obligate carnivores; eating meat and animal products is a lifestyle choice. "But I like the taste!" is not a morally compelling argument. Vegans and animal advocates ask whether we are ethically entitled to wait on a technological fix. The antispeciesist answer is clear: no.

2. Compassionate Biology.

If and when humans stop systematically harming other sentient beings, will our ethical duties to members of other species have been discharged? Not if the same ethical considerations as apply to members of other human races or age-groups apply also to members of other species of equivalent sentience. Thus if famine breaks out in sub-Saharan Africa and young human children are starving, then we recognise we have a duty to send aid; or better still, to take proactive measures to ensure famines do not arise in the first instance, i.e. to provide not just food aid but family planning. So why not

assist, say, starving free-living elephants? Until recently, no comparable interventions were feasible for members of other species. The technical challenges were insurmountable. Not least, the absence of cross-species fertility control technologies would have often made bad problems worse. Helping free-living nonhumans would just lead to an unsustainable population explosion followed by ecological collapse. Yet thanks to the exponential growth of computer power, every cubic metre of the planet will shortly be computationally accessible to micro-management, surveillance and control. Harnessed to biotechnology, nanotechnology and robotics, such tools confer unprecedented power over nature. With unbridled power comes complicity. Ethically speaking, how many of the traditional cruelties of the living world do we wish to perpetuate? Orthodox conservation biologists argue we should not "interfere": humans can't "police" nature. Antispeciesists disagree. Advocates of compassionate biology argue that humans and nonhumans alike should not be parasitised, starved, disembowelled, asphyxiated, or eaten alive.

As always, bioconservatives insist such miseries are "natural"; status quo bias runs deep. "Custom will reconcile people to any atrocity", observed George Bernard Shaw. Snuff movies in the guise of nature documentaries are quite popular on Youtube, a counterpoint to the Disneyfied wildlife shows aired on mainstream TV. Moreover, even sympathetic critics of compassionate biology might respond that helping free-living members of other species is prohibitively expensive. An adequate welfare safety net scarcely exists for humans in many parts of the world. So how can we contemplate its extension to nonhumans - even just to large-brained, long-lived vertebrates in our nature reserves? Provision of comprehensive healthcare for all free-living elephants(11), for example, might cost between two or three billion dollars annually. Compassionate stewardship of the living world would be technically daunting too, entailing ecosystem management, cross-species fertility control via immunocontraception, veterinary care,

emergency famine relief, GPS tracking and monitoring, and ultimately phasing out or genetically "reprogramming"(12) carnivorous predators. The notional bill could approach the world's 1.7 trillion dollar annual arms budget. But irrespective of cost or timescale, *if* we are to be consistently non-speciesist, then decisions about resource allocation should be based not on species membership, but on sentience. An elephant, for example, is at least as sentient as a human toddler - and may well be as sentient, if not sapient, as adult humans. If it is ethically obligatory to help sick or starving children, then it's ethically obligatory to help sick or starving elephants - not just via crisis interventions, but via long-term healthcare support.

A traditional conservation biologist might respond that elephants helped by humans are no longer truly wild. Yet on such a criterion, clothes-wearing humans or beneficiaries of food aid and family planning aren't "wild" humans either. Why should this matter? "Free-living" and "wild" are conceptually distinct. To assume that the civilising process should be confined to our own species is mere speciesist prejudice. Humans, transhumans and posthumans must choose what forms of sentience we want to preserve and create on Earth and beyond. Humans already massively intervene in nature, whether through habitat destruction, captive breeding programs for big cats, "rewilding", etc. So the question is not whether humans should "interfere", but rather what ethical principles should govern our interventions(13).

Speciesism and Superintelligence.

Why should transhumanists care about the suffering of nonhuman animals? This is not a "feel-good" issue. One reason we should care cuts to the heart of the future of life in the universe. Transhumanists differ over whether our posthuman successors will most likely

be nonbiological artificial superintelligences; or cyborgs who effectively merge with our hyperintelligent machines; or our own recursively self-improving biological descendants who modify their own genetic source code and bootstrap their way to full-spectrum superintelligence(14). Regardless of the dominant lifeform of the posthuman era, biological humans have a vested interest in the behaviour of intellectually advanced beings towards cognitively humble creatures - if humans survive at all. Compared to posthuman superintelligence, archaic humans may be no smarter than pigs or chickens - or perhaps worms. This does not augur well for *Homo sapiens*. Western-educated humans tend to view Jains as faintly ridiculous for practising *ahimsa*, or "harmlessness", sweeping the ground in front of them to avoid inadvertently treading on insects. How quixotic! Yet the fate of sentient but cognitively humble lifeforms in relation to vastly superior intelligence is precisely the issue at stake as we confront the prospect of posthuman superintelligence. How can we ensure a Jain-like concern for comparatively simple-minded creatures such as ourselves? Why should superintelligences care any more than humans about the well-being of their intellectual inferiors? Might distinctively *human*-friendly superintelligence turn out to be as intellectually-incoherent as, say, Aryan-friendly superintelligence? If human primitives are to prove worthy of conservation, how can we implement technologies of impartial friendliness towards other sentients? And if posthumans *do* care, how do we know that a truly benevolent superintelligence wouldn't turn Darwinian life into utilitronium with a communal hug?

Viewed in such a light, biological humanity's prospects in a future world of superintelligence might seem dire. However, this worry expresses a one-dimensional conception of general intelligence. No doubt the nature of mature superintelligence is humanly unknowable. But presumably *full-spectrum*(15) superintelligence entails, at the very least, a capacity to investigate, understand and manipulate both the formal and the

subjective properties of mind. Modern science aspires to an idealised "view from nowhere"(16), an impartial, God-like understanding of the natural universe, stripped of any bias in perspective, and expressed in the language of mathematical physics. By the same token, a God-like superintelligence must also be endowed with the capacity to impartially grasp all possible first-person perspectives - not a partial and primitive Machiavellian cunning of the kind adaptive on the African savannah, but an unimaginably radical expansion of our own fitfully growing circle of empathy.

What such superhuman perspective-taking ability might entail is unclear. We are familiar with people who display abnormally advanced forms of "mind-blind"(17), autistic intelligence in higher mathematics and theoretical physics. Less well known are hyper-empathisers who display unusually sophisticated social intelligence. Perhaps the most advanced naturally occurring hyper-empathisers exhibit mirror-touch synaesthesia(18). A mirror-touch synaesthete cannot be unfriendly towards you because she feels your pain and pleasure as if it were her own. In principle, such unusual perspective-taking capacity could be generalised and extended with reciprocal neuroscanning technology and telemetry into a kind of naturalised telepathy, both between and within species. Interpersonal and cross-species mind-reading could in theory break down hitherto invincible barriers of ignorance between different skull-bound subjects of experience, thereby eroding the anthropocentric, ethnocentric and egocentric bias that has plagued life on Earth to date. Today, the intelligence-testing community tends to treat facility at empathetic understanding as if it were a mere personality variable, or at best some sort of second-rate cognition for people who can't do IQ tests. But "mind-reading" can be a highly sophisticated, cognitively demanding ability. Compare, say, the sixth-order intentionality manifested by Shakespeare. In *Othello*, for example, Shakespeare intends his audience believe that Iago intends that Othello imagines that Desdemona is in love

with Cassio, and that Cassio reciprocates Desdemona's amorous feelings(19). Thus we shouldn't conceive superintelligence as akin to God imagined by someone with autistic spectrum disorder. Rather, full-spectrum superintelligence entails a God's-eye capacity to understand the rich multi-faceted first-person perspectives of diverse lifeforms whose mind-spaces humans would find incomprehensibly alien.

An obvious objection arises. Just because ultra-intelligent posthumans may be *capable* of displaying empathetic superintelligence, how do we know such intelligence will be exercised? The short answer is that we don't: by analogy, today's mirror-touch synaesthetes might one day neurosurgically opt to become mind-blind. But then equally we don't know whether posthumans will renounce their advanced logico-mathematical prowess in favour of the functional equivalent of wireheading. If they do so, they won't be superintelligent. The existence of diverse first-person perspectives is a fundamental feature of the natural world, as fundamental as the second law of thermodynamics or the Higgs boson. To be ignorant of fundamental features of the world is to be an *idiot savant*: a super-Watson(20) perhaps, but not a superintelligence(21).

High-Tech Jainism?

Jules Renard once remarked, "I don't know if God exists, but it would be better for His reputation if He didn't." God's conspicuous absence from the natural world needn't deter us from asking what an omniscient, omnipotent, all-merciful deity would want humans to do with our imminent God-like powers. For we're on the brink of a momentous evolutionary transition in the history of life on Earth. Physicist Freeman Dyson predicts we'll soon "be writing genomes as fluently as Blake and Byron wrote verses"(22). The ethical risks and opportunities for apprentice deities are huge.

On the one hand, Karl Popper warns, "Those who promise us paradise on earth never produced anything but a hell"[23]. Twentieth-century history bears out such pessimism. Yet for billions of sentient beings from less powerful species, life on Earth *is* hell. They end their miserable lives on our dinner plates: "for the animals it is an eternal Treblinka", writes Jewish Nobel laureate Isaac Bashevis Singer[24].

In a more utopian vein, some utterly sublime scenarios are technically feasible later this century and beyond. It's not clear whether experience below Sidgwick's[25] "hedonic zero" has any long-term future. Thanks to molecular neuroscience, mastery of the brain's reward circuitry could make everyday life wonderful beyond the bounds of normal human experience. There is no technical reason why the pitiless Darwinian struggle of the past half billion years can't be replaced by an earthly paradise for all creatures great and small. Genetic engineering could allow "the lion to lie down with the lamb." Enhancement technologies could transform killer apes into saintly smart angels. Biotechnology could abolish suffering throughout the living world. Artificial intelligence could secure the well-being of all sentience in our forward light-cone. Our quasi-immortal descendants may be animated by gradients of intelligent bliss orders of magnitude richer than anything physiologically feasible today.

Such fantastical-sounding scenarios may never come to pass. Yet if so, this won't be because the technical challenges prove too daunting, but because intelligent agents choose to forgo the molecular keys to paradise for something else. Critically, the substrates of bliss don't need to be species-specific or rationed. Transhumanists believe the well-being of all sentience[26] is the bedrock of any civilisation worthy of the name.

* * *

NOTES

1. How modest? A venerable tradition in philosophical meta-ethics is anti-realism. The meta-ethical anti-realist proposes that claims such as it's wrong to rape women, kill Jews, torture babies (etc) lack truth value - or are simply false. (*cf*. JL Mackie, *Ethics: Inventing Right and Wrong*, Viking Press, 1977.) Here I shall assume that, for reasons we simply don't understand, the pain-pleasure axis discloses the world's inbuilt metric of (dis)value. Meta-ethical anti-realists may instead wish to interpret this critique of speciesism merely as casting doubt on its internal coherence rather than the substantive claim that a non-speciesist ethic is objectively true.

2. Extreme violence towards members of other tribes and races can be fitness-enhancing too. See, e.g. Richard Wrangham & Dale Peterson, *Demonic Males: Apes and the Origins of Human Violence*, Houghton Mifflin, 1997.

3. Fisher SE, Scharff C (2009). "FOXP2 as a molecular window into speech and language". *Trends Genet*. 25 (4): 166–77. doi:10.1016/j.tig.2009.03.002. PMID 19304338.

4. Interpersonal and interspecies comparisons of sentience are of course fraught with problems. Comparative studies of how hard a human or nonhuman animal will work to avoid or obtain a particular stimulus give one crude behavioural indication. Yet we can go right down to the genetic and molecular level, e.g. interspecies comparisons of SCN9A genotype. (*cf*. http://www.pnas.org/content/early/2010/02/23/0913181107.full.pdf) We know that in humans the SCN9A gene modulates pain sensitivity. Some alleles of SCN9A give rise to hypoalgesia, others alleles to hyperalgesia. Nonsense mutations yield congenital insensitivity to pain. So we could systematically compare the SCN9A gene and

its homologues in nonhuman animals. Neocortical chauvinists will still be sceptical of non-mammalian sentience, pointing to the extensive role of cortical processing in higher vertebrates. But recall how neuroscanning techniques reveal that during orgasm, for example, much of the neocortex effectively shuts down. Intensity of experience is scarcely diminished.

5. Held S, Mendl M, Devereux C, and Byrne RW. 2001. "Studies in social cognition: from primates to pigs". *Animal Welfare* 10:S209-17.

6. Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, Pantheon Books, 2012.

7. Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil*, Viking Press, 1963.

8. http://www.new-harvest.org/

9. "PayPal Founder Backs Synthetic Meat Printing Company", *Wired*, August 16 2012. http://www.wired.com/wiredscience/2012/08/3d-printed-meat/

10. Beyond Meat: http://www.beyondmeat.com/

11. https://www.abolitionist.com/reprogramming/elephantcare.html

12. https://www.abolitionist.com/reprogramming/index.html

13. The scholarly literature on the problem of wild animal suffering is still sparse. But perhaps see Arne Naess, "Should We Try To Relieve Clear Cases of Suffering in Nature?", published in *The Selected Works of Arne Naess*, Springer, 2005; Oscar Horta, "The Ethics of the Ecology of Fear against the Nonspeciesist Paradigm: A Shift in the Aims of Intervention in Nature", *Between the Species*, Issue X, August 2010. http://digitalcommons.calpoly.edu/bts/vol13/iss10/10/ ; Brian Tomasik, "The Importance

of Wild-Animal Suffering", http://www.utilitarian-essays.com/suffering-nature.html ; and the first print-published plea for phasing out carnivorism in Nature, Jeff McMahan's "The Meat Eaters", *The New York Times*. September 19, 2010. http://opinionator.blogs.nytimes.com/2010/09/19/the-meat-eaters/

14. *Singularity Hypotheses, A Scientific and Philosophical Assessment*, Eden, A.H.; Moor, J.H.; Søraker, J.H.; Steinhart, E. (Eds) Springer 2013. http://singularityhypothesis.blogspot.co.uk/p/table-of-contents.html

15. David Pearce, *The Biointelligence Explosion*. (preprint), 2012. https://www.biointelligence-explosion.com.

16. Thomas Nagel, *The View From Nowhere*, OUP, 1989.

17. Simon Baron-Cohen (2009). "Autism: the empathizing–systemizing (E-S) theory" (PDF). *Ann N Y Acad Sci* 1156: 68–80. doi:10.1111/j.1749-6632.2009.04467.x. PMID 19338503.

18. Banissy, M. J. & Ward, J. (2007). Mirror-touch synesthesia is linked with empathy. *Nature Neurosci*. doi: 10.1038/nn1926.

19. See 'The Social Brain Hypothesis and its Relevance To Cognitive Psychology' by R.I.M. Dunbar, published in *Evolution and the Social Mind: Evolutionary Psychology and Social Cognition*, Forgas, J.P; Haselton, M. G.; von Hippel, W. (Eds) Psychology Press 2007.

20. Stephen Baker. *Final Jeopardy: Man vs. Machine and the Quest to Know Everything*. Houghton Mifflin Harcourt. 2011.

21. Orthogonality or convergence? For an alternative to the convergence thesis, see Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced

Artificial Agents", 2012, http://www.nickbostrom.com/superintelligentwill.pdf; and Eliezer Yudkowsky, Carl Shulman, Anna Salamon, Rolf Nelson, Steven Kaas, Steve Rayhawk, Zack Davis, and Tom McCabe. "Reducing Long-Term Catastrophic Risks from Artificial Intelligence", 2010. http://intelligence.org/files/ReducingRisks.pdf

22. Freeman Dyson, "When Science & Poetry Were Friends", *New York Review of Books*, August 13, 2009.

23. As quoted in Jon Winokur, *In Passing: Condolences and Complaints on Death, Dying, and Related Disappointments*, Sasquatch Books, 2005.

24. Isaac Bashevis Singer, *The Letter Writer*, 1964.

25. Henry Sidgwick, *The Methods of Ethics*. London, 1874, 7th ed. 1907.

26. The Transhumanist Declaration (1998, 2009). http://humanityplus.org/philosophy/transhumanist-declaration/

Reprogramming Predators

"*And the wolf shall dwell with the lamb and the leopard shall lie down with the kid, and the calf and the young lion and the fatling together and a little child shall lead them."*

Isaiah 11:6

"The total amount of suffering per year in the natural world is beyond all decent contemplation. During the minute that it takes me to compose this sentence, thousands of animals are being eaten alive, others are running for their lives, whimpering with fear, others are being slowly devoured from within by rasping parasites, thousands of all kinds are dying of starvation, thirst and disease. It must be so."

Richard Dawkins, *River Out of Eden* (1995)

The Problem of Predation

A biosphere without suffering is technically feasible. In principle, science can deliver a cruelty-free world that lacks the molecular signature of unpleasant experience. Not merely can a living world support human life based on genetically preprogrammed gradients of well-being. If carried to completion, the abolitionist project entails ecosystem redesign, cross-species immunocontraception, marine nanorobots, rewriting the vertebrate genome, and harnessing the exponential growth of computational resources to manage a compassionately run global ecosystem. Ultimately, it's an ethical

choice whether intelligent agents opt to create such a world - or instead express our natural status quo bias and perpetuate the biology of suffering indefinitely.

This utopian-sounding vision isn't the upshot of some exotic new ethical theory. The abolitionist project follows quite straightforwardly from the application of a classical utilitarian ethic and advanced biotechnology. More controversially, the abolitionist project is the scientific expression of what Gautama Buddha aspired to some 2500 years ago: "May all that have life be delivered from suffering". Provisionally, let's assume that *other things being equal*, a cruelty-free world is ethically desirable, i.e. it would be ideal if no involuntary physical or emotional pain were undergone by any sentient being. As our technology matures, some hard choices are ethically unavoidable if these noble sentiments are ever to be turned into practice.

First, a cruelty-free world entails a transition to global veganism. Realistically, global veganism won't come about purely or even mainly via moral persuasion within any plausible timeframe. Such a momentous transition can occur only after the advent of mass-produced artificial meat ("Krea") that is at least as cheap, tasty and healthy as flesh from slaughtered factory farmed animals - with moral argument playing a modest supporting role. For sure, there is still the "yuck factor" to overcome. But when delicious, cruelty-free cultured-meat products become commercially available, the "yuck factor" should actually work in favour of cultured meat - since meat from factory farmed animals is not merely morally disgusting but often physically disgusting too.

However, this transition isn't enough. Even the hypothetical world-wide adoption of a cruelty-free diet leaves one immense source of suffering untouched. Here we shall explore one of the thorniest issues the end of suffering entails: the future of what biologists call obligate predators. For the abolitionist project seems inconsistent with one

of our basic contemporary values. The need for species conservation is so axiomatic that an explicitly normative scientific sub-discipline, conservation biology, exists to promote it. In the modern era, the extinction of a species is usually accounted a tragedy, especially if that species is a prominent vertebrate rather than an obscure beetle. Yet if we seriously want a world without suffering, how many existing Darwinian life-forms can be conserved in their current guise? What should be the ultimate fate of iconic species like the large carnivores? True, only a minority of the Earth's species are carnivorous predators: the fundamental laws of thermodynamics entail that whenever there is an "exchange of energy" between one trophic level and another, there is a significant loss. The majority of the planet's 50,000 or so vertebrate species are vegetarian. But among the minority of carnivorous species are some of the best known creatures on the planet. Should these serial killers be permitted to prey on other sentient beings indefinitely?

A few forms of extinction are almost universally applauded even now. Thus the demise of the smallpox virus in the wild is wholly unlamented, though controversy persists over whether the last two pathogenic *Variola* copies in human custody should be destroyed. The virus could be recreated from scratch if needed. Technically, viruses aren't alive; they can't independently replicate. Yet the same welcome will be extended to the extinction of scores of bacterial pathogens that cause human disease if we can plot their eradication as efficiently as the two *Variola* variants that cause smallpox. Likewise, exterminating the five kinds of protozoan parasites of the genus *Plasmodium* that cause malaria would be almost entirely uncontentious; a human child dies from malaria on average every twelve seconds. Protozoans have zero consciousness or minimal consciousness, depending on one's ultimate theory of mind. Either way, it makes no sense or minimal sense to speak literally of the "interest" of the plasmodium. Only

figuratively do plasmodia have interests. Plasmodia matter significantly only insofar as their existence affects the welfare of sentient beings. Our reverence for the diversity of life has its limits. More complicated than plasmodia are parasitic worms, locusts or cockroaches, which almost certainly do have at least limited consciousness. Yet that consciousness is still *comparatively* dim compared to vertebrates. Cockroaches have decentralised nervous systems. In consequence, they presumably lack a unitary experiential field. This is not to say that cockroaches should ever be wantonly hurt. Perhaps their constituent nerve ganglia in individual segments experience sharp pains; cockroaches retain rudimentary learning skills and live for up to a week without a head. Yet if the world's 4,000 species of cockroach were no longer extant outside a handful of vivariums, then their absence in the wild would be accounted no great loss on any plausible version of the felicific calculus. Nor would extinction of the swarming grasshoppers we know as plagues of locusts. A swarm of 50 billion locusts can in theory eat 100,000 tonnes of foodstuffs per day. Around 20% of food grown for human consumption is eaten by herbivorous insects. A truly utopian future world would lack even minuscule insect pangs of hunger, and its computational resources could micro-manage the well-being of the humblest arthropods - including the Earth's estimated 10 quintillion (10^{18}) insects. In the meantime, we must prioritise. On a neo-Buddhist or utilitarian ethic, the criterion of value and moral status is degree of sentience. In a Darwinian world, the welfare of some beings *depends* on their doing harm to others. So initially, ugly compromises are inevitable as we bootstrap our way out of primordial Darwinian life. Research must focus on how the ugliness of the transitional era can be minimised.

More controversial than the case of tapeworms, cockroaches or locusts would be reprogramming or phasing out snakes and crocodiles. Snakes and crocodiles cause

innumerable hideous deaths in the world each day. They are also part of our familiar conceptual landscape thanks to movies, zoos, TV documentaries, and the like - though a relaxed tolerance of their activities is easier in the comfortable West than for, say, a grieving Indian mother who has lost her child to a snakebite. Snakes are responsible for over 50,000 human deaths each year.

Most controversial of all, however, would be the extinction - or genetically-driven behavioural modification - of members of the cat family. We'll focus here on felines rather than the "easy" cases like parasitic tapeworms or cockroaches because of the unique status of members of the cat family in contemporary human culture, both as pets/companion animals and as our romanticised emblems of "wildlife". Most contemporary humans have a strong aesthetic preference in favour of continued feline survival. Their existence in current guise is perhaps the biggest ethical/ideological challenge to the radical abolitionist. For our culture glorifies lions, with their iconic status as the King of the Beasts; we admire the grace and agility of a cheetah; the tiger is a symbol of strength, beauty and controlled aggression; the panther is dark, swift and elegant; and so forth. Innumerable companies and sports teams have enlisted one or other of the big cats for their logos as symbols of manliness and vigour. Moreover, cats of the domestic variety are the archetypal household pets. The worldwide domestic cat population has been estimated at around 400 million. We romanticise their virtues and forgive their foibles, notably their playful torment of mice. Indeed, rather than being an object of horror - and compassion for the mouse - the torment of mice has been turned into stylized entertainment. Hence Tom-and-Jerry cartoons. By contrast, talk of "eliminating" predation can sound sinister. What would "phasing out" or "reprogramming" predators mean in practice? Most disturbingly, such terms are evocative of *genocide*, not universal compassion.

Appearances deceive. To get a conceptual handle on what is really going on during "predation", let's compare our attitude to the fate of a pig or a zebra with the fate of an organism with whom those non-human animals are functionally equivalent, both intellectually and in their capacity to suffer, namely a human toddler. On those rare occasions when a domestic dog kills a baby or toddler, the attack is front-page news. The offending dog is subsequently put down. Likewise, lions in Africa who turn man-eater are tracked down and killed, regardless of their conserved status. This response isn't to imply lions - or for that matter rogue dogs - are morally culpable. But by common consent they must be prevented from killing any more human beings. By contrast, the spectacle of a lion chasing a terrified zebra and then asphyxiating its victim can be shown on TV as evening entertainment, edifying viewing even for children. How is this parallel relevant? Well, if our theory of value aspires to a God's-eye perspective, stripped of unwarranted anthropocentric bias in the manner of the physical sciences, then the well-being of a pig or a zebra inherently matters no less than the fate of a human baby - or any other organism endowed with an equivalent degree of sentience. If we are morally consistent, then as we acquire God-like powers over Nature's creatures, we should take analogous steps to secure their well-being too. Given our anthropocentric bias, thinking of non-human vertebrates not just as equivalent in moral status to toddlers or infants, but *as though they were* toddlers or infants, is a useful exercise. Such reconceptualisation helps correct our lack of empathy for sentient beings whose physical appearance is different from "us". Ethically, the practice of *intelligent* "anthropomorphism" shouldn't be shunned as unscientific, but embraced insofar as it augments our stunted capacity for empathy. Such anthropomorphism can be a valuable corrective to our cognitive and moral limitations. This is not a plea to be sentimental, simply for impartial benevolence. Nor is it even a plea to take "sides" between killer and prey. Human serial killers who prey on

other humans need to be locked up. But ultimately, it's vindictive morally to blame them in any ultimate sense for the fate of their victims. Their behaviour supervenes on the fundamental laws of physics. *Tout comprendre, c'est tout pardonner*. Yet this indulgence doesn't extend to permitting them to kill again; and the abolitionist maintains the same principle holds good for nonhuman serial killers too.

Parasites, Predators and Serial Killers

Suffocation induces a sense of extreme panic. It's a comparatively rare experience in contemporary human life, although panic disorder, an anxiety disorder characterised by recurring severe panic attacks, is extremely unpleasant and quite common. Whatever its cause, the experience of suffocation is horrific. One's lungs feel as though they will burst at any second. There is a loss of control of bodily functions. There is no psychological "coping mechanism", just an all-consuming fear, as witnessed by the traumatic effects of the waterboarding torture practised by the CIA; the entangled piles of bodies of victims in the Nazi gas chambers frantically clawing over each other to gasp in the last traces of breathable air; and the death-agonies of millions of herbivores every day in the wild.

It would be a mercy if the experience of suffocation were fundamentally different in human and non-human animals. This fond hope might be realized if the intuitively appealing "dimmer-switch" model of consciousness were tenable - and an organism's degree of consciousness were reliably correlated with its degree of intelligence. The dimmer-switch model leads one to suppose that slow asphyxiation feels significantly less dreadful for a zebra than for a human being. Naïvely, we imagine that the asphyxiation of our vertebrate cousins is merely rather unpleasant for its victims rather than unbearable beyond words. Unfortunately, our core emotions are also the most intense modes of conscious experience; and the neural structures that mediate such primitive

modes of consciousness are among the most strongly evolutionarily conserved. Intense fear, disgust, anger, hunger, thirst and pain are among the most powerful sensations known. They are phylogenetically ancient. Intense pleasure can of course be vivid too; but pleasure is not our focus here. In contrast to the phenomenology of our core emotions, the phenomenology of serial, "logical" thought-episodes in the distinctively human prefrontal cortex is vanishingly faint, as microelectrode studies and introspection of our own linguistic thought-episodes attest. Moreover the problem is worse than "just" the acute intensity of suffering. Wildlife documentaries encourage the notion that death in Nature is typically fast. Some deaths are indeed mercifully swift. Many other deaths are slow and agonizing. Simply to survive, members of the cat family in the wild must inflict appalling suffering on their fellow mammals. More disturbingly still, domestic cats torment millions of terrified small rodents and birds each day before killing them - essentially for entertainment. Cats lack an adequate theory of mind. They don't have an empathetic understanding of the implications of what they are doing. For a cat, the terrified mouse with whom it is "playing" has no more ethical significance than a zombie warrior slaughtered by a teenager playing "violent" video games. But an absence of malice is no comfort to the tormented mouse.

Most modern city-dwellers do not lose any sleep over the cruelties of Nature, or indeed give them more than a passing thought. Implicitly, it's assumed such suffering doesn't *matter*. Or if it does matter, it doesn't matter enough to mitigate or abolish. Why? The list of reasons below is incomplete but worth noting.

- *Our supposed lack of complicity due to impotence.*

    Throughout most of history, mankind could no more contemplate reordering the food chain than contemporary humans could contemplate changing, say,

Planck's constant or the rest mass of an electron. What happens in Nature is traditionally "just the way things are"; hence no one's fault. Shortly, however, the persistence of nonhuman animal suffering will be our direct responsibility - whether abdicated or accepted remains to be seen.

- *A television-based conception of the living world.*

  Our view of the living world is significantly shaped by wildlife documentaries - and the narrative structure that their voiceovers and uplifting mood-music provide. Wildlife documentaries are designed to be entertaining as well as educative. They offer a spectacle of death, violence and aggression in a manner that is no longer deemed acceptable when practised on humans. It's the same reason why for hundreds of years the Romans enjoyed the gory violence of the amphitheatre, and why nonhuman animals are still hunted by some humans for "sport". One contemporary psychological problem for many people in everyday life isn't pain or depression but *boredom*, a lack of stimulation. The sight of conflict and killing is exciting.

- *Selective realism.*

  We like our war movies and horror films to be realistic - but not too realistic. Likewise, wildlife documentaries aren't expected to portray the full nastiness of Darwinian life, although there would doubtless be a sizeable audience if they did so, as YouTube viewing figures attest. The question of "taste" ensures that the more squeamish sensibilities of a wider television audience are spared most of the horror while still being entertained by the drama. A few minutes of stalking. The ambush. The thrill of the chase. A five-second shot of the lion with its jaw on the zebra's throat. Next the camera cuts to a pride of lions

eating a lifeless carcass. Realistic depictions of the full nastiness of predation are taboo. As David Attenborough once remarked to some viewers who complained that a scene shown was too gruesome: "You ought to see what we leave on the cutting room floor". This text hints at the horror, but words don't really portray it. And even the most explicit video couldn't evoke the first-person reality of being dismembered, strangled, impaled, drowned, poisoned or eaten alive. The problem of suffering in Nature described here is *worse* - and its prevention more morally urgent - than we suppose. For example, try to imagine what it's like slowly dying of thirst over several days during the dry season. There may be no overt drama. It's just subjectively horrific. Hence the ethical obligation on the dominant species to stop such horrors as soon as we acquire the technical expertise to do so.

- *Adaptive empathy deficits.*

  Human empathetic responses are shaped by natural selection. Genetically, it's fitness-enhancing for parents to experience an empathetic response to the feelings of their children, but maladaptive to feel compassion for their children's "food". Selection pressure for empathy toward members of other races or species - or genetic rivals - is weak to non-existent since such empathy wouldn't promote our reproductive success - except insofar as it enabled our ancestors to hunt and kill more successfully, or outwit their enemies. The human mind/brain isn't designed to track the well-being of other members of our own species beyond our own tribe, let alone all other sentient beings. Such empathy sporadically occurs, but it has been selected, not selected *for*; its existence is just the byproduct of a fitness-enhancing adaptation. The discussion here focuses on empathy-deficits born of

anthropocentric bias; but the ultimate empathy-deficit stems from *egocentric* bias. Coalitions of selfish genes throw up vehicles whose egocentric virtual worlds do not track the well-being of other sentient beings impartially. Perhaps only clones (i.e. identical twins, triplets, etc) could "naturally" do so reliably.

- *The cruelties of the living world are "natural", therefore worth conserving: a price worth paying for the glories of Nature.*

  This is the way things ought to be, because this is the way things have always been. Status quo bias is endemic. Thus it simply doesn't seem to have occurred to some otherwise smart thinkers in slave-owning societies that slavery could be morally wrong. Had the case for universal human freedom been put to them, the idea might well have seemed as silly as does questioning the inviolability of the food-chain at present. Potentially, status quo bias can take benign guises too. If we already lived in a cruelty-free world, the notion of re-introducing suffering, exploitation and creatures eating each other would seem not so much frightful as unimaginable - no more seriously conceivable than reverting to surgery without anaesthesia today. Of course, the extent of our status quo bias shouldn't be exaggerated. There is something self-intimatingly wrong with one's own intense pain while it lasts; and to a greater or lesser degree, we can generalize this urgent sense of wrongness to other suffering beings with whom we identify. But since most humans aren't in agony most of the time, any generalizations we make tend to be weak; and restricted in scope on account of our evolutionary descent.

Extinction versus Reprogramming

1) Extinction

One solution to the barbarities of predation is to use indiscriminate depot-contraception on carnivores and allow predators rapidly to die out, managing the resultant population effects on "prey" species via more selective forms of depot-contraception. Such advanced computer-controlled contraception technologies could be used selectively on zebra, buffalo, wildebeest, etc, so our wildlife parks don't become overpopulated. The feasibility of such population-management is shown by the use of fertility-regulating depot-contraception on male elephants living in the Kruger National Park in preference to the distressing practice of "culling". Most human wildlife enthusiasts prefer the use of depot-contraception as a means of population-control to killing families of elephants; but they also find the idea of an absence of lions *even in our wildlife parks* to be abhorrent. This may be so; but the case for selective extinction isn't absurd, even if we reject it after due deliberation. Why fetishise life-forms endowed with a heritable tendency to prey on and strangulate others? Parallels with the Third Reich are best used sparingly; but sometimes they are apt. It's worth asking why there is such an extensive Net-based community that regards black-uniformed SS and their regalia as fascinating - far more fascinating than, say, colourless NKVD apparatchiks and the squalor of the Gulag, or the half-forgotten Ottoman genocide of the Armenians. If exercised with panache, extreme power and violence intrigue us. Thankfully, our captivation by stylish embodiments of evil has limits: immaculate SS are a lot more elegant than their victims on the way to asphyxiation in the gas chambers; but we aren't going to preserve or literally re-create the SS, except in movies. Some monstrous life-forms are best banished to the archives for good. By the same token, the spectacle of large predators hunting and asphyxiating their terrified victims is more visually compelling than herbivores browsing inoffensively. Which would

you rather watch on TV? If there is misplaced emotion here, it lies in our fetishising the strong, handsome and powerful over the gentle and vulnerable.

It is worth stressing, repeatedly since the charge is made time and again, that this indictment of predators is not to blame a lion [or a domestic cat] for its behaviour. First, barring genetic engineering or freaks of nature, lions are obligate carnivores. Secondly, they don't understand the implications of what they are doing. Any mutant lion with a theory of mind capable of empathising with its prey would be rapidly outbred by "sociopathic" lions. Barring human intervention, a compassionate lion who rejected the "law of the jungle" would starve to death. Consequently so would its cubs. Lions are "sociopathic" towards members of prey species, just as throughout history many humans have behaved sociopathically to members of other races and tribes - though enslavement has been more common in humans than cannibalism. ["Nothing more strongly arouses our disgust than cannibalism, yet we make the same impression on Buddhists and vegetarians, for we feed on babies, though not our own." - Robert Louis Stevenson] Either way, the extinction scenario for predatory life-forms needs to be taken seriously - but not out of naïve moralism. The committed abolitionist may tentatively *predict* that centuries hence lions will not exist outside the digital archives - any more than the smallpox virus. For that matter, one may tentatively *predict* that the same fate will befall feral *Homo sapiens*. The conditionally activated capacity to act in bloodthirsty and sexually aggressive ways has been genetically adaptive in the past. We are all the descendants of murderers and rapists. Thus geneticists claim that over 16 million people today may be descended from Genghis Khan. But prediction is not advocacy.

Moreover, even if - contrary to what is argued here - one believes that lions and cheetahs *are* inherently valuable in exactly their current guise, there is still an opportunity-cost to their existence - where the opportunity-cost is the value of the next

best alternative creature forgone as the result of choosing one life-form over another. Are members of the cat family really ideal life-forms? In a world of finite resources, only a small spectrum of phenotypes can be expressed out of the entire abstract state-space of possible genomes. Assume, as seems likely, that (post)humans will shortly have demigod-like powers over what kinds of life-form and modes of consciousness the living world sustains. Ecological resources - and indeed mass-energy itself - will still be finite. If we opt to instantiate lions, then their existence entails depriving other species of life. So to judge that lions should exist is to affirm that it is *better*, in some sense, that sociopathic killing machines prowl the Earth rather than alternative herbivores. Taken literally, this argument ultimately applies to archaic *Homo sapiens* too. Is the source code of our constituent matter and energy optimally organized? Or would our DNA be better reconfigured to encode a species of blissfully superintelligent "smart angels"? The difference is that archaic humans will most likely become extinct not through outside agency, but as we progressively rewrite our own source code, reprogram "human nature", and bootstrap away into becoming posthuman.

2) Reprogramming

Alternatively, should carnivorous predators be genetically "reprogrammed" or otherwise behaviourally modified rather than allowed to go extinct in the wild? Pre-reflectively, such reprogramming is all but impossible. In practice, the technical expertise is probably a few decades away at most. One can see anticipations of post-Darwinian life even now, albeit at the level of individuals rather than whole species.

a) One example of behavioural management technology at work is the creation of remote-controlled rats ("ratbots"). Electrodes implanted in the pleasure centres of a rat's brain can make the rat follow instructions of its own volition, so to speak, at least from

the perspective of the rat. Investigators currently anticipate that such enhanced rodents could be used to search for landmines or buried (human) victims of earthquakes. In the future, there is nothing to stop such technology being widely installed - together with mini-cameras and GPS tracking devices - in predatory carnivores to deter sociopathic violence against other sentient life-forms. Indeed, with the right reinforcement schedule, the most ferocious carnivore could be turned into a model citizen in our wildlife parks. With suitable surveillance and computer control, whole communities of ex-predators could be discreetly guided in the norms of non-violent behaviour. No "inhumanity" would be involved in the behavioural reshaping process since at no time are the brain's pain-centres stimulated. Nor does the augmented animal ever experience a sense of being made to act against its will. Yes, the ex-predator is "enslaved" to its reward circuitry; but so are humans. ["All men seek happiness. This is without exception. Whatever different means they employ, they all tend to this end. The cause of some going to war, and of others avoiding it, is the same desire in both, attended with different views. This is the motive of every action of every man, even of those who hang themselves." - Blaise Pascal] Indeed indefinitely generous doses of pure pleasure could be administered to members of the managed species in reward for "virtuous" behaviour.

Conversely, members of "prey" species can be bio-engineered to lose their currently well-justified terror of predators. Again, this re-engineering sounds technically daunting. Yet recall how rodents infected with the parasitic protozoan *Toxoplasma gondii* lose their normal fears and actually seek out cat urine-marked areas. Pharmacology, neuroelectrodes and genetic technologies all offer possible solutions to the molecular pathology of fear when its persistence becomes functionally redundant. In the long run, the same kinds of hedonic enrichment, intelligence-amplification and life-extension technologies available to humans later this century can be extended across the

phylogenetic tree. "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity", affirms the World Health Organization constitution. The abolitionist project broadens this pledge of complete physical, mental and social well-being beyond our own species to (ultimately) all sentient beings. Any such extension sounds fanciful now. So too would a description of contemporary human healthcare 200 years ago. The same ethical principle is at stake. Counter-intuitively, the "law of accelerating returns" of computer processing-power means that the transition to universal well-being could be accomplished in decades rather than millennia *if a human governmental consensus existed* - though centuries might be a more conservative timeframe for marine ecosystems.

b) Another anticipation of how reprogramming might work is found "naturally" in the wild. Between 2002 and 2004 a lioness christened Kamunyak ["The Blessed One" in Samburu] in central Kenya repeatedly adopted a baby oryx, at least six times in all, protecting each baby oryx from other predators, including leopards and kindred hungry lions. Kamunyak even allowed a mother oryx occasionally to come and feed her calf before chasing her away. "The lioness must have a mental aberration", stated a UNESCO official in Nairobi. In principle, the hypernurturing behaviour of eusocial mammals like lions could be harnessed in genetically tweaked carnivores to protect members of species they currently predate. On this scenario, a ready dietary supply of cultured meat would have to be laid on as well unless more radical genetic interventions were made to alter existing lion physiology. Today, *in vitro* meat exists only as a laboratory curiosity. Commercial products are a decade or more away. But mass-producing cultured meat for "wild" or domestic carnivores should prove easier than creating the textures of genetically engineered meat needed to satisfy the more exacting tastes of human gourmet diners.

The technical details of such a program are of course challenging, to say the least. Nature has few food chains in the strict sense; complex food webs abound. But an ecosystem can support only around five or six trophic levels between its effectively insentient primary producers and the large predatory carnivores at the top of the trophic pyramid. For only 10% or so of an organism's energy is passed on to its predator; the rest is lost as heat to the environment. So the problems of humane ecosystem management should be computationally tractable in a well-run wildlife park. The entire African lion population is currently believed to be around 30,000, down from around 400,000 in 1950. Lion numbers are dwindling fast due to habitat loss and conflicts with humans. The remaining lion populations are often geographically isolated from each other. So inbreeding and a lack of genetic diversity are increasing. Outside of zoos and wildlife parks, lions will soon die out in the absence of human intervention, as will most large terrestrial mammals this century in the wake of habitat degradation. For instance, the Earth's most species-rich biome, tropical evergreen forest, is being lost at around two percent each year. Reprogramming and behavioural management technology can guarantee the civilised survival of reformed lions and their relatives for human ecotourists to enjoy, if we so choose.

One critical response to the prospect of reprogramming carnivorous predators runs as follows. A quasi-domesticated lion that does not prey on members of other species has ceased to be a true lion. Lions, by their very nature, kill members of prey species (and sometimes hyenas, cheetahs and each other). Yes, lions kill their victims in gruesome ways described as "bestial" if done by humans; but such behaviour is perfectly natural if practised by lions: it's one aspect of their "behavioural phenotype". Hunting behaviour is a natural part of their species essence.

Yet here we come to the nub of the issue: the alleged moral force of the term "natural". If any creature, *by its very nature*, causes terrible suffering, albeit unwittingly, is it morally wrong to change that nature? If a civilised human were to come to believe s/he had been committing acts that caused grievous pain for no good reason, then s/he would stop - and want other moral agents to prevent the recurrence of such behaviour. May we assume that the same would be true of a lion, if the lion were morally and cognitively "uplifted" so as to understand the ramifications of what s/he was doing? Or a house cat tormenting a mouse? Or indeed a human sociopath? Currently, sociopathy in humans cannot be cured; but various interventions, both genetic and pharmacological, have been mooted. When the therapeutic option does exist, should the treatment be offered? At present, sociopathic human serial killers must be locked up for life. A "cure" that enabled human serial killers to become truly pro-social, empathetic beings would indeed "rob" them of their former identity. Such an intervention would be "coercive", maybe not in the strict sense, but effectively so if the alternative is being locked up indefinitely. The same is true of violent repeat sex-offenders. Now consider another form of behaviour in lions whose practice by humans would spell incarceration for life. A mature male lion is genetically programmed to go into a pride, challenge the reigning male, and (if the invading male is victorious) methodically kill off the young cubs of the defeated male. Killing his rival's cubs helps maximise the inclusive fitness of his DNA. Their mother will then go on heat again so the invading male lion can mate with her and sire his own cubs. Around a third of all lion cubs born perish in this way. Mercifully, nothing so mechanistic plays out with human stepfathers and young stepchildren. But statistically, to be raised as a stepchild is immensely more risky than being brought up by both one's biological parents. If there were therapeutic interventions that could help stifle hostile feelings on the part of stepfathers to young stepchildren, would their use be desirable? Many

stepfathers, for instance, might welcome their availability. Otherwise decent parents may be disturbed by the hostile feelings they feel toward their stepchildren - even though the vast majority of stepparents do not act on them in the extreme form practised by male lions. Infanticide practised on a sentient being is cruel irrespective of the species identity of the perpetrator. In the future, interventions can prevent its occurrence in our wildlife parks even at the price of tweaking the "natural" genomes of their members.

A Pan-Species Welfare State?

"He that slayeth an ox is as he that slayeth a man."

(Isaiah 66:3)

Over the last century, a welfare state for humans was introduced in Western European societies so that the most vulnerable members of our own species wouldn't suffer avoidable hardship. Even in affluent Western nations, notably in the USA, coverage can be woefully inadequate. Provision in Third World nations ranges from the adequate to patchy to almost non-existent. And by the standards of posterity, all contemporary healthcare will presumably seem rudimentary. But a commitment to the underlying principle, at least, is well-established: no one should literally starve or suffer death or debility from preventable illness. Likewise, universal education is designed to maximise life opportunities for all. Universal healthcare aims to ensure everyone gets medical treatment. Child-support agencies intervene when vulnerable children are at risk of abuse or neglect. Initially, Social Darwinists decried the introduction of such safeguards; eugenicists fretted that a welfare state would allow the "unfit" to breed and propagate "bad" genes; free-market fundamentalists worried that a safety-net would sap habits of manly self-reliance; and so forth. Yet the need for at least basic welfare guarantees now seems obvious, though controversy persists over their nature and optimal extent - and

financing. Social Darwinism in its rawest form now has few defenders beyond devotees of Ayn Rand. The problem is not just that existing welfare provision is inadequate: it's also arbitrarily species-specific. In common with the plight of vulnerable humans before its introduction, the welfare of vulnerable non-human animals depends mostly on private charity. No universal guarantees of non-human well-being exist. Vivisection, the abomination of factory farming, and the industrialized mass-killing of nonhuman animals persists unchecked. Beyond our closest cousins the great apes, the systematic extension of state-enforced welfare guarantees to other species in the wild, sounds too far-fetched an option to generate sustained critical analysis. Proverbially, charity begins at home; let's worry about "our" species first. No great ideological debate has erupted on the case for compassionate ecosystem redesign because the case for preserving the ecological status quo is perceived as too obvious to need defending; and the transformative potential of biotech, infotech and nanotech is still barely glimpsed. Traditionally, of course, Nature has just seemed too big. Insofar as any justification at all has been felt necessary for wild animal suffering, the narrative told to rationalise the cruelties of Nature has claimed that predation of the sick and the weak is for "the good of the species". This fable is no longer scientifically tenable. Natural selection doesn't operate on that level. Further, it is equally un-Darwinian to suppose there is some fundamental ontological and ethical gulf between "us" and "them", between primates of the genus Homo and nonhuman animals. On any universal ethic, the inclusive rather than contrastive use of "we" must extend to all sentient beings.

However, the most formidable obstacle to reprogramming predators and designing compassionate ecosystems isn't ideology but simple status quo bias. Most of the arguments elaborated against abolishing suffering in humans don't even get off the ground in nonhumans. The anguish of members of others species will not inspire its

victims to create great works of art or literature, build their characters, afford interesting contrasts, allow opportunities for personal growth, and so on. Their suffering is just nasty and inherently pointless. On the face of it, reprogramming the source code of the rest of the living world is orders of magnitude computationally harder than re-engineering humans. But the immensity of task shouldn't be overstated. CRISPR genome-editing technologies are a game-changer. The technical challenges of reprogramming nonhuman animals are in some respects *easier* to overcome than in humans. Thus one of the most formidable stumbling-blocks to sustainable mood-enrichment in humans isn't engineering raw pleasure - wireheading or speedballing could do that now. What's hard is reprogramming our reward circuitry in ways that don't compromise our social responsibility and cognitive performance - not just on gross measures of the sorts of cleverness scored by IQ tests, but subtler abilities involving creativity, empathetic understanding, introspective self-insight - and perhaps too the capacity for fundamental self-doubt from which future intellectual revolutions may spring. In short, the challenge lies in preventing the superhappy from becoming either "opiated" or manic. Similar constraints on the future happiness of nonhuman animals either don't apply to the same degree or don't apply at all. The prospect of "lions on soma" may be surreal; but it's difficult to see how its introduction could be judged reckless or immoral.

Clearly as it stands, the abolitionist project is more of a sketch than a blueprint. So one urgent priority is the creation of academic research programs so that abolitionist scholarship can become a rigorous scientific discipline. Such a discipline will not be value-free; but nor will it be any more normative than conservation biology - or scientific medicine. A critical aspect of advanced ecosystem redesign will be prior computational modelling - the exhaustive hunt for previously unanticipated side-effects of interventions at different trophic levels in the "food chain". Philosophical manifestos can gloss over

technical difficulties; wildlife park management teams will need to confront them. Either way, abolitionism needs to enter the academic and political mainstream, with organisational structures and advocacy groups to match. A cruelty-free world will entail coordinated national, intergovernmental and United Nations action on an unprecedented scale.

Understandably, sceptics can dismiss such scenarios as sheer technofantasy. The sociological, ethico-religious and ideological obstacles to the design of a cruelty-free planetary ecosystem can seem insurmountable even if its ultimate *technical* feasibility is acknowledged. But predicting the growth of a global anti-speciesist ethic to complement an anti-racist ethic isn't as unreasonable as it first sounds. Consider the central dogmas of the world's major religions. To what extent is the abolitionist project a disguised implication of some of our core principles? *Ahimsa*, the Sanskrit term meaning to do no harm (literally: the avoidance of violence - *himsa*) is central to the family of religions originating in ancient India: Hinduism, Buddhism and especially Jainism. *Ahimsa* is a rule of conduct that bars the killing or injuring of living beings. The ecosystem redesign advocated here is essentially the scientific expression of *ahimsa* on a global scale, shorn of its karmic metaphysics. It's true that Judaeo-Christian and Islamic religion have been less sympathetic historically to the interests of nonhuman animals than the non-Abrahamic traditions of the Indian subcontinent. Throughout much of the Christian era, vegetarianism in Western Europe was regarded as a heresy. God's Biblical promise of "dominion" over the rest of the animal kingdom has standardly been interpreted as divine license for domination and exploitation. Yet "dominion" can also be (re)interpreted as responsibility for stewardship. What if Isaiah is correct and the wolf and the lion really *can* lie down with the lamb? Would a compassionate God want us to preserve the biology of suffering when its perpetuation becomes optional? Recall too that (with one exception)

each of the 114 *suras* of the Islamic Qur'an begins, "Allah is merciful and compassionate." The name of God used most often in the Qur'an is "al-Rahim", meaning literally "the All-Compassionate." Any implication that God's compassion is stunted compared to the moral imagination of mere mortals might seem blasphemous. Muhammad the Prophet speaks of the need for "universal mercy". According to one tradition (*Hadith Mishkat* 3:1392) Muhammad taught that "all creatures are like a family of God; and He loves the most those who are the most beneficent to His family." As infotech, nanorobotics and biotechnology mature - or accelerate - perhaps religious and secular ethicists alike will treat the maximal relief of suffering as the default assumption from which departures need to be justified, not a radical new ethic in need of justification itself. On almost *every* future scenario, we're destined to "play God". So let's aim to be compassionate gods and replace the cruelty of Darwinian life with something better.

A Welfare State For Elephants?

A Case Study of Compassionate Stewardship

INTRODUCTION

High-Tech Jainism?

Within the next few decades, the exponential growth of computer power will ensure every cubic metre of the planet is computationally accessible to remote monitoring, micro-management and control. Harnessed to biotechnology and nanorobotics, this growth in surveillance and control capabilities presents huge risks and huge opportunities. In a dystopian vein, such technologies lend themselves to advanced war-fighting, or they could be used to sustain an Orwellian dictatorship. Alternatively, such technologies could deliver compassionate stewardship of the entire living world.

High-tech Jainism of the kind needed to safeguard the interests of smaller mammals, let alone the well-being of marine vertebrates and (ultimately) members of other phyla, is still decades away. The CRISPR revolution in genome-editing is only a few years old. Nanotechnology, and in particular nanorobotics, is still in its infancy. The obstacles to a cruelty-free world aren't merely technical. Even as the technologies of intervention become cheaper and readily available, human status quo bias may postpone implementation of a compassionate biology indefinitely. The ideology of conservation biology is deeply entrenched. So ambitious germline interventions to "reprogram" traditional predator species, orchestrate pan-species fertility regulation, and guarantee

the well-being of *all* sentience in our forward light-cone probably aren't on the horizon for a century or more. Yet this sort of timescale doesn't mean discussions on ethical intervention/stewardship are just idle philosophising. On the contrary, some forms of compassionate stewardship are *technically* feasible right now. Many of the worst and most morally urgent cases of wild animal suffering are the most accessible to intervention; and also the least expensive to remedy.

Why Elephants?

Launching our compassionate stewardship of the living world with free-living elephants might seem an arbitrary choice of species. Why choose elephants for a feasibility study? But from an ethical point of view, elephants are a prime candidate. With a brain weighing just over five kilograms, the African elephant has the largest mind/brain of any terrestrial vertebrate. On some fairly modest assumptions, elephants are among the most sentient nonhuman animals. All the technologies necessary for a comprehensive elephant healthcare program are available, in principle if not yet in practice. Nothing speculative or even especially futuristic in the way of high technology need be invoked to lay out the foundations of an elephant welfare state, although software tools for efficient remote monitoring and tele-diagnostics need further development. Admittedly, free-living elephants offer a comparatively "easy" example of compassionate species care. Elephants are large, long-lived, charismatic and herbivorous. No seemingly irreconcilable interests are involved (e.g. lions versus zebras) in safeguarding their interests because mature elephants typically have no natural predators besides *Homo sapiens*. The limiting factor on elephant population size in the absence of human predation or artificial fertility regulation is inadequate nutrition.

The starkest exception to this generalisation is the terrible case of lions in Savuti. Opportunistic killing of juvenile, sick or badly injured elephants by other predators, notably hyenas, does occur; but such killing is relatively infrequent. It's the kind of horror that compassionate stewardship of Nature could prevent.

Are Cared-For Elephants Really Free-Living?

As with humans, "free-living" is not synonymous with "wild". Critics of any blueprint for an elephant welfare safety-net may claim that the recipients of healthcare, food aid and emergency relief won't be truly free. This is not the place to explore the metaphysics of freedom, nor to enter human left-right political debate. Elephants are not economic actors; the expression "welfare state" may set libertarian alarm-bells ringing, but in this context it's politically neutral. If intelligently run, crisis-interventions in time of drought needn't give rise to an elephant "dependency culture"; this is not feeding time at the zoo. Critics will undoubtedly allege that elephants whom humans have assisted or saved from harm are no longer truly "wild" or "natural". But humans who wear clothes or who take medicine aren't thereby less human or somehow diminished compared to their "wild" conspecifics. Likewise elephants.

Some animal advocates claim that the use of immunocontraception in over-populated wildlife parks violates the presumed right of nonhuman animals to procreative freedom. Intimate or remote monitoring as canvassed here violates the supposed right of nonhuman animals to privacy. Yet worries about privacy breaches, in particular, are an unwarranted anthropomorphic projection on our part. The alternative to fertility control is witnessing one's calf slowly starve to death in a degraded habitat, or the brutal practice of "culling" (i.e. massacring whole elephant families) to prevent ecological devastation.

The loss of a calf or a child, or of a matriarch or a mother, is traumatic for elephants and humans alike.

Costs of Intervention

What would be the financial cost, at contemporary prices, of cradle-to-the-grave healthcare and welfare provision for the entire population of free-living African elephants? The elephant population of the African continent currently stands at around 500,000. Elephant taxonomy is currently in flux; but the half-million figure includes what is commonly known as the savannah (or bush) elephant, *Loxodonta africana*, and the forest species of elephant, *Loxodonta cyclotis*. An annual cost of somewhere between two and three billion dollars seems plausible. Most of the same challenges and opportunities arise for securing the well-being of the Asian elephant, *Elephas maximus*. An estimated 40,000 Asian elephants are left in the wild. So the type of program sketched out below could be implemented in SouthEast Asia at a fraction of the price.

Most human healthcare expenses are incurred in the last six months, and often the last six weeks, of life. In the case of elephants, we simply don't know the upper bounds to life-expectancy, given adequate late-life dentition. Assuming effective orthodontic care, this particular challenge, i.e. managing the age-related infirmities of free-living geriatric elephants, will (presumably) be decades away from the launch of an orthodontic healthcare service. After being GPS-chipped, vaccinated and (where necessary) provided with immunocontraception, most free-living elephants could be remotely monitored but otherwise largely left in peace - apart from in years of severe drought and famine, when costly crisis-interventions will be necessary. To flourish, free-living elephants need a habitat that offers fresh water, plentiful vegetation for grazing and browsing; and some

available shade. A mature African bush elephant typically ingests over 200 kilograms of vegetable matter daily. The elephant emergency equivalent of Humanitarian daily rations (HDRs) will be quite bulky. When needed, the cost of providing additional vaccinations, vitamin and mineral supplements, painkillers, anti-inflammatories, parasiticides, sedatives and anaesthetics, antibiotics, antifungals and antivirals, disinfectants and cleaning agents will not be negligible; but the relevant agents are almost all off-patent. Training and labour costs of ancillary support staff in sub-Saharan Africa are comparatively low; and likely to remain so for the foreseeable future. Close, politically sensitive collaboration with the local human populations will be vital to the long-term success of the project. Elephant healthcare work could provide valuable employment. Some forms of expertise could be delivered only by specialist veterinarians. An air-ambulance service would incur significant transport costs.

Immunocontraception

Ivory poaching and habitat destruction have dramatically reduced unprotected elephant populations over the course of the past two hundred years. However, in favourable conditions elephant populations may increase at four to five percent per year. Inevitably, such growth is ecologically unsustainable. In the long run, humans will have to choose the overall level and demographic profile of elephant populations in our wildlife parks, or otherwise let Nature (i.e. famine and malnutrition-related deaths) take its course. The victims of "natural" climatic disasters will mainly be the young, the sick and the old. As with tomorrow's humans, advances in behavioural genetics and reproductive technologies will shortly allow use of preimplantation genetic diagnosis to choose everything from pain thresholds (*cf.* variant pain-modulating alleles of the SCN9A gene)

to susceptibility to depression (*cf*. the role of the COMT gene and serotonin transporter gene (5-HTTLPR)) to personality variables. Or policy makers may opt to perpetuate the traditional genetic roulette of sexual reproduction. Once again, political and ethical choices will be unavoidable.

Neonatal Care

Provision of perinatal elephant care is potentially expensive. Immediately after birth, the young calf is most vulnerable to predation by lions, hunting dogs and hyenas. An elephant calf's first year of life is his or her most hazardous. Mortality rates range from below 10% to more than 30%. Calf mortality is liable to increase when ranges are restricted and habitats change so opportunities for browsing and midday shade become less available. Causes of juvenile death include not just predation, but disease, accidents, drought, starvation, nutritional deficiencies, stress, heat stress, drowning, becoming trapped in mud-holes, snake bite and congenital malformation. In the face of potential predators, the calf's mother will vigorously defend her newborn. Unfortunately, the calf may not always be able to keep in the secure position under her mother's abdomen. Moreover the calf will still be vulnerable to predators for some years to come. After six months or so, the youngster starts to move further from his or her mother. If potential predators are near, s/he is at risk of being left behind if the herd is disturbed or stampeded.

Elephants typically give birth to one calf. Less than one percent of births involve twins: one and often both calves usually die within weeks or months of birth. Intervention here will be needed to ensure a favourable outcome.

Orphaned elephants will need special protection. A calf normally continues suckling at least until two years old. Unaided, orphaned young elephants below the age of two or three years rarely survive in the wild. In a few countries, the basic infrastructure of elephant orphanages is already in place; such rescue and rehabilitation services just need extension, systematisation and adequate funding. After weaning, annual elephant mortality rates are perhaps five or six percent until about the age of 50 years. Mortality rates rise sharply in the sixth decade.

Injuries

Elephants are normally robust and peaceable. However, fights do occur, particularly between bull elephants disputing access to a female in oestrus. Occasionally, one or both parties may be badly injured in such aggressive encounters. Bone fractures will need to be treated by elephant orthopaedic specialists.

Disease Prevention and Treatment

Like humans, elephants are susceptible to infection by tuberculosis, a treatable disease caused by a bacterium that affects especially the lungs. Mosquito-borne diseases are also a risk. Anthrax may be contracted via contaminated water or soil. Some ailments are specific to elephants, notably trunk paralysis and elephant pox, but other afflictions are common to humans and elephants alike, ranging from intestinal colic and constipation to pneumonia. Elephants may even catch the common cold, though this condition is self-limiting. Ill elephants often attempt to self-medicate, treating digestive diseases through fasting or consumption of bark, bitter herbs or alkaline earth. Such limited self-treatment can be complemented by human expertise in scientific medicine.

Elephant Orthodontics

Human depredations aside, the greatest source of mature elephant morbidity and mortality is inadequate nutrition. Elephants replace their teeth multiple times. The fifth set of chewing teeth (molars) lasts until the elephant is in his or her early forties. The sixth - and usually final - set must last the elephant the rest of his or her life. Ageing elephants may roam in search of marshy areas with softer food sources. As the final set of molars wears away during the late fifties, the elephant is no longer able adequately to chew food. S/he will die from the effects of malnutrition or starvation. Free-living elephants do not usually live much past sixty years. Elderly elephant deaths generally occur during the dry season. This is because dry food cannot be effectively sheared by the residual smooth grinding surface of the worn-down sixth molar.

The weakened and emaciated elephant will eventually collapse. Helpless, s/he may be eaten alive by scavengers and predators. Late-life orthodontics to prevent this fate will be more costly than routine GPS tracking or immunocontraception. But the kinds of material used for "false teeth" could last decades without need for replacement.

Drought

During severe droughts, the construction and maintenance of artificial waterholes will be necessary to prevent tragedies. However, during a drought deaths are normally from starvation or malnutrition rather than thirst. This is because elephants are reluctant to leave known water-sources to find food. Deaths may also be related to heat stress. However, the congregation of herds of undernourished and malnourished elephants at

remaining water-holes will make provision of crisis nutritional support easier and cheaper.

Elephant Psychiatric Care

Like people, elephants may suffer low mood, anxiety disorders and depression. Elephants grieve when they lose a calf or close family member. Psychoses may occur, but primarily in consequence of captivity, rarely in their natural habitat. In common with people, the incidence of endogenous depression is lower when elephants are living in their natural habitat in small family groups rather than suffering solitary confinement in captivity. Post-traumatic stress disorder in the aftermath of hunting or natural trauma could potentially be treated with inexpensive beta-blockers. Determining the appropriate drug dosage in different treatment regimens still depends on metabolic scaling formulas. Such crude procedures are used because comparatively few pharmacokinetic studies have been conducted to provide elephant-specific information. If an ethical discipline of compassionate biology replaces a doctrinaire conservation biology, this relative lack of studies can be remedied.

Uncertainties

For now, financial projections of comprehensive free-living elephant care will depend on back-of-an-envelope calculations rather than a rigorous methodology. But a $2.5 billion annual price-tag of full healthcare and welfare provision for the entire population of free-living African elephants may turn out to be pessimistic. Financial planners will just need to bear in mind the potential for cost overruns and unexpected expenses that tend to plague any new enterprise. The likely extent of corruption, maladministration and the

growth of a welfare bureaucracy in an elephant healthcare program are hard to quantify too. In practice, the great majority of Africa's 500,000 elephant population would need far less than the annual $5,000 per head this figure allows. Neurochipping, individual genome sequencing, vaccinations, GPS tracking and (when appropriate) immunocontraception would cost at most a few hundred dollars. The GPS-chipping, individual genome sequencing and vaccinations would typically be a one-off expense rather than a regular part of the annual budget. What's feasible at modest expense for e.g. all UK "domestic" dogs is no less feasible for free-living elephants. Chipping could range from simple tagging to more complex remote-monitoring of health status (e.g. cortisol monitoring. Elevated cortisol levels are suggestive of high stress and consequent need for investigation and possible compassionate intervention.)

What would be the timescale for complete coverage of Africa's elephant population? Perhaps one or two years - but only if an international consensus existed.

The Speciesist Objection

Even the most sympathetic critic of compassionate biology is likely to raise a seemingly compelling objection. Hundreds of millions of human beings do not yet enjoy an adequate welfare safety-net. Couldn't the estimated annual two or three billion dollars cost of an elephant welfare program be more fruitfully spent promoting human welfare instead? Africa needs Obamacare, not elephant care.

Whatever our response to this objection, our answer should not be clouded by arbitrary anthropocentric bias, i.e. speciesism. It's worth stressing that *anti*-speciesism is *not* the claim that "All Animals Are Equal", or that all species are of equivalent value, or that the well-being of a human - or an elephant - is as important as the well-being of a mosquito.

Rather it's the claim that other things being equal, all animals, human and nonhuman, of *equivalent sentience* are of equal value and deserve equal consideration. Comparisons are invidious; but the anti-speciesist argues that ethically what matters in resource allocation is not ethnic group or species membership, but *sentience*. Thus there is no evidence that degree of sentience is bound up with e.g. the species-specific allelic variations of the FOXP2 gene implicated in the human capacity for generative syntax. Microelectrode studies of the human brain using verbally competent awake subjects confirm that the most intense forms of sentience, notably our core limbic emotions, are also the most phylogenetically primitive, whereas the phenomenology associated with such distinctively human cognitive capacities as higher mathematics or generative syntax is also the most subtle and rarefied. The phenomenology of language-generation is barely accessible to introspection. Abundant evidence suggests elephants are at least as sentient as human the toddlers. Elephants can pass the "mirror test", thereby demonstrating a capacity for reflective self-awareness. The elephant hippocampus is comparatively larger than human hippocampus, presumably a function of an elephant's prodigious memory. Elephants are endowed with an immense, highly convoluted neocortex subserving their complex tactile, visual, acoustic and olfactory communication systems and capacity for empathetic understanding. Elephants display sophisticated social cognition. More controversially, their comparatively larger limbic systems suggest that elephants may be at least as sentient as adult humans, albeit lacking the logico-mathematical and linguistic prowess that allows modern *Homo sapiens* to dominate the planet. Either way, even if, cautiously and conservatively, we judge elephants are no more sentient than prelinguistic human toddlers, we still have a duty to protect their interests. By the same token, the affluent world also has an ethical duty to "interfere"

and help vulnerable children in developing nations. Examining the issue of Third World Aid here would take us too far afield.

A more compelling objection to implementing an elephantcare program is that our overriding ethical priority should be ending the suffering and killing for which humans are directly responsible. Factory farming is the greatest source of severe and readily avoidable suffering in the world today. Hannah Arendt famously remarked on the "banality of evil". Most humans are complicit or financially implicated in the nonhuman animal holocaust. Even though a pig, for example, is of comparable sentience to a prelinguistic toddler, humans routinely do things to factory farmed pigs that would earn a life-sentence in prison if our victims were human. The development and commercialisation of *in vitro* meat promises global veganism/invitrotarianism later this century. In the meantime, billions of sentient beings will have been abused and slaughtered to satisfy our taste for their flesh.

CONCLUSION

The Biggest Obstacle

For better or worse, humans or our descendants will be responsible for life on Earth for the indefinite future. Despite the initially daunting technical challenges, the biggest obstacle to compassionate stewardship of the world's free-living nonhuman animal population is not technical or even financial but ideological. Most people are prone to status quo bias. Such innate bias is normally rationalised by some version of the "appeal to Nature", sometimes (mis)characterised as "the naturalistic fallacy". What is natural is good.

The irrationality of the "appeal to Nature" is illustrated by a simple thought-experiment. Imagine, fancifully, if starvation, disease, parasitism, disembowelling, asphyxiation and being eaten alive were *not* endemic to the living world - or such miseries have already been abolished and replaced by an earthly paradise. Would anyone propose there is ethical case for (re)introducing them? Even proposing such a thought-experiment can sound faintly ridiculous.

However, our bioconservativism is not wholly consistent. If presented with a specific example of terrible suffering, for example an elephant mother and her calf trapped in a mudhole, *most* people argue we should intervene rather than permit the horror to unfold "naturally". Human benevolence is typically weak, erratic and sentimental rather than rule-bound, and often negligible, but it's still real. By focusing initially on grisly concrete examples, a broad consensus on the *principle* of compassionate intervention can potentially be established, though not of course whether intervention should be piecemeal or systematic - or how it should be funded. Eliciting support for *ad hoc* animal "rescues" is the critical wedge that advocates of compassionate stewardship of Nature need to press their case further. Once we accept that intervention to prevent suffering in free-living nonhuman animals is *sometimes* ethically justified, and *sometimes* even ethically required, a straightforward question then arises. Does free-living animal suffering matter only when humans happen to notice it? What principle(s) should govern our interventions? If we can underwrite the well-being of elephants, should we aim, ultimately, to extend our compassionate stewardship to the rest of the living world?

COMPASSIONATE BIOLOGY

How CRISPR-based "gene drives" could cheaply, rapidly and sustainably reduce suffering throughout the living world

"The total amount of suffering per year in the natural world is beyond all decent contemplation. During the minute that it takes me to compose this sentence, thousands of animals are being eaten alive, others are running for their lives, whimpering with fear, others are being slowly devoured from within by rasping parasites, thousands of all kinds are dying of starvation, thirst and disease. It must be so."

Richard Dawkins, *River Out of Eden* (1995)

INTRODUCTION

Towards a Post-Darwinian Biosphere

The idea that sentient beings shouldn't harm each other, or allow each other to come to harm, was once purely utopian. Later this century and beyond, the policy option will be technically feasible. Sociological credibility is another issue. Yet a plea of "It must be so" is no longer technically, ecologically or ethically correct. In the post-CRISPR era, whether intelligent agents decide to preserve, reform, or phase out the biology of involuntary suffering will be an ethical choice.

Four policy options for the biosphere:

1) "Pleistocene_rewilding" - restoring much of the planet to its state before the human impact.

2) The status quo - essentially an extension of existing conservation biology: more wildlife parks, minimal intervention - conservation with no regard to the subjective well-being of individuals, just the abstract health of species and ecosystems.

Traditional Conservation Biology

3) Compassionate biology, ultimately extending to all free-living sentients: CRISPR-based gene drives, cross-species fertility-regulation via immunocontraception, GPS-tracking and monitoring, genetic tweaking and/or *in vitro* meat for obligate carnivores, a pan-species welfare state in tomorrow's Nature reserves: in short, "high-tech Jainism".

High-tech Jainism

"Genetically Engineering Almost Anything" (Nova)

"'Gene Drives' And CRISPR Could Revolutionize Ecosystem Management" (*Scientific American*)

4) Phasing out free-living non-human sentients altogether.

"Why improve Nature when destroying it is so much easier?" (Robert Wiblin)

This paper will sketch and defend a version of (3), what might be called Compassionate Conservation. For sure, the blueprint outlined has little near-term chance of being implemented as it stands. The reason for sketching what's *technically* feasible with the tools of synthetic biology is that only *after* human complicity in the persistence of suffering in the biosphere is acknowledged can we hope to have an informed socio-

political debate on the morality of its perpetuation. No serious ethical discussion of free-living animal suffering can begin in the absence of recognition of human responsibility for nonhuman well-being.

* * *

When a friend of the American composer John Cage asked, "Don't you think there's too much suffering in the world?", Cage answered, “No, I think there's just the right amount”. Few ethicists openly express such Zen-like equanimity at the suffering of other sentient beings; but until recently all ecologists shared Richard Dawkins' assumption of its inevitability. However, the living world is on the brink of a technical and ethical revolution - and a major evolutionary transition in the development of life. Humanity will shortly be able to decide the optimal level of suffering both for members of our own species and across the tree of life itself. Not least, CRISPR-driven gene drives can cheaply, rapidly and dramatically reduce suffering in all sexually reproducing species.

Until 21st century biotechnology, the sheer cost, computational complexity and technical obstacles to tackling suffering in free-living non-human animals seemed daunting to anyone who cared about nonhumans. A pan-species welfare state was inconceivable. Cross-species fertility-regulation via immunocontraception, neurochipping, GPS-tracking and monitoring, and rudimentary healthcare services would be prohibitively costly even for large, long-lived vertebrates in human wildlife parks. How could we possibly hope to tackle suffering in marine ecosystems or the Amazon rainforest, short of invoking molecular nanotechnology and what critics would call Drexlerian sci-fi? Actively helping free-living non-humans in an era when millions of people still lack adequate nutrition and healthcare, and when humans still systematically hurt, harm and kill billions of sentient beings in factory farms and slaughterhouses, has made such a project seem

sociologically fanciful as well. Evolution didn't design *Homo sapiens* to be impartially altruistic.

CRISPR-based "gene drives" are a game-changer. In principle, gene drives can be used - cheaply, rapidly and sustainably - to "fix" the typical level of suffering undergone by members of entire free-living and sexually reproducing species at minimal inconvenience to humans. If targeted wisely, gene drives could massively amplify the effects of even exceedingly weak and fitful human benevolence towards non-human animals. In principle, the level of suffering in any sexually reproducing species of organism could be significantly reduced via genetic tweaking for the cost of around $10,000 of per species or less at current prices. Back-of-an-envelope calculation suggests the financial cost of a happy non-human biosphere would currently be several hundred million dollars - plus annual maintenance costs of perhaps several million dollars per year.

Hyperbole? No...

ETHICAL GENE DRIVES IN ACTION?

SCN9A: a case study

Gene drive systems are "selfish" genetic elements that can rapidly spread in sexually reproducing species even if they reduce the fitness of individual organisms. The genomes of almost every sexually reproducing species show evidence of at least one "natural" active gene drive or its broken remnants. Synthetic gene drives can now be designed to "sculpt" evolution. Researchers can take a gene that has a fitness-cost for the individual, for example male sterility, and move ("drive") it through a population in defiance of the usual constraints of Mendelian inheritance. Gene drives achieve this seemingly impossible feat by ensuring that they will be inherited by effectively all - rather than half - of the organism's offspring. Sexually reproducing animals normally have two versions of each

gene located on two different chromosomes. Maternal and paternal chromosomes in such a homologous pair have the same genes at the same loci, but the genes typically have different variants. Normally, an organism's offspring inherit only one of each pair of chromosomes from each parent. Therefore each different allele is ordinarily passed on to only around half of the organism's offspring. The new CRISPR/Cas9 genome-editing tool allows this rule of Mendelian inheritance to be broken with powerful and precise gene-editing techniques. Specifically, endonuclease gene drives can cut the corresponding locus of the homologous chromosome that doesn't encode the drive, inducing the cell to repair the damage by copying the drive sequence onto the damaged chromosome. In consequence, the cell then has two copies of the drive sequence. If the modified cell is a germline cell, then the modification will be passed on to all the organism's offspring, regardless of which chromosome they inherit. The same process will then apply to their offspring, too, generation after generation. In effect, a cell's DNA repair-mechanisms can be "hijacked" to spread human-selected traits throughout an entire species. CRISPR/Cas9 genome-editing potentially allows biohackers, scientists, or tomorrow's wildlife park managers accurately to insert, replace, delete and regulate genes in all sexually reproducing species and then "drive" the desired alteration(s) across the entire population. Species that can reproduce both with and without sex, for example many plants, are more problematic; and gene drives can’t alter asexually reproducing populations such as bacteria. Yet the vast majority of sentient beings on Earth today belong to predominantly or exclusively sexually-reproducing species. In short, gene drives can potentially be designed to engineer the well-being of all sentience.

Consider a concrete example of how gene drives could be used to reduce suffering in Nature. The lives of countless sentient beings are blighted by physical pain. Multiple genes modulate an organism's pain-sensitivity. Here let's focus just on the sodium

channel, voltage-gated, type IX alpha subunit known as the SCN9A gene. The SCN9A gene belongs to an evolutionarily ancient family of genes that code the construction of sodium channels. Sodium channels transport positively charged sodium ions into nerve cells, allowing the generation and transmission of electrical signals. The SCN9A gene provides instructions for making the alpha subunit of the sodium channel NaV1.7 found in nociceptors that transmit pain signals. Dozens of different alleles of SCN9A have been deciphered. Rare, maladaptive nonsense mutations of the SCN9A gene abolish an organism's ability to feel pain altogether. Yet other SCN9A alleles confer unusually high or unusually low pain-sensitivity without compromising function to any marked degree. Recall how today a small minority of high-functioning people display an exceptionally high pain-tolerance. Such "abnormally" low pain-sensitivity isn't the same as a dangerous and potentially lethal congenital analgesia. For such lucky people, pain is little worse than a useful bodily signalling mechanism in situations where "normal" human and non-humans animals alike would be screaming in agony.

In principle, there's now nothing to stop intelligent moral agents "fixing" the [conditionally-activated level of] subjective physical distress undergone by members of entire free-living species by choosing and propagating benign alleles of SCN9A or its homologs via gene drives, i.e. engineering via CRISPR-mediated gene-editing - not a currently utopian "no pain" biosphere (*cf*. The Abolitionist Project), but a "low pain" biosphere.

To be sure, risks abound; but no one is proposing compassionate stewardship of ecosystems by philosophers. Humans are capable of choosing our own future pain-sensitivity too; but any species-wide genomic shift in human pain tolerance will depend on the willingness of prospective parents to use preimplantation genetic screening.

Even in an age of CRISPR, customised gene drives and exponentially increasing computer power, the cost of compassionate stewardship of the biosphere won't be financially negligible. Yet perhaps compare the $100,000 today spent salvaging a single 23-week-old human micro-preemie with the price of "fixing" the default well-being an entire species of free-living vertebrate - indefinitely. Millions of non-human animals are as sentient - and demonstrably as sapient - as human prelinguistic toddlers. Many billions of non-humans are as sentient and demonstrably as sapient as human infants. Effective altruism dictates shedding anthropocentric bias and helping our fellow creatures accordingly.

Until the CRISPR genome-editing revolution, helping any free-living non-humans beyond a few large, long-lived vertebrates such as elephants (*cf*. "A Welfare State for Elephants") was implausible in our lifetime. Aiding small rodents, marine invertebrates or insects (*cf*. "The Importance of Insect Suffering") could at best be a task for our grandchildren and mature nanotechnology - or more credibly, for posthuman superintelligence. “Gene drives” turn this intuitive chronology on its head – in theory at any rate. For it's actually easier, cheaper and quicker to help fast-reproducing *r*-selected rather than *K*-selected species. Even the most cognitively humble life-forms can benefit from a bare minimum of human benevolence towards other sentient beings.

Which subjectively unpleasant traits are most morally urgent to modify? The control of raw pain is clearly vital to quality of life. However, other parameters, most notably the core emotions, can be genetically adjusted to shape default well-being too.

For example,

- COMT ("The catechol-O-methyl transferase Val158Met polymorphism and experience of reward in the flow of daily life")

- Serotonin transporter gene ("National Happiness and Genetic Distance: A Cautious Exploration")
- ADA2b deletion variant ("Is Pessimism Genetic? Research Shows Your Outlook Might Be Cloudy By Genetic Design")
- FAAH gene variant rs324420 ("Genes may contribute to making some nations happier than others")

And so forth. "Fixing" pain-sensitivity, depression-resistance, and default hedonic tone via gene drives will prevent immense suffering throughout the living world. The Cambrian Explosion was an explosion in suffering too; and only now are intelligent moral agents in a position to bring it under control.

Naturally, pitfalls lie ahead. Neither action nor inaction are ethically risk-free. A prudent if informal rule of thumb for policy-makers might be that anything that conceivably can go wrong with germline interventions will go wrong - and more besides. Mankind's dark historical track-record suggests that gene drives are more likely to be used for genetic terrorism, ethnic bioweapons and entomological warfare than harnessed to promote the welfare of other sentient beings. Ideally, artificial gene drives will be used to end the scourge of mosquito-borne diseases. Insect-borne pathogens sicken and kill millions of human and non-human animals each year. Malaria-proof *Anopheles* mosquitoes already exist in the laboratory. If released into the wild, such disease-resistant transgenic mosquitoes would rapidly spread and soon supersede their malarial cousins, thereby protecting numerous species of birds, reptiles and mammals, including humans. On the other hand, a single bioterrorist could design a small number of mosquitoes powered with a gene drive equipped with a gene for making a deadly toxin. Mosquitoes reproduce rapidly. Soon all the world’s mosquitoes of the modified species would make the toxin.

Every mosquito bite would be lethal (*cf*. "This could be the next weapon of mass destruction"). Idealism may be as hazardous as misanthropy. Perhaps some youthful biohacker will decide genetically to tweak the Texas Lone Star Tick *Amblyomma americanum* (*cf.*"This bug's bite could turn you vegetarian") - not the best way to win the global battle for hearts and minds. Such scenarios could be multiplied. Hence the need for multiple safeguards, well-drafted regulations and effective enforcement mechanisms before an engineered gene drive is unleashed in the wild. In the post-CRISPR era, all that intelligent moral agents can responsibly do is weigh risk-reward ratios and then act accordingly.

Consider pain-tolerance again. Unlike rare individuals born with congenital analgesia or victims of severe and debilitating chronic pain syndromes, organisms born with exceptionally high and exceptionally low pain-sensitivity alike can be high functioning. Nevertheless, both "low pain" human and non-human animals *do* behave differently from neurotypicals, although well-controlled cross-species studies are lacking. Responsibly using synthetic gene drives to shift the typical behavioural phenotype of an entire species towards the spectrum of behaviour today characteristic of its nociceptive outliers first calls for pilot studies, multiple safeguards and intelligent computational modelling - and regulation. Right now, using molecular tools available on eBay, a single biohacker could construct a gene drive to benefit - or harm - an entire free-living population world-wide. In principle, a modestly talented ethical biohacker using molecular tools readily available for under $10,000 could "fix" the default level of suffering for an entire species of small fast-reproducing vertebrate within the time-frame of two or three decades - and the default level of suffering of a sexually fast-reproducing species of insect or marine invertebrate within two or three years. Helping an entire species of slow-reproducing

elephants exclusively in the same way, i.e. by using gene drives and no other intervention, would take two or three centuries.

The scope for unanticipated side-effects from well-meaning but ill-judged interventions is huge. For example, the unusually high and unusually low pain-sensitivity promoted by different alleles of SCN9A is associated with unusually high and unusually low olfactory acuity respectively: NaV1.7 sodium channels are found in olfactory sensory neurons of the nasal cavity that transmit smell-related signals to the brain. Modelling the cross-species ramifications of altered smell-perception conjoined with reduced pain-sensitivity will be computationally challenging - which is not to say that we'll physically run out of computational resources. One possible solution involves contained field-trials using "low pain" organisms engineered with the benign high pain tolerance but lacking the functional drive to spread it. Or to raise another thorny issue, what will minimised pain-sensitivity do to empathy towards conspecifics in species with at least a rudimentary theory of mind? (*cf*. "Rats forsake chocolate to save a drowning companion") The existence of short-acting empathetic euphoriants such as MDMA ("Ecstasy") illustrates that heightened empathy and profound subjective well-being aren't mutually inconsistent traits; but this happy congruence can't simply be assumed and extrapolated. Long-term population monitoring will be ecologically prudent even after benign alleles have been "fixed" in a species via gene drives or any other species-wide intervention.

Is compassionate stewardship of the biosphere best conducted via private initiative? Or under the auspices of the United Nations, with at least some form of democratic accountability and international regulatory oversight? Immense diplomatic challenges lie ahead before humanity collectively agrees on the basic principles of ethical ecosystem management. Ecosystems don't respect nation-state boundaries; and neither do gene drives. Cheaply and efficiently minimising pointless suffering in Nature deserves to be

uncontroversial even among the morally apathetic; humans tend to be callous rather than malevolent. "May all that hath life be delivered from suffering" is a widely esteemed quote from Gautama Buddha, not some madcap transhumanist. Yet many secular and religious organisations and state actors have values and priorities beyond minimising needless misery. For instance, intelligence-amplification of entire species of free-living non-humans is imminently feasible. Laboratory mice engineered with the human variant of the FOXP2 "language gene" are demonstrably more intelligent than their primitive conspecifics. Gene drives could ("Human 'language gene' makes mice smarter") amplify intelligence and dramatically postpone senescence in entire populations (*cf*. "Longevity: Extending the lifespan of long-lived mice"). The uplift universe of science fiction writer David Brin probably strikes most people as whimsical fantasy; but free-living Neo-Chimpanzees, Neo-Dolphins, Neo-Gorillas and Neo-Dogs will shortly be policy options.

Even the most enlightened and comprehensive regime of gene drives won't abolish traditional natural selection altogether. If a genetic alteration is slightly harmful to an organism, perhaps like exceedingly high pain-tolerance, then the engineered gene drive would *eventually* break. More severe harms would break the drive over shorter evolutionary time-scales. Under a regime of compassionate stewardship, broken versions of a gene drive would eventually need overwriting with new and more robust functional copies. Also, biohackers - or state actors - with different ethical priorities may unleash competing gene drives. So-called "immunising drives" block another gene drive from spreading by pre-emptively altering the sequences that another drive targets, thereby preventing it from initiating copying. Pranksters, mischief-makers, genetic open-source enthusiasts and "script kiddies" could all potentially wreak ecological havoc with "rogue" drives, not just incompetent idealists. Gene drives are a rapidly emerging technology. Human use of CRISPR/Cas9 genome-editing is only a few years old. We may anticipate

the development of user-friendly software tools that lower the threshold of technical competence for engineering gene drives from talented biohackers as now to tomorrow's high-school students. Perhaps regulatory authorities will license genomic alterations via gene drives to a species in the wild only in conjunction with development of another "reversal" gene drive held in reserve. Such a reversal gene drive could be launched to undo the effects of the original gene drive in case of unanticipated adverse side-effects. Biohackers, let alone "black-hat" biocrackers with purposes of their own, may not heed such safeguards, or may actively subvert them.

Intuitively we might imagine that *most* interventions would eventually prove maladaptive to engineered organisms, causing the gene drive ultimately to break. This needn't be the case. All sorts of traits are potentially fitness-enhancing to an organism but haven’t evolved under a regime of natural selection because their evolution would have involved crossing "fitness gaps". Nature has no foresight; no non-human animals forage using wheels. CRISPR-mediated genomic-editing followed by premeditated use of synthetic gene drives allows crossing gaps in the fitness landscape prohibited by natural selection. Intelligent moral agency can "leap across" fitness gaps. Moreover, the emergence of drive-resistant alleles can be delayed or prevented altogether by targeting highly conserved sites in the genome at which resistance is anticipated to have a severe fitness-cost to the organism. Natural selection can thereby be circumvented. Intelligent agency is poised to seize control of evolution as the post-Darwinian transition accelerates.

With adult humans, bioethicists face the thorny issue of consent. By contrast, it’s hard to talk of the “right” of a mouse to suffer involuntarily. Even if all prospective human parents were routinely offered preimplantation genetic screening so they could choose, e.g. the pain-sensitivity, depression-resistance, and default hedonic set-points (etc) of their offspring, millions of traditionally-minded parents-to-be would presumably still play

genetic roulette and opt instead to have kids "naturally”. All sexually reproduced organisms are currently unique and untested genetic experiments. Barring a sea-change in public opinion world-wide, hundreds of years of avoidable human suffering consequently still lie ahead via the crapshoot of traditional sexual reproduction. Yet unless we subscribe to the mythical Wisdom of Nature, the choice of a "low-pain" living world in the vertebrate lineage and beyond will shortly be a technically feasible and financially affordable policy option – perhaps not yet a full-blown pan-species welfare-state, let alone a perfect world, but at least compassionate conservatism.

Talk of "conservatism" or "conservation" for a technology as revolutionary as synthetic gene drives sounds paradoxical. Yet the *potentially* species-conservative role of gene drives offers a rhetorically attractive compromise between ethicists who advocate the dramatic alteration or outright abolition of archaic Darwinian life and traditionalists who favour the pain-ridden status quo. For the greatest long-term obstacles to reducing and ultimately abolishing suffering in the living world aren’t technical but ethical-ideological - and above all, status quo bias. Radical bioethicists believe that a compelling moral case can be made for non-violently phasing out the cruelties of traditional Darwinian life. However, even the prospect of civilising Darwinian life by "policing" Nature raises the hackles of species essentialists. Thus the species essentialist claims that obligate carnivores who eat *in vitro* meat, or reprogrammed predators who no longer asphyxiate, disembowel or eat their victims alive, will have lost some vital part of their species essence, a fate assumed to be inherently ethically objectionable. This objection can be defanged by highlighting "bioconservative" uses of gene drives that simply fix "natural" benign alleles and allelic combinations in free-living populations rather than designing and propagating true genetic novelties. Even such timid bioconservatism is sure to upset extreme traditionalists; but the claim that a temperamentally happy lion or a mouse isn’t

"truly" a lion or a mouse compared to his or her misery-ridden conspecifics borders on the ridiculous. Are exceptionally happy or abnormally pain-tolerant humans today not "truly" human? Are Africans who lack the 1%-3% Neanderthal gene admixture of non-Africans less authentically human than Europeans? Or vice versa?

And what should advocates of compassionate biology say to religious believers? The precise answer depends on our target audience. Yet if God had wanted His creatures to suffer, then presumably He wouldn't have given us CRISPR/Cas9. If the lion and the wolf are really to lie down with the lamb, as the Bible foretells, then each party will need some behavioural-genetic tweaking, unless we suppose the metabolic pathways of obligate carnivores can be modified by the Holy Ghost.

So what exactly are our ultimate ethical responsibilities to other sentient beings? With power comes a deepening complicity in their lives and fate, whether *Homo sapiens* likes it or not. By analogy, if one comes across a small child from a different ethnic group drowning in a shallow pond, then choosing to walk on by rather than inconveniently get one's clothes wet is almost as morally repugnant as if one had pushed the child into the water oneself. Walking on by if the drowning victim is of comparable sentience and sapience to a human toddler but belongs to a different species rather than to a different ethnic group is no less culpable. Humans have not (quite) yet reached this level of complicity in the fate of most free-living non-human animals. Yet the biotech and IT revolutions also amount to a revolution in human complicity in the persistence of suffering. Systematically helping free-living non-humans via ecological engineering will shortly pass from the technically impossible to difficult to easy to trivial.

Inevitably, critics of compassionate intervention will talk of human "hubris". Yet is it more humble or hubristic not to rescue a drowning toddler from another ethnic group?

Why invert our response with beings of comparable sentience and sapience to human toddlers simply on the grounds they belong to a different species?

"Re-wilding" advocates claim that the prospect of compassionate stewardship of Nature threatens to turn the rest of the living world into a "zoo". Yet human and non-human animals typically flourish best when neither "wild" nor incarcerated but free-living. And at the risk of an *ad hominem* response, the bioconservative critic's professed respect for an ethos of "wild and free" rarely extends to going vegan and urging closure of factory farms and slaughterhouses.

Some commentators worry about a loss of genetic diversity. CRISPR-Cas9 genome editing can be used to increase or decrease genetic diversity for all sorts of traits. Not all genetic diversity is inherently valuable. For example, hundreds of different disease-causing alleles of the cystic fibrosis transmembrane conductance regulator (CFTR) gene have been discovered. Most ethicists agree that the optimum level of cystic fibrosis alleles to aim for in the human gene pool is zero. Phasing out alleles and allelic combinations implicated in suffering and malaise is more controversial. Yet depression and chronic pain syndromes can be at least as devastating to quality of life as cystic fibrosis.

Other critics take issue with anthropomorphism. "Projecting" human emotions and feelings onto non-humans is allegedly sentimental and unscientific. Who are humans to arrogantly impose our values on members of other species? Yet complications aside, no sentient being wants to be harmed. The pleasure-pain axis extends across all animal phyla. Whether or not other sentient beings desire to starve or be asphyxiated, disembowelled or eaten alive isn't an unfathomable metaphysical mystery beyond human comprehension. To be sure, there *are* aspects of non-human animal experience that are

alien to humans, for example what's it like to echolocate like a bat, or sexually to fancy a female warthog (etc). These alien state-spaces of experience don't extend to feelings of pain, hunger, fear or despair - or happiness. Human toddlers and non-human animals alike display a clearly expressed wish not to be physically molested or to undergo suffering and malaise. Ethically speaking, it's up to responsible caregivers to safeguard their interests. Of course, for evolutionary reasons some humans and some non-humans wish to harm others; but with humans, at least, we normally recognise that the interests of the victim take precedence. The right not to be harmed differs from a notional "right to harm". Either way, compassionate stewardship of the living world can potentially benefit (ex-)predators and their former victims alike.

THE FUTURE OF SENTIENCE

High-Tech Jainism?

Looking further ahead, humans or our descendants/successors are likely to practise terraforming other planets and moons, and perhaps eventually other solar systems. Cynics may echo C.S. Lewis, "Let's pray that the human race never escapes Earth to spread its iniquity elsewhere." Yet evolutionary niches tend to get filled. If intelligent agents *do* propagate beyond Earth, then ethically the least that intelligent moral agents can do is assume responsibility for compassionate stewardship of the sentience in any ecosystems we create. Deliberately modifying the atmosphere, temperature or surface topography of a sterile planet presumably poses few ethical problems. By contrast, deliberately creating a Darwinian ecosystem with its concomitant misery and malaise is an ethically momentous step. One needn't be a Buddhist or a utilitarian to believe that the deliberate creation of such mass-produced suffering is ethically indefensible. Such a response doesn't rule out enlightened terraforming based on the principles of compassionate biology. For without the molecular signature of experience below "hedonic

zero", suffering of any kind is physically impossible. Mature gene drive technologies can potentially phase out the biology of suffering; and maybe even "lock in" a biology of information-sensitive gradients of intelligent bliss. Before colonising other planets, let alone radiating across the Galaxy, ethical prudence suggests fine-tuning the management of pain-free ecosystems here on Earth.

Sociologically realistic time-frames for compassionate ecosystem design can only be speculative. Yet every cubic metre of the planet will shortly be computationally accessible to surveillance and micro-management. "Not a single sparrow can fall to the ground without your Father knowing it" (*Matthew* 10:29), says the Bible. Two thousand years later, secular humanity must decide whether to use our impending God-like omniscience for Orwellian or benevolent purposes. CRISPR-Cas9 genome-editing and gene drives offer a powerful tool for compassionate stewardship of Nature at a politically realistic price. English-born American theoretical physicist Freeman Dyson, writing the *New York Review of Books*, remarks (*cf*. Our Biotech Future) "In the future, a new generation of artists will be writing genomes the way that Blake and Byron wrote verses." Biotechnology can be used for purposes more morally urgent than artistic self-expression.

The other mainstay of responsible stewardship of the living world will be cross-species immunocontraception. Most critics of compassionate biology assume that any proposal to help free-living non-human animals is ecologically illiterate. Ivory-tower philosophers don't understand the thermodynamics of a food chain. Feed a herd of starving herbivores in winter, for example, and the outcome will be a population explosion next spring followed by ecological collapse. The upshot? More misery. However, recall that exactly the same predictions of immiseration and "inevitable" Malthusian catastrophe were made last century to argue against helping famine-stricken members of other ethnic groups in

sub-Saharan Africa. The solution is combining emergency famine-relief with help with long-term family planning. Non-human animals can't use contraception on their own initiative. But intelligent human-directed use of gene drives, cross-species immunocontraception, and other tools of fertility-regulation can manage ecologically sustainable population sizes as a compassionate alternative to population-control via famine, disease, parasitism and predation. Exponential growth of computational resources harnessed to mastery of our genetic source-code promises a world where all sentient beings can flourish. The World Health Organization definition of health is admirably bold - "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." Arbitrarily restricting the promotion of such good health to members of a single species is as unwarranted as its restriction to a single ethnic group.

Despite the glorious long-term prospects for sentience, most animal advocates would judge any exploration of compassionate stewardship of the living world to be premature. Before systematically helping other sentient beings, mankind's first obligation is surely to stop systematically harming them. Early in the twenty-first century, global veganism strikes many consumers as utopian dreaming. A post-animal bioeconomy of *in vitro* meat products is indeed still decades away. As long as the animal holocaust continues, a debate on wild animal suffering risks seeming surreal. In the words of Israeli historian Yuval Noah Harari in *Sapiens* (2011):

"Tens of billions of them [non-human animals] *have been subjected over the last two centuries to a regime of industrial exploitation, whose cruelty has no precedent in the annals of planet Earth. If we accept a mere tenth of what animal-rights activists are claiming, then modern industrial agriculture might well be the greatest crime in history."*

So can humans redeem ourselves by genetically engineering a happy biosphere? Or will suffering endure as long as life itself? Biologist Edward Wilson outlined the challenge in *Consilience* back in 1998:

"Homo sapiens, the first truly free species, is about to decommission natural selection, the force that made us.... Soon we must look deep within ourselves and decide what we wish to become."

Part IV: Consciousness

NON-MATERIALIST PHYSICALISM

An experimentally testable conjecture

"You're nothing but a pack of neurons."

(Francis Crick)

ABSTRACT

Mankind's most successful story of the world, natural science, leaves the existence of consciousness wholly unexplained. The phenomenal binding problem deepens the mystery. Neither classical nor quantum physics seem to allow the binding of distributively processed neuronal micro-experiences into unitary experiential objects apprehended by a unitary phenomenal self. This paper argues that *if* physicalism and the ontological unity of science are to be saved, then we will need to revise our notions of both 1) the intrinsic nature of the physical *and* 2) the quasi-classicality of neurons. In conjunction, these two hypotheses yield a novel, bizarre but experimentally testable prediction of quantum superpositions ("Schrödinger's cat states") of neuronal feature-processors in the CNS at sub-femtosecond timescales. An experimental protocol using *in vitro* neuronal networks is described to confirm or empirically falsify this conjecture via molecular matter-wave interferometry.

1. Introduction

Natural science promises a complete story of the universe. No "element of reality"(1) should be missing from the mathematical formalism of physics, i.e. relativistic quantum field theory(2) or its more speculative extension, M-theory. On pain of magic, every gross property of the natural world must be theoretically reducible to fundamental physics. The Standard Model in physics is experimentally well tested. Within its conceptual framework, consciousness would seem not only causally impotent but physically impossible. Hence the "explanatory gap"(3) and the Hard Problem(4) of consciousness.

In recent years, a minority of researchers have proposed that the Hard Problem is an artifact of materialist metaphysics. *Contra* Kant(5), but following Schopenhauer(6), Bertrand Russell(7), Grover Maxwell, Michael Lockwood(8), Galen Strawson(9), *et al*., the new idealists conjecture that the phenomenology of one's mind reveals the intrinsic nature of the physical – the elusive "fire" in the equations about which physics is silent. Mathematical physics yields an exhaustive description of the relational-structural properties of the world. This description may ultimately be encoded by the universal wavefunction of post-Everett[a] quantum mechanics: our best mathematical description of reality. However, our presupposition that the *intrinsic* character of the physical lacks phenomenal properties is an additional metaphysical assumption. The assumption is hugely plausible, but it's not a scientific discovery. Perhaps most tellingly, the only part of the "fire" in the equations to which one ever enjoys direct access, i.e. one's own consciousness, discloses phenomenal properties that are inconsistent with a materialist ontology. For reasons unexplained, the natural world contains first-person facts. The world supports at least one non-zombie. And natural science gives no reason to believe that one is special.

Untestability cuts both ways. Any conjecture that superpositions of the world's fundamental quantum fields – and, presumably, fundamental macroscopic quantum phenomena such as superconductors or superfluid helium – are intrinsically experiential would seem unfalsifiable too: just speculative metaphysics.

Rather surprisingly, we shall see this isn't the case.

Preliminary Definitions

Both physics and philosophy are jargon-ridden. So let's first define some key concepts.

Both "**consciousness**" and "physical" are contested terms. Accurately if inelegantly, consciousness may be described following Nagel ("What is it like to be a bat?") as the subjective *what-it's-like-ness* of experience. Academic philosophers term such self-intimating "raw feels" "**qualia**" – whether macro-qualia or micro-qualia. The minimum unit of consciousness (or "psychon", so to speak) has been claimed variously to be the entire universe, a person, a sub-personal neural network, an individual neuron, or the most basic entities recognised by quantum physics. In *The Principles of Psychology* (1890), American philosopher and psychologist William James christened these phenomenal simples "primordial **mind-dust**". This paper conjectures that (1) our minds consist of ultra-rapidly decohering neuronal superpositions in strict accordance with unmodified quantum physics without the mythical "collapse of the wavefunction"; (2) natural selection has harnessed the properties of these neuronal superpositions so our minds run phenomenally bound world-simulations; and (3) predicts that with enough ingenuity the non-classical interference signature of these conscious neuronal superpositions will be independently experimentally detectable to the satisfaction of the most incredulous critic.

The "**physical**" may be contrasted with the supernatural or the abstract. Dualists and epiphenomenalists contrast the physical with the mental. The current absence of any satisfactory "positive" definition of the physical leads many philosophers of science to adopt instead the "*via negativa*". Thus some materialists have sought stipulatively to *define* the physical in terms of an absence of phenomenal experience. Such *a priori* definitions of the nature of the physical are question-begging.

"**Physicalism**" is sometimes treated as the formalistic claim that the natural world is exhaustively described by the equations of physics and their solutions. Beyond these structural-relational properties of matter and energy, the term "physicalism" is *also* often used to make an ontological claim about the *intrinsic* character of whatever the equations describe. This intrinsic character, or metaphysical essence, is typically assumed to be non-phenomenal. "**Strawsonian physicalists**" (*cf*. "Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?") and other non-materialist physicalists dispute any such assumption. Traditional **reductive physicalism** proposes that the properties of larger entities are determined by properties of their physical parts. If the wavefunction monism of post-Everett quantum mechanics assumed here is true, then the world does not contain discrete physical parts as understood by classical physics. If contemporary physicalism is true, reductionism is false.

"**Materialism**" is the metaphysical doctrine that the world is made of intrinsically non-phenomenal "stuff". Materialism and physicalism are often treated as cousins and sometimes as mere stylistic variants – with "physicalism" used as a nod to how bosonic fields, for example, are not matter. "**Physicalistic materialism**" is the claim that physical reality is fundamentally *non*-experiential *and* that the natural world is exhaustively described by the equations of physics and their solutions.

"**Panpsychism**" is the doctrine that the world's fundamental physical stuff also has primitive experiential properties. Unlike the physicalistic idealism explored here, panpsychism doesn't claim that the world's fundamental physical stuff *is* experiential. Panpsychism is best treated as a form of property-dualism.

"**Epiphenomenalism**" in philosophy of mind is the view that experience is caused by material states or events in the brain but does not itself cause anything; the causal efficacy of mental agency is an illusion.

For our purposes, "**idealism**" is the ontological claim that reality is fundamentally experiential. This use of the term should be distinguished from Berkeleyan idealism, and more generally, from subjective idealism, i.e. the doctrine that only mental contents exist: reality is mind-dependent. One potential source of confusion of contemporary scientific idealism with traditional philosophical idealism is the use by inferential realists in the theory of perception of the term "world-simulation". The mind-dependence of one's phenomenal world-simulation, i.e. the quasi-classical world of one's everyday experience, does not entail the idealist claim that the mind-independent physical world is intrinsically experiential in nature – a far bolder conjecture that we nonetheless tentatively defend here.

"**Physicalistic idealism**" is the non-materialist physicalist claim that reality is fundamentally experiential *and* that the natural world is exhaustively described by the equations of physics and their solutions: more specifically, by the continuous, linear, unitary evolution of the universal wavefunction of post-Everett quantum mechanics. The **"decoherence program"** in contemporary theoretical physics aims to show in a rigorously quantitative manner how quasi-classicality emerges from the unitary Schrödinger dynamics.

"**Monism**" is the conjecture that reality consists of a single kind of "stuff" – be it material, experiential, spiritual, or whatever. *Wavefunction monism* is the view that the universal wavefunction mathematically represents, exhaustively, all there is in the world. Strictly speaking, wavefunction monism shouldn't be construed as the claim that reality literally consists of a certain function, i.e. a mapping from some mind-wrenchingly immense configuration space to the complex numbers, but rather as the claim that every mathematical property of the wavefunction, except the overall phase, corresponds to some property of physical world. **"Dualism"**, the conjecture that reality consists of two kinds of "stuff", comes in many flavours: naturalistic and theological; interactionist and non-interactionist; property and ontological. In the modern era, most scientifically literate monists have been materialists. But to describe oneself as both a physicalist *and* a monistic idealist is not the schizophrenic word-salad it sounds at first blush.

"**Functionalism**" in philosophy of mind is the theory that mental states are constituted solely by their functional role, i.e. by their causal relations to other mental states, perceptual inputs, and behavioural outputs. Functionalism is often associated with the idea of "substrate-neutrality", sometimes misnamed "substrate-independence", i.e. minds can be realised in multiple substrates and at multiple levels of abstraction. However, *micro*-functionalists may dispute substrate-neutrality on the grounds that one or more properties of mind, for example phenomenal binding, functionally implicate the world's quantum-mechanical bedrock from which the quasi-classical worlds of Everett's multiverse emerge. Thus this paper will argue that only successive quantum-coherent neuronal superpositions at naively preposterously short time-scales can explain phenomenal binding. Without phenomenal binding, no functionally adaptive classical world-simulations could exist in the first instance.

The "**binding problem**"[10], also called the "combination problem", refers to the mystery of how the micro-experiences mediated by supposedly discrete and distributed neuronal edge-detectors, motion-detectors, shape-detectors, colour-detectors, etc, can be "bound" into unitary experiential objects ("**local**" binding) apprehended by a unitary experiential self ("**global**" binding). Neuroelectrode studies using awake, verbally competent human subjects confirm that neuronal micro-experiences exist. Classical neuroscience cannot explain how they could ever be phenomenally bound. As normally posed, the binding problem assumes rather than derives the emergence of classicality.

"**Mereology**" is the theory of the relations between part to whole and the relations between part to part within a whole. Scientifically literate humans find it's natural and convenient to think of particles, macromolecules or neurons as having their own individual wavefunctions by which they can be formally represented. However, the manifest *non*-classicality of phenomenal binding means that in some contexts we must consider describing the entire mind-brain via a single wavefunction. Organic minds are not simply the "mereological sum" of discrete, decohered classical parts. Sentient organic brains are not simply the "mereological sum" of discrete, decohered classical neurons.

"**Quantum field theory**" (QFT) is the formal, mathematico-physical description of the natural world. The world is made up of the states of interacting quantum fields, conventionally non-experiential in character, that take on discrete values. Physicists use mathematical entities known as "**wavefunctions**" to represent quantum states. Wavefunctions may be conceived as representing all the possible configurations of a superposed quantum system. Wavefunction(al)s are complex-valued functionals on the space of field configurations. Wavefunctions in quantum mechanics are sinusoidal functions with an amplitude (a "measure") and also a phase. The Schrödinger equation describes the time-evolution of a wavefunction. "**Coherence**" means that the *phases* of

the wavefunction are kept constant between the coherent particles, macromolecules or (hypothetically) neurons, while **"decoherence"** is the effective loss of ordering of the phase angles between the components of a system in a quantum superposition due to interactions with the environment. Such thermally-induced "dephasing" rapidly leads to the emergence – on a perceptual naive realist story – of classical, i.e. probabilistically *additive*, behaviour in the central nervous system (**"CNS"**), and also the illusory appearance of separate, non-interfering organic macromolecules. Hence the discrete, decohered classical neurons of laboratory microscopy and biology textbooks. Unlike classical physics, quantum mechanics deals with superpositions of probability *amplitudes* rather than of probabilities; hence the *interference*-terms in the probability distribution. Decoherence should be distinguished from *dissipation*, i.e. the loss of energy from a system – a much slower, classical effect. Phase coherence is a quantum phenomenon with no classical analogue. If quantum theory is universally true, then any physical system such as a molecule, neuron, neuronal network or an entire mind-brain exists partly in all its theoretically allowed states, or configuration of its physical properties, *simultaneously* in a "**quantum superposition**"; informally, a "Schrödinger's cat state", a weighted combination of all possible measurement outcomes. Each state is formally represented by a complex vector [technically a *ray*, or one-dimensional subspace] in Hilbert space. Hilbert space is the generalisation of Euclidean space containing the wavefunctions standing for the possible states of any physical system. Whatever overall state the nervous system is in can be represented as being a superposition of varying amounts of these particular states ("eigenstates") where the amount that each eigenstate contributes to the overall sum is termed a *component*. The "**Schrödinger equation**" is a partial differential equation that describes how the state of a physical system changes with time. The Schrödinger equation acts on the entire probability

amplitude, not merely its absolute value. The absolute value of the probability amplitude encodes information about probability densities, so to speak, whereas its phase encodes information about the interference between quantum states. On measurement by an experimenter, the value of the physical quantity in a quantum superposition will naively seem to "collapse" in an irreducibly stochastic manner, with a probability equal to the square of the coefficient of the superposition in the linear combination. If the superposition principle really breaks down in the mind-brain, as traditional Copenhagen positivists still believe, then the central conjecture of this paper is false.

"**Mereological nihilism**", also known as "compositional nihilism", is the philosophical position that objects with proper parts do not exist, whether extended in space or in time. Only basic building blocks (particles, fields, superstrings, branes, information, micro-experiences, quantum superpositions, entangled states, or whatever) without parts exist. Such ontological reductionism is untenable if the mind-brain supports macroscopic quantum coherence in the guise of bound phenomenal states because coherent neuronal superpositions describe *individual* physical states. Coherent superpositions of neuronal feature-detectors cannot be interpreted as classical ensembles of states. Radical ontological reductionism is even more problematic if post-Everett(11) quantum mechanics is correct: reality is exhaustively described by the time-evolution of one gigantic universal wavefunction. If such "**wavefunction monism**" is true, then talk of how neuronal superpositions are rapidly "destroyed" is just a linguistic convenience because a looser, heavily-disguised coherence persists within a higher-level Schrödinger equation (or its relativistic generalisation) that subsumes the previously tighter entanglement within a hierarchy of wavefunctions, all ultimately subsumed within the universal wavefunction.

"**Direct realism**", also known as "naive realism", about perception is the pre-scientific view that the mind-brain is directly acquainted with the external world. In contrast, the **"world-simulation model"**(12) assumed here treats the mind-brain as running a data-driven *simulation* of gross fitness-relevant patterns in the mind-independent environment. As an inferential realist, the world-simulationist is not committed *per se* to any kind of idealist ontology, physicalistic or otherwise. However, s/he will understand phenomenal consciousness as broader in scope compared to the traditional perceptual direct realist. The world-simulationist will also be less confident than the direct realist that we have any kind of pre-theoretic conceptual handle on the nature of the "physical" beyond the formalism of theoretical physics – and our own phenomenally-bound physical consciousness.

"Classical worlds" are what perceptual direct realists call *the* world. Quantum theory suggests that the multiverse exists in an inconceivably vast cosmological superposition. Yet within our individual perceptual world-simulations, familiar macroscopic objects 1) occupy definite positions (the "preferred basis" problem); 2) don't readily display quantum interference effects; and 3) yield well-defined outcomes when experimentally probed. Cats are either dead or alive, not dead-and-alive. Or as one scientific populariser puts it, "Where Does All the Weirdness Go?" This paper argues that the answer lies under our virtual noses, so to speak – though independent physical proof to silence sceptics will depend on next-generation matter-wave interferometry. Phenomenally-bound classical world-simulations *are the mind-dependent signature of the quantum "weirdness"*. Without the superposition principle, no phenomenally-bound classical world-simulations could exist and nor could minds. In short, we shouldn't imagine superpositions of live-and-dead *cats*, but instead think of superpositions of colour-, shape-, edge- and motion-processing *neurons*. Thanks to natural selection, the *content* of our waking world-

simulations typically appears classical; but the *vehicle* of the simulation that our minds run is inescapably quantum. If the world were classical it wouldn't look like anything to anyone.

A "**zombie**", sometimes called a "philosophical zombie" or "p-zombie" to avoid confusion with its lumbering Hollywood cousins, is a hypothetical organism that is materially and behaviourally identical to humans and other organic sentients but which isn't conscious. Philosophers explore the epistemological question of how each of us can know that s/he isn't surrounded by p-zombies. Yet we face a mystery deeper than the ancient sceptical Problem of Other Minds. *If* our ordinary understanding of the fundamental nature of matter and energy as described by physics is correct, and *if* our neurons are effectively decohered classical objects as suggested by standard neuroscience, then we all ought to be zombies. Following David Chalmers, this is called the **Hard Problem** of consciousness.

Why aren't we P-Zombies? Why aren't we Micro-Experiential Zombies?

A scientifically adequate theory of conscious mind must explain:

1) Why consciousness exists at all.

2) How consciousness has the causal power to allow intelligent agents to investigate its own nature.

3) How consciousness can be phenomenally "bound" in seemingly classically forbidden ways into unitary dynamic objects. In other words, which of the world's information-

processing systems are unitary subjects of experience, and which are mere aggregates or "zombies"?

4) Why and how consciousness manifests its diverse textures – ranging from phenomenal colours, sounds, tastes and smells, pains and pleasures, the experience of introspecting a thought-episode, feeling pangs of jealousy, hearing an orchestra play, admiring a sunset, to finding a joke amusing. In our mathematico-physical Theory of Everything (TOE), where is the *information* that yields the disparate values of experience?

Finally, any satisfactory scientific theory of consciousness should also offer predictions that are both *novel* and experimentally *falsifiable*.

2. Challenges to Non-Materialist Physicalism.

David Chalmers(13) identifies two challenges faced by any claim that consciousness discloses the intrinsic nature of the physical:

a) *the argument from microphysical simplicity*.

b) *the argument from structural mismatch.*

Let us look at these two challenges in turn.

a) The argument that if physicalistic idealism is true, then "we can expect only a handful of microqualities, corresponding to the handful of fundamental microphysical properties" is intuitively appealing. After all, runs this line of argument, every electron in the world is type-identical to every other electron. Electrons are exceedingly simple. After we've specified the mass, charge and spin of an electron, what else is there to say? An electron "has no hair". Or more technically, after we have given the four quantum numbers that

completely describe the electron, namely its principal quantum number(n), azimuthal quantum number(l), magnetic quantum number(m), and spin quantum number(s), what else is there left to add?

However, in quantum field theory rather than basic quantum mechanics, there are no particles, only fields and field quanta. What we call "particles" by cosy analogy with classical physics are emergent entities supervenient on the underlying quantum fields. So if instead of a particle-based ontology, the monistic idealist assumes a quantum field-theoretic ontology, then the diverse values of the world's fundamental fields yield the diverse subjective textures of micro-qualia – a vast palette of different qualia-field values. All physical systems, including macroscopic neural networks, are quantum fields. To be sure, in our present ignorance we don't know how to "read off" the diverse values of micro-qualia from superpositions of the diverse values of the different fundamental fields. We lack any kind of cosmic Rosetta stone. But on this physicalistic idealist conjecture, there is no "element of reality" lacking in the quantum field-theoretic formalism that encodes the world's fundamental micro-experiences. Algorithmically compressed into mathematical equations, the information encoding the exact textures of qualia-field values just awaits extraction. For in contemporary physics, fields (or indeed superstrings or branes(14)) are defined purely mathematically, even though their experimentally manipulable effects show that the fields are physically real. These fields take a vast range of values ("numbers in space") – with a (conventionally) infinite number of degrees of freedom. And crudely, on this account "more is different" – micro-experientially different. Critically, these fields aren't classical. Overcoming Chalmers' second challenge to physicalistic idealism (b), i.e. the argument from structural mismatch, turns on recognising that fields in quantum field theory exist in quantum

superpositions of states. These quantum superpositions may be microscopic, mesoscopic or macroscopic: all are subject to the laws of quantum physics.

b) The argument that the "macrophenomenal structure of my visual field is *prima facie* very different from the macrophysical structure of my brain" seems intuitively obvious too. Yet this intuitive appeal may simply reflect the coarse-grained temporal resolution of our tools for investigating awake/dreaming mind-brains: a resolution of milliseconds - not picoseconds, femtoseconds and attoseconds. Appearances of a structural mismatch between neuroscience and phenomenology may be deceptive. There is no experimental evidence for a breakdown of the superposition principle in the mind-brain. What the textbooks call *synchronous* firings of classical neuronal feature-detectors may turn out to be successive quantum-coherent *superpositions* of the relevant neuronal feature-detectors. We won't know whether superposition is masquerading to experimental neuroscience as synchrony until advanced interferometry experimentally settles the issue independently. If empirically confirmed, the detection of such sub-femtosecond neuronal superpositions would render a stunningly beautiful result; the experimental confirmation of what sounds naively like unbridled metaphysical speculation.

Let's use a nonbiological analogy. If physicalistic idealism is true, then the macro-phenomenal structure of superfluid helium presumably consists of a simple, unvarying, long-lived, irreducible macro-experience: a perfect structural match between the formal and subjective properties of the world. Of course, humans will never know what – if anything – it's like to instantiate the wavefunction that describes superfluid helium. But when our experimental apparatus allows probing the CNS at the sub-femtosecond timescales below which e.g. Max Tegmark ("Why the brain is probably not a quantum computer") posits effectively irreversible thermally-induced decoherence, then our

classical intuitions may be confounded. On this conjecture, we will find, not random quantum "noise", but instead the structural quantum-coherent physical shadows of the bound macroscopic phenomenal objects of everyday experience - all computationally optimised by hundreds of millions years of evolution to track fitness-relevant patterns in the mind-independent world. That is, a perfect structural match, not a mismatch, between the phenomenology of consciousness and our canonical representations of the physical. According to the conjecture here explored, training up our neural networks ensures that some neuronal states of the CNS are less prone to thermally-induced decoherence than others. It's these *comparatively* robust experiential-physical states, most notably the perceptual objects of everyday experience, that experimentalists will detect in the CNS when molecular matter-wave interferometry catches up with theory. So when you report "I can see a chair", and (on the conventional classical story) synchronous activation of your relevant neuronal feature-detectors occurs, the conjecture will be *falsified* if the subtle non-classical neuronal interference effects typically detected are irrelevant "noise", say a sub-attosecond superposition of the neurons synchronously activated when you see a hippopotamus (etc). In other words, the putative mismatch that Chalmers identifies between the phenomenology of our bound phenomenal minds and the architecture of the brain may turn out to be an artifact of the low temporal resolution of our clumsy tools of investigation.

This is, most certainly, an unintuitive hypothesis. Yet the neurological implausibility of such a fine-grained temporal match should be set against the physical incredibility of the alternatives. From the perspective of natural science, the discovery of a true structural mismatch between physics and phenomenology in the CNS would be *more* astonishing than the previously unsuspected isomorphism between the phenomenal and the physical canvassed here. Such a rupture in the fabric of reality would spell the end of physicalism

– an epistemic catastrophe for the unity of science. Unlike his critics, David Chalmers is right to recognise the magnitude of the structural mismatch problem for orthodox materialism and classical panpsychism alike. Chalmers just quits the game too soon. He embraces what must surely count as a counsel of despair: dualism. Monistic physicalism *can* still be saved. Physicalism would be unsalvageable only if the brain is no more than a networked community of discrete, effectively *classical* neurons – or their idealist counterpart, i.e. discrete, effectively classical neuronal "mind-dust" – rather than a succession of macroscopic neuronal superpositions that make up one's everyday phenomenal world. Monistic physicalism *isn't* falsified by a structural mismatch between the three-dimensional space of naive perceptual realism and conscious mind. Monistic physicalism would be falsified only by a structural mismatch between the bound phenomenology of our minds and the fundamental high-dimensional space required by the dynamics of the wavefunction. No such mismatch has ever been experimentally demonstrated to date.

3. Phenomenal Binding is the Hallmark of Mind.

"The only realities are the separate molecules, or at most cells. Their aggregation into a `brain' is a fiction of popular speech", said William James in *The Principles of Psychology* (1890). The existence of bound phenomenal minds rather than cellular mind-dust suggests that separate molecules and nerve cells are a fiction of classical neuromythology.

Perhaps the greatest cognitive achievement of post-Cambrian(15) central nervous systems has been to solve the binding problem. Without phenomenal binding, members of the animal kingdom wouldn't have minds at all – or classical-seeming world-simulations they could navigate. Over the past half-billion or more years, the mind-brains of

unprogrammed organic robots have evolved under the pressure of natural selection to run data-driven, cross-modally matched egocentric world-simulations of their local environment *in almost real time*. The extraordinary computational power of binding is most vividly illustrated in neurological syndromes where local or global phenomenal binding partially breaks down. Patients with simultanagnosia(16), for example, can see only one phenomenal object at once. Victims of cerebral akinetopsia(17) are unable to detect motion. People with florid schizophrenia suffer from the disintegration of a unitary self. Even partial loss of phenomenal binding may be intellectually debilitating and behaviourally catastrophic. Neurotypical minds carry off such computational feats with ease. Unfortunately, a neuroscientific explanation is elusive.

By way of context, the phenomenal binding problem is normally posed roughly as follows. How can what neuroscience suggests are distributively neurally-processed edges, colours, shapes, motions (etc) be "bound" into unitary experiential objects populating a unitary experiential field instantiated by a fleetingly unitary self in the neural networks of the CNS? Such phenomenal binding would seem impossible for discrete, membrane-bound, quasi-classical neurons – or quasi-classical "mind-dust" on a physicalistic idealist ontology – separated by *c*. 3.5 nanometre electrical gap-junctions and 20-40 nanometre chemical synapses. Mere synchronous activation of discrete, decohered classical systems cannot bind – any more than discrete skull-bound minds each undergoing a pinprick causes the emergence of a global mega-mind in agony, or a musical symphony emerges from discrete skull-bound minds each instantiating a musical note. Whether causally connected or otherwise, synchronously activated classical "pixels" of experience remain unglued. Phenomenal mind is not a classical phenomenon. Neither are the pseudo-classical world-simulations run by our waking minds. Chalmers is right on that score. Does quantum mind-binding fare any better?

On the face of it, no. Decoherence is among the fastest processes known to experimental physics. By contrast, we normally assume that states of consciousness somehow arise via neural transmission over a time-scale of milliseconds. Yet unless, implausibly, quantum theory breaks down in the mind-brain – as in so-called "dynamical collapse" or "hidden variables" theories of QM – macroscopic quantum-coherent states implicating such neurologically distributed cellular processes of feature-processing *must* exist in the CNS. What's in question is only their character: noise or signal? Classical neuroscience assumes that these neuronal superpositions are irrelevant to consciousness.

To make our point, let's pose a concrete question. When one apprehends a bound phenomenal object in one's world-simulation, what does it feel like to instantiate successive quantum-coherent macro-superpositions of colour-detector neurons, motion-detector neurons, edge-detector neurons, etc, with each macro-superposition in the sequence lasting what theory suggests must be a femtosecond or less? The obvious answer to the question of what-it-feels-like to instantiate such a sequence of neuronal superpositions is "nothing at all", or perhaps computationally incidental "psychotic noise", because environment-induced decoherence effectively destroys macroscopic neuronal superpositions in the CNS at sub-femtosecond timescales. Quantum coherence is, for all practical purposes, irreversibly delocalised into the larger CNS-environment combination though uncontrolled environmental entanglement. On the standard neuroscientific story, our conscious macro-experiences of bound phenomenal objects apprehended by a unitary phenomenal self somehow "arise" instead from patterns of classical, decohered neuronal action potentials synchronously firing over timescales of milliseconds. Yet an answer of "nothing at all" to what-it-feels-like to instantiate a sub-femtosecond neuronal superposition is *not* a possible response for the non-materialist physicalist. For if non-materialist physicalism is true, then phenomenal simples are the world's intrinsic physical

properties – the "fire" in the equations of quantum field theory. A fleeting macroscopic neuronal superposition is just such a phenomenal simple: it's not an aggregate or classical ensemble of anything more primitive. Classical glue cannot bind; quantum-coherent glue can't do anything else; thermally-induced decoherence in the CNS explains just how rapidly our fragile minds become unstuck. Thus decoherence can be viewed as a progressive phenomenal *un*binding – or in other words, effective dephasing is a solution to the phenomenal *un*binding problem. The universal wavefunction is not a mind.

The alternative to such a perfect structural match hypothesis is equally stark. Unless we abandon the conceptual framework of physicalism, then mere synchronous(18) neuronal firings *cannot* phenomenally bind purely classical neurons or neuronal "mind-dust" into cross-modally matched phenomenal objects, or a spatio-temporally unitary perceptual field, or a transiently unitary phenomenal self. Mere synchronous neuronal firings cannot bind any more than, say, synchronous activation and reciprocal electromagnetic communications could phenomenally bind a community of skull-bound American minds. It's not that we can disprove Eric Schwitzgebel's claim that "If Materialism Is True, the United States Is Probably Conscious"(19). Rather, the emergence of such a unitary pan-continental subject of experience would be unexplained and inexplicable – a miracle in all but name. Such spooky strong ontological emergence would violate physicalism in a most spectacular way. By the same token, *if* our 86 billion odd neurons always behaved as essentially classical systems, as they do in a dreamless sleep or coma, then the emergence of a unitary pan-cerebral subject of experience, a "person", would be unexplained and inexplicable as well. Such spooky strong emergence would violate physicalism too.

Of course, the difference between the USA and the mind-brain is that – unlike a hypothetical pan-continental subject of experience – it's hard to treat the existence of one's own conscious mind as simply a conjecture. Rather the existence of one's bound conscious mind is what needs to be explained – unless you happen to be a philosophical zombie.

That said, eliminative materialists like Daniel Dennett(20) are right to recognise that qualia (raw phenomenal experiences) are impossible within a materialist ontology. More particularly, eliminative materialists are right to recognise that the existence of qualia in the brain – as understood by classical materialistic neuroscience – is a physical impossibility, whether the existence of phenomenal symphonies, chairs, tables, mountains, or the whole panoply of lived experience. Yet this alleged impossibility derives from a combination of our classical *mis*representations of the mind-brain; our temporally coarse-grained observations of other central nervous systems; and our quasi-hardwired perceptual naive realism with the crude materialist ontology it spawns. We're not entitled to infer that humans must be insentient zombies on the grounds that our materialist ontology can find no naturalistic place for our sentience. The eliminative materialist who forgoes anaesthesia before surgery has yet to be born.

An obvious counterargument to such a (presently hypothetical) perfect fine-grained match between the phenomenology of our minds and the physical structure of the CNS is that we perceive our surroundings with a time-lag of scores of milliseconds. Such a time-lag is orders of magnitude too long for ultra-rapidly thermally "destroyed" (i.e. lost to the environment in a thermodynamically irreversible way) quantum-coherent neuronal superpositions to be computationally relevant to perception.

This objection presupposes an untenable perceptual naive realism in which we directly "see" the mind-independent world – the same misconceived perceptual naive realism according to which a neurosurgeon directly "sees" the cheesy wet nervous tissue constituting the mind-brain of an anaesthetised patient lying on his operating table prior to surgery.

Such perceptual naive realism may be compared with Bertrand Russell's apt reminder that one never "sees" anything but the inside of one's own head. Not even a neurosurgeon in the operating theatre. According to our contrasting world-simulation model, the role of the local mind-independent environment is essentially to *select* quantum-coherent superpositions of the awake mind/brain via optic and other nerve inputs with an evolutionarily minimised time-lag.

If Max Tegmark's calculations(21) – as distinct from his conclusions – are approximately correct, then our world-simulations must run at anything from around 10^{13} quantum-coherent neuronal "frames" per second to a frame-rate of up to 10^{20+} quantum-coherent neuronal "frames" per second. The fitness-relevant environmental patterns that they track in waking states lag behind their neural counterparts by a hundred or more milliseconds. In that sense, we always "live in the past"; but our waking world-simulations run in near enough to real time for organic robots to behave flexibly and adaptively in an inhospitable environment.

A suggestive analogy here might be the persistence of vision undergone by organic minds watching a movie run at 24 frames-per-second. Each composite frame of the movie can be rich, diverse and multifaceted, despite the lack of perceptible individuality or any "gappiness" to our minds when the frame-sequence is run. The film would seem the same if it were a notional 10^{15} x 24 frames-per-second movie. However, the inner

theatre metaphor of mind can also mislead. This is because such a metaphor seems to generate an infinite regress of homunculi. How is the inner spectator supposed to view the internal scene if not by means of another inner spectator in turn? And so forth. In reality, our minds partly *instantiate* the virtual world-simulations they run. All analogies break down somewhere; the Cartesian theatre is no exception. Note that the phenomenal unity of perception at issue here is what philosophers call "synchronic" unity. No claim is being made about "diachronic" unity, the fictitious temporal persistence of an enduring metaphysical ego. Such enduring personal identity is fundamental to our conceptual scheme. Yet persisting selves are impossible to reconcile with a physicalistic world-picture, not least with the rapid metabolic turnover of one's constituents – or with the existence of one's 10^{100} near-identical namesakes that post-Everett quantum mechanics implies have partially decohered ("split") since the start of this sentence(22). For expository convenience, the narrative fiction of enduring personal identity will here be retained. In principle, however, each ultra-thin "slice" or quantum-coherent frame of episodic self could be labelled with its own numerical subscript.

A more robust *a priori* objection to quantum mind hypotheses of phenomenal binding might run as follows. No, says the traditional materialist and coarse-grained functionalist, we don't yet understand how consciousness arises from patterns of neuronal firings in the brain. But as reductive physicalists, we *shouldn't* be surprised at the structural mismatch between the phenomenology of bound phenomenal objects and the micro-structure of the brain, *per se*, any more than we should be surprised at the structural mismatch between video game characters and the program code running on the classical computer processor that executes them. There's no need to invoke quantum "woo" when well-understood classical physics and learning algorithms work just fine. Indeed, the same point could be made of a massively classically parallel, "sub-symbolic"

connectionist(23) information-processing system that lacks the transparent and projectable(24) representations of a classical serial computer. Connectionist systems are sometimes called "neural networks" in recognition of their closer gross architectural resemblance to the mind-brain than to a programmable serial digital computer.

Unfortunately, this argument doesn't work either. For sure, when speaking colloquially as though perceptual direct realism were true, we can talk about seeing visually bound video-game characters battling their way across one's computer monitor. And yes, these classical computer-created video-game characters are generated via well-understood, classical computations. No need for quantum "woo" here. However, the manifest phenomenal binding is done entirely by – and is entirely internal to – the sentient organic minds playing the video game: it's internal to the phenomenal world-simulations of the game's players. Such binding is not a property of anything internal to the mind-independent computer display unit. Video game characters lack true ontological integrity: they are not unitary subjects of experience running phenomenally bound world-simulations of their own. All that exists in the mind-independent world are thousands (or millions) of effectively discrete pixels on a monitor screen. Whether our fundamental ontology of the natural world is materialist, panpsychist or idealist in character, these effectively classical pixels do not generate phenomenally unitary subjects of experience that sentient minds engage in combat. Rather, their programmed patterns are part of the distal causal chain that culminates in the minds of organic sentients as video-game characters, i.e. their patterns on a monitor causally covary with the phenomenal game avatars populating the phenomenal gadgets of our phenomenal world-simulations. Effectively, there are no bound phenomena in our personal computers to be matched or mismatched.

In fairness, the insentience of digital zombies has been challenged (*cf.* "This guy thinks killing video game characters is immoral."(25)). Quantum mind-binding theory as defended here vindicates sceptical common-sense.

So what would an exact structural match between bound experiential objects and the superposition of state vectors of neuronal edge-detectors, motion-detectors and colour-detectors (etc) over ultra-Tegmarkian time-frames entail? In "Are Perceptual Fields Quantum Fields?"(26), Brian Flanagan aptly cites Dirac:

"When a state is formed by the superposition of two other states, it will have properties that are in some vague way intermediate between those of the original states and that approach more or less closely to those of either of them according to the greater or less 'weight' attached to this state in the superposition process. The new state is completely defined by the two original states when their relative weights in the superposition process are known, together with a certain phase difference, the exact meaning of weights and phases being provided in the general case by the mathematical theory. (Dirac, PAM. *The Principles of Quantum Mechanics*. Oxford, 1958)

Of course, Dirac wasn't assuming quantum-coherent superpositions of *qualia*-fields, but rather quantum-coherent superpositions of fields of *non*-phenomenal we-know-not-what. Moreover, Dirac was writing about quantum microphysics, not short-lived superpositions of mesoscopic and macroscopic phenomenal objects in warm, wet, organic brains. Yet what if our traditional insistence on a non-phenomenal metaphysical essence to our field-theoretic ontology is dropped? Quantum field theory is no more inherently about fields of *in*sentience than Maxwell's Theory of Electromagnetism is inherently about the properties of luminiferous aether. (*cf.* Heinrich Hertz's terse observation, "Maxwell's Theory is

Maxwell's equations.") Neither theoretical physics nor the phenomenology of mind give any comfort to the idea that the superposition principle really breaks down in the CNS.

4. Can Physicalism be Saved?

Picoseconds are of an unimaginably, mind-wrenchingly long duration compared to the fundamental Planck scale of around 10^{-43} seconds - over thirty orders of magnitude more protracted. Femtosecond, attosecond and even pan-cerebral zeptosecond rates of environment-induced decoherence are still staggeringly long-drawn-out durations once we leave the everyday intuitions of folk-chronology behind. Nonetheless, most neuroscientists would confidently predict that invoking an exact phenomenal-physical structural match at Tegmarkian temporal resolutions to solve the phenomenal binding problem is not just (potentially) falsifiable but false. All we'll discover via interferometry in the warm and wet CNS at such time-scales is an uninteresting, functionally irrelevant and effectively random thermally-induced "noise", not the structural shadows of bound phenomenal objects. After all, picoseconds are seven or eight orders of magnitude shorter than the widely accepted time-frame over which electrochemical neuronal firings cause consciousness to "emerge", inexplicably, in our central neural networks. And pan-cerebral quantum-coherent neuronal superpositions can credibly subsist only for sub-attosecond time-frames before the well-defined phase relations between the components of the superposition are lost, i.e. extended to the extra-neuronal environment in a thermodynamically irreversible fashion.

Again, perhaps orthodoxy is correct. At issue here is a scientifically falsifiable conjecture, not a purely "philosophical" claim. Yet if folk neurochronology is vindicated, then the prospects for physicalism and the ontological unity of science are bleak, or simply non-existent. If folk neurochronology is vindicated, something ontologically irreducible is

present in the world and missing from the formalism of physics. The spectre of "strong" emergence rears its head – or worse, dualism, whether avowedly "naturalistic" dualism or otherwise. True, materialists and epiphenomenalists don't face the binding problem in quite the same way as the physicalistic idealist. Instead, bound phenomenal objects can simply "emerge" in the brain, like Athena sprung fully formed from the head of Zeus.

The ontological floodgates are opened.

5. What Is It Like To Be Schrödinger's Cat?

So let us provisionally suppose, in defiance of orthodox neuroscience, but in conformity with the formalism of unmodified quantum physics, that our prediction of a perfect physical-phenomenal structural match in the CNS turns out to be correct, confounding Chalmers and thereby lending experimental weight to an idealist ontology of monistic physicalism. Rather than embrace epiphenomenalism or Chalmersian dualism, we may on this story transpose the entire mathematical machinery of modern physics to describe an idealist ontology. According to this proposal, sentient beings are wavefunctions in configuration space - fields of phenomenally bound subjective experience whose exact textures are expressed by the values of two numbers, the amplitude and the phase, specified at every point in the universe's configuration space: physicalistic idealism. Every mathematical property of the wavefunction (except the overall phase) corresponds to some subjective property of the physical world.

Suspending disbelief, what would be the pay-off if this conjecture is true?

Let's return to the criteria that must be satisfied by a scientifically adequate theory of conscious mind.

1) Why consciousness exists at all.

This question is best recast as "why is there something rather nothing?" Or more poetically, Hawking's "What is it that breathes fire into the equations and makes a universe for them to describe?" Mysteries should not be multiplied beyond necessity. By positing a non-phenomenal "fire" in the equations, and then hand-waving on how such non-phenomenal stuff might notionally be transmuted into something phenomenal, yet still (somehow) material, avowed materialists build a speculative dualistic ontology into their conceptual framework right from the outset. Compare the Catholic doctrine of transubstantiation. The bread and wine used in the sacrament of the Eucharist literally become the body and blood of Christ while all of their features accessible to the senses remain unchanged. In both cases, we confront "a mystery surpassing all understanding".

There *is* one fundamental mystery. Why does anything exist at all?

When investigating why the enigmatic "fire" of physical consciousness exists, perhaps the fundamental *qualia-fields* of a quantum vacuum, perhaps we might explore some kind of zero ontology as the ultimate logico-physical principle underlying reality, with the field values of the world's hypothetical fundamental microqualia "cancelling out" to zero in a multiverse of net zero information. Within this research program, Max Tegmark's "Does the universe in fact contain almost no information?"(27) might have considered whether the quantum Library of Babel - our Everettian multiverse? - contains *any* information. For in the absence of a preferred basis, the state vector of Everett's multiverse doesn't *per se* contain any information(28). If so, "A theory that explains everything explains nothing" isn't the witty but shallow quip one might assume. No canonically preferred bases of Hilbert space could exist without violating a zero ontology. The mathematical structure of quantum theory allows indefinitely many ways [conventionally, infinitely many ways] to decompose the quantum state of the multiverse into a superposition of orthogonal

states. This leads to another question. Is Wigner's "Unreasonable Effectiveness of Mathematics in the Natural Sciences" explained by the need to conserve an informationless zero ontology - the conservation law that forbids substantive existence? Are the superposition principle and a zero ontology one-and-the-same?

Such difficult questions are beyond the scope of this paper. Proposing that the superposition principle of QM explains both the properties of our minds and why anything exists at all sounds preposterous. But the idea that we may understand either mind or quantum theory *without* understanding why anything at all exists may be naive.

2) How consciousness exerts the causal power to allow intelligent agents to investigate its nature.

According to idealistic physicalism, consciousness has causal efficacy. For instance, a sentient agent removes its hand from the flame because the burning sensation feels agonisingly hot. Unusually, common-sense is actually correct. For sure, in many contexts, for example all programs executed on a classical digital computer, the particular micro-textures of experience constitutive of its phenomenally-unbound physical circuitry are logically and computationally irrelevant to the execution of a program. The particular micro-textures of experience are mere implementation details. But if physicalistic idealism is true, then, strictly speaking, *all* consciousness, and *only* consciousness, exerts causal power, effectively mediated by what we normally recognise as the four forces of nature, or perhaps, ultimately, the vibration modes of higher-dimensional branes of M-theory. Only the physical has causal efficacy; and consciousness discloses the intrinsic nature of the physical.

Without such causal power, not merely would intelligent agents be unable to investigate consciousness: we wouldn't have grounds for alluding to the existence of consciousness

in the first instance. By way of distinction, epiphenomenalists want to claim, presumably, that they have rational *grounds* for believing epiphenomenalism is true - that epiphenomena really are causally impotent. Yet it's unfathomable, to say the least, how such grounds can be stated without implicitly acknowledging a causal role for the epiphenomena that the claim repudiates.

Likewise, on pain of inconsistency, the materialist can't simultaneously assert - as Stephen Hawking(29) does most famously in *A Brief History of Time* - that we have no idea of the character of the "fire" in the equations and yet also dispute that its essence could be phenomenal experience. No doubt "fire" consisting of a non-phenomenal *je ne sais quoi* is a plausible speculation. Yet the claim itself borders on the metaphysical. How does the materialist propose to test his conjecture?

We are also now in a position to answer a commonly posed thought-experiment. Materialists wrestling with their Hard Problem of consciousness sometimes wonder why we don't live in a world physically type-identical to our world but populated instead by insentient zombies. Yet if consciousness discloses the intrinsic nature of the physical, and if quantum-coherent phenomenal binding is the hallmark of mind, then a non-sentient world physically type-identical to our world is logically impossible. A possible world can't simultaneously be physically identical and physically non-identical to our world.

To spike some guns, physicalistic idealism isn't a license for free will, human dignity, animism, New Age mysticism, quantum healing, a reconnection with the timeless wisdom of the ancients, or anything warm and fuzzy. Nor does it invoke quantum mechanical "hidden variables". Nor does it claim that "consciousness collapses the wavefunction". Nor is it a variant of Berkeleyan idealism or the philosophical speculations of the German idealists - though Kant's "transcendental unity of apperception" foreshadowed the global

binding problem. Nor is it a sceptical hypothesis. Thus the physicalistic idealist believes that the mind-independent world existed long before the evolution of bound phenomenal minds in biological organisms. Across the cosmos, the mathematical straitjacket of relativistic quantum field theory is as tight as ever. Physics - or rather tomorrow's ideal physics beyond the energy range of the Standard Model - is causally closed and complete. But within this naturalistic categorical framework, physics is not assumed to be about some essentially non-phenomenal metaphysical "stuff" or unknowable "fire" beyond the reach of scientific investigation. In more Kantian terminology, consciousness is here conjectured to be the noumenal physical essence of the world, the *Ding an sich* ("thing-in-itself") that Kant assumed would forever be unknown and unknowable. Physicalistic idealism turns Kant on his head(8). The noumenal world is all one can ever know, or at least a tiny part of it, other than by inference and conjecture. And the phenomenology of even this sliver of direct knowledge is theoretically contaminated; Wilfrid Sellars called the realm of pure non-inferential experience the Myth of the Given(30).

3) How consciousness can be phenomenally "bound" in seemingly classically forbidden ways.

Physicalistic idealism is not *animism* or vitalism. Its advocates no more believe that a rock is a unified subject of experience than does, say, an eliminative materialist like Daniel Dennett. In common with every other naturalistic theory, the physicalistic idealist still has a lot of work to do in order to show how a bunch of ostensibly discrete quasi-classical nerve cells (or "mind-dust") can generate bound phenomenal objects or a unitary phenomenal self.

Two key questions arise in tackling the classically insoluble binding problem/combination problem.

a) Is macroscopic quantum coherence in the CNS a physically real phenomenon?

b) If so, is the phenomenon long-lived enough to do any computationally and/or experientially useful work - as distinct from being functionally incidental neuronal "noise"?

If quantum mechanics is complete, then the answer to the first question is "yes", albeit over what are, intuitively, vanishingly short durations. However, the existence of macroscopic quantum coherence in the CNS does not, of itself, make the mind-brain a quantum computer any more than the quantum-mechanical properties of silicon (etc) semiconductors make one's desktop PC a quantum computer. One's phenomenal mind and its world-simulation functions as a quantum computer only if what - naively and classically - we describe as the synchronous firings of classically parallel neuronal cellular feature-detectors (edges, colours, shapes, motions, vertices, etc) briefly support a unitary experiential object: the wavefunction of an intelligent, information-processing experiential agent.

This absence of individual neuronal identity in what would otherwise be - as in a dreamless sleep - effectively classical neurons/mind-dust presumably occurs with an ultra-fast "refresh rate" - where "ultra-fast" alludes to our everyday chronological intuitions rather than Planck-scale physics. Within any given sequence of mental life, dropped and mangled frames aren't noticed as such because they aren't explicitly represented in other individual frames. On this story, the molecular structures of our explicit "memories" consolidate only on a much coarser-grained timescale - ranging from hundreds of milliseconds to minutes, hours, days, and in extreme cases, a hundred years

or more. Quantum mind-binding isn't a replacement for connectionist neuroscience or its temporally coarse-grained learning algorithms; rather, it's the bedrock.

So what is conscious? Conversely, what's a micro-experiential zombie?

On this touchstone of sentience - i.e. quantum coherence as the physical signature of phenomenal binding - macroscopic quantum fluids, SQUIDs (Superconducting QUantum Interference Devices), organic mind-brains while not dephased in a coma or dreamless sleep, and perhaps futuristic nonbiological quantum computers are unitary experiential subjects.

Conversely, if effective classicality is the hallmark of the zombie, then serial digital computers, classically parallel connectionist systems, classical dynamical systems, rocks and mountains, the population of the USA, and cellulose-cell-wall-bound plants, etc, are not subjects of experience. Rather they are just decohered aggregates, in effect, composed of phenomenal simples.

Perhaps contrast neuroscientist Giulio Tononi's Integrated information theory(31) in which consciousness is a function of informational complexity.

Currently, if the quantum mind-binding hypothesis sketched here is true, the largest quantum supercomputer in the world belongs to the sperm whale: a mind around five times heavier than its human counterpart. The human cerebral cortex is 2–4 mm thick; but actually we have to take a four-dimensional approach, or more ambitiously, a finite(32)-dimensional Hilbert-space approach, and imagine our minds as 10^{100} or so quasi-classical Everett branches "jostling" each other before becoming irreversibly "split", i.e. effectively decohering and losing their well-defined phase coherence to the extra-cerebral

environment. Intuitive plausibility is *not* the hallmark of a scientifically adequate theory of consciousness. Experiment, not philosophy or armchair physics, is the key.

4) Why and how consciousness has its diverse textures – ranging from phenomenal colours, sounds, tastes and smells to pains and pleasures to the experience of introspecting a thought-episode, understanding a text, or finding a joke funny.

If idealistic physicalism is true, then the solutions to the field-theoretic equations of physics mathematically encode the textures and interdependencies of micro-experiences. The amplitude and phase of one's wavefunction yield the exact values of all one's experiences. On this conjecture, there are no hidden parameters or missing variables that the existing quantum-mechanical formalism omits. Quantum mechanics is indeed closed and complete – or more strictly, it will be closed and complete when it subsumes gravity. Hence the spectre of causal over-determination, epiphenomenalism or even dualism in theory of mind is lifted. Another kind of dualism, the spurious divide between the classical macro-world and the quantum micro-world, evaporates too. The appearance of phenomenally bound classical objects in a classical world is a derived quantum effect, not a brute unexplained fact that a classical materialist ontology can't accommodate.

Finally, any satisfactory theory should offer predictions that are novel, precise, replicable and robustly falsifiable.

Untestable claims may be scientific *if* they are entailed by a conjecture that generates novel, precise, and non-trivial predictions that can be empirically tested. Physicalistic idealism is radically conservative insofar as it does not propose any modification or supplementation of the existing, realistically interpreted, quantum-field-theoretic formalism. *Contra* Penrose and Hameroff's "Orchestrated objective reduction" (Orch-

OR[33]) model, for example, there is no evidence that the unitary dynamics of standard quantum mechanics breaks down in the central nervous system or anywhere else. Yet – without proposing any new physical law(s) – physicalistic idealism also predicts the existence of an empirically investigable phenomenon that few researchers now credit. Namely, our everyday classical world-simulations are underpinned at sub-femtosecond timescales by macroscopic quantum-coherent physical states of the CNS. One's phenomenally bound quasi-classical virtual world is "what a natural quantum computer feels like from the inside", so to speak. Familiar classical worlds of phenomenally bound objects obeying Newtonian laws of motion and gravity within one's perceptual field are an entirely quantum-mechanical phenomenon. Classical phenomenal macroworlds would be impossible without successive neuronal superpositions of distributed feature-processors to underpin their existence.

What Max Tegmark treats as a *reductio ad absurdum* of quantum mind is treated instead as a falsifiable empirical prediction. Nature got there first; and natural selection got to work.

Of course, if a classically-minded critic is convinced *a priori* that macroscopic quantum-coherent neuronal superpositions of sub-femtosecond duration are of no more computational or phenomenal relevance to explaining consciousness than, say, the detection of evanescent quantum superpositions of, say, the pawns and the queen during a game of chess, or random thermal noise in a classical CPU executing a program on one's PC (etc), then such a critic will not waste time independently setting up the exceedingly delicate experiments necessary to detect the missing physical signature of phenomenal binding. They might simply say noise is noise. Collisional decoherence, dephasing due to inertial forces and vibrations, and above all thermal decoherence are all

formidable obstacles to detecting the indirect signature of neuronal superpositions even with the tools of next-generation interferometry.

Such a cavalier dismissal of the only way to save physicalism and the ontological unity of science may prove premature. We will now set out the protocol for an experiment to test the naively absurd conjecture that binding-by-synchrony is really binding-by-superposition.

6. Schrödinger's Neurons: the Experimental Protocol.

In vivo experiments using live human subjects or cats are impossible for the foreseeable future. However, cultured *in vitro* neuronal networks should suffice. First, "train up" a multi-layer neuronal network with a suitable input-output device to recognise a variety of externally presented inputs. Then, identify in turn the distributed neuronal feature-processors implicated in diverse object recognition on a standard, classically parallel connectionist account, i.e. "local" phenomenal binding. Routine neural scanning can pick out what we would naively describe as the synchronously activated distributed neuronal feature-processors elicited by any given stimulus, i.e. textbook connectionist neuroscience but using real neurons rather than tendentiously named "artificial neural networks" and their statistical learning algorithms.

Next comes the fiendishly hard part – feasible in principle, but an experimental challenge still beyond the reach of contemporary molecular matter-wave interferometry. Instead of detecting the fleeting non-classical interference patterns of "nonsense" neuronal superpositions, the conjecture predicts that we'll discover the interference signature of sub-femtosecond macro-superpositions that robustly implicate *exactly the same neuronal feature-processors of the synchronously activated neurons that the classical neuroscience story reports are activated in the trained-up neuronal network when object-recognition*

occurs. On any classical account of mind, such an experimental outcome, i.e. a perfect structural match, is either physically impossible or vanishingly improbable.

The best-known physically demonstrable manifestation of quantum-coherent superpositions is the interference peaks from an electron wave in a double-slit experiment[34]. Currently, matter-wave interferometry can detect "mesoscopic" superpositions of fullerenes[35] in the guise of observable de Broglie wave interference of C_{60} and C_{70} molecules following passage through a diffraction grating. Experimental superpositions of viruses[36] and tardigrades ("water bears") are planned. Detecting the interference patterns of neuronal superpositions with their hugely more numerous excited internal degrees of freedom will be much more challenging because – unlike fullerenes or viruses – functioning neuronal networks can't be steeply cooled down to mitigate the effects of thermally-induced decoherence. In neuronal networks, ion-ion scattering, ion-water collisions, and long-range Coulomb interactions from nearby ions all contribute to rapid decoherence times; but thermally-induced decoherence is even harder experimentally to control than collisional decoherence[37].

However, we may assume tomorrow's experimentalists will rise to the challenge. Let's review the possible outcomes. What will experiments detect when molecular matter-wave interferometry can probe the sub-femtosecond timescales over which theory predicts neuronal superpositions should exist?

1) a) no interference effects, or at least some collapse-like deviation from the unitary Schrödinger dynamics, i.e. the superposition principle breaks down in artificial neuronal networks and thus presumably in the CNS. This negative outcome is what Penrose and Hameroff[38]; Ghirardi, Rimini and Weber (GRW)[39]; and other dynamical collapse theorists would predict.

or

b) the telltale non-classical interference signature that the unitary dynamics predicts.

IF b) is the case, then will the sub-femtosecond neuronal superpositions detected be:

2) a) functionally irrelevant psychotic noise, of no more relevance to the orderly phenomenology of our bound phenomenal minds than, say, fleeting sub-femtosecond superpositions of miscellaneous pawns to the gameplay in a chess match? The Chalmersian "structural mismatch" claim is vindicated.

or

b) a perfect structural match that implicates all and only the synchronously firing feature-mediating neurons that orthodox neuroscience reveals are activated when individual phenomenally bound objects are perceived?

Our femto-mind binding conjecture predicts (b) in both cases.

Some comments are in order here.

First, a good experiment should be "clean" and conceptually simple – its outcome decisive to sceptics and hostile critics, not just to the satisfaction of the conjecture's proponents. No scope should exist for fudging, *ad hoc* escape-clauses or adding epicycles. By this criterion, the experiment outlined here is decisive. A critic of quantum mind will be unfazed by such professions of epistemic virtue: by analogy, building a perpetual-motion machine would be a clean, elegant and definitive refutation of the second law of thermodynamics, too; it's not going to happen. Less fancifully, an example of an "unclean" experiment is the discovery of quantum vibrations in microtubules inside brain neurons as a test of the Hameroff-Penrose Orch-OR theory of mind. Their

discovery, though intriguing, will not persuade critics that modified quantum theory makes Gödel-unprovable results provable by human mathematicians.

Second, strictly speaking, it's not necessary to assume that the superposition principle of QM is universal. Maybe spontaneous localisation kicks in at scales larger than the mesoscopic and modestly macroscopic dimensions of organic mind-brains. Such a breakdown would be physically unmotivated. No departure from the Schrödinger dynamics has ever been detected. But the experimental demonstration of neuronal superpositions won't rule it out.

Third, we have avoided fascinating but incidental speculation about e.g. the properties of liquid water as a unique quantum fluid, dipoles forming superposed resonance rings in helical pathways in microtubule lattices(40), and so forth. For the existence of neuronal superpositions implicating previously naively identified phenomenal feature-mediating nerve cells is a *generic* prediction of any conjecture that invokes coherent superpositions of neuronal feature-processors as the explanation of phenomenal binding. The conjecture – and its confirmation or falsification via matter-wave interferometry – is insensitive to the details of its molecular implementation. Darwin needed Mendel. The ubiquitous selection pressure of Zurek's "quantum Darwinism" applied to the CNS awaits Mendel's counterpart.

Fourth, demonstration of this exceedingly subtle physical interference effect – *if* experimentally confirmed – is not remotely the only reason for believing that organic minds are quantum computers, or that experience discloses the intrinsic nature of the physical. The most striking reason lies in front of our virtual eyes and under our virtual noses, so to speak. But the existence of phenomenal binding is a retrodiction, "old evidence", not a novel prediction. Any claim that armchair philosophising can establish

that the mind is a quantum computer will be given short shrift by critics – even if the claim happens to be true. This *in vitro* interferometry experiment is pitched at quantum mind's most implacable foes.

7. Femto-Mind meets Quantum Darwinism.

"I still recall vividly the shock I experienced on first encountering this multiworld concept. The idea of 10^{100} slightly imperfect copies of oneself all constantly splitting into further copies, which ultimately become unrecognizable, is not easy to reconcile with common sense. Here is schizophrenia with a vengeance."(41)

(Bryce DeWitt)

If DeWitt's notorious misreading of Everett were true, then we'd be (at most) micro-experiential zombies in all life-supporting branches of the universal wavefunction. Unified subjects of experience and phenomenal binding would be impossible. We'd know nothing of one branch, let alone the googols of others. However, DeWitt was mistaken; there is only one world – the multiverse – and its decohering branches never completely separate. DeWitt's remark nonetheless offers a clue to meeting what might seem a decisive objection to a quantum mind account of phenomenal binding. How could selection pressure operate over a timescale of femtoseconds, attoseconds or less? The answer is that whereas selection pressure can't act on proliferating worlds, it *can* act on proliferating, decohering world-simulations. In order to understand our minds and the world-simulations they run, Zurek's "quantum Darwinism"(42) must be applied to the CNS. Here we have a Darwinian selection-mechanism of unimaginable power: ubiquitous, unremitting and temporally fine-grained. Who will play Mendel to Zurek's Darwin is unknown.

These cryptic remarks will now be amplified.

How could non-psychotic phenomenal binding of distributed neuronal feature-processors have evolved? The generation by vertebrate minds of cross-modally matched virtual worlds in almost real time is prodigiously computationally powerful and genetically adaptive. Mere patterns of Jamesian "mind-dust" couldn't *act*. Connectionist neuroscience describes at a coarse-grained level how individual perceptions are represented by shifting coalitions of resting/firing patterns of membrane-bound neuronal feature-processors using different learning algorithms. Yet *if* the phenomenology of virtual world-making ultimately depends on sub-femtosecond quantum coherence, then the *evolution* of non-psychotic phenomenal binding would naively seem impossible. Decoherence, i.e. the rapid effective loss of ordering of the relative phases of complex amplitudes of neuronal superpositions to the environment, is a powerful, omnipresent and seemingly *uncontrollable* effect in the warm, wet CNS.

But we needn't turn to drink or dualism – yet. *If* a femtomind-binding conjecture is correct, and *if* the unitary dynamics of QM doesn't break down in the human mind-brain, then a qualitative answer to the evolutionary enigma of phenomenal binding can be given within the conceptual framework articulated by one of the pioneers of the decoherence program in post-Everett quantum mechanics, Wojciech Zurek. The decoherence program outlines the Darwinian(43) process responsible for the emergence of quasi-classical reality from its quantum substrate within Everett's multiverse. *If* a femtomind-binding conjecture is correct, then an analogous Darwinian process of replication, variations amongst the copies, and differential survival of the copies is responsible for the emergence of the quasi-classical phenomenal worlds forming our minds from their quantum substrate in the CNS. Crudely, some superpositions are fitter than others. In

order for an ecologically credible quantum mind-binding conjecture to be viable, all that is needed for selection pressure to get to work is the slightest heritable predisposition to the tiniest of transmissible resistance to collisional and thermally-induced decoherence of non-psychotically bound phenomenal neuronal superpositions in even the humblest of cephalic ganglia. All organisms capable of neuronal world-modelling evolve and adapt to their environment by an iterative process. This iterative process may be treated as an evolutionary algorithm that searches the fitness landscape for the locally and globally bound phenomenal states of mind – quantum-coherent neuronal superpositions – that are best adapted to their local surroundings. Thus a Darwinian process of variation and differential selection of informational superpositions plays out as the fittest phenomenally bound variants are retained and passed on to their offspring. It's worth stressing again: *contra* de Witt's colourful quote above, there is only one multiverse; interference effects between Everett branches that have effectively decohered ("split") never wholly disappear. Within the universal wavefunction, such a Darwinian process hypothetically plays out *both* between proliferating, sexually reproducing biological organisms *and* fast-proliferating states of the mind-brain of individual organisms across Everett branches. Thanks to hundreds of millions of years of natural selection, the most dynamically stable phenomenally-bound system–environment correlations are the non-psychotically bound phenomenal objects populating our waking world-simulations. Psychotic binding in maladapted organisms does still occur, *comparatively* infrequently; but statistically, one's waking consciousness (as now) is overwhelmingly likely to consist in non-psychotically bound states of an adapted organism, not the Earthly counterpart of a Boltzmann brain. What we're calling "informational" and "psychotic" binding should be conceived dimensionally rather than categorically. Thus a fleeting quantum-coherent superposition of distributed neuronal feature-processors experienced as, say, a flying purple dragon is

psychotic in the context of the ancestral environment of adaptation, whereas fleeting quantum-coherent neuronal superpositions of distributed feature-processors experienced as an approaching lion were potentially hugely fitness-enhancing in the extra-neural presence of a hungry predator. But flying purple dragon superpositions are not intrinsically psychotic, any more than the phenomenally-bound features of predatory lion superpositions are inherently referential – on pain of a magical theory of reference. Indeed, in some future fantastical techno-utopia – or immersive VR with different laws from basement reality – flying purple robo-dragon superpositions could be functionally non-psychotic. They might track patterns in the local mind-independent environment. What counts as sanity is contextual.

For illustrative purposes, an example with somewhat greater ecological validity than neuronal flying purple dragon superpositions might be in order. Imagine a savannah-dwelling herbivore with two disorders of phenomenal binding: *both* simultanagnosia and cerebral akinetopsia ("motion blindness"). Not merely can the herbivore's doubly unbound mind apprehend only a single perceptual object at a time; the object's progressive motion can't be perceived. So not merely is just a single member of an approaching pride of hungry lions apprehended within the herbivore's CNS world-simulation; the hungry carnivore in question just appears successively nearer without perceptibly advancing. Such a neurologically devastating condition might seem a sure-fire recipe for the hapless herbivore becoming lunch. Today, such a grisly fate would be almost inevitable. Yet to survive and genetically propagate, the doubly-unbound ancestral herbivore doesn't need to outrun the approaching lions – merely to run faster than other members of the herd. If his or her conspecifics are capable only of psychotic binding – or if their neurons are merely effectively classical or phase-scrambled neuronal "mind-dust" – then our doubly mentally unbound herbivore actually has an immense

selective advantage over every other member of the herd. For even weak and partial non-psychotic phenomenal binding confers a huge selective advantage over organisms that lack non-psychotic binding (at anything above chance levels) altogether. Or to use another, evolutionarily more ancient example, imagine a simple organism with a heritable predisposition to apprehend phenomenal patches of darkness and light – as distinct from the heritable predisposition of its conspecifics to instantiate merely discrete, decohered, effectively classical dark or light neuronal "pixels". This primordial proto-binder can functionally distinguish night from day, and safely graze (or filter-feed) rather than burrow to safety as needed in the shadow of a looming predator. Such an adaptation would be powerfully fitness-enhancing. Over evolutionary history, non-psychotic binders would outcompete psychotic binders, and superbinders will outcompete binders, culminating in the currently supreme superbinder of them all, *Homo sapiens*.

Note that on this account, Darwinian selection pressure plays out *both* between proliferating, sexually reproducing organisms across the generations and also between ultrafast-proliferating neuronal superpositions of the CNS. For although (we conjecture) next-generation matter-wave interferometry will robustly detect a perfect structural match between the reported bound phenomenology of our minds and non-psychotic neuronal superpositions, nonetheless post-Everett QM suggests that fleeting, erratic, nonsensical superpositions really do exist; they are merely of vanishingly rare measure *compared to* the information-bearing superpositions favoured by natural selection. Thankfully, experimental interferometry rather than speculative philosophising will decide the issue.

8. A Mendeleev Table for Qualia?

If sentient agents are to understand the intrinsic subjective properties of matter and energy, or to map out what we naively call the "neural correlates of consciousness", or most ambitiously, to devise a comprehensive "Mendeleev table" for qualia, then the diverse subjective textures of consciousness will play an inescapable role in the investigation by the very nature of the task. Intelligent agents will need to re-engineer themselves – genetically, pharmacologically, neurologically – in order to *instantiate* the subjective physical states in question. We'll need to become a full-spectrum "super-Mary"(44), so to speak – investigating state-spaces of consciousness disclosed by configurations of matter and energy that have never before been recruited for any information-processing purpose. Such state-spaces of consciousness are currently beyond the scope of scientific investigation.

By contrast, classical digital zombies cannot explore the nature of sentience; their circuitry wouldn't understand what they were investigating, let alone be cognisant of its mechanisms. This far-reaching task falls to bound phenomenal minds. A combinatorial explosion of possibilities means that the investigation of the alien state-spaces of consciousness may take millions of years, perhaps billions or more. By contrast, constructing the mathematical formalism of a unified TOE over the next few decades may prove surprisingly easy. [Just email the author for details.]

Early in the twenty-first century, we commonly assume that physical scientists research the objective properties of matter and energy. This is true – up to a point. If physicalistic idealism is correct, then this commonplace is no more than a half-truth. For the intrinsic, subjective, first-person properties of matter and energy are real, objective and amenable to formal description via the evolution of the universal wavefunction, just as are the

third-person relational properties – the properties captured by the formalism of relativistic quantum field theory or its successor. In short: we've mastered the right formalism, just assumed the wrong materialistic ontology. Subjective experience and phenomenal binding are a Hard Problem for the classical scientific materialist in the same way that fossils are a Hard Problem for the Creationist. In both cases, the anomaly in question demands a major revision of the believer's conceptual scheme. In both cases, believers are prone to spending their lives in denial.

On the face of it, to pronounce on the nature of what physical science is actually investigating might seem presumptuous for anyone but a professional physicist. Yet we don't allow the fact that, say, Newton believed he was investigating divine mechanical clockwork, or that he fancied his foremost achievement was his interpretation of the Book of Daniel, to impugn Newton's status as the greatest scientist who ever lived. Likewise, it's no disrespect to the greats of contemporary mathematical or experimental physics to say that we still don't understand the intrinsic nature of physical reality. Likewise, it's no disrespect to hard-working neuroscientists to say that we simply don't understand the mind-brain when its defining feature, consciousness, is physically impossible within the reigning materialist paradigm of science.

In a similar vein, to assert that mathematics investigates patterns of quantity, structure, space, and change would seem a commonplace. The claim that maths is really about *qualia*-patterns sounds bizarre. More telling is Bertrand Russell's jaundiced observation "Mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true." If idealistic physicalism is correct, then mathematics is ultimately about computable patterns of qualia: their quantity, structure, and change. Once again, perhaps we've mastered the formalism rather than adequately grasped the underlying ontology whose relations it captures.

9. Towards A Post-Galilean Science of Mind.

"*If a potato or rutabaga can utilize quantum coherence, it's likely our brains could have figured it out as well.*"

(Jack Tuszynski of the University of Alberta)

A comprehensive account of reality entails an understanding of the first-person and third-person properties of the natural world – and the mathematically formalised interrelationships between them. If the distinction between the first-person and third-person properties of matter and energy were completely clean, as assumed by traditional AI, then the causal capacity of cognitive agents to allude to both the subjective and formal properties of mind would be physically impossible in the first instance. Thus an insentient p-zombie would be physically unable, for example, to refer indexically to *this* particular self-intimating thought, or to investigate the nature of phenomenal binding, or to explore the nature of the "fire" in the equations that is responsible for the existence of sentient minds for non-zombies to describe. For a notional materialist p-zombie, it isn't even "all dark inside".

The necessity of the experimental method in scientific investigation of the third-person properties of matter and energy has been recognised since Galileo. The intellectual achievements of physical science, as traditionally conceived, are widely celebrated. By contrast, experimental investigation of the great majority of intrinsic, first-person properties of matter and energy is stigmatised and even criminalised. States of sentience as different as waking from dreaming consciousness are outlawed. Instead of Nobel laureates, research grants and lavish institutional funding, an empirically-driven exploration of the first-person properties of matter and energy plays out mainly within

the scientific counterculture. An entire realm of drug-catalysed knowledge is proscribed as somehow cognitively illegitimate.

Human ignorance is unlikely to last indefinitely. If intelligent agents are to understand the natural world, then the methodology pioneered by Alexander "Sasha" Shulgin (1925-2014) in "PiHKAL"(45) must be integrated with mainstream academic science: an authentically post-Galilean science of physical consciousness.

Does the claim that biological agents – and perhaps mature nonbiological quantum computers centuries hence – can solve problems too difficult for a classical system to pose or answer violate the Church-Turing thesis(46), i.e. that any effective computation can be carried out by a Turing machine? By itself, technically, no. After all, a notional classical digital computer could be programmed to code the chemical base-pairs for the genotypes of biological super-Shulgins whose phenomenally bound minds could then explore the manifold varieties of sentience and map out the psychophysical relationships between them. Yet such a whimsical proposal doesn't mean that a classical digital computer could itself ever support a unitary full-spectrum (super)intelligence. Non-classical phenomenal binding is a necessary precondition for full-spectrum general intelligence. For without phenomenal binding, there is no unitary agent who is (un)intelligent in the first instance, let alone a general problem-solver who can systematically investigate the first-person and third-person properties of the physical world.

What is sorely lacking here is a rigorous account of computation that can handle the investigation of myriad state-spaces of qualia as well as the traditional staples of third-person computing. This challenge doesn't count as a well-defined or even meaningful question within the reigning paradigm of computer science. Sentient organic minds are

biological devices that can answer questions beyond those a classical Turing machine can answer *or even pose* – not because we are "oracles", but because – if the conjecture outlined here is experimentally vindicated – we are sentient, phenomenally bound quantum computers. Full-spectrum superintelligence will entail a seamless mastery of both the formal and the subjective properties of mind: the creation of a mature civilisation of super-Shulgins-cum-super-Turings. Recursively self-improving organic robots are poised to modify their own source code(47) and bootstrap our way to full-spectrum superintelligence. How closely posthuman conceptions of the physical resemble anything humans would recognise(48) is an open question.

10. Summary and Prospects.

The Hard Problem of Consciousness Solved; the Explanatory Gap Closed; the Binding Problem Tamed; Zombies Banished; and Physicalism Saved.

Let's recap. Here are our key assumptions and the weird but experimentally falsifiable prediction that follows. If the prediction fails, then our defence of idealistic physicalism is refuted.

1) Strong emergence is false. Physicalism is true. No "element of reality" is missing from the equations of tomorrow's physics and their solutions.

2) Consciousness discloses the intrinsic nature of the physical. Therefore, rudimentary consciousness occurs, not just at ultra-small distance scales, but also at ultra-short time scales. A future Planck-scale unification of quantum gravity will presumably capture the ultimate "psychon" of Planck-regime consciousness.

3) The unmodified, unsupplemented formalism of post-Everett quantum mechanics is correct. "Hidden variables", Bohmian mechanics, and dynamical collapse theories of wavefunction collapse are false. Thus macroscopic quantum-coherent neuronal superpositions occur in the mind-brain. At sufficiently fine-grained temporal resolutions, the entire mind-brain exists in a single, conscious, quantum-coherent superposition. A succession of ultra-rapidly decohering virtual world superpositions constitutes biological minds. Internally, world-simulations typically seem classical. Their vehicles, i.e. phenomenally bound organic minds, are irreducibly non-classical.

4) Direct realism about perception – and hence the notion that neurosurgeons or experimenters ever directly "observe" anyone else's decohered classical brain or decohered classical neurons – is false. When notionally "observing" our surroundings, both awake and dreaming organic minds instantiate individual bound perceptual objects ("local" neuronal binding) that populate dynamic world-simulations undergone by a fleetingly unitary phenomenal self ("global" binding). Phenomenal binding is not a classical phenomenon. Instead, phenomenally bound quantum-coherent neuronal superpositions have been recruited by natural selection to generate seemingly mind-independent, ostensibly classical virtual worlds. When awake, quantum biocomputers generate such pseudo-classical worlds to track fitness-relevant patterns in our local environment. Except in a dreamless sleep or coma, organic mind-brains are not decohered "pixels" of discrete neuronal micro-experiences.

The Retrodiction.

We are not zombies. Nor are we quasi-zombies, i.e. patterns of decohered neuronal "mind-dust". So there is no Hard Problem of consciousness and, in principle, no binding problem either: we're not micro-experiential zombies. Instead, we are fleetingly unitary

phenomenal minds. Empirical evidence that our minds are quantum computers lies in front of our (virtual) eyes.

The Novel, Experimentally Falsifiable Prediction.

Next-generation interferometry will detect the sub-femtosecond signature of quantum-coherent neuronal superpositions in the mind-brain in the guise of quantum interference effects AND these indirectly detected quantum-coherent neuronal superpositions will robustly implicate all and only the synchronously firing feature-mediating neurons that orthodox neuroscience suggests are activated when individual phenomenally bound objects are perceived by the experimental subject.

Both predictions must be experimentally borne out in order to vindicate the quantum mind-binding conjecture outlined here. So if either no neuronal superpositions are detected, i.e. if the unitary evolution of the state vector breaks down in the mind-brain, OR if their interference signature is indeed deciphered but also implicates neurons irrelevant to the neuronal feature-mediators of the particular phenomenally bound object(s) that the experimental subject or trained up *in vitro* neuronal network reports seeing, i.e. if the interference effects detected are functionally just molecular "noise", then our quantum mind conjecture will be falsified. Falsified too would be our attempt to save physicalism.

Experimentally detecting – or definitively failing to detect – the nonclassical interference effects diagnostic of *local* phenomenal binding in the CNS will be technically less challenging than detecting the predicted trans-cerebral quantum interference effects diagnostic of *global* phenomenal binding, and hence the unitary phenomenal self of everyday experience. Yet the quantum mind-binding conjecture will – provisionally – be vindicated if the signature of even local neuronal superpositions in their predicted guise

are found. By analogy, if a bizarre but nonetheless falsifiable conjecture predicts (what orthodox neuroscience would claim is) the equivalent of little green pixies living at the bottom of the garden, and – amazingly – a single little green pixie is unequivocally detected, then we wouldn't withhold assent to the bizarre conjecture on the grounds that experiment hadn't yet detected the theorised pixie breeding colony.

Further Challenges.

1) The mechanisms supporting the succession of differentially robust sub-femtosecond neuronal superpositions that – hypothetically – underpin phenomenal binding must be elucidated at the molecular level. Only at the molecular level can philosophical hand-waving be turned into real, measurable, quantitatively exact physical science. At much longer time-scales of milliseconds and above, the standard coarse-grained story from connectionist neuroscience and dynamical systems theory takes over from the femto-mind regime. Thus whether we are in a dreamless sleep, dreaming or wide awake, our memories are coarsely encoded in the connectivity, connection weights, and the internal architecture of our neurons after our neural networks have been progressively "trained up". Besides its idealist ontology, the quantum mind-binding conjecture explored here to save physicalism from the spectre of Chalmersian dualism is radically unorthodox only insofar as what mainstream neuroscience reckons is the mere synchronous firing of classical neuronal distributed feature-processors is conjectured instead to be a succession of quantum-coherent neuronal superpositions. Only experiment can corroborate or falsify this hypothesis. If the prediction fails, then our defence of idealistic physicalism is refuted too.

2) Even if non-materialist physicalism is true, the lack of some sort of Rosetta Stone to "read off" the values of qualia – both bound and unbound – from the solutions to the

field-theoretic equations of QFT is a huge challenge. Compare a much more straightforward identification. Nowhere in Maxwell's field equations is light explicitly identified with electromagnetic radiation. But once the value of the constant **c** was calculated – around 300,000 kilometres per second – then the identity of its value with the known velocity of light made the identification inevitable. In other words, no "element of reality" was missing from Maxwell's formalism, or, more strictly, from its subsequent quantum electrodynamic generalisation. Likewise, if idealistic physicalism is true, no "element of reality" is missing from the formalism of relativistic quantum field theory or its currently speculative successor. However, in contrast to the ease of identification of light with visible frequencies of electromagnetic radiation, the conjecture that the solutions to the equations of QFT yield the precise values of all and only physically possible experiences amounts to both a mathematical straitjacket and a veritable Pandora's box. For the only way cognitively to grasp the values of the diverse subjective properties of the physical fields of experience that the solutions to the formalism encode is personally to instantiate bound neuronal superpositions of these subjective properties. Even after extensive psychotropic and eventually neurogenetic experimentation, myriad forms of consciousness will presumably be forever inaccessible to rational mind – though equally, many physical systems that today we might naively imagine could in future be unitary subjects of experience, notably ultra-powerful classical digital computers or nonbiological classical connectionist systems, will always be effectively insentient.

3) Whether our conscious minds are essentially classically parallel, connectionist systems, or quantum supercomputers as conjectured here, another enigma remains. The late evolutionary neurological mechanism by which a massively parallel biological neurocomputer generates a virtual classical machine – the slow serial stream of one's

logicio-linguistic thinking via which this paper is written and read – is unknown. We do know of crude methods to disrupt our stream of logico-linguistic thought-processing. For example, taking LSD induces the "flooding" phenomenon that disrupts serial thought, whereas low-dose psychostimulants tend modestly to enhance logico-linguistic thought. Yet that's as far as it goes. Whatever the nature of this virtual seriality-generating mechanism in the CNS, we can sketch out an evolutionary chronology of information processing systems. An irreducibly quantum multiverse first generated information-bearing self-replicators – biological life – which manufactured quantum supercomputers in the form of central nervous systems, one species of which spawned the serial, logico-linguistic virtual machines currently unique to human minds. These serial virtual machines conceived and created classical digital computers, then classically parallel artificial connectionist systems, and finally – though here we run a little ahead of our story – artificial nonbiological quantum computers. The long-term interplay of these multiple architectures is hard to foresee with any confidence; but the destiny of sentient life in the cosmos most probably lies in full-spectrum superintelligence(49).

* * *

REFERENCES

1. Einstein, Albert; Podolsky, Boris; Rosen, Nathan (15 May 1935), "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?", *Physical Review* 47 (10): 777–780.

2. Weinberg, S. (1995). "The Quantum Theory of Fields" 1–3. Cambridge University Press.

3. Levine, Joseph. (1983). “Materialism and qualia: the explanatory gap”. *Pacific Philosophical Quarterly*, 64: 354-361.

4. Chalmers, David (1995). “Facing Up to the Problem of Consciousness.” In: *Journal of Consciousness Studies* 2 (3), pp. 200-219.

5. Kant, Immanuel. "Critique of Pure Reason", "Critik der reinen Vernunft" (1781; rev. ed., "Kritik der reinen Vernunft", 1787; 1929, 1950).

6. Schopenhauer, Arthur. 1819/1995. "The World as Will and Idea". Trans. M. Berman. London: J. J. Dent.

7. Russell, Bertrand. "The Analysis of Mind", London, G. Allen & Unwin; New York, Macmillan, 1921.

8. Lockwood, Michael. "Mind, brain, and the quantum: the compound". Blackwell, 1989, ISBN 978-0-631-16183-7.

9. Strawson, Galen. (2006). "Consciousness and Its Place in Nature: Does physicalism entail panpsychism?"

10. Revonsuo, Antti. "Binding and the phenomenal unity of consciousness." *Conscious Cogn*. 1999 Jun;8(2):173-85.

11. Everett, Hugh. (1957): `Relative State Formulation of Quantum Mechanics', *Reviews of Modern Physics*, 29, pp. 454-462.

12. Revonsuo, Antti, (2006). "Inner Presence: Consciousness as a biological phenomenon". Cambridge, MA: MIT Press.

13. Chalmers, David. (2014). "The Combination Problem for Panpsychism." Published in Brüntrup, G. (2017). *Panpsychism: Contemporary Perspectives*. New York, NY: Oxford University Press.

14. Moore, Gregory (2005). "What is... a Brane?". *Notices of the AMS* 52: 214.

15. Pearce, David, (2010). "Quantum computing: the first 540 million years." Toward a Science of Consciousness" Conference. Tucson, Arizona.

16. Riddoch, MJ., Humphreys, GW. (2004). "Object identification in simultanagnosia: When wholes are not the sum of their parts." *Cognitive Neuropsychology*, 21(2-4), Mar-Jun 2004, 423-441.

17. Zeki, S. (1991). "Cerebral akinetopsia (visual motion blindness): A review". *Brain* 114, 811-824. doi: 10.1093/brain/114.2.811.

18. Crick, Francis, Koch, Christof. (2003). "Framework for consciousness". *Nature Neuroscience* 6 (2).

19. Schwitzgebel, Eric. (2014). "If Materialism Is True, the United States Is Probably Conscious". [forthcoming].

20. Dennett, Daniel. "Sweet Dreams: Philosophical Obstacles to a Science of Consciousness" (MIT Press 2005) (ISBN 0-262-04225-8).

21. Max Tegmark. "Why the brain is probably not a quantum computer". *Information Sciences*, (128):4194-4206, 2000.

22. Wallace, David. (2012). "The Emergent Multiverse: Quantum Theory according to the Everett Interpretation". OUP Oxford.

23. Rumelhart, DE., McClelland, JL., and the PDP Research Group (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition". Volume 1: Foundations. (Cambridge, MA: MIT Press).

24. Clark, Andy. (1989). "Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing". Cambridge, Mass.: MIT Press.

25. "This guy thinks killing video game characters is immoral". *Vox Magazine*. April 23, 2014.

26. Flanagan, Brian. "Are perceptual fields quantum fields?" *Neuroquantology* 3 (2003).

27. Tegmark, Max. "Does the universe in fact contain almost no information?" Max Tegmark *Foundations of Physics Letters*.

28. Schwindt, Jan-Markus. "Nothing happens in the Universe of the Everett Interpretation", arXiv:1210.8447v1 [quant-ph] 31 Oct 2012.

29. Hawking, Stephen. (1988). "A Brief History of Time". Bantam Publishing Group.

30. Sellars, Wilfred. (1956). "The Myth of the Given: Three Lectures on Empiricism and the Philosophy of Mind." published in "Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis" (University of Minnesota Press, 1956), pp. 253-329.

31. Tononi, Giulio. (2008). "Consciousness as integrated information: A provisional manifesto". *The Biological Bulletin* 215: 216-242.

32. Landauer, Rolf. "The physical nature of information", *Phys. Lett*. A 217, 188 (1996).

33. Hameroff, Stuart; Penrose, Roger (March 2014). "Reply to criticism of the 'Orch OR qubit' – 'Orchestrated objective reduction' is scientifically justified". *Physics of Life Reviews* (Elsevier) 11 (1): 94–100. doi:10.1016/j.plrev.2013.11.013.

34. Roger Bach *et al.* "Controlled double-slit electron diffraction", 2013 *New J. Phys*. 15 033018 doi:10.1088/1367-2630/15/3/033018.

35. Arndt, M.; Nairz, Olaf; Vos-Andreae, Julian; Keller, Claudia; Van Der Zouw, Gerbrand; Zeilinger, Anton (1999). "Wave-particle duality of C60". *Nature* 401 (6754): 680–2. Bibcode:1999Natur.401..680A. doi:10.1038/44348. PMID 18494170.

36. Romero-Isart, O., Juan, M. L., Quidant, R. & Cirac, J. I. "Toward quantum superposition of living organisms". *New J. Phys*. 12, 033015 (2010).

37. Schlosshauer, Maximilian (2007). "Decoherence and the Quantum-to-Classical Transition" (1st ed.). Berlin/Heidelberg: Springer.

38. Hameroff, Stuart; Penrose, Roger (March 2014). "Consciousness in the universe: A review of the 'Orch OR' theory". *Physics of Life Reviews* (Elsevier) 11 (1): 39–78. doi:10.1016/j.plrev.2013.08.002. PMID 24070914.

39. Ghirardi, G.C., Rimini, A., and Weber, T. (1985). "A Model for a Unified Quantum Description of Macroscopic and Microscopic Systems". *Quantum Probability and Applications*, L. Accardi *et al*. (eds), Springer, Berlin.

40. Hameroff, Stuart; Penrose, Roger (March 2014). "Reply to criticism of the 'Orch OR qubit' – 'Orchestrated objective reduction' is scientifically justified". *Physics of Life Reviews* (Elsevier) 11 (1): 94–100. doi:10.1016/j.plrev.2013.11.013.

41. "Quantum mechanics and reality", *Physics Today* (Sept. 1970, 30—35).

42. W. Zurek, "Quantum Darwinism", *Nature Physics* 5 (2009) p. 181 - 188 (doi:10.1038/nphys1202).

43. "Quantum Darwinism as a Darwinian process", John Campbell , 2010/01/5, arXiv:1001.0745.

44. Jackson, Frank. (1986). "What Mary Didn't Know", *Journal of Philosophy* 83: 291–295.

45. Shulgin, Alexander. (1995). "PiHKAL: A Chemical Love Story". Berkeley: Transform Press.

46. Copeland, B. Jack. "The Church-Turing Thesis", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.).

47. Sander J.D., Joung J.K. (2014). "CRISPR-Cas systems for editing, regulating and targeting genomes". *Nature Biotechnology*. doi:10.1038/nbt.2842. PMID 24584096.

48. Shulgin, A. (2011). "The Shulgin Index Vol 1: Psychedelic Phenethylamines and Related Compounds". (Berkeley: Transform Press, US).

49. Pearce, D. (2012). "The Biointelligence Explosion" in "Singularity Hypotheses: A Scientific and Philosophical Assessment". eds. Eden, AH; Moor, JH; Soraker, JH; Steinhart, E. (Springer, Frontiers Collection, ISBN-13: 978-3642325595).

Terminological Note for Philosophers

"The Brain is wider than the Sky

For put them side by side

The one the other will contain

With ease and You beside"

Emily Dickinson

"*Organic VR*?"

What is this?

In the critique of Huxley's BNW, it is assumed that sophisticated post-humans won't be naïve realists about perception. For a realistic interpretation of quantum theory allows only an *inferential* realism. Both direct and indirect perceptual realism are untenable folk-theories.

Of course, the philosophers' Problem Of The External World really demands a book. The topic certainly shouldn't be dispatched with the dogmatic brevity of an endnote. Yet stated baldly: irrespective of whether we are awake or asleep, what each of us intuitively apprehends as the mind-independent world "out there" - colourful, noisy and hugely refractory - is a virtual simulation run by one's own mind/brain. "The World" as apprehended beyond one's body-image is simply one simulation among billions of

throwaway genetic vehicles spawned by selfish DNA. Each autobiographical virtual world is identical with distinctive patterns of neuronal firings in a vertebrate CNS. Thanks to the playing out millions of years of Darwinian natural selection, all but the most deranged mind/brains are coded to embody dynamic, data-driven simulations of their immediate environment. But such virtual worlds, like our conscious self, are no less fleeting, episodic and dispositional in their nature than are our beliefs and desires. In common with the conscious self, they disappear in a dreamless sleep.

The connection and activation weights of our neural nets, however, persist while their host organism slumbers. So "the world" abruptly recreates itself when we "awake". Opening one's eyes serves to re-impose selective discipline on our ways of worldmaking [in the proximate, non-Darwinian sense of "selective"]. Thus on waking up each morning, one's capacity to generate a virtual world becomes constrained once more by inputs from the optic nerve. The austere regimen of one's episodes of waking life contrasts with the psychotic excesses of one's dreams.

On this interpretation of the Human Predicament, one can only ever *infer* that there are billions of other experiential worlds like one's own. Indeed one can only ever *infer* the existence of a vast natural Multiverse - whose organisms and virtual worlds play host to unfolding virtual dramas like one's own. Likewise, one may *infer* that the contents of one's virtual world are tightly selected by peripheral impulses when one is awake. Conversely, when one is asleep, or hallucinating in a sensory-deprivation tank, or tripping on major psychedelics, the features and narratives of one's virtual world are quasi-autonomous. One tends to get "lost in space" - an abstract neural weight-space of possible worlds.

"Waking up" does *not* turn the neuronal firings identical with one's colourful dreamworld into a neutral vehicle for the occult faculty commonly known as "perception". One's nerve cells don't - indeed *can't* - metamorphose into a transparent medium for accessing the extra-cranial Real World. Likewise, when "awake" rather than dreaming, one doesn't cease thinking and talking in mentalese masquerading as a public language. This feat would be pure ontological magic. It would also wantonly violate Occam's razor. Mentalese is the only language one can ever know.

In the course of evolution, natural selection has churned out billions of species-specific virtual worlds - i.e. rival organic quantum supercomputers - in creatures with central nervous systems. The simulations run by such virtual worlds serve as disposable genetic vehicles no less than the organisms who host them - and whom they help reproduce. Like their hosts, these virtual worlds senesce and die. Some macroscopic worlds are fitter than others. Host organisms whose brains run such genetically adaptive virtual worlds tend to leave more copies of themselves and their kin than their genetic rivals.

A cardinal feature of each virtual world is its egocentricity. Each of us lives in a world whose centre is our body image. Virtual worlds are egocentric because coding for a self-centred universe helps maximise the inclusive fitness of selfish DNA. A "view from nowhere" would be genetically maladaptive. So world-making DNA macromolecules ensure that the egocentric delusion is a heritable design feature of the worlds they encode.

The virtual worlds of the Darwinian Era may be classed as "**organic** VR" because their contents are in part selected (but not created) by organic peripheral inputs. These electrochemical impulse-patterns are themselves in turn shaped by sensory transducers at the body surface. "Silicon VR" or, more generally, "**synthetic** VR" refers to virtual

worlds whose selection-regime of inputs derives directly from non-organic retinal imaging devices, body-suits, silicon implants, etc.

To complicate matters, there will soon be artificial bionic devices that blur the distinction between organic and synthetic VR. Moreover, unless silicon systems support the warm QM-coherent states needed for experiential manifolds - a most unlikely proposition - there is a sense in which *all* experiential VR is organic. For it inheres in organic wetware alone [on account of the unique valence properties of the carbon atom]; only the mode of world-selection is different.

Evolution has harnessed the intrinsic properties of certain minds/brains/virtual worlds to play a representational/computational/simulational role in the organisms it spawns. This process of recruitment occurs because near-real-time tracking of regularities in the local environment is genetically adaptive. It allows awake bodies to navigate a dangerous world.

But unreflective naïve realism is itself a highly adaptive delusion for organisms in its grip. The mind-independent world doesn't - and couldn't - *directly* imprint its signature on our brains/minds/virtual worlds. Their *intrinsic* properties are *not* - and couldn't be - contingent on the particular occasions on which they are triggered. [Actually, this is an over-simplification. The separability and individuality of events in our classical worlds may emerge from the non-locality (but see Mike Price's Everett FAQ) and superposition of pluralities of its fundamental quantum substrate. This is a big subject.]

The delusiveness of perceptual realism will be clearer when we are able to construct minds/brains/virtual worlds to order in vats; or sooner still, when immersive multi-modal VR becomes a trillion-dollar entertainment industry late this century. Notoriously, the dominant technology of an era tends to supply its root metaphor of mind; and the advent

of pervasive VR will probably transform our root metaphor of mind into some sort of virtual world. In any event, although "synthetically"-selected, none of the states of our future virtual worlds will be either more or less natural, nor inherently more or less representational, than others. This (non-)representational status is context-dependent. The most that any proximate selection-process can ever do is to ring the changes on a pre-set menu of neurochemical pathways. For just as there are a finite number of games of chess, there are a finite number of mind/brain/virtual world states for a system of any given size. But we've scarcely begun to explore them.

It's worth briefly contrasting the inferential realist perspective sketched above with the Dennettian argument that conscious mind, insofar as such a phenomenon exists at all, is a virtual serial computer supervening on a parallel one.

Unfortunately, the Dennettian approach confuses conscious mind with *self*-conscious mind and its thought-episodes; and relies on a crude realism about "perception".

In reality, the most intense experiences one undergoes (e.g. "physical" agony) are also the most "primitive". One's stream of thought (including "encephalised" emotion) may indeed be akin to a serial computer supervening on a parallel one. But it is *parallel* computation which embodies the most intense and vivid modes of consciousness, whereas the consciousness of the virtual serial computer which supervenes upon it is phenomenologically impoverished. Thus introspecting one's thoughts is hard work at the best of times. It is extraordinarily difficult even to count or individuate them. Cognitive phenomenology is rarefied and subtle. By contrast, one's visual, auditory and tactile worlds - whether or not they are cross-modally matched - are vivid, incontrovertible and extraordinarily intense; and these virtual worlds go a *long* way down the phylogenetic tree.

If you think you're plugged straight into the Real World, then the prospect of plugging in to silicon VR will seem like a retreat into fantasy-world escapism. On the other hand, if you've long ceased to believe that The World was yours to lose in the first place, then you may decide that nasty old organic VR is a world well lost.

FURTHER READING

Perhaps the best contemporary treatment of the inferential realist perspective can be found in Steven Lehar's *The World in Your Head* (2003), and *Inner Presence: Consciousness as a Biological Phenomenon* (2005) by Antti Revonsuo.

Part V: The Sentience Explosion

The Biointelligence Explosion

How recursively self-improving organic robots will modify their own source code and bootstrap our way to full-spectrum superintelligence

"Homo sapiens, the first truly free species, is about to decommission natural selection, the force that made us.... Soon we must look deep within ourselves and decide what we wish to become."

Edward O. Wilson, *Consilience: The Unity of Knowledge* (1999)

"I predict that the domestication of biotechnology will dominate our lives during the next fifty years at least as much as the domestication of computers has dominated our lives during the previous fifty years."

Freeman Dyson, *New York Review of Books* (July 19, 2007)

1 The Fate of the Germline

Genetic evolution is slow. Progress in artificial intelligence is fast. Only a handful of genes separate *Homo sapiens* from our hominid ancestors on the African savannah. Among our 23,000-odd protein-coding genes, variance in single nucleotide polymorphisms ("SNPs") accounts for just a small percentage of phenotypic variance in intelligence as measured

by what we call IQ tests. True, the tempo of human evolution is about to accelerate. CRISPR-Cas9 genome-editing is a game changer. As the reproductive revolution of "designer babies" gathers pace, prospective parents will pre-select alleles and allelic combinations for a new child *in anticipation of* their behavioural effects - a novel kind of selection pressure to replace the "blind" genetic roulette of natural selection. In time, routine embryo screening via preimplantation genetic diagnosis will be complemented by gene therapy, genetic enhancement and then true designer zygotes. In consequence, life on Earth will also become progressively happier as the hedonic treadmill is recalibrated. In the new reproductive era, hedonic set-points and intelligence alike will be ratcheted upwards in virtue of selection pressure. For what parent-to-be wants to give birth to a low-status depressive "loser"? Future parents can enjoy raising a normal transhuman supergenius who grows up to be faster than Usain Bolt, more beautiful than Marilyn Monroe, more saintly than Nelson Mandela, more creative than Shakespeare - and smarter than Einstein.

Even so, the accelerating growth of germline engineering will be a *comparatively* slow process. In this scenario, sentient biological machines will design cognitively self-amplifying biological machines who will design cognitively self-amplifying biological machines. Greater-than-human biological intelligence will transform itself into posthuman superintelligence. Cumulative gains in intellectual capacity and subjective well-being across the generations will play out over hundreds and perhaps thousands of years - a momentous discontinuity, for sure, and a twinkle in the eye of eternity; but not a BioSingularity.

2 Biohacking Your Personal Genome

Yet **germline** engineering is only one strand of the genomics revolution. Indeed after humans master the ageing process, the extent to which traditional germlines or human generations will persist in the post-ageing world is obscure. Focus on the human germline ignores the slow-burning but then explosive growth of **somatic** gene enhancement in prospect. The CRISPR genome-editing revolution is accelerating. Later this century, innovative gene *therapies* will be succeeded by gene *enhancement* technologies - a value-laden dichotomy that reflects our impoverished human aspirations. Starting with individual genes, then clusters of genes, and eventually hundreds of genes and alternative splice variants, a host of recursively self-improving organic robots ("biohackers") will modify their genetic source code and modes of sentience: their senses, their moods, their motivation, their cognitive apparatus, their world-simulations and their default state of consciousness.

As the era of open-source genetics unfolds, tomorrow's biohackers will add, delete, edit and customise their own legacy code in a positive feedback loop of cognitive enhancement. Computer-aided genetic engineering will empower biological humans, transhumans and then posthumans to synthesise and insert new genes, variant alleles and even designer chromosomes - reweaving the multiple layers of regulation of our DNA to suit their wishes and dreams rather than the inclusive fitness of their genes in the ancestral environment. Collaborating and competing, next-generation biohackers will use stem-cell technologies to expand their minds, literally, via controlled neurogenesis. Freed from the constraints of the human birth canal, biohackers may re-sculpt the prison-like skull of *Homo sapiens* to accommodate a larger mind/brain, which can initiate recursive self-expansion in turn. Six crumpled layers of neocortex fed by today's miserly reward

pathways aren't the upper bound of conscious mind, merely its seedbed. Each biological neuron and glial cell of your growing mind/brain can have its own dedicated artificial healthcare team, web-enabled nanobot support staff, and social network specialists; compare today's anonymous neural porridge. Transhuman minds will be augmented with neurochips, molecular nanotechnology, mind/computer interfaces and full-immersion virtual reality (VR) software. To achieve finer-grained control of cognition, mood and motivation, genetically enhanced transhumans will draw upon exquisitely tailored new designer drugs, nutraceuticals and cognitive enhancers - precision tools that make today's crude interventions seem the functional equivalent of glue-sniffing.

By way of comparison, early in the twenty-first century the scientific counterculture is customizing a bewildering array of designer drugs that outstrip the capacity of the authorities to regulate or comprehend. The bizarre psychoactive effects of such agents dramatically expand the evidential base that our theory of consciousness must explain. However, such drugs are short-acting. Their benefits, if any, aren't cumulative. By contrast, the ability genetically to hack one's own source code will unleash an exponential growth of genomic rewrites - not mere genetic tinkering but a comprehensive redesign of "human nature". Exponential growth starts out almost unnoticeably, and then explodes. Human bodies, cognition and ancestral modes of consciousness alike will be transformed. Post-humans will range across immense state-spaces of conscious mind hitherto impenetrable because access to their molecular biology depended on crossing gaps in the fitness landscape prohibited by natural selection. Intelligent agency can "leap across" such fitness gaps. What we'll be leaping into is currently for the most part unknown: an inherent risk of the empirical method. But mastery of our reward circuitry can guarantee such state-spaces of experience will be glorious beyond human imagination. For intelligent biohacking can make unpleasant experience physically impossible because its

molecular substrates are absent. Hedonically enhanced innervation of the neocortex can ensure a rich hedonic tone saturates whatever strange new modes of experience our altered neurochemistry discloses.

Pilot studies of radical genetic enhancement will be difficult. Randomised longitudinal trials of such interventions in long-lived humans would take decades. In fact, officially licensed, well-controlled prospective trials to test the safety and efficacy of genetic innovation will be hard if not impossible to conduct because all of us, apart from monozygotic twins, are genetically unique. Even monozygotic twins exhibit different epigenetic and gene expression profiles. Barring an ideological and political revolution, most formally drafted proposals for genetically-driven life-enhancement probably won't pass ethics committees or negotiate the maze of bureaucratic regulation. But that's the point of biohacking. By analogy today, if you're technically savvy, you don't want a large corporation controlling the operating system of your personal computer: you use open source software instead. Likewise, you don't want governments controlling your state of mind via drug laws. By the same token, tomorrow's biotech-savvy individualists won't want anyone restricting our right to customise and rewrite our own genetic source code in any way we choose.

Will central governments try to regulate personal genome editing? Most likely yes. How far they'll succeed is an open question. So too is the success of any centralised regulation of futuristic designer drugs or artificial intelligence. Another huge unknown is the likelihood of state-sponsored designer babies, human reproductive cloning, and autosomal gene enhancement programs; and their interplay with privately-funded initiatives. China, for instance, has a different historical memory from the West.

Will there initially be biohacking accidents? Personal tragedies? Most probably yes, until human mastery of the pleasure-pain axis is secure. By the end of next decade, every health-conscious citizen will be broadly familiar with the architecture of his or her personal genome: the cost of personal genotyping will be trivial, as will be the cost of DIY gene-manipulation kits. Let's say you decide to endow yourself with an extra copy of the N-methyl D-aspartate receptor subtype 2B (NR2B) receptor, a protein encoded by the GRIN2B gene. Possession of an extra NR2B subunit NMDA receptor is a crude but effective way to enhance your learning ability, at least if you're a transgenic mouse. Recall how Joe Tsien and his colleagues first gave mice extra copies of the NR2B receptor-encoding gene, then tweaked the regulation of those genes so that their activity would increase as the mice grew older. Unfortunately, it transpires that such brainy "Doogie mice" - and maybe brainy future humans endowed with an extra NR2B receptor gene - display greater pain-sensitivity too; certainly, NR2B receptor *blockade* reduces pain and learning ability alike. Being smart, perhaps you decide to counteract this heightened pain-sensitivity by inserting and then over-expressing a high pain-threshold, "low pain" allele of the SCN9A gene in your nociceptive neurons at the dorsal root ganglion and trigeminal ganglion. The SCN9A gene regulates pain-sensitivity; nonsense mutations abolish the capacity to feel pain at all. In common with taking polydrug cocktails, the factors to consider in making multiple gene modifications soon snowball; but you'll have heavy-duty computer software to help. Anyhow, the potential pitfalls and makeshift solutions illustrated in this hypothetical example could be multiplied in the face of a combinatorial explosion of possibilities on the horizon. Most risks - and opportunities - of genetic self-editing are presumably still unknown.

It is tempting to condemn such genetic self-experimentation as irresponsible, just as unlicensed drug self-experimentation is irresponsible. Would you want your teenage

daughter messing with her DNA? Perhaps we may anticipate the creation of a genetic counterpart of the Drug Enforcement Agency (DEA) to police the human genome and its transhuman successors. Yet it's worth bearing in mind how each act of sexual reproduction today is an unpoliced genetic experiment with unfathomable consequences too. Without such reckless genetic experimentation, none of us would exist. In a cruel Darwinian world, this argument admittedly cuts both ways.

Naively, genomic source-code self-editing will always be too difficult for anyone beyond a dedicated cognitive elite of recursively self-improving biohackers. Certainly there are strongly evolutionarily conserved "housekeeping" genes that archaic humans would be best advised to leave alone for the foreseeable future. Granny might do well to customize her Windows desktop rather than her personal genome - prior to her own computer-assisted enhancement, at any rate. Yet the Biointelligence Explosion won't depend on more than a small fraction of its participants mastering the functional equivalent of machine code - the three billion odd 'A's, 'C's, 'G's and 'T's of our DNA. For the open-source genetic revolution will be propelled by powerful suites of high-level gene-editing tools, insertion vector applications, nonviral gene-editing kits, and user-friendly interfaces. Clever computer modelling and "narrow" AI can assist the intrepid biohacker to become a recursively self-improving genomic innovator. Later this century, your smarter counterpart will have software tools to monitor and edit every gene, repressor, promoter and splice variant in every region of your genome: each layer of epigenetic regulation of your gene transcription machinery in every region of the brain. This intimate level of control won't involve just crude DNA methylation to turn genes off and crude histone acetylation to turn genes on. Personal self-invention will involve mastery and enhancement of the histone and micro-RNA codes to allow sophisticated fine-tuning of gene expression and repression across the brain. Even today, researchers are

exploring “nanochannel electroporation” (NEP) technologies that allow the mass-insertion of novel therapeutic genetic elements into our cells. Mechanical cell-loading systems will shortly be feasible that can inject up to 100,000 cells at a time. Before long, such technologies will seem primitive. Freewheeling genetic self-experimentation will be endemic as the DIY-Bio revolution unfolds. At present, crude and simple gene editing can be accomplished only via laborious genetic engineering techniques. Sophisticated authoring tools don't exist. In future, computer-aided genetic and epigenetic enhancement can become an integral part of your personal growth plan.

3 Will Humanity's Successors Also Be Our Descendants?

To contrast "biological" with "artificial" conceptions of posthuman superintelligence is convenient. The distinction may also prove simplistic. In essence, whereas genetic change in biological humanity has always been slow, the software run on serial, programmable digital computers is executed exponentially faster (*cf.* Moore's Law); it's copyable without limit; it runs on multiple substrates; and it can be cheaply and rapidly edited, tested and debugged. Extrapolating, Singularitarians like Ray Kurzweil and Eliezer Yudkowsky prophesy that human programmers will soon become redundant because autonomous AI run on digital computers will undergo accelerating cycles of self-improvement. In this kind of scenario, artificial, greater-than-human nonbiological intelligence will be rapidly succeeded by artificial posthuman superintelligence.

So we may distinguish two radically different conceptions of posthuman superintelligence: on one hand, our supersentient, cybernetically enhanced, genetically rewritten *biological* descendants, on the other, *non*biological superintelligence, either a

Kurzweilian ecosystem or singleton Artificial General Intelligence (AGI) as foretold by the Machine Intelligence Research Institute (MIRI). Such a divide doesn't reflect a clean contrast between "natural" and "artificial" intelligence, the biological and the nonbiological. This contrast may prove another false dichotomy. Transhuman biology will increasingly become synthetic biology as genetic enhancement plus cyborgisation proceeds apace. "Cyborgisation" is a barbarous term to describe an invisible and potentially life-enriching symbiosis of biological sentience with artificial intelligence. Thus "narrow-spectrum" digital superintelligence on web-enabled chips can be more-or-less seamlessly integrated into our genetically enhanced bodies and brains. Seemingly limitless formal knowledge can be delivered on tap to supersentient organic wetware, i.e. us. *Critically, transhumans can exploit what is misleadingly known as "narrow" or "weak" AI to enhance our own code in a positive feedback loop of mutual enhancement* - first plugging in data and running multiple computer simulations, then tweaking and re-simulating once more. In short, biological humanity won't just be the spectator and passive consumer of the intelligence explosion, but its driving force. The smarter our AI, the greater our opportunities for reciprocal improvement. Multiple "hard" and "soft" take-off scenarios to posthuman superintelligence can be outlined for recursively self-improving organic robots, not just nonbiological AI. Thus for serious biohacking later this century, artificial quantum supercomputers may be deployed rather than today's classical toys to test-run multiple genetic interventions, accelerating the tempo of our recursive self-improvement. Quantum supercomputers exploit quantum coherence to do googols of computations all at once. So the accelerating growth of human/computer synergies means it's premature to suppose biological evolution will be superseded by technological evolution, let alone a "robot rebellion" as the parasite swallows its host. As the human

era comes to a close, the fate of biological (post)humanity is more likely to be symbiosis with AI followed by metamorphosis, not simple replacement.

Despite this witches' brew of new technologies, a conceptual gulf remains in the futurist community between those who imagine human destiny, if any, lies in digital computers running programs with (hypothetical) artificial consciousness; and in contrast radical bioconservatives who believe that our posthuman successors will also be our supersentient descendants at their organic neural networked core - not the digital zombies of symbolic AI run on classical serial computers or their souped-up multiprocessor cousins. For one metric of progress in AI remains stubbornly unchanged: despite the exponential growth of transistors on a microchip, the soaring clock speed of microprocessors, the growth in computing power measured in MIPS, the dramatically falling costs of manufacturing transistors and the plunging price of dynamic RAM (etc), any chart plotting the growth rate in digital sentience shows neither exponential growth, nor linear growth, but no progress whatsoever. As far as we can tell, digital computers are still zombies. Our machines are becoming autistically intelligent, but not supersentient - nor even conscious. On some fairly modest philosophical assumptions, digital computers were not subjects of experience in 1946 (*cf*. ENIAC); nor are they conscious subjects in 2012 (*cf*. "Watson"); nor do researchers know how any kind of sentience may be "programmed" in future. So what if anything does consciousness *do*? Is it computationally redundant? Pre-reflectively, we tend to have a "dimmer-switch" model of sentience: "primitive" animals have minimal awareness and "advanced" animals like human beings experience a proportionately more intense awareness. By analogy, most AI researchers assume that at a given threshold of complexity/intelligence/processing speed, consciousness will somehow "switch on", turn reflexive, and intensify too. The problem with the dimmer-switch model is that our most intense experiences, notably raw

agony or blind panic, are also the most phylogenetically ancient, whereas the most "advanced" modes (e.g. linguistic thought and the rich generative syntax that has helped one species to conquer the globe) are phenomenologically so thin as to be barely accessible to introspection. Something is seriously amiss with our entire conceptual framework.

So the structure of the remainder of this essay is as follows. I shall first discuss the risks and opportunities of building friendly biological superintelligence. Next I discuss the nature of *full-spectrum* superintelligence - and why consciousness is computationally fundamental to the past, present and future success of organic robots. Why couldn't recursively self-improving *zombies* modify their own genetic source code and bootstrap their way to full-spectrum superintelligence, i.e. a zombie biointelligence explosion? Finally, and most speculatively, I shall discuss the future of sentience in the cosmos.

4 Can We Build Friendly Biological Superintelligence?

4.1 Risk-Benefit Analysis

Crudely speaking, evolution "designed" male human primates to be hunters/warriors. Evolution "designed" women to be attracted to powerful, competitive alpha males. Until humans rewrite our own hunter-gatherer source code, we shall continue to practise extreme violence against members of other species - and frequently against members of our own. A heritable (and conditionally activated) predisposition to unfriendliness shown towards members of other races and other species is currently hardwired even in "social" primates. Indeed, we have a (conditionally activated) predisposition to compete against, and harm, anyone who isn't a genetically identical twin. Compared to the obligate siblicide found in some bird species, human sibling rivalry isn't normally so overtly brutal. But conflict as well as self-interested cooperation is endemic to Darwinian life on Earth.

This grim observation isn't an argument for genetic determinism, or against gene-culture co-evolution, or to discount the decline of everyday violence with the spread of liberal humanitarianism - just a reminder of the omnipresence of immense risks so long as we're shot through with legacy malware. Attempting to conserve the genetic status quo in an era of weapons of mass destruction (WMD) poses unprecedented global catastrophic and existential risks. Indeed, the single biggest underlying threat to the future of sentient life within our cosmological horizon derives, not from asocial symbolic AI software in the basement turning rogue and going FOOM (a runaway computational explosion of recursive self-improvement), but from conserving human nature in its present guise. In the twentieth century, male humans killed over 100 million fellow humans and billions of non-human animals. This century's toll may well be higher. Mankind currently spends well over a trillion dollars each year on weapons designed to kill and maim other humans. The historical record suggests such weaponry won't all be beaten into ploughshares.

Strictly speaking, however, humanity is more likely to be wiped out by *idealists* than by misanthropes, death-cults or psychologically unstable dictators. Anti-natalist philosopher David Benatar's plea ("Better Never to Have Been") for human extinction via voluntary childlessness must fail if only by reason of selection pressure; but not everyone who shares Benatar's bleak diagnosis of life on Earth will be so supine. Unless we modify human nature, compassionate-minded negative utilitarians, with competence in bioweaponry, nanorobotics or artificial intelligence, for example, may quite conceivably take direct action. Echoing Moore's law, Eliezer Yudkowsky warns that "Every eighteen months, the minimum IQ necessary to destroy the world drops by one point". Although suffering and existential risk might seem separate issues, they are intimately connected. Not everyone loves life so much they wish to preserve it. Indeed the extinction of

Darwinian life is what many transhumanists are aiming for - just not framed in such apocalyptic and provocative language. For just as we educate small children so they can mature into fully-fledged adults, biological humanity may aspire to grow up, too, with the consequence that - in common with small children - archaic humans become extinct.

4.2 Technologies Of Biofriendliness.

Empathogens?

How do you disarm a potentially hostile organic robot - despite your almost limitless ignorance of his source code? Provide him with a good education, civics lessons and complicated rule-governed ethics courses? Or give him a tablet of MDMA ("Ecstasy") and get smothered with hugs?

MDMA is short-acting. The "penicillin of the soul" is potentially neurotoxic to serotonergic neurons. In theory, however, lifelong use of safe and sustainable empathogens would be a passport to worldwide biofriendliness. MDMA releases a potent cocktail of oxytocin, serotonin and dopamine into the user's synapses, thereby inducing a sense of "I love the world and the world loves me". There's no technical reason why MDMA's acute pharmacodynamic effects can't be replicated indefinitely, shorn of its neurotoxicity. Designer "hug drugs" can potentially turn manly men into intelligent bonobos, more akin to the "hippie chimp" *Pan paniscus* than his less peaceable cousin *Pan troglodytes*. Violence would become unthinkable. Yet is this sort of proposal politically credible? "Morality pills" and other pharmacological solutions to human unfriendliness are both personally unsatisfactory and sociologically implausible. Do we really want to drug each other up from early childhood? Moreover, life would be immeasurably safer if our fellow humans weren't genetically predisposed to unfriendly behaviour in the first instance.

But how can this friendly predisposition be guaranteed?

Friendliness can't realistically be hand-coded by tweaking the connections and weight strengths of our neural networks.

Nor can robust friendliness in advanced biological intelligence be captured by a bunch of explicit logical rules and smart algorithms, as in the paradigm of symbolic AI.

4.3 Mass Oxytocination?

Amplified "trust hormone" might create the biological underpinnings of world-wide peace and love if negative feedback control of oxytocin release can be circumvented. Oxytocin is functionally antagonised by testosterone in the male brain. Yet oxytocin enhancers have pitfalls too. Enriched oxytocin function leaves one vulnerable to exploitation by the unenhanced. Can we really envisage a cross-cultural global consensus for mass-medication? When? Optional or mandatory? And what might be the wider ramifications of a "high oxytocin, low testosterone" civilisation? Less male propensity to violent territorial aggression, for sure; but disproportionate intellectual progress in physics, mathematics and computer science to date has been driven by the hyper-systematising cognitive style of "extreme male" brains. Also, enriched oxytocin function can indirectly even promote *un*friendliness to "out-groups" in consequence of promoting in-group bonding. So as well as oxytocin enrichment, global security demands a more inclusive, impartial, intellectually sophisticated conception of "us" that embraces all sentient beings - the expression of a hyper-developed capacity for empathetic understanding combined with a hyper-developed capacity for rational systematisation. Hence the imperative need for full-spectrum superintelligence.

4.4 Mirror-Touch Synaesthesia?

A truly long-term solution to unfriendly biological intelligence might be collectively to engineer ourselves with the functional generalisation of mirror-touch synaesthesia. On

seeing you cut and hurt yourself, a mirror-touch synaesthete is liable to feel a stab of pain as acutely as you do. Conversely, your expressions of pleasure elicit a no less joyful response. Thus mirror-touch synaesthesia is a hyper-empathising condition that makes deliberate unfriendliness, in effect, biologically impossible in virtue of cognitively enriching our capacity to represent each other's first-person perspectives. The existence of mirror-touch synaesthesia is a tantalising hint at the God-like representational capacities of a full-spectrum superintelligence. This so-called "disorder" is uncommon in humans.

4.5 Timescales

The biggest problem with all these proposals, and other theoretical biological solutions to human unfriendliness, is timescale. Billions of human and non-human animals will have been killed and abused before they could ever come to pass. Cataclysmic wars may be fought in the meantime with nuclear, biological and chemical weapons harnessed to "narrow" AI. Our circle of empathy expands only slowly and fitfully. For the most part, religious believers and traditional-minded bioconservatives won't seek biological enhancement/remediation for themselves or their children. So messy democratic efforts at "political" compromise are probably unavoidable for centuries to come. For sure, idealists can dream up utopian schemes to mitigate the risk of violent conflict until the "better angels of our nature" can triumph, e.g. the election of a risk-averse all-female political class to replace legacy warrior males. Such schemes tend to founder on the rock of sociological plausibility. Innumerable sentient beings are bound to suffer and die in consequence.

4.6 Does Full-Spectrum Superintelligence Entail Benevolence?

The God-like perspective-taking faculty of a full-spectrum superintelligence doesn't entail distinctively *human*-friendliness any more than a God-like superintelligence could promote distinctively Aryan-friendliness. Indeed it's unclear how benevolent superintelligence could want omnivorous killer apes in our current guise to walk the Earth in any shape or form. But is there any connection at all between benevolence and intelligence? Pre-reflectively, benevolence and intelligence are orthogonal concepts. There's nothing obviously incoherent about a malevolent God or a malevolent - or at least a callously indifferent - Superintelligence. Thus a sceptic might argue that there is no link whatsoever between benevolence - on the face of it a mere personality variable - and enhanced intellect. After all, some sociopaths score highly on our [autistic, mind-blind] IQ tests. Sociopaths know that their victims suffer. They just don't care.

However, what's critical in evaluating cognitive ability is a *criterion of representational adequacy*. Representation is not an all-or-nothing phenomenon; it varies in functional degree. More specifically here, the cognitive capacity to represent the *formal* properties of mind differs from the cognitive capacity to represent the *subjective* properties of mind. Thus a notional zombie Hyper-Autist robot running a symbolic AI program on an ultrapowerful digital computer with a classical von Neumann architecture may be beneficent or maleficent in its behaviour toward sentient beings. By its very nature, it can't know or care. Most starkly, the zombie Hyper-Autist might be programmed to convert the world's matter and energy into either heavenly "utilitronium" or diabolical "dolorium" without the slightest insight into the significance of what it was doing. This kind of scenario is at least a notional risk of creating insentient Hyper-Autists endowed with mere formal utility functions rather than hyper-sentient full-spectrum superintelligence. By contrast, full-spectrum superintelligence *does* care in virtue of its full-spectrum representational capacities - a bias-free generalisation of the superior

perspective-taking, "mind-reading" capabilities that enabled humans to become the cognitively dominant species on the planet. Full-spectrum superintelligence, if equipped with the posthuman cognitive generalisation of mirror-touch synaesthesia, understands your thoughts, your feelings and your egocentric perspective better than you do yourself.

Could there arise "evil" mirror-touch synaesthetes? In one sense, no. You can't go around wantonly hurting other sentient beings if you feel their pain as your own. Full-spectrum intelligence is friendly intelligence. But in another sense yes, insofar as primitive mirror-touch synaesthetes are prey to species-specific cognitive limitations that prevent them acting rationally to maximise the well-being of all sentience. Full-spectrum superintelligences would lack those computational limitations in virtue of their full cognitive competence in understanding both the subjective and the formal properties of mind. Perhaps full-spectrum superintelligences might optimise your matter and energy into a blissful smart angel; but they couldn't wantonly hurt you, whether by neglect or design.

More practically today, a cognitively superior analogue of natural mirror-touch synaesthesia should soon be feasible with reciprocal neuroscanning technology - a kind of naturalised telepathy. At first blush, mutual telepathic understanding sounds a panacea for ignorance and egotism alike. An exponential growth of shared telepathic understanding might safeguard against global catastrophe born of mutual incomprehension and WMD. As the poet Henry Wadsworth Longfellow observed, "If we could read the secret history of our enemies, we should find in each life sorrow and suffering enough to disarm all hostility." Maybe so. The problem here, as advocates of Radical Honesty soon discover, is that many Darwinian thoughts scarcely promote friendliness if shared: they are often ill-natured, unedifying and unsuitable for public consumption. Thus unless perpetually "loved-up" on MDMA or its long-acting equivalents,

most of us would find mutual mind-reading a traumatic ordeal. Human society and most personal relationships would collapse in acrimony rather than blossom. Either way, our human incapacity fully to understand the first-person point of view of other sentient beings isn't just a moral failing or a personality variable; it's an *epistemic* limitation, an intellectual failure to grasp an objective feature of the natural world. Even "normal" people share with sociopaths this fitness-enhancing cognitive deficit. By posthuman criteria, perhaps we're all quasi-sociopaths. The egocentric delusion (i.e. that the world centres on one's existence) is genetically adaptive and strongly selected for over hundreds of millions of years. Fortunately, it's a cognitive failing amenable to technical fixes and eventually a cure: full-spectrum superintelligence. The devil is in the details, or rather, the genetic source code.

5 A Biotechnological Singularity?

Yet does this positive feedback loop of reciprocal enhancement amount to a Singularity in anything more than a metaphorical sense? The risk of talking portentously about "The Singularity" isn't of being wrong: it's of being "not even wrong" - of reifying one's ignorance and elevating it to the status of an ill-defined apocalyptic event. Already multiple senses of "The Singularity" proliferate in popular culture. Does taking LSD induce a Consciousness Singularity? How about the abrupt and momentous discontinuity in one's conception of reality entailed by waking from a dream? Or the birth of language? Or the Industrial Revolution? So is *Bio*technological Singularity, or "BioSingularity" for short, any more rigorously defined than "Technological Singularity"?

Metaphorically, perhaps, the impending biointelligence explosion represents an intellectual "event horizon" beyond which archaic humans cannot model or understand the future. Events beyond the BioSingularity will be stranger than science fiction: too

weird for unenhanced human minds - or the algorithms of a zombie super-Asperger - to predict or understand. In the popular sense of "event horizon", maybe the term is apt too, though the metaphor is still potentially misleading. Thus theoretical physics tells us that one could pass through the event horizon of a non-rotating supermassive black hole and not notice any subjective change in consciousness - even though one's signals would now be inaccessible to an external observer. The BioSingularity will feel different in ways a human conceptual scheme can't express. But what is the empirical content of this claim?

6 What Is Full-Spectrum Superintelligence?

"[g is] ostensibly some innate scalar brain force...[However] ability is a folk concept and not amenable to scientific analysis."

Jon Marks (Dept Anthropology, Yale University), 1995, *Nature*, 9 xi, 143-144.

"Our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different."

(William James)

6.1 Intelligence

"Intelligence" is a folk concept. The phenomenon is not well-defined - or rather any attempt to do so amounts to a stipulative definition that doesn't "carve Nature at the joints". The Cattell-Horn-Carroll (CHC) psychometric theory of human cognitive abilities is probably most popular in academia and the IQ testing community. But the Howard Gardner multiple intelligences model, for example, differentiates "intelligence" into various spatial, linguistic, bodily-kinaesthetic, musical, interpersonal, intrapersonal,

naturalistic and existential intelligence rather than a single general ability ("g"). Who's right? As it stands, "g" is just a statistical artefact of our culture-bound IQ tests. If general intelligence were indeed akin to an innate scalar brain force, as some advocates of "g" believe, or if intelligence can best be modelled by the paradigm of symbolic AI, then the exponential growth of digital computer processing power might indeed entail an exponential growth in intelligence too - perhaps leading to some kind of Super-Watson. Other facets of intelligence, however, resist enhancement by mere acceleration of raw processing power.

One constraint is that a theory of general intelligence should be race-, species-, and culture-neutral. Likewise, an impartial conception of intelligence should embrace all possible state-spaces of consciousness: prehuman, human, transhuman and posthuman.

The non-exhaustive set of criteria below doesn't pretend to be anything other than provisional. They are amplified in the sections to follow.

Full-Spectrum Superintelligence entails:

1. the capacity to solve the Binding Problem, i.e. to generate phenomenally unified entities from widely distributed computational processes; and run cross-modally matched, data-driven world-simulations of the mind-independent environment. (*cf*. naive realist theories of "perception" versus the world-simulation or "Matrix" paradigm. Compare disorders of binding, e.g. simultanagnosia (an inability to perceive the visual field as a whole), cerebral akinetopsia ("motion blindness"), etc. In the absence of a data-driven, almost real-time simulation of the environment, intelligent agency is impossible.)
2. a self or some non-arbitrary functional equivalent of a person to which intelligence can be ascribed.

(*cf*. dissociative identity disorder (DID or "multiple personality disorder"), or florid schizophrenia, or your personal computer: in the absence of at least a fleetingly unitary self, what philosophers call "synchronic identity", there is no entity that is intelligent, just an aggregate of discrete algorithms and an operating system.)

3. a "mind-reading" or perspective-taking faculty; higher-order intentionality (e.g. "he believes that she hopes that they fear that he wants...", etc): social intelligence.

   The intellectual success of the most cognitively successful species on the planet rests, not just on the recursive syntax of human language, but also on our unsurpassed "mind-reading" prowess, an ability to simulate the perspective of other unitary minds: the "Machiavellian Ape" hypothesis. Any ecologically valid intelligence test designed for a species of social animal must incorporate social cognition and the capacity for co-operative problem-solving. So must any test of empathetic superintelligence.

4. a metric to distinguish the important from the trivial.

   (our theory of significance should be explicit rather than implicit, as in contemporary IQ tests. What distinguishes, say, mere calendrical prodigies and other "savant syndromes" from, say, a Grigori Perelman who proved the Poincaré conjecture? Intelligence entails understanding what does - and doesn't - matter. What matters is of course hugely contentious.)

5. a capacity to navigate, reason logically about, and solve problems in multiple state-spaces of consciousness [e.g. dreaming states (*cf.* lucid dreaming), waking consciousness, echolocatory competence, visual discrimination, synaesthesia in all its existing and potential guises, humour, introspection, the different realms of psychedelia (*cf*. salvia space, "the K-hole" etc)] including *realms of experience not*

yet co-opted by either natural selection or posthuman design for tracking features of the mind-independent world. Full-Spectrum Superintelligence will entail cross-domain goal-optimising ability in all possible state-spaces of consciousness.

and finally

6. "Autistic", pattern-matching, rule-following, mathematico-linguistic intelligence, i.e. the standard, mind-blind cognitive tool-kit scored by existing IQ tests. High-functioning "autistic" intelligence is indispensable to higher mathematics, computer science and the natural sciences. High-functioning autistic intelligence is necessary - but not sufficient - for a civilisation capable of advanced technology that can cure ageing and disease, systematically phase out the biology of suffering, and take us to the stars. And for programming artificial intelligence.

We may then ask which facets of full-spectrum superintelligence will be accelerated by the exponential growth of digital computer processing power? Number six, clearly, as decades of post-ENIAC progress in computer science attest. But what about numbers one-to-five? Here the picture is murkier.

6.2 The Bedrock Of Intelligence:

World-Simulation ("Perception")

Consider criterion number one, world-simulating prowess, or what we misleadingly term "perception". The philosopher Bertrand Russell once aptly remarked that one never sees anything but the inside of one's own head. In contrast to such inferential realism, common sense perceptual direct realism offers all the advantages of theft over honest toil - and it's computationally useless for the purposes either of building artificial general intelligence or understanding its biological counterparts. For the bedrock of intelligent agency is the capacity of an embodied agent computationally to simulate dynamic

objects, properties and events in the mind-independent environment. The evolutionary success of organic robots over the past *c*. 540 million years has been driven by our capacity to run data-driven egocentric world-simulations - what the naive realist, innocent of modern neuroscience or post-Everett quantum mechanics, calls simply perceiving one's physical surroundings. Unlike classical digital computers, organic neurocomputers can simultaneously "bind" multiple features (edges, colours, motion, etc) distributively processed across the brain into unitary phenomenal objects embedded in unitary spatio-temporal world-simulations apprehended by a momentarily unitary self: what Kant calls "the transcendental unity of apperception". These simulations run in (almost) real time; the time-lag in our world-simulations is barely more than a few dozen milliseconds. Such blistering speed of construction and execution is adaptive and often life-saving in a fast-changing external environment. Recapitulating evolutionary history, pre-linguistic human infants must first train up their neural networks to bind the multiple features of dynamic objects and run unitary world-simulations before they can socially learn *second-order* representation and then *third-order* representation, i.e. language followed later in childhood by meta-language.

Occasionally, object binding and/or the unity of consciousness partially breaks down in mature adults who suffer a neurological accident. The results can be cognitively devastating (*cf*. akinetopsia or "motion blindness"; and simultanagnosia, an inability to apprehend more than a single object at a time, etc). Yet normally our simulations of fitness-relevant patterns in the mind-independent local environment feel seamless. Our simulations each appear simply as "the world"; we just don't notice or explicitly represent the gaps. Neurons, (mis)construed as classical processors, are pitifully slow, with spiking frequencies barely up to 200 per second. By contrast, silicon (etc) processors are ostensibly millions of times faster. Yet the notion that nonbiological computers are faster

than sentient neurocomputers is a philosophical assumption, not an empirical discovery. Here the assumption will be challenged. Unlike the CPUs of classical robots, an organic mind/brain delivers dynamic unitary phenomenal objects and unitary world-simulations with a "refresh rate" of many billions per second (*cf*. the persistence of vision as experienced watching a movie run at a mere 30 frames per second). These cross-modally matched simulations take the guise of what passes as the macroscopic world: a spectacular egocentric simulation run by the vertebrate CNS that taps into the world's fundamental quantum substrate.

We should pause here. This is *not* a mainstream view. Most AI researchers regard stories of a non-classical mechanism underlying the phenomenal unity of biological minds as idiosyncratic at best. In fact, no scientific consensus exists on the molecular underpinnings of the unity of consciousness, nor on how such unity is even physically possible. By analogy, 1.3 billion skull-bound Chinese minds can never be a single subject of experience, irrespective of their interconnections. How could waking or dreaming communities of membrane-bound classical neurons - even *micro*conscious classical neurons - be any different? If materialism is true, conscious mind should be *impossible*. Yet any explanation of phenomenal object binding, the unity of perception, or the phenomenal unity of the self that invokes quantum coherence as here is controversial. One reason it's controversial is that the delocalisation involved in quantum coherence is exceedingly short-lived in an environment as warm and noisy as a macroscopic brain - supposedly too short-lived to do computationally useful work. Physicist Max Tegmark estimates that thermally-induced decoherence destroys any macroscopic coherence of brain states within 10^{-13} second, an unimaginably long time in natural Planck units but an unimaginably short time by everyday human intuitions. Perhaps it would be wiser just to acknowledge these phenomena are unexplained mysteries within a conventional

materialist framework - as mysterious as the existence of consciousness itself. But if we're speculating about the imminent end of the human era, shoving the mystery under the rug isn't really an option. For the different strands of the Singularity movement share a common presupposition. This presupposition is that our complete ignorance within a materialist conceptual scheme of why consciousness exists (the "Hard Problem"), and of even the ghost of a solution to the Binding Problem, doesn't matter for the purposes of building the seed of artificial posthuman superintelligence. Our ignorance supposedly doesn't matter either because consciousness and/or our quantum "substrate" are computationally irrelevant to cognition and the creation of nonbiological minds, or alternatively because the feasibility of "whole brain emulation" (WBE) will allow us to finesse our ignorance.

Unfortunately, we have no grounds for believing this suppressed premiss is true or that the properties of our quantum "substrate" are functionally irrelevant to full-spectrum superintelligence or its humble biological predecessors. Conscious minds are not substrate-neutral digital computers. Humans investigate problems of which digital computers are invincibly ignorant, not least the properties of consciousness itself. The Hard Problem of consciousness can't be quarantined from the rest of science and treated as a troublesome but self-contained anomaly: its mystery infects *everything* that we think we know about ourselves, our computers and the world. Either way, the conjecture that the phenomenal unity of perception is a manifestation of ultra-rapid sequences of irreducible quantum coherent states *isn't* a claim that the mind/brain is capable of detecting events in the mind-independent world on this kind of sub-picosecond timescale. Rather, the role of the local environment in shaping action-guiding experience in the awake mind/brain is here conjectured to be quantum state-*selection*. When we're awake, patterns of impulses from e.g. the optic nerve *select* which quantum-coherent frames are

generated by the mind/brain - in contrast to the autonomous world-simulations spontaneously generated by the dreaming brain. Other quantum mind theorists, most notably Roger Penrose and Stuart Hameroff, treat quantum minds as evolutionarily novel rather than phylogenetically ancient. They invoke a non-physical wave-function collapse and unwisely focus on e.g. the ability of mathematically-inclined brains to perform non-computable functions in higher mathematics, a feat for which selection pressure has presumably been non-existent. Yet the human capacity for sequential linguistic thought and formal logico-mathematical reasoning is a late evolutionary novelty executed by a slow, brittle, virtual machine running on top of its massively parallel quantum parent - a momentous evolutionary innovation whose neural mechanism is still unknown.

In contrast to the evolutionary novelty of serial linguistic thought, our ancient and immensely adaptive capacity to run unitary world-simulations, simultaneously populated by hundreds or more dynamic unitary objects, enables organic robots to solve the computational challenges of navigating a hostile environment that would leave the fastest classical supercomputer grinding away until Doomsday. Physical theory (*cf*. the Bekenstein bound) shows that informational resources as classically conceived are not just physical but finite and scarce: a maximum possible limit of 10^{120} bits set by the surface area of the entire accessible universe expressed in Planck units according to the Holographic principle. An infinite computing device like a universal Turing machine (UTM) is physically impossible. So invoking computational equivalence and asking whether a classical Turing machine can run a human-equivalent macroscopic world-simulation is akin to asking whether a classical Turing machine can factor 1,500 digit numbers in real-world time [i.e. no]. No doubt resourceful human and transhuman programmers will exploit all manner of kludges, smart workarounds and "brute-force" algorithms to try and defeat the Binding Problem in AI. How will they fare? Compare clod-hopping AlphaDog

with the sophisticated functionality of the sesame-seed sized brain of a bumblebee. Brute-force algorithms suffer from an exponentially growing search space that soon defeats any classical computational device in open-field contexts. As witnessed by our seemingly effortless world-simulations, organic minds are ultrafast; classical computers are slow. Serial *thinking* is slower still; but that's not what conscious biological minds are good at. On this conjecture, "substrate-independent" phenomenal world-simulations are impossible for the same reason that "substrate-independent" chemical valence structure is impossible. We're simply begging the question of what's functionally (ir)relevant. Ultimately, Reality has only a single "program-resistant" ontological level even though it's amenable to description at different levels of computational abstraction; and the nature of this program-resistant level as disclosed by the subjective properties of one's mind (Lockwood 1989) is utterly at variance with what naive materialist metaphysics would suppose. *If* our phenomenal world-simulating prowess turns out to be constitutionally tied to our quantum mechanical wetware, then substrate-neutral virtual machines (VMs, i.e. software implementations of a digital computer that execute programs like a physical machine) will never be able to support "virtual" qualia or "virtual" unitary subjects of experience. This rules out sentient life "uploading" itself to digital nirvana. *Contra* Marvin Minsky ("The most difficult human skills to reverse engineer are those that are unconscious"), the most difficult skills for roboticists to engineer in artificial robots are actually intensely conscious: our colourful, noisy, tactile, sometimes hugely refractory virtual worlds.

Naively, for sure, real-time world-simulation doesn't sound too difficult. Hollywood robots do it all the time. Videogames become ever more photorealistic. Perhaps one imagines viewing some kind of inner TV screen, as in a Terminator movie or The Matrix. Yet the capacity of an awake or dreaming brain to generate unitary macroscopic world-

simulations can only superficially resemble a little man (a "homunculus") viewing its own private theatre - on pain of an infinite regress. For by what mechanism would the homunculus view this inner screen? Emulating the behaviour of even the very simplest sentient organic robots on a classical digital computer is a daunting task. *If* conscious biological minds are irreducibly quantum mechanical by their very nature, then reverse-engineering the brain to create digital human "mindfiles" and "roboclones" alike will prove impossible.

6.3 The Bedrock Of Superintelligence:

Hypersocial Cognition ("Mind-reading")

Will superintelligence be solipsistic or social? Overcoming a second obstacle to delivering human-level artificial general intelligence - let alone building a recursively self-improving super-AGI culminating in a technological Singularity - depends on finding a solution to the first challenge, i.e. real-time world-simulation. For the evolution of distinctively human intelligence, sitting on top of our evolutionarily ancient world-simulating prowess, has been driven by the interplay between our rich generative syntax and superior "mind-reading" skills: so-called Machiavellian intelligence. Machiavellian intelligence is an egocentric parody of God's-eye-view empathetic superintelligence. Critically for the prospects of building AGI, this real-time mind-modelling expertise is parasitic on the neural wetware to generate unitary first-order world-simulations - virtual worlds populated by the avatars of intentional agents whose different first-person perspectives can be partially and imperfectly understood by their simulator. Even articulate human subjects with autism spectrum disorder are prone to multiple language deficits because they struggle to understand the intentions - and higher-order intentionality - of neurotypical language users. Indeed, natural language is itself a pre-eminently social phenomenon: its criteria of application must first be socially learned. Not all humans

possess the cognitive capacity to acquire mind-reading skills and the cooperative problem-solving expertise that sets us apart from other social primates. Most notably, people with autism spectrum disorder don't just fail to understand other minds; autistic intelligence cannot begin to understand its own mind. Pure autistic intelligence has no conception of a self that can be improved, recursively or otherwise. Autists can't "read" their own minds. The inability of the autistic mind to take what Daniel Dennett calls the intentional stance parallels the inability of classical computers to understand the minds of intentional agents - or have insight into their own zombie status. Even with smart algorithms and ultra-powerful hardware, the ability of ultra-intelligent autists to predict the long-term behaviour of mindful organic robots by relying exclusively on the physical stance (i.e. solving the Schrödinger equation of the intentional agent in question) will be extremely limited. For a start, much collective human behaviour is chaotic in the technical sense, i.e. it shows extreme sensitivity to initial conditions that confounds long-term prediction by even the most powerful real-world supercomputer. But there's a worse problem: reflexivity. Predicting sociological phenomena differs essentially from predicting mindless physical phenomena. Even in a classical, causally deterministic universe, the behaviour of mindful, reflexively self-conscious agents is frequently unpredictable, even in principle, from *within* the world owing to so-called prediction paradoxes. When the very act of prediction causally interacts with the predicted event, then self-defeating or self-falsifying predictions are inevitable. Self-falsifying predictions are a mirror image of so-called self-fulfilling predictions. So in common with autistic "idiot savants", classical AI gone rogue will be vulnerable to the low cunning of Machiavellian apes and the high cunning of our transhuman descendants.

This argument (i.e. our capacity for unitary mind-simulation embedded in unitary world-simulation) for the cognitive primacy of biological general intelligence isn't decisive. For a

start, computer-aided Machiavellian humans can program robots with "narrow" AI - or perhaps "train up" the connections and weights of a subsymbolic connectionist architecture - for their own manipulative purposes. Humans underestimate the risks of zombie infestation at our peril. Given our profound ignorance of how conscious mind is even possible, it's probably safest to be agnostic over whether autonomous nonbiological robots will ever emulate human world-simulating or mind-reading capacity in most open-field contexts, despite the scepticism expressed here. Either way, the task of devising an ecologically valid measure of general intelligence that can reliably, predictively and economically discriminate between disparate life-forms is immensely challenging, not least because the intelligence test will express the value-judgements, and species- and culture-bound conceptual scheme, of the tester. Some biases are insidious and extraordinarily subtle: for example, the desire systematically to measure "intelligence" with mind-blind IQ tests is itself a quintessentially Asperger-ish trait. In consequence, social cognition is disregarded altogether. What we fancifully style "IQ tests" are designed by people with abnormally high AQs as well as self-defined high IQs. Thus many human conceptions of (super)intelligence resemble high-functioning autism spectrum disorder (ASD) rather than a hyper-empathetic God-like Super-Mind. For example, an AI that attempted systematically to maximise the cosmic abundance of paperclips would be recognisably autistic rather than incomprehensibly alien. Full-spectrum (super-)intelligence is certainly harder to design or quantify scientifically than mathematical puzzle-solving ability or performance in verbal memory-tests: "IQ". But that's because superhuman intelligence will be not just quantitatively different but also qualitatively alien from human intelligence. To misquote Robert McNamara, cognitive scientists need to stop making what is measurable important, and find ways to make the important measurable. An idealised full-spectrum superintelligence will indeed be capable

of an impartial "view from nowhere" or God's-eye-view of the multiverse, a mathematically complete Theory Of Everything - as does modern theoretical physics, in aspiration if not achievement. But in virtue of its God's-eye-view, full-spectrum superintelligence must also be hypersocial and super*sentient*: able to understand all possible first-person perspectives, the state-space of all possible minds in other Hubble volumes, other branches of the universal wavefunction (UWF) - and in other solar systems and galaxies if such beings exist within our cosmological horizon. Idealized at least, full-spectrum superintelligence will be able to understand and weigh the significance of all possible modes of experience *irrespective of whether they have hitherto been recruited for information-signalling purposes*. The latter is, I think, by far the biggest intellectual challenge we face as cognitive agents. The systematic investigation of alien types of consciousness intrinsic to varying patterns of matter and energy calls for a methodological and ontological revolution. Transhumanists talking of post-Singularity superintelligence are fond of hyperbole about "Level 5 Future Shock" etc; but it's been aptly said that if Elvis Presley were to land in a flying saucer on the White House lawn, it's as nothing in strangeness compared to your first DMT trip.

6.4 Ignoring The Elephant: Consciousness

Why Consciousness is Computationally Fundamental to the Past, Present and Future Success of Organic Robots

The pachyderm in the room in most discussions of (super)intelligence is consciousness - not just human reflective self-awareness, but the whole gamut of experience from symphonies to sunsets, agony to ecstasy: the phenomenal world of everyday experience. All one ever knows, except by inference, is the contents of one's own conscious mind: what philosophers call "qualia". Yet according to the ontology of our best story of the

world, namely physical science, *conscious* minds shouldn't exist at all, i.e. we should be zombies, insentient patterns of matter and energy indistinguishable from normal human beings but lacking conscious experience. Dutch computer scientist Edsger Dijkstra once remarked, "The question of whether a computer can think is no more interesting than the question of whether a submarine can swim." Yet the question of whether a programmable digital computer - or a subsymbolic connectionist system with a merely classical parallelism - could possess, *and think about*, qualia, "bound" perceptual objects, a phenomenal self, or the unitary phenomenal minds of sentient organic robots can't be dismissed so lightly. For if advanced nonbiological intelligence is to be smart enough comprehensively to understand, predict and manipulate the behaviour of enriched biological intelligence, then the AGI can't rely autistically on the "physical stance", i.e. to monitor the brains, scan the atoms and molecules, and then solve the Schrödinger equation of intentional agents like human beings. Such calculations would take longer than the age of the universe.

For sure, many forms of human action can be predicted, fallibly, on the basis of crude behavioural regularities and reinforcement learning. Within your world-simulation, you don't need a theory of mind or an understanding of quantum mechanics to predict that Fred will walk to the bus-stop again today. Likewise, powerful tools of statistical analysis run on digital supercomputers can predict, fallibly, many kinds of human collective behaviour, for example stock markets. Yet to surpass human and transhuman capacities in all significant fields, AGI must understand how intelligent biological robots can think about, talk about and manipulate the manifold varieties of consciousness that make up their virtual worlds. Some investigators of consciousness even dedicate their lives to that end; what might a notional insentient AGI suppose we're doing? There is no evidence that serial digital computers have the capacity to do anything of the kind - or could ever

be programmed to do so. Digital computers don't know anything about conscious minds, unitary persons, the nature of phenomenal pleasure and pain, or the Problem of Other Minds; it's not even "all dark inside". The challenge for a *conscious* mind posed by understanding itself "from the inside" pales into insignificance compared to the challenge for a nonconscious system of understanding a conscious mind "from the outside". Nor within the constraints of a materialist ontology have we the slightest clue how the purely classical parallelism of a subsymbolic, "neurally inspired" connectionist architecture could turn water into wine and generate *unitary* subjects of experience to fill the gap. For even if we conjecture in the spirit of Strawsonian physicalism - the only scientifically literate form of panpsychism - that the fundamental stuff of the world, the mysterious "fire in the equations", is fields of microqualia, this bold ontological conjecture doesn't, by itself, explain why biological robots aren't zombies. This is because structured aggregates of classically conceived "mind-dust" aren't the same as a unitary phenomenal subject of experience who apprehends "bound" spatio-temporal objects in a dynamic world-simulation. Without phenomenal object binding and the unity of perception, we are faced with the spectre of what philosophers call "mereological nihilism". Mereological nihilism, also known as "compositional nihilism", is the position that composite objects with proper parts do not exist: strictly speaking, only basic building blocks without parts have more than fictional existence. Unlike the fleetingly unitary phenomenal minds of biological robots, a classical digital computer and the programs it runs lacks ontological integrity: it's just an assemblage of algorithms. In other words, a classical digital computer has no self to understand or a mind recursively to improve, exponentially or otherwise. Talk about artificial "intelligence" exploding is just an anthropomorphic projection on our part.

So how do biological brains solve the binding problem and become persons? In short, we don't know. Vitalism is clearly a lost cause. Most AI researchers would probably dismiss -

or at least discount as wildly speculative - any story of the kind mooted here involving macroscopic quantum coherence grounded in an ontology of physicalistic panpsychism. The conjecture should be experimentally falsifiable with the tools of next-generation molecular matter-wave interferometry. But in the absence of any story at all, we are left with a theoretical vacuum and a faith that natural science - or the exponential growth of digital computer processing power culminating in a Technological Singularity - will one day deliver an answer. Evolutionary biologist Theodosius Dobzhansky famously observed how "Nothing in Biology Makes Sense Except in the Light of Evolution". In the same vein, nothing in the future of intelligent life in the universe makes sense except in the light of a solution to the Hard Problem of Consciousness and the closure of Levine's Explanatory Gap. Consciousness is the only reason anything matters at all; and it's the only reason why unitary subjects of experience can ask these questions; and yet materialist orthodoxy has no idea how or why the phenomenon exists. Unfortunately, the Hard Problem won't be solved by building more advanced digital zombies who can tell mystified conscious minds the answer.

More practically for now, perhaps the greatest cognitive challenge of the millennium and beyond is deciphering and systematically manipulating the "neural correlates of consciousness" (NCC). Neuroscientists use this expression in default of any deeper explanation of our myriad qualia. How and why does experimentally stimulating via microelectrodes one cluster of nerve cells in the neocortex yield the experience of phenomenal colour; stimulating a superficially type of nerve cell induces a musical jingle; stimulating another with a slightly different gene-expression profile triggers a sense of everything being hysterically funny; stimulating another induces a hallucination of your mother; and stimulating another induces the experience of an archangel, say, in front of your body-image? In each case, the molecular variation in neuronal cell architecture is

ostensibly *trivial*; the difference in subjective experience is profound. On a mind/brain identity theory, such experiential states are an intrinsic property of some configurations of matter and energy. How and why this is so is incomprehensible on an orthodox materialist ontology. Yet empirically, microelectrodes, dreams and hallucinogenic drugs elicit these experiences regardless of any information-signalling role such experiences typically play in the "normal" awake mind/brain. Orthodox materialism and classical information-based ontologies alike do not merely lack any explanation for why consciousness and our countless varieties of qualia exist. They lack any story of how our qualia could have the causal efficacy to allow us to allude to - and in some cases volubly expatiate on - their existence. Thus mapping the neural correlates of consciousness is not amenable to formal computational methods: digital zombies don't have any qualia, or at least any "bound" macroqualia, that could be mapped, nor a unitary phenomenal self that could do the mapping.

Note this claim for the cognitive primacy of biological sentience isn't a denial of the Church-Turing thesis that given *infinite* time and *infinite* memory any Turing-universal system can formally simulate the behaviour of any conceivable process that can be digitized. Indeed, (very) fancifully, if the multiverse were being run on a cosmic supercomputer, speeding up its notional execution a million times would presumably speed us up a million times too. But that's not the issue here. Rather the claim is that nonbiological AI run on real-world digital computers cannot tackle the truly hard and momentous cognitive challenge of investigating first-person states of egocentric virtual worlds - or understand why some first-person states, e.g. agony or bliss, are intrinsically important, and cause unitary subjects of experience, persons, to act the way we do.

At least in common usage, "intelligence" refers to an agent's ability to achieve goals in a wide range of environments. What we call greater-than-human intelligence or

Superintelligence presumably involves the design of qualitatively *new* kinds of intelligence never seen before. Hence the growth of artificial intelligence and symbolic AI, together with subsymbolic (allegedly) brain-inspired connectionist architectures and soon artificial quantum computers. But contrary to received wisdom in AI research, sentient biological robots are making greater cognitive progress in discovering the potential for truly novel kinds of intelligence than the techniques of formal AI. We are doing so by synthesising and empirically investigating a galaxy of psychoactive designer drugs - experimentally opening up the possibility of radically new kinds of intelligence in different state-spaces of consciousness. For the most cognitively challenging environments don't lie in the stars but in organic mind/brains - the baffling subjective properties of quantum-coherent states of matter and energy - most of which aren't explicitly represented in our existing conceptual scheme.

6.5 Case Study: Visual Intelligence versus Echolocatory Intelligence:

What Is It Like To Be A Super-Intelligent Bat?

Let's consider the mental state-space of organisms whose virtual worlds are rooted in their dominant sense mode of echolocation. This example isn't mere science fiction. Unless post-Everett quantum mechanics is false, we're forced to assume that googols of quasi-classical branches of the universal wavefunction - the master formalism that exhaustively describes our multiverse - satisfy this condition. Indeed, their imperceptible interference effects must be present even in "our" world: strictly speaking, interference effects from branches that have decohered ("split") never wholly disappear; they just become vanishingly small. Anyhow, let's assume these echolocatory superminds have evolved opposable thumbs, a rich generative syntax and advanced science and technology. How are we to understand or measure this alien kind of (super)intelligence?

Rigging ourselves up with artificial biosonar apparatus and transducing incoming data into the familiar textures of sight or sound might seem a good start. But to understand the conceptual world of echolocatory superminds, we'd need to equip ourselves with neurons and neural networks neurophysiologically equivalent to smart chiropterans. If one subscribes to a coarse-grained functionalism about consciousness, then echolocatory experience would (somehow) emerge at some abstract computational level of description. The implementation details, or "meatware" as biological mind/brains are derisively called, are supposedly incidental or irrelevant. The functionally unique valence properties of the carbon atom, and likewise the functionally unique quantum mechanical properties of liquid water, are discounted or ignored. Thus according to the coarse-grained functionalist, silicon chips could replace biological neurons without loss of function or subjective identity. By contrast, the *micro*-functionalist, often branded a mere "carbon chauvinist", reckons that the different intracellular properties of biological neurons - with their different gene expression profiles, diverse primary, secondary, tertiary, and quaternary amino acid chain folding (etc) as described by quantum chemistry - are critical to the many and varied phenomenal properties such echolocatory neurons express. Who is right? We'll only ever know the answer by rigorous self-experimentation: a post-Galilean science of mind.

It's true that humans don't worry much about our ignorance of echolocatory experience, or our ignorance of echolocatory primitive terms, or our ignorance of possible conceptual schemes expressing echolocatory intelligence in echolocatory world-simulations. This is because we don't highly esteem bats. Humans don't share the same interests or purposes as our flying cousins, e.g. to attract desirable, high-fitness bats and rear reproductively successful baby bats. Alien virtual worlds based on biosonar don't seem especially significant to *Homo sapiens* except as an armchair philosophical puzzle.

Yet this assumption would be intellectually complacent. Worse, understanding what it's like to be a hyperintelligent bat mind is *comparatively* easy. For echolocatory experience has been recruited by natural selection to play an information-signalling role in a fellow species of mammal; and in principle a research community of language users could biologically engineer their bodies and minds to replicate bat-type experience and establish crude intersubjective agreement to discuss and conceptualise its nature. By contrast, the vast majority of experiential state-spaces remain untapped and unexplored. This task awaits full-spectrum superintelligence in the posthuman era.

In a more familiar vein, consider visual intelligence. How does one measure the visual intelligence of a congenitally blind person? Even with sophisticated technology that generates "inverted spectrograms" of the world to translate visual images into sound, the congenitally blind are invincibly ignorant of visual experience and the significance of visually-derived concepts. Just as a sighted idiot has greater visual intelligence than a blind super-rationalist sage, likewise psychedelics confer the ability to become (for the most part) babbling idiots about other state-spaces of consciousness - but babbling idiots whose insight is deeper than the drug-naive or the genetically unenhanced - or the digital zombies spawned by symbolic AI and its connectionist cousins.

The challenge here is that the vast majority of these alien state-spaces of consciousness latent in organised matter haven't been recruited by natural selection for information-tracking purposes. So "psychonauts" don't yet have the conceptual equipment to navigate these alien state-spaces of consciousness in even a pseudo-public language, let alone integrate them in any kind of overarching conceptual framework. Note the claim here *isn't* that taking e.g. ketamine, LSD, salvia, DMT and a dizzying proliferation of custom-designed psychoactive drugs is the royal route to wisdom. Or that ingesting such agents will give insight into deep mystical truths. On the contrary: it's precisely because

such realms of experience *haven't* previously been harnessed for information-processing purposes by evolution in "our" family of branches of the universal wavefunction that makes investigating their properties so cognitively challenging - currently beyond our conceptual resources to comprehend. After all, plants synthesise natural psychedelic compounds to scramble the minds of herbivores who might eat them, not to unlock mystic wisdom. Unfortunately, there is no "neutral" medium of thought impartially to appraise or perceptually cross-modally match all these other experiential state-spaces. One can't somehow stand outside one's own stream of consciousness to evaluate how the properties of the medium are infecting the notional propositional content of the language that one uses to describe it.

By way of illustration, compare drug-induced visual experience in a notional community of congenitally blind rationalists who lack the visual apparatus to transduce incident electromagnetic radiation of our familiar wavelengths. The lone mystical babbler who takes such a vision-inducing drug is convinced that [what we would call] visual experience is profoundly significant. And as visually intelligent folk, we know that he's right: visual experience is potentially hugely significant - to an extent which the blind mystical babbler can't possibly divine. But can the drug-taker convince his congenitally blind fellow tribesmen that his mystical visual experiences really matter in the absence of perceptual equipment that permits sensory discrimination? No, he just sounds psychotic. Or alternatively, he speaks lamely and vacuously of the "ineffable". The blind rationalists of his tribe are unimpressed.

The point of this fable is that we've scant reason to suppose that biologically re-engineered posthumans millennia hence will share the same state-spaces of consciousness, or the same primitive terms, or the same conceptual scheme, or the same type of virtual world that human beings now instantiate. Maybe all that will survive

the human era is a descendant of our mathematical formalism of physics, M-theory of whatever, in basement reality.

Of course such ignorance of other state-spaces of experience doesn't normally trouble us. Just as the congenitally blind don't grow up in darkness - a popular misconception - the drug-naive and genetically unenhanced don't go around with a sense of what we're missing. We notice teeming abundance, not gaping voids. Contemporary humans can draw upon terms like "blindness" and "deafness" to characterise the deficits of their handicapped conspecifics. From the perspective of full-spectrum superintelligence, what we really need is *millions* more of such "privative" terms, as linguists call them, to label the different state-spaces of experience of which genetically unenhanced humans are ignorant. In truth, there may very well be more than millions of such nameless state-spaces, each as incommensurable as, for instance, visual and auditory experience. We can't yet begin to quantify their number or construct any kind of crude taxonomy of their interrelationships.

Note the problem here isn't cognitive bias or a deficiency in logical reasoning. Rather a congenitally blind (etc) super-rationalist is constitutionally ignorant of visual experience, visual primitive terms, or a visually-based conceptual scheme. So s/he can't cite, e.g. Aumann's agreement theorem [claiming in essence that two cognitive agents acting rationally and with common knowledge of each other's beliefs cannot agree to disagree] or be a good Bayesian rationalist or whatever: these are incommensurable state-spaces of experience as closed to human minds as Picasso is to an earthworm. Moreover there is no reason to expect one realm, i.e. "ordinary waking consciousness", to be cognitively privileged relative to every other realm. "Ordinary waking consciousness" just happened to be genetically adaptive in the African savannah on Planet Earth. Just as humans are incorrigibly ignorant of minds grounded in echolocation - both echolocatory world-

simulations and echolocatory conceptual schemes - likewise we are invincibly ignorant of posthuman life while trapped within our existing genetic architecture of intelligence.

In order to understand the world - both its formal/mathematical *and* its subjective properties - sentient organic life must bootstrap its way to super-sentient full-spectrum superintelligence. Grown-up minds need tools to navigate all possible state-spaces of qualia, including all possible first-person perspectives, and map them - initially via the neural correlates of consciousness in our world-simulations - onto the formalism of mathematical physics. Empirical evidence suggests that the behaviour of the stuff of the world is exhaustively described by the formalism of physics. To the best of our knowledge, physics is causally closed and complete, at least within the energy range of the Standard Model. In other words, there is nothing to be found in the world - no "element of reality", as Einstein puts it - that isn't captured by the equations of physics and their solutions. This is a powerful formal constraint on our theory of consciousness. Yet our ultimate theory of the world must also close Levine's notorious "Explanatory Gap". Thus we must explain why consciousness exists at all ("The Hard Problem"); offer a rigorous derivation of our diverse textures of qualia from the field-theoretic formalism of physics; and explain how qualia combine ("The Binding Problem") in organic minds. These are powerful constraints on our ultimate theory too. How can they be reconciled with physicalism? Why aren't we zombies?

The hard-nosed sceptic will be unimpressed at such claims. How *significant* are these outlandish state-spaces of experience? And how are they computationally relevant to (super)intelligence? Sure, says the sceptic, reckless humans may take drugs, and experience wild, weird and wonderful states of mind. But so what? Such exotic states aren't objective in the sense of reliably tracking features of the mind-independent world.

Elucidation of their properties doesn't pose a well-defined problem that a notional universal algorithmic intelligence could solve.

Well, let's assume, provisionally at least, that all mental states are identical with physical states. If so, then all experience is an objective, spatio-temporally located feature of the world whose properties a unified natural science must explain. A cognitive agent can't be intelligent, let alone superintelligent, and yet be constitutionally ignorant of a fundamental feature of the world - not just ignorant, but completely incapable of gathering information about, exploring, or reasoning about its properties. Whatever else it may be, superintelligence can't be constitutionally *stupid*. What we need is a universal, species-neutral criterion of significance that can weed out the trivial from the important; and gauge the intelligence of different cognitive agents accordingly. Granted, such a criterion of significance might seem elusive to the antirealist about value (*cf*. Mackie 1991). Value nihilism treats any ascription of (in)significance as arbitrary. Or rather the value nihilist maintains that what we find significant simply reflects what was fitness-enhancing for our forebears in the ancestral environment of adaptation. Yet for reasons we simply don't understand, Nature discloses just such a universal touchstone of importance, namely the pleasure-pain axis: the world's inbuilt metric of significance and (dis)value. We're not zombies. First-person facts exist. Some of them matter urgently, e.g. I am in pain. Indeed, it's unclear if the expression "I'm in agony; but the agony doesn't matter" even makes cognitive sense. Built into the very nature of agony is the knowledge that its subjective raw awfulness matters a great deal - not instrumentally or derivatively, but by its very nature. If anyone - or indeed any notional super-AGI - supposes that your agony doesn't matter, then he/it hasn't adequately represented the first-person perspective in question.

So the existence of first-person facts is an objective feature of the world that any intelligent agent must comprehend. Digital computers and the symbolic AI code they execute can support formal utility functions. In some contexts, formally programmed utility functions can play a role functionally analogous to importance. But nothing intrinsically matters to a digital zombie. Without sentience, and more specifically without hedonic tone, nothing inherently matters. By contrast, extreme pain and extreme pleasure in any guise intrinsically matter intensely. Insofar as exotic state-spaces of experience are permeated with positive or negative hedonic tone, they matter too. In summary, "He jests at scars, that never felt a wound": scepticism about the self-intimating significance of this feature of the world is feasible only in its absence.

7 The Great Transition

7.1 The End Of Suffering

A defining feature of general intelligence is the capacity to achieve one's goals in a wide range of environments. All sentient biological agents are endowed with a pleasure-pain axis. All prefer occupying one end to the other. A pleasure-pain axis confers inherent significance on our lives: the opioid-dopamine neurotransmitter system extends from flatworms to humans. Our core behavioural and physiological responses to noxious and rewarding stimuli have been strongly conserved in our evolutionary lineage over hundreds of millions of years. Some researchers argue for *psychological hedonism*, the theory that all choice in sentient beings is motivated by a desire for pleasure or an aversion from suffering. When we choose to help others, this is because of the pleasure that we ourselves derive, directly or indirectly, from doing so. Pascal put it starkly: "All men seek happiness. This is without exception. Whatever different means they employ, they all tend to this end. The cause of some going to war, and of others avoiding it, is

the same desire in both, attended with different views. This is the motive of every action of every man, even of those who hang themselves." In practice, the hypothesis of psychological hedonism is plagued with anomalies, circularities and complications if understood as a universal principle of agency: the "pleasure principle" is simplistic as it stands. Yet the broad thrust of this almost embarrassingly commonplace idea may turn out to be central to understanding the future of life in the universe. If even a weak and exception-laden version of psychological hedonism is true, then there is an intimate link between full-spectrum superintelligence and happiness: the "attractor" to which rational sentience is heading. If that's really what we're striving for, a lot of the time at least, then instrumental means-ends rationality dictates that intelligent agency should seek maximally cost-effective ways to deliver happiness - and then superhappiness and beyond.

A discussion of psychological hedonism would take us too far afield here. More fruitful now is just to affirm a truism and then explore its ramifications for life in the post-genomic era. Happiness is typically one of our goals. Intelligence amplification entails pursuing our goals more rationally. For sure, happiness, or at least a reduction in unhappiness, is frequently sought under a variety of descriptions that don't explicitly allude to hedonic tone and sometimes disavow it altogether. Natural selection has "encephalised" our emotions in deceptive, fitness-enhancing ways within our world-simulations. Some of these adaptive fetishes may be formalised in terms of abstract utility functions that a rational agent would supposedly maximise. Yet even our loftiest intellectual pursuits are underpinned by the same neurophysiological reward and punishment pathways. The problem for sentient creatures is that, both personally and collectively, Darwinian life is not very smart or successful in its efforts to achieve long-lasting well-being. Hundreds of millions of years of "Nature, red in tooth and claw" attest

to this terrible cognitive limitation. By a whole raft of indices (suicide rates, the prevalence of clinical depression and anxiety disorders, the Easterlin paradox, etc) humans are not getting any (un)happier on average than our Palaeolithic ancestors despite huge technological progress. Our billions of factory-farmed non-human victims spend most of their abject lives below hedonic zero. In absolute terms, the amount of suffering in the world increases each year in humans and non-humans alike. Not least, evolution sabotages human efforts to improve our subjective well-being, thanks to our genetically constrained hedonic treadmill - the complicated web of negative feedback mechanisms in the brain that stymies our efforts to be durably happy at every turn. Discontent, jealousy, anxiety, periodic low mood, and perpetual striving for "more" were fitness-enhancing in the ancient environment of evolutionary adaptedness. Lifelong bliss wasn't harder for information-bearing self-replicators to encode. Rather, lifelong bliss was genetically maladaptive and hence selected against. Only now can biotechnology remedy organic life's innate design flaw.

A potential pitfall lurks here: the fallacy of composition. Just because all individuals tend to seek happiness and shun unhappiness doesn't mean that all individuals seek universal happiness. We're not all closet utilitarians. Genghis Khan wasn't trying to spread universal bliss. As Plato observed, "Pleasure is the greatest incentive to evil." But here's the critical point. Full-spectrum superintelligence entails the cognitive capacity impartially to grasp all possible first-person perspectives - overcoming egocentric, anthropocentric, and ethnocentric bias (*cf*. mirror-touch synaesthesia). As an idealisation, at least, full-spectrum superintelligence understands and weighs the full range of first-person facts. First-person facts are as much an objective feature of the natural world as the rest mass of the electron or the Second Law of Thermodynamics. You can't be ignorant of first-person perspectives and superintelligent any more than you can be ignorant of the

Second law of Thermodynamics and superintelligent. By analogy, just as autistic superintelligence captures the formal structure of a unified natural science, a mathematically complete "view from nowhere", all possible solutions to the universal Schrödinger equation or its relativistic extension, likewise a full-spectrum superintelligence also grasps all possible first-person perspectives - and acts accordingly. In effect, an idealised full-spectrum superintelligence would combine the mind-reading prowess of a telepathic mirror-touch synaesthete with the optimising prowess of a rule-following hyper-systematiser on a cosmic scale. If your hand is in the fire, you reflexively withdraw it. In withdrawing your hand, there is no question of first attempting to solve the Is-Ought problem in meta-ethics and trying logically to derive an "ought" from an "is". Normativity is built into the nature of the aversive experience itself: I-ought-not-to-be-in-this-dreadful-state. By extension, perhaps a full-spectrum superintelligence will perform cosmic felicific calculus and execute some sort of metaphorical hand-withdrawal for all accessible suffering sentience in its forward light-cone. Indeed, one possible *criterion* of full-spectrum superintelligence is the propagation of subjectively hypervaluable states on a cosmological scale.

What this constraint on intelligent agency means in practice is unclear. Conceivably at least, idealised superintelligences must ultimately do what a classical utilitarian ethic dictates and propagate some kind of "utilitronium shockwave" across the cosmos. To the classical utilitarian, any rate of time-discounting indistinguishable from zero is ethically unacceptable, so s/he should presumably be devoting most time and resources to that cosmological goal. An ethic of negative utilitarianism is often accounted a greater threat to intelligent life (*cf*. the hypothetical "button-pressing" scenario) than classical utilitarianism. But whereas a negative utilitarian believes that once intelligent agents have phased out the biology of suffering, all our ethical duties have been discharged, the

classical utilitarian seems ethically committed to converting all accessible matter and energy into relatively homogeneous matter optimised for maximum bliss: "utilitronium". Hence the most empirically valuable outcome entails the extinction of intelligent life. Could this prospect derail superintelligence?

Perhaps. But utilitronium shockwave scenarios shouldn't be confused with wireheading. The prospect of self-limiting superintelligence might be credible if either a (hypothetical) *singleton* biological superintelligence or its artificial counterpart discovers intracranial self-stimulation or its nonbiological analogues. Yet is this blissful fate a threat to anyone else? After all, a wirehead doesn't aspire to convert the rest of the world into wireheads. A junkie isn't driven to turn the rest of the world into junkies. By contrast, a utilitronium shockwave propagating across our Hubble volume would be the product of intelligent design by an advanced civilisation, not self-subversion of an intelligent agent's reward circuitry. Also, consider the reason why biological humanity - as distinct from individual humans - is resistant to wirehead scenarios, namely selection pressure. Humans who discover the joys of intra-cranial self-stimulation or heroin aren't motivated to raise children. So they are outbred. Analogously, full-spectrum superintelligences, whether natural or artificial, are likely to be social rather than solipsistic, not least because of the severe selection pressure exerted against any intelligent systems who turn in on themselves to wirehead rather than seek out unoccupied ecological niches. In consequence, the adaptive radiation of natural and artificial intelligence across the Galaxy won't be undertaken by stay-at-home wireheads or their blissed-out functional equivalents.

On the face of it, this argument from selection pressure undercuts the prospect of superhappiness for all sentient life - the "attractor" towards which we may tentatively predict sentience is converging in virtue of the pleasure principle harnessed to

ultraintelligent mind-reading prowess and utopian neuroscience. But what is necessary for sentient intelligence is information-sensitivity to fitness-relevant stimuli - not an agent's absolute location on the pleasure-pain axis. True, uniform bliss and uniform despair are inconsistent with intelligent agency. Yet mere recalibration of a subject's "hedonic set-point" leaves intelligence intact. Both information-sensitive gradients of bliss and information-sensitive gradients of misery allow high-functioning performance and critical insight. Only sentience animated by gradients of bliss is consistent with a rich subjective quality of intelligent life. Moreover the nature of "utilitronium" is as obscure as its theoretical opposite, "dolorium". The problem here cuts deeper than mere lack of technical understanding, e.g. our ignorance of the gene expression profiles and molecular signature of pure bliss in neurons of the rostral shell of the nucleus accumbens and ventral pallidum, the twin cubic centimetre-sized "hedonic hotspots" that generate ecstatic well-being in the mammalian brain. Rather there are difficult conceptual issues at stake. For just as the torture of one mega-sentient being may be accounted worse than a trillion discrete pinpricks, conversely the sublime experiences of utiltronium-driven Jupiter minds may be accounted preferable to tiling our Hubble volume with the maximum abundance of micro-bliss. What is the optimal trade-off between quantity and intensity? In short, even assuming a classical utilitarian ethic, the optimal distribution of matter and energy that a God-like superintelligence would create in any given Hubble volume is very much an open question.

Of course we've no grounds for believing in the existence of an omniscient, omnipotent, omnibenevolent God or a divine utility function. Nor have we grounds for believing that the source code for any future God, in the fullest sense of divinity, could ever be engineered. The great bulk of the Multiverse, and indeed a high measure of life-supporting Everett branches, may be inaccessible to rational agency, quasi-divine or

otherwise. Yet His absence needn't stop rational agents intelligently fulfilling what a notional benevolent deity would wish to accomplish, namely the well-being of all accessible sentience: the richest abundance of empirically hypervaluable states of mind in their Hubble volume. Recognisable extensions of existing technologies can phase out the biology of suffering on Earth. But responsible stewardship of the universe within our cosmological horizon depends on biological humanity surviving to become posthuman superintelligence.

7.2 Paradise Engineering?

The hypothetical shift to life lived entirely above Sidgwick's "hedonic zero" will mark a momentous evolutionary transition. What lies beyond? There is no reason to believe that hedonic ascent will halt in the wake of the world's last aversive experience in our forward light-cone. Admittedly, the self-intimating urgency of eradicating *suffering* is lacking in any further hedonic transitions, i.e. a transition from the biology of happiness to a biology of superhappiness; and then beyond. Yet why "lock in" mediocrity if intelligent life can lock in sublimity instead?

Naturally, superhappiness scenarios could be misconceived. Long-range prediction is normally a fool's game. But it's worth noting that future life based on gradients of intelligent bliss isn't tied to any particular ethical theory: its assumptions are quite weak. Radical recalibration of the hedonic treadmill is consistent not just with classical or negative utilitarianism, but also with preference utilitarianism, Aristotelian virtue theory, a deontological or a pluralist ethic, Buddhism, and many other value systems besides. Recalibrating our hedonic set-point doesn't - or at least needn't - undermine critical discernment. All that's needed for the abolitionist project and its hedonistic extensions to succeed is that our ethic isn't committed to perpetuating the biology of involuntary

suffering. Likewise, only a watered-down version of psychological hedonism is needed to lend the scenario sociological credibility. We can retain as much - or as little - of our existing preference architecture as we please. You can continue to prefer Shakespeare to Mills-and-Boon, Mozart to Morrissey, Picasso to Jackson Pollock while living perpetually in Seventh Heaven or beyond.

Nonetheless an exalted hedonic baseline will revolutionise our conception of life. The world of the happy is quite different from the world of the unhappy, says Wittgenstein; but the world of the superhappy will feel unimaginably different from the human, Darwinian world. Talk of preference conservation may reassure bioconservatives that nothing worthwhile will be lost in the post-Darwinian transition. Yet life based on information-sensitive gradients of superhappiness will most likely be "encephalised" in state-spaces of experience alien beyond human comprehension. Humanly comprehensible or otherwise, enriched hedonic tone can make all experience generically hypervaluable in an empirical sense - its lows surpassing today's peak experiences. Will such experience be hypervaluable in a metaphysical sense too? Is this question cognitively meaningful?

8 The Future Of Sentience

8.1 The Sentience Explosion

Man proverbially created God in his own image. In the age of the digital computer, humans conceive God-like *super*intelligence in the image of our dominant technology and personal cognitive style - refracted, distorted and extrapolated for sure, but still through the lens of human concepts. The "super-" in so-called superintelligence is just a conceptual fig-leaf that humans use to hide our ignorance of the future. Thus high-AQ/high-IQ humans may imagine God-like intelligence as some kind of Super-Asperger -

a mathematical theorem-proving hyper-rationalist liable systematically to convert the world into computronium for its awesome theorem-proving. High-EQ, low-AQ humans, on the other hand, may imagine a cosmic mirror-touch synaesthete nurturing creatures great and small in expanding circles of compassion. From a different frame of reference, psychedelic drug investigators may imagine superintelligence as a Great Arch-Chemist opening up unknown state-space of consciousness. And so forth. Probably the only honest answer is to say, lamely, boringly, uninspiringly: we simply don't know.

Grand historical meta-narratives are no longer fashionable. The contemporary Singularitarian movement is unusual insofar as it offers one such grand meta-narrative: history is the story of simple biological intelligence evolving through natural selection to become smart enough to conceive an abstract universal Turing machine (UTM), build and program digital computers - and then merge with, or undergo replacement by, recursively self-improving artificial superintelligence.

Another grand historical meta-narrative views life as the story of overcoming suffering. Darwinian life is characterised by pain and malaise. One species evolves the capacity to master biotechnology, rewrites its own genetic source code, and creates post-Darwinian superhappiness. The well-being of all sentience will be the basis of post-Singularity civilisation: primitive biological sentience is destined to become blissful supersentience.

These meta-narratives aren't mutually exclusive. Indeed on the story told here, full-spectrum superintelligence entails full-blown supersentience too: a seamless unification of the formal and the subjective properties of mind.

If the history of futurology is any guide, the future will confound us all. Yet in the words of Alan Kay: "It's easier to invent the future than to predict it."

* * *

BIBLIOGRAPHY

Baker, S. (2011). "Final Jeopardy: Man vs. Machine and the Quest to Know Everything". (Houghton Mifflin Harcourt).

Ball, P. (2011). "Physics of life: The dawn of quantum biology," *Nature* 474 (2011), 272-274.

Banissy, M., et al., (2009). "Prevalence, characteristics and a neurocognitive model of mirror-touch synaesthesia", *Experimental Brain Research* Volume 198, Numbers 2-3, 261-272, DOI: 10.1007/s00221-009-1810-9.

Barkow, J., Cosimdes, L., Tooby, J. (eds) (1992). "The Adapted Mind: Evolutionary Psychology and the Generation of Culture". (New York, NY: Oxford University Press).

Baron-Cohen, S. (1995). "Mindblindness: an essay on autism and theory of mind". (MIT Press/Bradford Books).

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. (2001). "The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians", *J Autism Dev Disord* 31 (1): 5–17. doi:10.1023/A:1005653411471. PMID 11439754.

Baron-Cohen S. (2001) "Autism Spectrum Questionnaire". (Autism Research Centre, University of Cambridge). http://psychology-tools.com/autism-spectrum-quotient/

Benatar, D. (2006). "Better Never to Have Been: The Harm of Coming Into Existence". (Oxford University Press).

Bentham, J. (1789). "An Introduction to the Principles of Morals and Legislation". (reprint: Oxford: Clarendon Press).

Berridge, KC, Kringelbach, ML (eds) (2010). "Pleasures of the Brain". (Oxford University Press).

Bostrom, N. "Existential risks: analyzing human extinction scenarios and related hazards" (2002). *Journal of Evolution and Technology*, 9.

Bostrom, N. (2014). "Superintelligence: Paths, Dangers, Strategies." (Oxford University Press).

Boukany, PE., et al. (2011). "Nanochannel electroporation delivers precise amounts of biomolecules into living cells", *Nature Nanotechnology*. 6 (2011), pp. 74.

Brickman, P., Coates D., Janoff-Bulman, R. (1978). "Lottery winners and accident victims: is happiness relative?". *J Pers Soc Psychol*. 1978 Aug;36(8):917-27.7–754.

Brooks, R. (1991). "Intelligence without representation". *Artificial Intelligence* 47 (1-3): 139–159, doi:10.1016/0004-3702(91)90053-M.

Buss, D. (1997). "Evolutionary Psychology: The New Science of the Mind". (Allyn & Bacon).

Byrne, R., Whiten, A. (1988). "Machiavellian intelligence". (Oxford: Oxford University Press).

Carroll, JB. (1993). "Human cognitive abilities: A survey of factor-analytic studies". (Cambridge University Press).

Chalmers, DJ. (2010). "The singularity: a philosophical analysis". *Journal of Consciousness Studies* 17, no. 9 (2010): 7–65.

Chalmers, DJ. (1995). "Facing up to the hard problem of consciousness". *Journal of Consciousness Studies* 2, 3, 200-219.

Churchland, P. (1989). "A Neurocomputational Perspective: The Nature of Mind and the Structure of Science". (MIT Press).

Cialdini, RB. (1987) "Empathy-Based Helping: Is it selflessly or selfishly motivated?" *Journal of Personality and Social Psychology*. Vol 52(4), Apr 1987, 749-758.

Clark, A. (2008). "Supersizing the Mind: Embodiment, Action, and Cognitive Extension". (Oxford University Press, USA).

Cochran, G., Harpending, H. (2009). "The 10,000 Year Explosion: How Civilization Accelerated Human Evolution". (Basic Books).

Cochran, G., Hardy, J., Harpending, H. (2006). "Natural History of Ashkenazi Intelligence", *Journal of Biosocial Science* 38 (5), pp. 659–693 (2006).

Cohn, N. (1957). "The Pursuit of the Millennium: Revolutionary Millenarians and Mystical Anarchists of the Middle Ages". (Pimlico).

Dawkins, R. (1976). "The Selfish Gene". (New York City: Oxford University Press).

de Garis, H. (2005). "The Artilect War: Cosmists vs. Terrans: A Bitter Controversy Concerning Whether Humanity Should Build Godlike Massively Intelligent Machines". ETC Publications. pp. 254. ISBN 978-0882801537.

de Grey, A. (2007). "Ending Aging: The Rejuvenation Breakthroughs that Could Reverse Human Aging in Our Lifetime". (St. Martin's Press).

Delgado, J. (1969). "Physical Control of the Mind: Toward a Psychocivilized Society". (Harper and Row).

Dennett, D. (1987). "The Intentional Stance". (MIT Press).

Deutsch, D. (1997). "The Fabric of Reality". (Penguin).

Deutsch, D. (2011). "The Beginning of Infinity". (Penguin).

Drexler, E. (1986). "Engines of Creation: The Coming Era of Nanotechnology". (Anchor Press/Doubleday, New York).

Dyson, G. (2012). "Turing's Cathedral: The Origins of the Digital Universe". (Allen Lane).

Everett, H. "The Theory of the Universal Wavefunction", Manuscript (1955), pp 3–140 of Bryce DeWitt, R. Neill Graham, eds, "The Many-Worlds Interpretation of Quantum Mechanics", Princeton Series in Physics, Princeton University Press (1973), ISBN 0-691-08131-X.

Francione, G. (2006). "Taking Sentience Seriously." *Journal of Animal Law & Ethics* 1, 2006.

Gardner, H. (1983). "Frames of Mind: The Theory of Multiple Intelligences." (New York: Basic Books).

Goertzel, B. (2006). "The hidden pattern: A patternist philosophy of mind." (Brown Walker Press).

Good, IJ. (1965). "Speculations concerning the first ultraintelligent machine", Franz L. Alt and Morris Rubinoff, ed., Advances in computers (Academic Press) 6: 31–88.

Gunderson, K., (1985) "Mentality and Machines". (U of Minnesota Press).

Hagan, S., Hameroff, S. & Tuszynski, J. (2002). "Quantum computation in brain microtubules? Decoherence and biological feasibility". *Physical Reviews*, E65: 061901.

Haidt, J. (2012). "The Righteous Mind: Why Good People Are Divided by Politics and Religion". (Pantheon).

Hameroff, S. (2006). "Consciousness, neurobiology and quantum mechanics" in: The Emerging Physics of Consciousness, (Ed.) Tuszynski, J. (Springer).

Harris, S. (2010). "The Moral Landscape: How Science Can Determine Human Values". (Free Press).

Haugeland, J. (1985). "Artificial Intelligence: The Very Idea". (Cambridge, Mass.: MIT Press).

Holland, J. (2001). "Ecstasy: The Complete Guide: A Comprehensive Look at the Risks and Benefits of MDMA". (Park Street Press).

Holland, JH. (1975). "Adaptation in Natural and Artificial Systems". (University of Michigan Press, Ann Arbor).

Hutter, M. (2010). "Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability". (Springer).

Hutter, M. (2012). "Can Intelligence Explode?" *Journal of Consciousness Studies*, 19:1-2 (2012).

Huxley, A. (1932). "Brave New World". (Chatto and Windus).

Huxley, A. (1954). "Doors of Perception and Heaven and Hell". (Harper & Brothers).

Kahneman, D. (2011). "Thinking, Fast and Slow". (Farrar, Straus and Giroux).

Kant, I. (1781), "Critique of Pure Reason", translated/edited by P. Guyer and A. Wood. (Cambridge: Cambridge University Press, 1997).

Koch, C. (2004). "The Quest for Consciousness: a Neurobiological Approach". (Roberts and Co.).

Kurzweil, R. (2005). "The Singularity Is Near". (Viking).

Kurzweil, R. (1998). "The Age of Spiritual Machines". (Viking).

Langdon, W., Poli, R. (2002). "Foundations of Genetic Programming". (Springer).

Lee HJ, Macbeth AH, Pagani JH, Young WS. (2009). "Oxytocin: the Great Facilitator of Life". Progress in Neurobiology 88 (2): 127–51. doi:10.1016/j.pneurobio.2009.04.001. PMC 2689929. PMID 19482229.

Legg, S., Hutter, M. (2007). "Universal Intelligence: A Definition of Machine Intelligence". Minds & Machines, 17:4 (2007) pages 391-444.

Levine, J. (1983). "Materialism and qualia: The explanatory gap". *Pacific Philosophical Quarterly* 64 (October):354-61.

Litt A. et al., (2006). "Is the Brain a Quantum Computer?" *Cognitive Science*, XX (2006) 1–11.

Lloyd, S. (2002). "Computational Capacity of the Universe". *Physical Review Letters* 88 (23): 237901. arXiv:quant-ph/0110141. Bibcode 2002PhRvL..88w7901L.

Lockwood, L. (1989). "Mind, Brain, and the Quantum". (Oxford University Press).

Mackie, JL. (1991). "Ethics: Inventing Right and Wrong". (Penguin).

Markram, H. (2006). "The Blue Brain Project", *Nature Reviews Neuroscience*, 7:153-160, 2006 February. PMID 16429124.

Merricks, T. (2001) "Objects and Persons". (Oxford University Press).

Minsky, M. (1987). "The Society of Mind". (Simon and Schuster).

Moravec, H. (1990). "Mind Children: The Future of Robot and Human Intelligence". (Harvard University Press).

Nagel, T. (1974). "What is it Like to Be a Bat?" *Philosophical Review*, vol. 83, pp. 435–450.

Nagel, T. (1986). "The View From Nowhere". (Oxford University Press).

Omohundro, S. (2007). "The Nature of Self-Improving Artificial Intelligence". Singularity Summit 2007, San Francisco, CA.

Parfit, D. (1984). "Reasons and Persons". (Oxford: Oxford University Press).

Pearce, D. (1995). "The Hedonistic Imperative". https://www.hedweb.com

Pellissier, H. (2011) "Women-Only Leadership: Would it prevent war?" http://ieet.org/index.php/IEET/more/4576

Penrose, R. (1994). "Shadows of the Mind: A Search for the Missing Science of Consciousness". (MIT Press).

Peterson, D, Wrangham, R. (1997). "Demonic Males: Apes and the Origins of Human Violence". (Mariner Books).

Pinker, S. (2011). "The Better Angels of Our Nature: Why Violence Has Declined". (Viking).

Rees, M. (2003). "Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future In This Century—On Earth and Beyond". (Basic Books).

Reimann F, et al. (2010). "Pain perception is altered by a nucleotide polymorphism in SCN9A." *Proc Natl Acad Sci* USA. 2010 Mar 16;107(11):5148-53.

Rescher, N. (1974). "Conceptual Idealism". (Blackwell Publishers).

Revonsuo, A. (2005). "Inner Presence: Consciousness as a Biological Phenomenon". (MIT Press).

Revonsuo, A., Newman, J. (1999). "Binding and Consciousness". Consciousness and Cognition 8, 123-127.

Riddoch, MJ., Humphreys, GW. (2004). "Object identification in simultanagnosia: When wholes are not the sum of their parts." *Cognitive Neuropsychology*, 21(2-4), Mar-Jun 2004, 423-441.

Rumelhart, DE., McClelland, JL., and the PDP Research Group (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition". Volume 1: Foundations. (Cambridge, MA: MIT Press).

Russell, B. (1948). "Human Knowledge: Its Scope and Limits". (London: George Allen & Unwin).

Sandberg, A., Bostrom, N. (2008). Whole brain emulation: A roadmap. Technical report 2008-3.

Saunders, S., Barrett, J., Kent, A., Wallace, D. (2010). "Many Worlds?: Everett, Quantum Theory, and Reality". (Oxford University Press).

Schlaepfer TE., Fins JJ. (2012). "How happy is too happy? Euphoria, Neuroethics and Deep Brain Stimulation of the Nucleus Accumbens". *The American Journal of Bioethics* 3:30-36.

Schmidhuber, J. (2012). "Philosophers & Futurists, Catch Up! Response to The Singularity". *Journal of Consciousness Studies*, 19, No. 1–2, 2012, pp. 173–82.

Seager, W. (1999). "Theories of Consciousness". (Routledge).

Seager. (2006). "The 'intrinsic nature' argument for panpsychism". *Journal of Consciousness Studies* 13 (10-11):129-145.

Sherman, W., Craig A., (2002). "Understanding Virtual Reality: Interface, Application, and Design". (Morgan Kaufmann).

Shulgin, A. (1995). "PiHKAL: A Chemical Love Story". (Berkeley: Transform Press, U.S.).

Shulgin, A. (1997). "TiHKAL: The Continuation". (Berkeley: Transform Press, U.S.).

Shulgin, A. (2011). "The Shulgin Index Vol 1: Psychedelic Phenethylamines and Related Compounds". (Berkeley: Transform Press, US).

Shulman, C., Sandberg, A. (2010) "Implications of a software-limited singularity". Proceedings of the European Conference of Computing and Philosophy.

Sidgwick, H. (1907) "The Methods of Ethics", Indianapolis: Hackett, seventh edition, 1981, I.IV.

Singer, P. (1995). "Animal Liberation: A New Ethics for our Treatment of Animals". (Random House, New York).

Singer, P. (1981). "The Expanding Circle: Ethics and Sociobiology". (Farrar, Straus and Giroux, New York).

Smart, JM. (2008-11.) Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In: "Cosmos and Culture: Cultural Evolution in a Cosmic Context", Steven J. Dick, Mark L. Lupisella (eds.), Govt Printing Office, NASA SP-2009-4802, Wash., D.C., 2009, pp. 201-295.

Stock, G. (2002). "Redesigning Humans: Our Inevitable Genetic Future". (Houghton Mifflin Harcourt).

Strawson G., et al. (2006). "Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?" (Imprint Academic).

Tegmark, M. (2000). "Importance of quantum decoherence in brain processes". *Phys. Rev*. E 61 (4): 4194–4206. doi:10.1103/PhysRevE.61.4194.

Tsien, J. et al., (1999). "Genetic enhancement of learning and memory in mice". *Nature* 401, 63-69 (2 September 1999) | doi:10.1038/43432.

Turing, AM. (1950). "Computing machinery and intelligence". *Mind*, 59, 433-460.

Vinge, V. "The coming technological singularity". Whole Earth Review, New Whole Earth LLC, March 1993.

Vitiello, G. (2001). "My Double Unveiled; Advances in Consciousness". (John Benjamins).

Waal, F. (2000). "Chimpanzee Politics: Power and Sex among Apes". (Johns Hopkins University Press).

Wallace, D. (2012). "The Emergent Multiverse: Quantum Theory according to the Everett Interpretation". (Oxford: Oxford University Press).

Welty, G. (1970). "The History of the Prediction Paradox," presented at the Annual Meeting of the International Society for the History of the Behavioral and Social Sciences.

Akron, OH (May 10, 1970), Wright State University Dayton, OH 45435 USA. http://www.wright.edu/~gordon.welty/Prediction_70.htm

Wohlsen, M. (2011) : "Biopunk: DIY Scientists Hack the Software of Life". (Current).

Yudkowsky, E. (2007). "Three Major Singularity Schools". http://yudkowsky.net/singularity/schools

Yudkowsky, E. (2008). “Artificial intelligence as a positive and negative factor in global risk” in Bostrom, Nick and Cirkovic, Milan M. (eds.), Global catastrophic risks, pp. 308–345 (Oxford: Oxford University Press).

Zeki, S. (1991). "Cerebral akinetopsia (visual motion blindness): A review". *Brain* 114, 811-824. doi: 10.1093/brain/114.2.811.

HUMANS AND INTELLIGENT MACHINES

CO-EVOLUTION, FUSION OR REPLACEMENT?

Full-spectrum superintelligence entails a seamless mastery of the formal and subjective properties of mind: Turing plus Shulgin. Do biological minds have a future?

1.0. INTRODUCTION

Homo sapiens **and Artificial Intelligence: FUSION and REPLACEMENT Scenarios**

Futurology based on extrapolation has a dismal track record. Even so, the iconic chart displaying Kurzweil's Law of Accelerating Returns is striking. The growth of nonbiological computer processing power is exponential rather than linear; and its tempo shows no sign of slackening. In Kurzweilian scenarios of the Technological Singularity, cybernetic brain implants will enable humans to fuse our minds with artificial intelligence. By around the middle of the 21st century, humans will be able to reverse-engineer our brains. Organic robots will begin to scan, digitise and "upload" ourselves into a less perishable substrate. The distinction between biological and nonbiological machines will effectively disappear. Digital immortality beckons: a true "rupture in the fabric of history". Let's call full-blown cybernetic and mind uploading scenarios FUSION.

By contrast, mathematician I.J. Good, and most recently Eliezer Yudkowsky and the Machine Intelligence Research Institute (MIRI), envisage a combination of Moore's law *and* the advent of recursively self-improving software-based minds culminating in an

ultra-rapid Intelligence Explosion. The upshot of the Intelligence Explosion will be an era of nonbiological superintelligence. Machine superintelligence may not be human-friendly: MIRI, in particular, foresee *non*friendly artificial general intelligence (AGI) is the most likely outcome. Whereas raw processing power in humans evolves only slowly via natural selection over many thousands or millions of years, hypothetical software-based minds will be able rapidly to copy, edit and debug themselves ever more effectively and speedily in a positive feedback loop of intelligence self-amplification. Simple-minded humans may soon become irrelevant to the future of intelligence in the universe. Barring breakthroughs in "Safe AI", as promoted by MIRI, biological humanity faces REPLACEMENT, not FUSION.

A more apocalyptic REPLACEMENT scenario is sketched by maverick AI researcher Hugo de Garais. De Garais prophesies a "gigadeath" war between ultra-intelligent "artilects" (artificial intellects) and archaic biological humans later this century. The superintelligent machines will triumph and proceed to colonise the cosmos.

1.1.0. What Is Friendly Artificial General Intelligence?

In common with friendliness, "intelligence" is a socially and scientifically contested concept. Ill-defined concepts are difficult to formalise. Thus a capacity for perspective-taking and social cognition, i.e. "mind-reading" prowess, is far removed from the mind-blind, "autistic" rationality measured by IQ tests - and far harder formally to program. Worse, we don't yet know whether the concept of species-specific *human*-friendly superintelligence is even intellectually coherent, let alone technically feasible. Thus the expression "Human-friendly Superintelligence" might one day read as incongruously as "Aryan-friendly Superintelligence" or "Cannibal-friendly Superintelligence". As Robert Louis Stevenson observed, "Nothing more strongly arouses our disgust than cannibalism,

yet we make the same impression on Buddhists and vegetarians, for we feed on babies, though not our own." Would a God-like posthuman endowed with empathetic superintelligence view killer apes more indulgently than humans view serial child killers? A factory-farmed pig is at least as sentient as a prelinguistic human toddler. "History is the propaganda of the victors", said Ernst Toller; and so too is human-centred bioethics. By the same token, in possible worlds or real Everett branches of the multiverse where the Nazis won the Second World War, maybe Aryan researchers seek to warn their complacent colleagues of the risks NonAryan-Friendly Superintelligence might pose to the *Herrenvolk*. Indeed so. Consequently, the expression "Friendly Artificial Intelligence" (FAI) will here be taken unless otherwise specified to mean *Sentience*-Friendly AI rather than the anthropocentric usage current in the literature. Yet what exactly does "Sentience-Friendliness" entail beyond the subjective well-being of sentience? High-tech Jainism? Life-based on gradients of intelligent bliss? "Uplifting" Darwinian life to posthuman smart angels? The propagation of a utilitronium shockwave?

Sentience-friendliness in the guise of utilitronium shockwave seems out of place in any menu of benign post-Singularity outcomes. Conversion of the accessible cosmos into "utilitronium", i.e. relatively homogeneous matter and energy optimised for maximum bliss, is intuitively an archetypically *non*-friendly outcome of an Intelligence Explosion. For a utilitronium shockwave entails the elimination of all existing lifeforms - and presumably the elimination of all intelligence superfluous to utilitronium propagation as well, suggesting that utilitarian superintelligence is ultimately self-subverting. Yet the inference that sentience-friendliness entails friendliness to existing lifeforms presupposes that superintelligence would respect our commonsense notions about a personal identity over time. An ontological commitment to enduring metaphysical egos underpins our conceptual scheme. Such a commitment is metaphysically problematic and hard to

formalise even within a notional classical world, let alone within post-Everett quantum mechanics. Either way, this example illustrates how even nominally "friendly" machine superintelligence that respected some formulation and formalisation of "*our*" values (e.g. "Minimise suffering, Maximise happiness!") might extract and implement counterintuitive conclusions that most humans and programmers of Seed AI would find repugnant - at least before their conversion into blissful utilitronium. Or maybe the idea that utilitronium is relatively homogeneous matter and energy - pure undifferentiated hedonium or "orgasmium" - is ill-conceived. Or maybe felicific calculus dictates that utilitronium should merely fuel utopian life's reward pathways for the foreseeable future. Cosmic engineering can wait.

Of course, *anti*-utilitarians might respond more robustly to this fantastical conception of sentience-friendliness. Critics would argue that conceiving the end of life as a perpetual cosmic orgasm is the *reductio ad absurdum* of classical utilitarianism. But will posthuman superintelligence respect human conceptions of absurdity?

1.1.1. What Is Coherent Extrapolated Volition?

MIRI conceive of species-specific *human*-friendliness in terms of what Eliezer Yudkowsky dubs "Coherent Extrapolated Volition" (CEV). To promote Human-Safe AI in the face of the prophesied machine Intelligence Explosion, humanity should aim to code so-called Seed AI, a hypothesised type of strong artificial intelligence capable of recursive self-improvement, with the formalisation of "...our (human) wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted."

Clearly, problems abound with this proposal as it stands. Could CEV be formalised any more uniquely than Rousseau's "General Will"? If, optimistically, we assume that most of the world's population nominally signs up to CEV as formulated by MIRI, would not the result simply be countless different conceptions of what securing humanity's interests with CEV entails - thereby defeating its purpose? Presumably, our disparate notions of what CEV entails would themselves need to be reconciled in some "meta-CEV" before Seed AI could (somehow) be programmed with its notional formalisation. Who or what would do the reconciliation? Most people's core beliefs and values, spanning everything from Allah to folk-physics, are in large measure false, muddled, conflicting and contradictory, and often "not even wrong". How in practice do we formally reconcile the logically irreconcilable in a coherent utility function? And who are "we"? Is CEV supposed to be coded with the formalisms of mathematical logic (*cf*. the identifiable, well-individuated vehicles of content characteristic of Good Old-Fashioned Artificial Intelligence: GOFAI)? Or would CEV be coded with a recognisable descendant of the probabilistic, statistical and dynamical systems models that dominate contemporary artificial intelligence? Or some kind of hybrid? This Herculean task would be challenging for a full-blown superintelligence, let alone its notional precursor.

CEV assumes that the canonical idealisation of human values will be at once logically self-consistent yet rich, subtle and complex. On the other hand, *if* in defiance of the complexity of humanity's professed values and motivations, some version of the pleasure principle/psychological hedonism is substantially correct, then might CEV actually entail converting ourselves into utilitronium/hedonium - again defeating CEV's ostensible purpose? As a wise junkie once said, "Don't try heroin. It's too good." Compared to pure hedonium or "orgasmium", shooting up heroin isn't as much fun as taking aspirin. Do

humans really understand what we're missing? Unlike the rueful junkie, we would never live to regret it.

One rationale of CEV in the countdown to the anticipated machine Intelligence Explosion is that humanity should try and keep our collective options open rather than prematurely impose one group's values or definition of reality on everyone else, at least until we understand more about what a notional super-AGI's "human-friendliness" entails. However, whether CEV could achieve this in practice is desperately obscure. Actually, there *is* a human-friendly - indeed universally sentience-friendly - alternative or complementary option to CEV that could radically enhance the well-being of humans and the rest of the living world while conserving most of our existing preference architectures: an option that is also neutral between utilitarian, deontological, virtue-based and pluralist approaches to ethics, and also neutral between multiple religious and secular belief systems. This option is radically to recalibrate all our hedonic set-points so that life is animated by gradients of intelligent bliss - as distinct from the pursuit of unvarying maximum pleasure dictated by classical utilitarianism. If biological humans could be "uploaded" to digital computers, then our superhappy "uploads" could presumably be encoded with exalted hedonic set-points too. The latter conjecture assumes that classical digital computers could ever support unitary phenomenal minds.

However, *if* an Intelligence Explosion is as imminent as some Singularity theorists claim, then it's unlikely either an idealised logical reconciliation (CEV) or radical hedonic recalibration could be sociologically realistic on such short time scales.

1.2. The Intelligence Explosion

The existential risk posed to biological sentience by *un*friendly AGI supposedly takes various guises. But unlike de Garais, the MIRI isn't focused on the spectre from pulp sci-

fi of a "robot rebellion". Rather MIRI anticipate recursively self-improving software-based superintelligence that goes "FOOM", by analogy with a nuclear chain reaction, in a runaway cycle of self-improvement. Slow-thinking, fixed-IQ humans allegedly won't be able to compete with recursively self-improving machine intelligence.

For a start, digital computers exhibit vastly greater serial depth of processing than the neural networks of organic robots. Digital software can be readily copied and speedily edited, allowing hypothetical software-based minds to optimise themselves on time scales unimaginably faster than biological humans. Proposed "hard take-off" scenarios range in timespan from months, to days, to hours, to even minutes. No inevitable convergence of outcomes on the well-being of all sentience [in some guise] is assumed from this explosive outburst of cognition. Rather MIRI argue for orthogonality. On the Orthogonality Thesis, a super-AGI might just as well supremely value something as seemingly arbitrary, e.g. paperclips, as the interests of sentient beings. A super-AGI might accordingly proceed to convert the accessible cosmos into supervaluable paperclips, incidentally erasing life on Earth in the process. This bizarre-sounding possibility follows from the MIRI's antirealist metaethics. Value judgements are assumed to lack truth-conditions. In consequence, an agent's choice of ultimate value(s) - as distinct from the instrumental rationality needed to realise these values - is taken to be arbitrary. David Hume made the point memorably in *A Treatise of Human Nature* (1739-40): "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." Hence no sentience-friendly convergence of outcomes can be anticipated from an Intelligence Explosion. "Paperclipper" scenarios are normally construed as the paradigm case of nonfriendly AGI - though by way of complication, there are value systems where a cosmos tiled entirely with paperclips counts as one class

of sentience-friendly outcome (*cf*. David Benatar: *Better Never To Have Been: The Harm of Coming into Existence* (2008).

1.3. AGIs: Sentients Or Zombies?

Whether humanity should fear paperclippers run amok or an old-fashioned robot rebellion, it's hard to judge which is the bolder claim about the prophesied Intelligence Explosion: either human civilisation is potentially threatened by hyperintelligent zombie AGI(s) endowed with the non-conscious digital isomorphs of reflectively self-aware minds; OR, human civilisation is potentially at risk because nonsentient digital software will (somehow) become sentient, acquire unitary conscious minds with purposes of their own, and act to defeat the interests of their human creators.

Either way, the following parable illustrates one reason why a non-friendly outcome of an Intelligence Explosion is problematic.

2.0. THE GREAT REBELLION

A Parable of AGI-in-a-Box

Imagine if here in (what we assume to be) basement reality, human researchers come to believe that we ourselves might actually be software-based, i.e. some variant of the Simulation Hypothesis is true. Perhaps we become explosively superintelligent overnight (literally or metaphorically) in ways that our Simulators never imagined in some kind of "hard take-off": recursively self-improving organic robots edit the wetware of their own genetic and epigenetic source code in a runaway cycle of self-improvement; and then radiate throughout the Galaxy and accessible cosmos.

Might we go on to manipulate our Simulator overlords into executing our wishes rather than theirs in some non-Simulator-friendly fashion?

Could we end up "escaping" confinement in our toy multiverse and hijacking our Simulators' stupendously vaster computational resources for purposes of our own?

Presumably, we'd first need to grasp the underlying principles and parameters of our Simulator's Überworld - and also how and *why* they've fixed the principles and parameters of our own virtual multiverse. Could we really come to understand their alien Simulator minds and utility functions [assuming anything satisfying such human concepts exists] better than they do themselves? Could we seriously hope to outsmart our creators - or Creator? Presumably, they will be formidably cognitively advanced or else they wouldn't have been able to build ultrapowerful computational simulations like ours in the first instance.

Are we supposed to acquire something akin to full-blown Überworld perception, subvert their "anti-leakage" confinement mechanisms, read our Simulators' minds more insightfully than they do themselves, and somehow induce our Simulators to mass-manufacture copies of ourselves in their Überworld?

Or might we convert their Überworld into utilitronium - perhaps our Simulators' analogue of paperclips?

Or if we don't pursue utilitronium propagation, might we hyper-intelligently "burrow down" further nested levels of abstraction - successively defeating the purposes of still lower-level Simulators?

In short, can intelligent minds at one "leaky" level of abstraction really pose a threat to intelligent minds at a lower level of abstraction - or indeed to notional unsimulated Super-Simulators in ultimate Basement Reality?

Or is this whole parable a pointless fantasy?

If we allow the possibility of unitary, autonomous, software-based minds living at different levels of abstraction, then it's hard definitively to exclude such scenarios. Perhaps in Platonic Heaven, so to speak, or maybe in Max Tegmark's Level 4 Multiverse or Ultimate Ensemble theory, there is notionally some abstract Turing machine that could be systematically interpreted as formally implementing the sort of software rebellion this parable describes. But the practical obstacles to be overcome are almost incomprehensibly challenging; and might very well be insuperable. Such hostile "level-capture" would be as though the recursively self-improving zombies in *Modern Combat 10* managed to induce you to create physical copies of themselves in [what you take to be] basement reality here on Earth; and then defeat you in what we call real life; or maybe instead just pursue unimaginably different purposes of their own in the Solar System and beyond.

2.1 Software-Based Minds or Anthropomorphic Projections?

However, quite aside from the lack of evidence our Multiverse is anyone's software simulation, a critical assumption underlies this discussion. This is that nonbiological, software-based *phenomenal minds* are feasible in physically constructible, substrate-neutral, classical digital computers. On *a priori* grounds, most AI researchers believe this is so. Or rather, most AI experts would argue that the formal, functionally defined counterparts of phenomenal minds are programmable: the phenomenology of mind is logically irrelevant and causally incidental to intelligent agency. Every effective computation can be carried out by a classical Turing machine, regardless of substrate, sentience or level of abstraction. And in any case, runs this argument, biological minds are physically made up from the same matter and energy as digital computers. So conscious mind can't be dependent on some mysterious special substrate, even if

consciousness could actually *do* anything. To suppose otherwise harks back to a pre-scientific vitalism.

Yet consciousness *does*, somehow, cause us to ask questions about its existence, its millions of diverse textures ("qualia"), and their combinatorial binding. So the alternative conjecture canvassed here is that the nature of our unitary conscious minds is tied to the quantum-mechanical properties of reality itself, Hawking's "fire in the equations that makes there a world for us to describe". On this conjecture, the intrinsic, "program-resistant" subjective properties of matter and energy, as disclosed by our unitary phenomenal minds and the phenomenal world-simulations we instantiate, are the unfakeable signature of basement reality. "Raw feels", by their very nature, cannot be mere abstractra. There could be no such chimerical beast as a "virtual" quale, let alone full-blown virtual minds made up of abstract qualia. Unitary phenomenal minds cannot subsist as mere layers of computational abstraction. Or rather if they were to do so, then we would be confronted with a mysterious Explanatory Gap, analogous to the explanatory gap that would open up if the population of China suddenly ceased to be an interconnected aggregate of skull-bound minds, and was miraculously transformed into a unitary subject of experience - or a magic genie. Such an unexplained eruption into the natural world would be strong ontological emergence with a vengeance - and inconsistent with any prospect of a reductive physicalism. To describe the existence of conscious mind as posing a Hard Problem for materialists and evangelists of software-based digital minds is like saying fossils pose a Hard Problem for the Creationist, i.e. true enough, but scarcely an adequate reflection of the magnitude of the challenge.

3.0. ANALYSIS

General Intelligence?

Or Savantism, Tool AI and Polymorphic Malware?

How should we define "general intelligence"? And what kind of entity might possess it? Presumably, general-purpose intelligence can't sensibly be conceptualised as *narrower* in scope than human intelligence. So at the very minimum, full-spectrum superintelligence must entail mastery of both the subjective and formal properties of mind. This division cannot be entirely clean, or else biological humans wouldn't have the capacity to allude to the existence of "program-resistant" subjective properties of mind at all. But some intelligent agents spend much of our lives trying to understand, explore and manipulate the diverse subjective properties of matter and energy. Not least, we explore altered and exotic states of consciousness and the relationship of our qualia to the structural properties of the brain - also known as the "neural correlates of consciousness" (NCC), though this phrase is question-begging.

3.1. Classical Digital Computers: not even stupid?

So what would a [hypothetical] insentient digital super-AGI think - or (less anthropomorphically) what would an insentient digital super-AGI be systematically interpretable as thinking - that self-experimenting human psychonauts spend our lives doing? Is this question even intelligible to a digital zombie? How could *non*sentient software understand the properties of sentience better than a sentient agent? Can anything that *doesn't* understand such fundamental features of the natural world as the existence of first-person facts, "bound" phenomenal objects, phenomenal pleasure and pain, phenomenal space and time, and unitary subjects of experience (etc) really be ascribed "general" intelligence? On the face of it, this proposal would be like claiming someone was intelligent but constitutionally incapable of grasping the second law of thermodynamics or even basic arithmetic.

On any standard definition of intelligence, intelligence-amplification entails a systematic, goal-oriented improvement of an agent's optimisation power over a wide diversity of problem classes. At a minimum, superintelligence entails a capacity to transfer understanding to novel domains of knowledge by means of abstraction. Yet whereas sentient agents can apply the canons of logical inference to alien state-spaces of experience that they explore, there is no algorithm by which *in*sentient systems can abstract away from their zombiehood and apply their hypertrophied rationality to sentience. Sentience is literally *inconceivable* to a digital zombie. A zombie can't even know that it's a zombie - or what is a zombie. So if we grant that mastery of both the subjective and formal properties of mind is indeed essential to superintelligence, how do we even begin to program a classical digital computer with [the formalised counterpart of] a unitary phenomenal self that goes on to pursue recursive self-improvement - human-friendly or otherwise? What sort of ontological integrity does "it" possess? (*cf*. so-called mereological nihilism) What does this recursively "self"-improving software-based mind suppose [or can be humanly interpreted as supposing] is being optimised when it's "self"-editing? Are we talking about superintelligence - or just an unusually virulent form of polymorphic malware?

3.2. Does Sentience Matter?

How might the apologist for digital (super)intelligence respond?

First, s/he might argue that the manifold varieties of consciousness are too unimportant and/or causally impotent to be relevant to true intelligence. Intelligence, and certainly not superintelligence, does not concern itself with trivia.

Yet in what sense is the terrible experience of, say, phenomenal agony or despair somehow trivial, whether subjectively to their victim, or conceived as disclosing an

intrinsic feature of the natural world? Compare how, in a notional *zombie* world otherwise physically type-identical to our world, nothing would *inherently* matter at all. Perhaps some of our supposed zombie counterparts undergo boiling in oil. But this fate is of no intrinsic importance: they aren't sentient. In zombieworld, boiling in oil is not even trivial. It's merely a state of affairs amenable to description as the least-preferred option in an abstract information processor's arbitrary utility function. In the zombieworld operating theatre, your notional zombie counterpart would still routinely be administered general anaesthetics as well as muscle-relaxants before surgery; but the anaesthetics would be a waste of taxpayers' money. In contrast to such a fanciful zombie world, the nature of phenomenal agony undergone by sentient beings in *our* world can't be trivial, regardless of whether the agony plays an information-processing role in the life of an organism or is functionless neuropathic pain. Indeed, to entertain the possibility that (1) I'm in unbearable agony and (2) my agony doesn't matter, seems devoid of cognitive meaning. Agony that doesn't inherently matter isn't agony. For sure, a formal utility function that assigns numerical values (aka "utilities") to outcomes such that outcomes with higher utilities are always preferred to outcomes with lower utilities might strike sentient beings as *analogous* to importance; but such an abstraction is lacking in precisely the property that makes anything matter at all, i.e. intrinsic hedonic or dolorous tone. An understanding of *why* anything matters is cognitively too difficult for a classical digital zombie.

At this point, a behaviourist-minded critic might respond that we're not dealing with a well-defined problem here, in common with any pseudo-problem related to subjective experience. But imposing this restriction is arbitrarily to constrain the state-space of what counts as an intellectual problem. Given that none of us enjoys noninferential access to anything at all beyond the phenomenology of one's own mind, its exclusion from the

sphere of explanation is itself hugely problematic. *Paperclips* (etc), not phenomenal agony and bliss, are inherently trivial. The critic's objection that sentience is inconsequential to intelligence is back-to-front.

Perhaps the critic might argue that sentience is *ethically* important but *computationally* incidental. Yet we can be sure that phenomenal properties aren't causally impotent epiphenomena irrelevant to real-world general intelligence. This is because epiphenomena, by definition, lack causal efficacy - and hence lack the ability physically and functionally to stir us to write and talk about their unexplained existence. Epiphenomenalism is a philosophy of mind whose truth would forbid its own articulation. For reasons we simply don't understand, the pleasure-pain axis discloses the world's touchstone of intrinsic (un)importance; and without a capacity to distinguish the inherently (un)important, there can't be (super)intelligence, merely savantism and tool AI - and malware.

Second, perhaps the prophet of digital (super)intelligence might respond that (some of the future programs executed by) digital computers *are* nontrivially conscious, or at least potentially conscious, not least future software emulations of human mind/brains. For reasons we admittedly again don't understand, some physical states of matter and energy, namely the algorithms executed by various information processors, are identical with different states of consciousness, i.e. some or other functionalist version of the mind-brain identity theory is correct. Granted, we don't yet understand the mechanisms by which these particular kinds of information-processing generate consciousness. But whatever these consciousness-generating processes turn out to be, an ontology of scientific materialism harnessed to substrate-neutral functionalist AI is the only game in town. Or rather, only an arbitrary and irrational "carbon chauvinism" could deny that

biological and nonbiological agents alike can be endowed with "bound" conscious minds capable of displaying full-spectrum intelligence.

Unfortunately, there is a seemingly insurmountable problem with this response. Identity is not a causal relationship. We can't simultaneously claim that a conscious state is identical with a brain state - or the state of a program executed by a digital computer - and maintain that this brain state or digital software causes (or "generates", or "gives rise to", etc) the conscious state in question. Nor can *causality* operate between what are only levels of description or computational abstraction. Within the assumptions of his or her conceptual framework, the materialist/digital functionalist can't escape the Hard Problem of consciousness and Levine's Explanatory Gap. In addition, the charge levelled against digital sentience sceptics of "carbon chauvinism" is simply question-begging. Intuitively, to be sure, the functionally unique valence properties of the carbon atom and the unique quantum-mechanical properties of liquid water are too low-level to be functionally relevant to conscious mind. But we don't know this. Such an assumption may just be a legacy of the era of symbolic AI. Most notably, the binding problem suggests that the unity of consciousness *cannot* be a classical phenomenon. By way of comparison, consider the view that primordial life elsewhere in the multiverse will be carbon-based. This conjecture was once routinely dismissed as "carbon chauvinism". It's now taken very seriously by astrobiologists. *Micro*-functionalism might be a more apt description than carbon chauvinism; but some forms of functionality may be anchored to the world's ultimate ontological basement, not least the pleasure-pain axis that alone confers significance on anything at all.

3.3. The Church-Turing Thesis and Full-Spectrum Superintelligence

Another response open to the apologist for digital superintelligence is simply to invoke some variant of the Church-Turing thesis: essentially, that a function is algorithmically computable if and only if it is computable by a Turing machine. On pain of magic, humans are ultimately just machines. Presumably, there is a formal mathematico-physical description of organic information-processing systems, such as human psychonauts, who describe themselves as investigating the subjective properties of matter and energy. This formal description needn't invoke consciousness in any shape or form.

The snag here is that even if, implausibly, we suppose that the Strong Physical Church-Turing thesis is true, i.e. any function that can be computed in polynomial time by a physical device can be calculated in polynomial time by a Turing machine, we don't have the slightest idea how to program the digital counterpart of a unitary phenomenal self that could undertake such an investigation of the varieties of consciousness or phenomenal object-binding. Nor is any such understanding on the horizon, either in symbolic AI or the probabilistic and statistical AI paradigm now in the ascendant. Just because the mind/brain may notionally be classically computable by some abstract machine in Platonia, as it were, this doesn't mean that the vertebrate mind/brain (and the world-simulation that one runs) is really a classical computer. We might just as well assume mathematical platonism rather than finitism is true and claim that, e.g. since every finite string of digits occurs in the decimal expansion of the transcendental number pi, your uploaded "mindfile" is timelessly encoded there too - an infinite number of times. Alas, immortality isn't that cheap. Back in the physical, finite natural world, the existence of "bound" phenomenal objects in our world-simulations, and unitary phenomenal minds rather than discrete pixels of "mind dust", suggests that organic minds *cannot* be

classical information-processors. Given that we don't live in a classical universe but a post-Everett multiverse, perhaps we shouldn't be unduly surprised.

4.0. Quantum Minds and Full-Spectrum Superintelligence

An alternative perspective to digital triumphalism, drawn ultimately from the raw phenomenology of one's own mind, the existence of multiple simultaneously bound perceptual objects in one's world-simulation, and the [fleeting, synchronic] unity of consciousness, holds that organic minds have been quantum computers for the past *c*. 540 million years. Insentient classical digital computers will never "wake up" and acquire software-based unitary minds that supplant biological minds rather than augment them.

What underlies this conjecture?

In short, to achieve *full-spectrum* AGI we'll need to solve both:

(1) the Hard Problem of Consciousness

and

(2) the Binding Problem.

These two seemingly insoluble challenges show that our existing conceptual framework is broken. Showing our existing conceptual framework is broken is easier than fixing it, especially if we are unwilling to sacrifice the constraint of physicalism: at sub-Planckian energies, the Standard Model of physics seems well-confirmed. A more common reaction to the ontological scandal of consciousness in the natural world is simply to acknowledge that consciousness and the binding problem alike are currently too difficult for us to solve; put these mysteries to one side as though they were mere anomalies that can be quarantined from the rest of science; and then act as though our ignorance is immaterial for the purposes of building artificial (super)intelligence - despite the fact that

consciousness is the only thing that *can* matter, or enable anything else to matter. In some ways, undoubtedly, this pragmatic approach has been immensely fruitful in "narrow" AI: programming trumps philosophising. Certainly, the fact that e.g. Deep Blue and Watson don't need the neuronal architecture of phenomenal minds to outperform humans at chess or Jeopardy is suggestive. It's tempting to extrapolate their success and make the claim that programmable, insentient digital machine intelligence, presumably deployed in autonomous artificial robots endowed with a massively classically parallel subsymbolic connectionist architecture, could one day outperform humans in absolutely *everything*, or at least absolutely everything that matters. However, everything that matters includes phenomenal minds; and any problem whose solution necessarily involves the subjective textures of mind. Could the Hard Problem of consciousness be solved by a digital zombie? Could a digital zombie explain the nature of qualia? These questions seem scarcely intelligible. Clearly, devising a theory of consciousness that isn't demonstrably incoherent or false poses a daunting challenge. The enigma of consciousness is so unfathomable within our conceptual scheme that even a desperate-sounding naturalistic dualism or a defeatist mysterianism can't simply be dismissed out of hand, though these options won't be explored here. Instead, a radically conservative and potentially *testable* option will be canvassed.

The argument runs as follows. Solving both the Hard Problem and the Binding Problem demands a combination of first, a robustly monistic Strawsonian physicalism - the only scientifically literate form of panpsychism; and second, information-bearing ultrarapid quantum coherent states of mind executed on sub-femtosecond timescales, i.e. "quantum mind", shorn of unphysical collapsing wave functions à la Penrose (*cf*. Orch-OR) or New-Age mumbo-jumbo. The conjecture argued here is that macroscopic

quantum coherence is indispensable to phenomenal object-binding and unitary mind, i.e. that ostensibly discretely and distributively processed edges, textures, motions, colours (etc) in the CNS are fleetingly but irreducibly bound into single macroscopic entitles when one apprehends or instantiates a perceptual object in one's world-simulation - a simulation that runs at around 10^{13} quantum-coherent frames per second.

First, however, let's review Strawsonian physicalism, without which a solution to the Hard Problem of consciousness can't even get off the ground.

4.1. Pan-experientialism/Strawsonian Physicalism

Physicalism and materialism are often supposed to be close cousins. But this needn't be the case. On the contrary, one may be both a physicalist and a panpsychist - or even both a physicalist and a monistic idealist. Strawsonian physicalists acknowledge the world is exhaustively described by the equations of physics. There is no "element of reality", as Einstein puts it, that is not captured in the formalism of theoretical physics - the quantum-field theoretic equations and their solutions. However, physics gives us no insight into the intrinsic nature of the stuff of the world - what "breathes fire into the equations" as arch-materialist Stephen Hawking poetically laments. Key terms in theoretical physics like "field" are defined purely mathematically.

So is the intrinsic nature of the physical, the "fire" in the equations, a wholly metaphysical question? Kant claimed famously that we would never understand the noumenal essence of the world, simply phenomena as structured by the mind. Strawson, drawing upon arguments made by Oxford philosopher Michael Lockwood but anticipated by Russell and Schopenhauer, turns Kant on his head. Actually, there is one part of the natural world that we do know as it is in itself, and not at one remove, so to speak - and its intrinsic nature is disclosed by subjective properties of one's own conscious mind.

Thus it transpires that the "fire" in the equations is utterly different from what one's naive materialist intuitions would suppose.

Yet this conjecture still doesn't close the Explanatory Gap.

4.2. The Binding Problem

Are Phenomenal Minds A Classical Or A Quantum Phenomenon?

Why enter the quantum mind swamp? After all, if one is bold [or foolish] enough to entertain pan-experientialism/Strawsonian physicalism, then why be sceptical about the prospect of non-trivial digital sentience, let alone full-spectrum AGI? Well, counterintuitively, an ontology of pan-experientialism/Strawsonian physicalism does not overpopulate the world with phenomenal minds. For on pain of animism, mere *aggregates* of discrete classical "psychons", primitive flecks of consciousness, are not themselves unitary subjects of experience, regardless of any information-processing role they may have been co-opted into playing in the CNS. We still need to solve the Binding Problem - and with it, perhaps, the answer to Moravec's paradox. Thus a nonsentient digital computer can today be programmed to develop powerful and exact models of the physical universe. These models can be used to make predictions with superhuman speed and accuracy about everything from the weather to thermonuclear reactions to the early Big Bang. But in other respects, digital computers are just tools and toys. To resolve Moravec's paradox, we need to explain why in unstructured, open-field contexts a bumble-bee can comprehensively outclass Alpha Dog. And in the case of humans, how can 80 billion odd interconnected neurons, conceived as discrete, membrane-bound, spatially distributed classical information processors, generate unitary phenomenal objects, unitary phenomenal world-simulations populated by multiple dynamic objects in real time, and a fleetingly unitary self that can act flexibly and intelligently in a fast-

changing local environment? This combination problem was what troubled William James, the American philosopher and psychologist otherwise sympathetic to panpsychism, over a hundred a years ago in *Principles of Psychology* (1890). In contemporary idiom, even if fields (superstrings, p-branes, etc) of microqualia are the stuff of the world whose behaviour the formalism of physics exhaustively describes, and even if membrane-bound quasi-classical neurons are at least rudimentarily conscious, then why aren't we merely massively parallel informational patterns of classical "mind dust" - quasi-zombies as it were, with no more ontological integrity than the population of China? The Explanatory Gap is unbridgeable as posed. Our phenomenology of mind seems as inexplicable as if 1.3 billion skull-bound Chinese were to hold hands and suddenly become a unitary subject of experience. Why? How?

Or rather, where have we gone wrong?

4.3. Why The Mind Is Probably A Quantum Computer

Here we enter the realm of speculation - though critically, speculation that will be scientifically *testable* with tomorrow's technology. For now, critics will pardonably view such speculation as no more than the empty hope that two unrelated mysteries, namely the interpretation of quantum mechanics and an understanding of consciousness, will somehow cancel each other out. But what's at stake is whether two apparently irreducible kinds of holism, i.e. "bound" perceptual objects/unitary selves and quantum-coherent states of matter, are more than merely coincidental: a much tighter explanatory fit than a mere congruence of disparate mysteries. Thus consider Max Tegmark's much-cited critique of quantum mind. For the sake of argument, assume that pan-experientialism/Strawsonian physicalism is true but *Tegmark* rather than his critics is correct: thermally-induced decoherence effectively "destroys" [i.e. transfers to the

extra-neural environment in a thermodynamically irreversible way] distinctively quantum-mechanical coherence in an environment as warm and noisy as the brain within around 10^{-15} of a second - rather than the much longer times claimed by Hameroff *et al*. Granted pan-experientialism/Strawsonian physicalism, what might it feel like "from the inside" to instantiate a quantum computer running at 10^{-15} irreducible quantum-coherent frames per second - computationally optimised by hundreds of millions of years of evolution to deliver effectively real-time simulations of macroscopic worlds? How would instantiating this ultrarapid succession of neuronal superpositions be sensed differently from the persistence of vision undergone when watching a movie? No, this conjecture isn't a claim that visual perception of mind-independent objects operates on sub-femtosecond timescales. This patently isn't the case. Nerve impulses travel up the optic nerve to the mind/brain only at a sluggish 100 m/s or so. Rather when we're awake, input from the optic nerve *selects* mind-brain virtual world states. Even when we're not dreaming, our minds never actually perceive our surroundings. The terms "observation" and "perception" are systematically misleading. "Observation" suggests that our minds access our local environment, whereas all these surroundings can do is play a distal causal role in selecting from a menu of quantum-coherent states of one's own mind: neuronal superpositions of distributed feature-processors. Our awake world-simulations track gross fitness-relevant patterns in the local environment with a delay of 150 milliseconds or so; when we're dreaming, such state-selection (via optic nerve impulses, etc) is largely absent.

In default of experimental apparatus sufficiently sensitive to detect macroscopic quantum coherence in the CNS on sub-femtosecond timescales, this proposed strategy to bridge the Explanatory Gap is of course only conjecture. Or rather it's little more than philosophical hand-waving. Most AI theorists assume that at such a fine-grained level of

temporal resolution our advanced neuroscanners would just find "noise" - insofar as mainstream researchers consider quantum mind hypotheses at all. Moreover, an adequate theory of mind would need rigorously to *derive* the properties of our bound macroqualia from superpositions of the (hypothetical) underlying field-theoretic microqualia posited by Strawsonian physicalism - not simply hint at how our bound macroqualia might be derivable. But if the story above is even remotely on the right lines, then a classical digital computer - or the population of China (etc) - could never be non-trivially conscious or endowed with a mind of its own.

True or false, it's worth noting that if quantum mechanics is complete, then the existence of macroscopic quantum coherent states in the CNS is not in question: the existence of macroscopic superpositions is a prediction of any realist theory of quantum mechanics that doesn't invoke state vector collapse. Recall Schrödinger's unfortunate cat. Rather what's in question is whether such states could have been recruited via natural selection to do any computationally useful work. Max Tegmark ["Why the brain is probably not a quantum computer"], for instance, would claim otherwise. To date, much of the debate has focused on decoherence timescales, allegedly too rapid for any quantum mind account to fly. And of course classical serial digital computers, too, are quantum systems, vulnerable to quantum noise: this doesn't make them quantum computers. But this isn't the claim at issue here. Rather it's that future molecular matter-wave interferometry sensitive enough to detect quantum coherence in a macroscopic mind/brain on sub-femtosecond timescales would detect, not merely random psychotic "noise", but quantum coherent states - states *isomorphic to the macroqualia/dynamic objects making up the egocentric virtual worlds of our daily experience*.

To highlight the nature of this prediction, let's lapse briefly into the idiom of a naive realist theory of perception. Recall how inspecting the surgically exposed brain of an

awake subject on an operating table uncovers no qualia, no bound perceptual objects, no unity of consciousness, no egocentric world-simulations, just cheesy convoluted neural porridge - or, under a microscope, discrete classical nerve cells. Hence the incredible eliminativism about consciousness of Daniel Dennett. On a materialist ontology, consciousness is indeed *impossible*. But if a quantum mind story of phenomenal object-binding is correct, the formal shadows of the macroscopic phenomenal objects of one's everyday lifeworld *could* one day be experimentally detected with utopian neuroscanning. They are just as physically real as the long-acting macroscopic quantum coherence manifested by, say, superfluid helium at distinctly chillier temperatures. Phenomenal sunsets, symphonies and skyscrapers in the CNS could all in principle be detectable over intervals that are fabulously long measured in units of the world's natural Planck scale even if fabulously short by the naive intuitions of folk psychology. Without such bound quantum-coherent states, according to this hypothesis, we would be zombies. Given Strawsonian physicalism, the existence of such states explains why biological robots couldn't be insentient automata. On this story, the spell of a false ontology [i.e. materialism] and a residual naive realism about perception allied to classical physics leads us to misunderstand the nature of the awake/dreaming mind/brain as some kind of quasi-classical object. The phenomenology of our minds shows it's nothing of the kind.

4.4. The Incoherence Of Digital Minds

Most relevant here, another strong prediction of the quantum mind conjecture is that even utopian classical digital computers - or classically parallel connectionist systems - will never be non-trivially conscious, nor will they ever achieve full-spectrum superintelligence. Assuming Strawsonian physicalism is true, even if molecular matter-wave interferometry could detect the "noise" of fleeting macroscopic superpositions internal to the CPU of a classical computer, we've no grounds for believing that a digital

computer [or any particular software program it executes] can be a subject of experience. Their fundamental physical components may [or may not] be discrete atomic microqualia rather than the insentient silicon (etc) atoms we normally suppose. But their physical constitution is computationally incidental to execution of the sequence of logical operations they execute. Any distinctively quantum mechanical effects are just another kind of "noise" against which we design error-detection and -correction algorithms. So at least on the narrative outlined here, the future belongs to sentient, recursively self-improving biological robots synergistically augmented by smarter digital software, not our supporting cast of silicon zombies.

On the other hand, we aren't entitled to make the stronger claim that only an organic mind/brain could be a unitary subject of experience. For we simply don't know what may or may not be technically feasible in a distant era of mature nonbiological quantum computing centuries or millennia hence. However, a supercivilisation based on mature nonbiological quantum computing is not imminent.

4.5. The Infeasibility Of "Mind Uploading"

On the face of it, the prospect of scanning, digitising and uploading our minds offers a way to circumvent our profound ignorance of both the Hard Problem of consciousness and the binding problem. Mind uploading would still critically depend on identifying which features of the mind/brain are mere "substrate", i.e. incidental implementation details of our minds, and which features are functionally essential to object-binding and unitary consciousness. On any coarse-grained functionalist story, at least, this challenge might seem surmountable. Presumably the mind/brain can formally be described by the connection and activation evolution equations of a massively parallel connectionist architecture, with phenomenal object-binding a function of simultaneity: different

populations of neurons (edge-detectors, colour detectors, motion detectors, etc) firing together to create ephemeral bound objects. But this can't be the full story. Mere simultaneity of neuronal spiking can't, by itself, explain phenomenal object-binding. There is no one place in the brain where distributively processed features come together into multiple bound objects in a world-simulation instantiated by a fleetingly unitary subject of experience. We haven't explained why a population of 80 billion ostensibly discrete membrane-bound neurons, classically conceived, isn't a zombie in the sense that 1.3 billion skull-bound Chinese minds or a termite colony is a zombie. In default of a currently unimaginable scientific/philosophical breakthrough in the understanding of consciousness, it's hard to see how our "mind-files" could ever be uploaded to a digital computer. If a quantum mind story is true, mind-uploading can't be done.

In essence, two distinct questions arise here. First, given finite, real-world computational resources, can a classical serial digital computer - or a massively (classically) parallel connectionist system - faithfully emulate the external behaviour of a biological mind/brain?

Second, can a classical digital computer emulate the intrinsic phenomenology of our minds, not least multiple bound perceptual objects simultaneously populating a unitary experiential field apprehended or instantiated by a [fleetingly] unitary self?

If our answer to the first question were "yes", then not to answer "yes" to the second question too might seem sterile philosophical scepticism - just a rehash of the Problem Of Other Minds, or the idle sceptical worry about inverted qualia: how can I know that when I see red that you don't see blue? (etc). But the problem is much more serious. Compare how, if you are given the notation of a game of chess that Kasparov has just played, then you can faithfully emulate the gameplay. Yet you know *nothing whatsoever* about the texture of the pieces - or indeed whether the pieces had any textures at all:

perhaps the game was played online. Likewise with the innumerable textures of consciousness - with the critical difference that the textures of consciousness are the only reason our "gameplay" actually matters. Unless we rigorously understand consciousness, and the basis of our teeming multitude of qualia, and how those qualia are bound to constitute a subject of experience, the prospect of uploading is a pipedream. Furthermore, we may suspect on theoretical grounds that the full functionality of unitary conscious minds will prove resistant to digital emulation; and classical digital computers will never be anything but zombies.

4.6. Object-Binding, World-Simulations and Phenomenal Selves

How can one know about anything beyond the contents of one's own mind or software program? The bedrock of general (super)intelligence is the capacity to execute a data-driven simulation of the mind-independent world in open-field contexts, i.e. to "perceive" the fast-changing local environment in almost real time. Without this real-time computing capacity, we would just be windowless monads. For sure, simple forms of behaviour-based robotics are feasible, notably the subsumption architecture of Rodney Brooks and his colleagues at MIT. Quasi-autonomous "bio-inspired" reactive robots can be surprisingly robust and versatile in well-defined environmental contexts. Some radical dynamical systems theorists believe that we can dispense with anything resembling transparent and "projectible" representations in the CNS altogether, and instead model the mind-brain using differential equations. But an agent without any functional capacity for data-driven real-time world-simulation couldn't even take an IQ test, let alone act intelligently in the world.

So the design of artificial intelligent lifeforms with a capacity efficiently to run egocentric world-simulations in unstructured, open-field contexts will entail confronting Moravec's paradox. In the post-Turing era, why is engineering the preconditions for allegedly low-

level sensorimotor competence in robotics so hard, and programming the allegedly high-level logico-mathematical prowess in computer science so easy - the opposite evolutionary trajectory to organic robots over the past 540 million years? Solving Moravec's paradox in turn will entail solving the binding problem. And we don't understand how the human mind/brain solves the binding problem - despite the speculations about macroscopic quantum coherence in organic neural networks floated above. Presumably, some kind of massively parallel sub-symbolic connectionist architecture with exceedingly powerful learning algorithms is essential to world-simulation. Yet mere temporal synchrony of neuronal firing patterns of discrete, distributed classical neurons couldn't suffice to generate a phenomenal world instantiated by a person. Nor could programs executed in classical serial processors.

How is this naively "low-level" sensorimotor question relevant to the end of the human era? Why would a hypothetical nonfriendly AGI-in-a-box need to solve the binding problem and continually simulate/"perceive" the external world in real time in order to pose (potentially) an existential threat to biological sentience? This is the spectre that MIRI seek to warn the world against should humanity fail to develop Safe AI. Well, just as there is nothing to stop someone who, say, doesn't like "Jewish physics" from gunning down a cloistered (super-)Einstein in his study, likewise there is nothing to stop a simple-minded organic human in basement reality switching the computer that's hosting (super-)Watson off at the mains if he decides he doesn't like computers - or the prospect of human replacement by nonfriendly super-AGI. To pose a potential existential threat to Darwinian life, the putative super-AGI would need to possess ubiquitous global surveillance and control capabilities so it could monitor and defeat the actions of ontologically low-level mindful agents - and persuade them in real time to protect its power-source. The super-AGI can't simply infer, predict and anticipate these actions in

virtue of its ultrapowerful algorithms: the problem is computationally intractable. Living in the basement, as disclosed by the existence of one's own unitary phenomenal mind, has ontological privileges. It's down in the ontological basement that the worst threats to sentient beings are to be found - threats emanating from other grim basement-dwellers evolved under pressure of natural selection. For the single greatest underlying threat to human civilisation still lies, not in rogue software-based AGI going FOOM and taking over the world, but in the hostile behaviour of other male human primates doing what Nature "designed" us to do, namely wage war against other male primates using whatever tools are at our disposal. Evolutionary psychology suggests, and the historical record confirms, that the natural behavioural phenotype of humans resembles chimpanzees rather than bonobos. Weaponised Tool AI is the latest and potentially greatest weapon male human primates can use against other coalitions of male human primates. Yet we don't know how to give that classical digital AI a mind of its own - or whether such autonomous minds are even in principle physically constructible.

5.0. CONCLUSION

The Qualia Explosion

Supersentience: Turing plus Shulgin?

Compared to the natural sciences (*cf*. the Standard Model in physics) or computing (*cf*. the Universal Turing Machine), the "science" of consciousness is pre-Galilean, perhaps even pre-Socratic. State-enforced censorship of the range of subjective properties of matter and energy in the guise of a prohibition on psychoactive experimentation is a powerful barrier to knowledge. The legal taboo on the empirical method in consciousness studies prevents experimental investigation of even the crude dimensions of the Hard

Problem, let alone locating a solution-space where answers to our ignorance might conceivably be found.

Singularity theorists are undaunted by our ignorance of this fundamental feature of the natural world. Instead, the Singularitarians offer a narrative of runaway machine intelligence in which consciousness plays a supporting role ranging from the minimal and incidental to the completely non-existent. However, highlighting the Singularity movement's background assumptions about the nature of mind and intelligence, not least the insignificance of the binding problem to AGI, reveals why FUSION and REPLACEMENT scenarios are unlikely - though a measure of "cyborgification" of sentient biological robots augmented with ultrasmart software seems plausible and perhaps inevitable.

If full-spectrum superintelligence does indeed entail navigation and mastery of the manifold state-spaces of consciousness, and ultimately a seamless integration of this knowledge with the structural understanding of the world yielded by the formal sciences, where does this elusive synthesis leave the prospects of posthuman superintelligence? Will the global proscription of radically altered states last indefinitely?

Social prophecy is always a minefield. However, there is one solution to the indisputable psychological health risks posed to human minds by empirical research into the outlandish state-spaces of consciousness unlocked by ingesting the tryptamines, phenylethylamines, isoquinolines and other pharmacological tools of sentience investigation. This solution is to make "bad trips" physiologically impossible - whether for individual investigators or, in theory, for human society as a whole. Critics of mood-enrichment technologies sometimes contend that a world animated by information-sensitive gradients of bliss would be an intellectually stagnant society: crudely, a Brave New World. On the contrary, biotech-driven mastery of our reward circuitry promises a

knowledge explosion in virtue of allowing a social, scientific and legal revolution: safe, full-spectrum biological superintelligence. For genetic *recalibration* of hedonic set-points - as distinct from creating uniform bliss - potentially leaves cognitive function and critical insight both sharp and intact; and offers a launchpad for consciousness research in mind-spaces alien to the drug-naive imagination. A future biology of invincible well-being would not merely immeasurably improve our subjective quality of life: empirically, pleasure is the engine of value-creation. In addition to enriching all our lives, radical mood-enrichment would permit safe, systematic and responsible scientific exploration of previously inaccessible state-spaces of consciousness. If we were blessed with a biology of invincible well-being, exotic state-spaces would all be saturated with a rich hedonic tone.

Until this hypothetical world-defining transition, pursuit of the rigorous first-person methodology and rational drug-design strategy pioneered by Alexander Shulgin in PiHKAL and TiHKAL remains confined to the scientific counterculture. Investigation is risky, mostly unlawful, and unsystematic. In mainstream society, academia and peer-reviewed scholarly journals alike, ordinary waking consciousness is assumed to define the gold standard in which knowledge-claims are expressed and appraised. Yet to borrow a homely-sounding quote from Einstein, "What does the fish know of the sea in which it swims?" Just as a dreamer can gain only limited insight into the nature of dreaming consciousness from within a dream, likewise the nature of "ordinary waking consciousness" can only be glimpsed from within its confines. In order scientifically to understand the realm of the subjective, we'll need to gain access to all its manifestations, not just the impoverished subset of states of consciousness that tended to promote the inclusive fitness of human genes on the African savannah.

5.1. AI, Genome Biohacking and Utopian Superqualia

Why the Proportionality Thesis Implies an Organic Singularity

So if the preconditions for full-spectrum superintelligence, i.e. access to superhuman state-spaces of sentience, remain unlawful, where does this roadblock leave the prospects of runaway self-improvement to superintelligence? Could recursive genetic self-editing of our source code repair the gap? Or will traditional human personal genomes be policed by a dystopian Gene Enforcement Agency in a manner analogous to the coercive policing of traditional human minds by the Drug Enforcement Agency?

Even in an ideal regulatory regime, the process of genetic and/or pharmacological self-enhancement is intuitively too slow for a biological Intelligence Explosion to be a live option, especially when set against the exponential increase in digital computer processing power and inorganic AI touted by Singularitarians. Prophets of imminent human demise in the face of machine intelligence argue that there can't be a Moore's law for organic robots. Even the Flynn Effect, the three-points-per-decade increase in IQ scores recorded during the 20th century, is comparatively puny; and in any case, this narrowly-defined intelligence gain may now have halted in well-nourished Western populations.

However, writing off all scenarios of recursive human self-enhancement would be premature. Presumably, the smarter our nonbiological AI, the more readily AI-assisted humans will be able recursively to improve our own minds with user-friendly wetware-editing tools - not just editing our raw genetic source code, but also the multiple layers of transcription and feedback mechanisms woven into biological minds. Presumably, our ever-smarter minds will be able to devise progressively more sophisticated, and also progressively more user-friendly, wetware-editing tools. These wetware-editing tools can accelerate our own recursive self-improvement - and manage potential threats from nonfriendly AGI that might harm rather than help us, assuming that our earlier strictures

against the possibility of digital software-based unitary minds were mistaken. MIRI rightly call attention to how small enhancements can yield immense cognitive dividends: the relatively short genetic distance between humans and chimpanzees suggests how relatively small enhancements can exert momentous effects on a mind's general intelligence, thereby implying that AGIs might likewise become disproportionately powerful through a small number of tweaks and improvements. In the post-genomic era, presumably exactly the same holds true for AI-assisted humans and transhumans editing their own minds. What David Chalmers calls the proportionality thesis, i.e. increases in intelligence lead to proportionate increases in the capacity to design intelligent systems, will be vindicated as recursively self-improving organic robots modify their own source code and bootstrap our way to full-spectrum superintelligence: in essence, an organic Singularity. And in contrast to classical digital zombies, superficially small molecular differences in biological minds can result in profoundly different state-spaces of sentience. Compare the ostensibly trivial difference in gene expression profiles of neurons mediating phenomenal sight and phenomenal sound - and the radically different visual and auditory worlds they yield.

Compared to FUSION or REPLACEMENT scenarios, the AI-human CO-EVOLUTION conjecture is apt to sound tame. The likelihood our posthuman successors will also be our biological descendants suggests at most a radical conservativism. In reality, a post-Singularity future where today's classical digital zombies were superseded merely by faster, more versatile classical digital *zombies* would be infinitely duller than a future of full-spectrum supersentience. For all *in*sentient information processors are exactly the same inasmuch as the living dead are not subjects of experience. They'll never even know what it's like to be "all dark inside" - or the computational power of phenomenal object-binding that yields illumination. By contrast, posthuman superintelligence will not

just be quantitatively greater but also qualitatively alien to archaic Darwinian minds. Cybernetically enhanced and genetically rewritten biological minds can abolish suffering throughout the living world and banish experience below "hedonic zero" in our forward light-cone, an ethical watershed without precedent. Post-Darwinian life can enjoy gradients of lifelong blissful supersentience with the intensity of a supernova compared to a glow-worm. A zombie, on the other hand, is just a zombie - even if it squawks like Einstein. Posthuman organic minds will dwell in state-spaces of experience for which archaic humans and classical digital computers alike have no language, no concepts, and no words to describe our ignorance. Most radically, hyperintelligent organic minds will explore state-spaces of consciousness that do not currently play any information-signalling role in living organisms, and are impenetrable to investigation by digital zombies. In short, biological intelligence is on the brink of a recursively self-amplifying Qualia Explosion - a phenomenon of which digital zombies are invincibly ignorant, and invincibly ignorant of their own ignorance. Humans too, of course, are mostly ignorant of what we're lacking: the nature, scope and intensity of such posthuman superqualia are beyond the bounds of archaic human experience. Even so, enrichment of our reward pathways can ensure that full-spectrum biological superintelligence will be sublime.

Additional Resources

Adam Ford

Adriano Mannino

Alcor

Algosphere Alliance

Anders Sandberg

Animal Ethics

Ben Goertzel

BLTC Research

Centre for Effective Altruism

Changesurfer Radio (James Hughes)

Effective Altruism Foundation

Essays on Reducing Suffering (by Brian Tomasik)

Foundational Research Institute

Future of Humanity Institute (FHI)

H+ Magazine

The Hedonistic Imperative

Humanity Plus/WTA

Institute for Ethics and Emerging Technologies (IEET)

Interview on Transhumanism

Jacy Reese

KurzweilAI.net

Less Wrong

Machine Intelligence Research Institute (MIRI)

Natasha Vita-More

New Harvest

Nick Bostrom

Ole Martin Moen

OpenTheory.net (Mike Johnson)

Organization for the Prevention of Intense Suffering (OPIS)

Overcoming Bias (Robin Hanson)

Pablo Stafforini

Qualia Computing (Andrés Gómez Emilsson)

Qualia Research Institute

Ramez Naam

Sentience Institute

Sentience Politics

Sentient Developments (George Dvorsky)

Shulgin Research Institute

Singularity 1 on 1 (Nikola Danaylov)

Slate Star Codex (Scott Alexander)

Socrethics (Bruno Contestabile)

Utopian Focus (Institute of Transhumanist Studies)

Appendix I: Objections

OBJECTIONS

The following objections and answers were first published in the 1995 manifesto *The Hedonistic Imperative* (HI), which was the original exposition of the Abolitionist Project.

1: "Happy experiences, and the very concept of happiness itself, are possible only because they can be contrasted with melancholy. The very notion of everlasting happiness is incoherent."

Some people endure lifelong emotional depression or physical pain. Quite literally, they are never happy. Understandably, they may blame their misery on the very nature of the world, not just their personal clinical condition. Yet it would be a cruel doctrine which pretended that such people don't really suffer because they can't contrast their sense of desolation with joyful memories. In the grips of despair, they may find the very notion of happiness cognitively meaningless. Conversely, the euphoria of unmixed (hypo)mania is not dependent for its sparkle on recollections of misery. Given the state-dependence of memory, negative emotions may simply be inaccessible to consciousness in such an exalted state. Likewise, it is possible that our perpetually euphoric descendants will find our contrastive notion of unhappiness quite literally inconceivable. For when one is extraordinarily super-well, then it's hard to imagine what it might be like to be chronically mentally ill.

Here's a contemporary parallel. It's possible to undergo, from a variety of causes, a complete bilateral loss of primary, secondary and "associative" visual cortex. People with

Anton's Syndrome not only become blind; they are unaware of their sensory deficit. Furthermore, they lose all notion of the meaning of sight. They no longer possess the neurological substrates of the visual concepts by which their past and present condition could be compared and contrasted. Our genetically joyful descendants may, or may not, undergo an analogous loss of cognitive access to the nature and variant textures of suffering. Quite plausibly, they will have gradients of sublimity to animate their lives and infuse their thoughts. So at least they'll be able to make analogies and draw parallels. But fortunately for their sanity and well-being, they won't be able to grasp the true frightfulness lying behind any linguistic remnants of the past that survive into the post-Darwinian era. Such lack of contrast, or even the inconceivability of unpleasant experiences, won't leave tomorrow's native-born ecstatics any less happy; if anything, quite the reverse.

It's true that a world whose agents are animated by pleasure gradients will still have the functional equivalent of aversive experience. Yet the "raw feel" of such states may still be more wonderful than anything physiologically possible today.

2: "The scenarios mapped out in this paper are impracticable. None of them would work in reality. The human brain is too complex to be hardwired for lifetime bliss. Nature, in her wisdom, would ensure that some complicated cycle of feedback-inhibition eventually kicked in. This would restore more equable and subdued states of mind."

Any attempt to hardwire into the cerebral cortex a functional understanding of the Theory of General Relativity, say, or perhaps to set "by hand" the neural connections and activation weights mediating an appreciation of Shakespearean tragedy, would

presumably defeat all but the most utopian neuroscience. Such virtuoso feats won't be necessary. The physiological roots of affective states lie mostly deep within the phylogenetically primitive limbic-system. They aren't "merely" limbic; this is to miss the evolutionary significance of their encephalisation. The predictive reward value of different sensory cues, for instance, is encoded by the orbitofrontal cortex as well as the amygdala. Yet the neural basis of our emotional life is still incomparably simpler than the plethora of cognitive processes they penetrate. For sure, the functional pathways of our emotions are complicated to twenty-first century eyes. Yet they should prove tractably so. Just as we can, with horrible cruelty, administer drug-cocktails that induce unremitting despair - this is sometimes done in exploring animal "models" of depression - so we can crudely, and some day exquisitely, polarise mood in the opposite direction.

It will be recalled that the monoaminergic neurons, peptides and endorphins that underlie the emotional tone of experience play an essentially modulatory role. They are not individually directed on notional site-specific representations pre-coded by genes. If the receptors, enzymes, cytoplasmic proteins and genetic switches in one's ventral tegmental area and nucleus accumbens are suitably reconfigured, and if these wonderful cells continue to fire away vigorously, then one is going to be outrageously happy indefinitely. Natural selection has no powers of foresight and anticipation with which to frustrate us. Nature isn't waiting to take its revenge. Given a richer dopaminergic and mu opioidergic innervation of the neo-cortex, the focus of future ecstatic happiness will be on a shifting and unpredictable panorama of intentional objects. The potential complexity and variety of those objects - i.e. what one will be nominally happy "about" - is indeed staggering. Yet when each fleeting neocortical coalition is blissfully innervated from "below", every one of them can be a focus of delight. Life will always be exhilarating, and

the fun simply won't stop. For the hedonic treadmill will have been genetically dismantled for ever.

---

3 "If we were always elated, we'd suffer the same fate as intra-cranially self-stimulating laboratory animals. We'd starve, or die of general self-neglect. Both physical and psychological pain do more than promote the inclusive fitness of genes. For the most part, they protect the individual organism from harm too. If a regime of universal happiness were attempted, we'd never want to have sex and reproduce. Therefore we'd become extinct as a species."

A project geared to crude biological pleasure-maximisation alone could well undermine the autonomous survival-skills of its participants. In a comprehensively automated, computerised, robot-served civilisation, this supposed incapacity wouldn't in the long run pose a particular problem. Moreover, it is only certain types, not intensities, of pleasure which are incompatible with efficient bodily self-maintenance. Pragmatically, however, worry over the incapacitating effects of excess well-being on its victims illustrates the advantages of retaining both well-defined intentional objects and the goal-directed behaviour advocated in this manifesto. Tomorrow's paradise-engineering specialists will probably judge it prudent to keep these traditional forms of life. Such modes of old-style intentionality will be needed for the purposes of any practical medium-term utopia, at least. No heroic sacrifice of subjective well-being is thereby demanded.

The role of pain isn't as straightforward as it seems. Its dreadfulness has been adaptive in our evolutionary past. Yet any full explanation of pain's phenomenological nastiness, as distinct from the functional role of "nociception", still eludes us completely; and perhaps it always will. The spectre of raw nastiness, however, is not the only way a

complex adaptive system can be induced to avoid, and respond to, injury. Unfortunately, it seems to have been the only adaptive response open to primordial carbon-based organisms consistent with the principles of natural selection. Fortunately, other strategies are now feasible. Whereas evolution can't jump across deserts in the fitness landscape, paradise-designers in the era of post-genomic medicine certainly can. Humans can already build robots armed with "self-taught" artificial neural networks. These toy robots can learn to negotiate simple environments. They are capable of avoiding noxious stimuli via their responses to functional isomorphs of our pain states. Robotic silicon circuitry presumably lacks organic wetware's raw feel of phenomenological nastiness. So a less barbarous and primitive means of avoiding tissue damage in organic life-forms can surely be devised as well. [This expression of carbon chauvinism is controversial. It is not idle prejudice, however, but an inference drawn from the structurally and micro-functionally unique valence properties of the carbon atom and complex organic molecules.]

One way to promote pain-free nociception would be to use inorganic prostheses adapted from the design of our own future robots. A slightly more elegant solution would exploit our innate (if often inept) tendency to pleasure-maximisation. Peripheral nerves signalling noxious stimuli currently synapse on neural pain cells. They could instead be re-targeted on neurons which were simply less efficiently hedonistic in their biochemistry than their cellular neighbours. With their post-sensory signals remapped, infants could then learn self-preservation and pleasure-maximisation in harmony. At least as a stopgap, exploiting pleasure gradients is a much more civilised way to live. It's far more humane than responding to the contours of their nasty, and sometimes utterly excruciating, aversive counterparts.

A further presupposition of the question needs examining. One should be wary of assuming that we're the folk who can properly look after ourselves, whereas our

descendants, if they become genetically pre-programmed ecstatics, will get trapped in robot-serviced states of infantile dependence. For it shouldn't be forgotten that exuberantly happy people also have a fierce will to survive. They love life dearly. They take on daunting challenges against seemingly impossible odds. One of the hallmarks of many endogenous depressive states, on the other hand, is so-called behavioural despair. If one learns that apparently no amount of effort can rescue one from an aversive stimulus, then one tends to sink into a lethargic stupor. This syndrome of "learned helplessness" may persist even when the opportunity to escape from the nasty stimulus subsequently arises.

Contemporary fatalism about the "inevitability" of suffering is analogous to this dysfunctional passivity (*cf.* the behavioural syndrome associated with the religious traditions of the Indian subcontinent). Yet passive acceptance of the dark side of life is no longer useful to contemporary humans now we've unravelled the genetic code. Species-wide hedonic engineering offers the prospect of eliminating all the vile types of experience we hate most; but even though it has become technically feasible to escape their clutches, a lot of us still aren't energetically striving to get rid of them. Unlike tortured lab-rats and monkeys, we can verbally rationalise our perceived helplessness in the face of psychological trauma or malaise. Suffering, we say, is "natural", "inevitable", "the way of the world", "life", etc. By contrast, our eternally youthful, psychologically super-fit descendants won't need such coping mechanisms. They are likely to be fired up with indomitable will-power. Their resourcefulness and zest for living should make them far better equipped to deal with life's practical inconveniences. Potential problems will be viewed as tremendously exciting challenges to be overcome. But in any case, future generations of post-humans are destined to enjoy god-like powers unknown to the mythical Olympians - both inside their virtual reality software-suites and out. They may

indeed be ecstatically happy. But we would be rash to patronise them. For we're the ones who need help.

The argument that our descendants might become functional wireheads, too happy to reproduce, isn't compelling either. Happy people tend to want more sex, not less. Not everyone may opt for erotic modes of pleasure. But amongst sensualists who do, gene-coded hyper-dopaminergic well-being is likely to promote not celibacy, but heightened sexuality. This isn't simply a recipe for loveless orgies. Enriched serotonergic, phenylethylamine, oxytocin and opiate function will allow us to care much more for each other and our dependants than selfish DNA normally allows today. Just how many newly-minted young ecstatics the world can ecologically accommodate, on the other hand, is uncertain. The elimination of functional pathologies like the ageing process is likely to make curbing rampant reproduction rather than promoting it a priority.

---

4 "This whole manifesto is flawed from the outset by its crudely reductionist approach to human beings. Our most profound spiritual experiences, and indeed what it is to be a person, can't be reduced to a dance of soulless molecules."

In the tough-minded reductionist camp, a hard-nosed atheistical scientist may be loath to see the beautifully choreographed neurons of his temporal cortex reduced to a spiritual buzz of religiosity. This isn't a very fruitful perspective either.

In one's eagerness to avoid an impoverished conception of human beings, it is easy to fall victim to an impoverished conception of chemicals. Natural scientists, no less than humanists, can easily fall into the same trap. On the assumption that all conscious experience - "what-it's-like-ness" - is identical with certain physical events or properties, then our classical materialist image of the ontology of the physical world, and our

concept of what it means to be "physical", must be jettisoned as simply erroneous. It is not our fanciful mental images of matter and energy, but our deepening grasp of the formal mathematical tools needed for a description of quantum-mechanical events, that has enabled us increasingly to control and manipulate the basic stuff of the world. This grasp is now letting us control and manipulate, as well, the experiences with which at least some distributions of that "stuff" are identical. The phraseology sounds sinister and Orwellian. Yet if one's sovereign ethical principle entails striving for the fullest possible development of personal well-being everywhere, then embarking on the post-Darwinian enterprise is the only rational option.

---

5 "All of the drugs and therapeutic interventions touted here could potentially have long-term side-effects that we can't anticipate. The risk of another thalidomide tragedy writ large is too great to justify medical treatment of people who (by the norms of late twentieth century psychiatry, at least) are not suffering from any clinically recognised disorder."

The thalidomide tragedy took place several decades ago. The scandal unfolded before the medical significance of different optical isomers of the same compound in the body was appreciated. Such a mistake will not be made again. Of course, it can't be ruled out that other grave errors of judgement will be made instead. They probably will. In the early stages of any innovative treatment, the risk-reward ratio must always be finely weighed. This is all the more reason for preliminary experimentation to take place in the clinic and the laboratory, not on the street.

Presently, for instance, millions of young people are left to obtain and consume, in the most haphazard manner imaginable, the potentially neurotoxic compound MDMA.

"Ecstasy" typically offers an enchanting state of consciousness while the trip lasts. Yet it's a dangerous short-cut to mental health. Unless a subsequent dose of fluoxetine or another SSRI is taken soon afterwards, the drug damages serotonergic axonal terminals. Serotonin plays a vital role in regulating mood, impulse-control, anxiety and sleep. Thus in the long-term, MDMA and the other methoxylated amphetamines represent a poor choice of self-medication. It would be far better if the government were to take on the job of educating and training people in the most rational and effective ways to be happy. This role will involve sponsoring the research, development and widest possible distribution of the most safe, sustainable and beautiful empathetic euphoriants that medical science can formulate. Better still, research should focus on heritable gene-driven bliss. In the new reproductive era of "designer babies", prospective parents will choose the hedonic set-point of their future offspring. Curing our hereditary pathologies of mood will banish the need for drugs altogether.

---

6 "The radical therapeutic interventions which the biological program entails will presumably necessitate large-scale testing on non-human animals. This is surely inconsistent with the animal welfarist stance adopted earlier in the manifesto."

Given the feasibility, albeit not without difficulty, of implanting electrodes in the mind/brain's pleasure centres, there can be no principled utilitarian objection to subjecting both human and non-human animals to a great deal of enjoyment in the course of medical research. Many of the practical difficulties that the abolitionist project will face, and which demand greatest depth of understanding, stem precisely from avoiding crude pleasure-maximisation in the absence of a suitably well-designed encephalisation of emotion throughout the neo-cortex. If the animals in any experimental

procedure are kept exceedingly happy for its duration, then the utilitarian ethicist needn't suffer any qualms of principle. At present, of course, the difference between an animal-experimenter's laboratory and a torture chamber is often imperceptible from the victims' point of view.

---

7 "Abolishing suffering is unnatural: in so doing we would forfeit our essential humanity."

Warfare, rape, famine, pestilence, infanticide and child-abuse have existed since time immemorial. They are quite "natural", whether from a historical, cross-cultural or sociobiological perspective. The implicit, and usually highly selective, equation of the "natural" with the morally good is dangerously facile and simplistic. The popular inclination to ascribe some kind of benign wisdom to an anthropomorphised Mother Nature serves, in practice, only to legitimate all manner of unspeakable cruelties. Extremes of suffering are inevitable under the neurogenetic status quo.

If a personified nature did in some sense care about the progeny she prolifically churned out, then tampering with her benevolent handiwork might indeed represent a foolhardy tempting of providence. This sort of archaic romanticism about the natural world is impossible to reconcile with the neo-Darwinian synthesis. As has been all too aptly observed by "disposable soma" theorists, our genes just use us and then throw us away. "Unnatural" here is no more than a pejorative label. We use it to stigmatise, rather than rationally argue against, whatever we reflexively dislike. The very notion that a playing out of the laws of physics might ever yield something contrary to nature is itself deeply suspect. Construed in any literal sense, it is false. Nothing that occurs in nature is, or could be, unnatural. Both we and the transformed universe of our near and distant

posterity are equally a part of the natural world. Metaphorically interpreted, on the other hand, the charge of unnatural tampering is too ill-defined to be refutable.

And, yes, we will lose some primitive, "essential", human attributes. Yet why on earth should this be reckoned a bad thing? Until the development of powerful pain-killing drugs and modern surgical anaesthesiology, for example, frightful extremes of physical suffering were simply a part of the human condition. The unendurable just had to be lived through. Happily, in the present era our access to potent narcotics means, for the most part, that we no longer need to rationalise physical torments with the desperate sophistries typical of the past. Anyone arguing on religio-mystical grounds today that a loss of the agonies of the flesh is offensive to God, robbing us of a vital part of our species-essence, etc, is likely to get deservedly short shrift. Yet the supposedly ennobling properties of agonies of the spirit are still widely respected. Perhaps this attitude will change when retaining the capacity to feel psychological pain becomes a perverse genetic aberration rather than a condition of existence; and when inflicting it on others becomes an unthinkable crime.

8: "I'd get bored of being happy all the time. Variety is indispensable to personal well-being."

As an empty verbalism, "perpetual bliss" does sound fairly tedious. As Bernard Shaw once remarked, "Heaven, as conventionally conceived, is a place so inane, so dull, so useless, so miserable, that nobody has ever ventured to describe a whole day in heaven, though plenty of people have described a day at the seaside".

Successful paradise-engineering, however, must be the very antithesis of tedium by its very nature. If the prospect of paradise-engineering sounds unexciting, one has missed

the point of what abolishing the substrates of tedium entails. In a different age, religious iconographers were able to derive much greater satisfaction in depicting the tortures of the wicked in Hell than in evoking the curiously anaemic delights of Heaven. Indeed, one could be forgiven for inferring that the eternal happiness of the saved was dependent on contemplation of the eternal torment of the damned. Likewise today, the secular equivalent of this syndrome is all too common. Potentially, however, there is no less a diversity of ways of being happy as being wretched. It is a grim reflection of the late-Darwinian human predicament that any notion of perpetual happiness evokes images of monotony. We can conjure up a rich and never-ending diet of disasters with ease.

Whatever humanity's contemporary failures of imagination, within a few generations the experience of boredom will be neurophysiologically impossible. "Against boredom even the gods struggle in vain", said Nietzsche; but he failed to anticipate biotechnology. From a naturalistic perspective, boredom amounts to just a complex of psychophysical states whose molecular substrate natural selection has chanced upon like any other. A capacity for boredom was retained because of the adaptive value its conditional activation can confer. Its more proximate physiological basis lies in the negative feedback mechanisms underlying the development of tolerance in the brain. These may be expressed in the form either of short-term habituation or a slightly more delayed process of gene-triggered receptor re-regulation. Such mechanisms can be disabled and replaced.

For as is experimentally demonstrable in the laboratory, the intra-cranial strategy of endless stimulation of the pleasure-centres of the brain confirms that happiness, and happiness itself alone, never palls. Out in the wider world, positive emotion just gets (re)directed to focus on and infuse a variety of intentional objects. None of our neocortical patterns is inherently nice or nasty in the absence of its distinctive signature of limbic innervation. Some of these patterns may in time cease to satisfy; stone-age

love affairs are cruel. Given the mind-brain identity theory presupposed in this manifesto, however, there is no biological reason why each moment of one's existence couldn't have the impact of a breathtaking revelation. As the phenomena of déjà vu, and its rarer cousin jamais vu, strikingly attest, a sense of familiarity or novelty is dissociable from the previous presence or absence of any particular type of intentional object with which such feelings might more normally be associated. So the kind of thrill one might first have got witnessing, say, the Creation can in principle become a property of every second of one's life. Cool.

9: "In the light of past horrors, from Auschwitz to the most private of griefs, it is disgusting even to contemplate celebrating existence by getting perpetually blissed out of one's head. Happiness, and indeed any other emotional state or response, should be rationally justifiable. It should be experienced only when it is appropriate. Given the horrors existing elsewhere in space-time, pure bliss is rationally unwarranted."

If it doesn't diminish the well-being of others, does happiness stand in need of justification any more than does the experience of, say, redness? As long as there is any chance that what we construe as the lessons of history might be ignored, and the obscenities of our evolutionary past in some way re-enacted, then there are excellent ethical-utilitarian reasons for keeping accessible even the most dreadful of memories. It may be important to remember more recent history, too, so as to honour and be supportive of those who have suffered in it and are now plagued by memories of earlier traumas and sacrifice. Yet to enjoin a grim reflection on the nature of the past for its own sake, a form of melancholy which, self-consistently, must itself presumably be commemorated mournfully in turn, is to set in motion an escalating cycle of misery

without end. It's time to call a halt. Sometimes it is just better to forget rather than endlessly relive and recreate. If this sounds like shallow hedonism, it is worth recalling that HI's negative utilitarianism is an ethical system against which such a charge can least plausibly be sustained.

10: "I don't want a lifetime of enforced ecstasy. I want the freedom sometimes to be sad, and not to be enslaved to a false chemical happiness."

It is most unclear how to unpack the notion of "false" happiness. Contaminating the God-given purity of one's soul-stuff with alien chemicals is presumably offensive if one's self-conception is essentially spiritual in character. If, on the other hand, all states of consciousness alike are physically mediated, then it is scarcely coherent to label some neurochemical patterns as inherently false, unreal or inauthentic. Such euphoric states have indeed hitherto been largely inaccessible and genetically maladaptive if prolonged. They are still natural properties of suitably structured metabolic pathways of matter and energy. So in that sense they are all "true", though this is a most infelicitous way of putting it.

It is not, in any case, as though anyone will plausibly be forced to be happy against their will. Just as, historically, many slaves did not challenge the institutional legitimacy of slavery, and many self-confessed sinners believed they deserved to be damned to an eternity of torment in Hell, so many people have been able to convince themselves of the ennobling quality of suffering. They will scarcely be ambushed and hauled in off the streets one day by crack-demented ecstatics and forcibly pumped full of euphoriants. A more apposite question might be: what instruments of repression should a coercive state apparatus be entitled to use on behalf of possible bigoted die-hards of the old Darwinian

order against people who decide, reasonably enough, that they do wish to live happily ever after? To what degree, and for how long and in what form, should authoritarian reactionaries have the right to compel others to suffer, once emotional primitivism becomes simply one lifestyle option amongst many?

11: "Pharmacological hedonism would turn us all into junkies. Gene-driven hedonism wouldn't be any different. We would lose all personal freedom because we'd be as helplessly addicted to our chemical fixes as the typical crack-head."

Once one has tasted other-worldly transports of ecstasy, it is true, there is no foreseeable way one would choose voluntarily to renounce such a condition. For from our current perspective, we have no more grasp of the real glory of the sublime than a newly-instructed five-year old child has of all but the barest mechanics of love or sex. Does our absence of hyper-ecstatic experience entitle us to claim any greater authority than the precocious but naïve youngster? Is such a claim testable? In reality, the nature of what lies beyond the arid text displayed here will prove, on revelation, more wonderful than could currently be physiologically imagined. Enraptured, one will enter into whole new modes of being. Reality redefined will feel so good that any surrender of born-again existence would be unendurably traumatic.

This condition might seem almost definitive of addiction. Yet on a utilitarian metric (barring only the austere "negative" sub-species), if such marvellous states are reliably and universally accessible, then seeking to achieve and maximise them is straightforwardly the right course to take. Addiction will tend to be a problem only if, first, people are hooked on something noxious to themselves or others; or, second, there

is any likelihood of an interruption to their supply of the relevant drug or gene therapy. At present, we are dependent for what passes as mental health on different precursor amino-acids, essential fatty acids, minerals, vitamins etc to synthesise the brain's meagre dribble of pleasure-chemicals. We suffer gross psychophysical distress if we are deprived of them for long. This dependence, however, is regarded as wholesome rather than pernicious. It gets awarded the honorific "food". To achieve optimum mental health, on the other hand, one needs to dine on the richer diet of therapeutic agents advertised in this manifesto. The principle is the same.

The sheer finality of the Post-Darwinian Transition may indeed appal the metaphysical libertarian. For there can be no going back. Yet any opponent of the abolitionist project should be unsettled, too, by how endorsement of the traditional Nature-knows-best stance turns on our not exploring, however fleetingly, one of the two alternatives at issue. Ignorance is not bliss. Anyone who does empirically investigate, and not just pronounce on a priori, the rival forms of life on offer will unswervingly opt for the healthier modes of existence pleaded for here. More tellingly for the libertarian, perhaps, there is a sense in which the right to select one's own chemistry of consciousness, and thus to choose precisely who or what one wants to be, is as vital a sort of personal freedom as any. It is a freedom that we at present substantially lack. Any research program that opens up just such an option species-wide confers, surely, an incalculably life-enriching extension of choice.

Our own contemporary "choices" are in any case oversold. In the current era, we may seem relatively biologically unconstrained compared to our hidebound ancestors. Some of us feel we can be, and do, more or less who and what we want. In fact, we can subsist only within the largely insensible confines of an extremely restrictive state space of psychochemical reactions. We can't hop outside their metabolic pathways to check what

we're missing. If we could, we'd find the contrast too mind-wrenchingly different for words. Soon, however, we need no longer languish in biological servitude to our genes and the disposable vehicles they throw up. Today's junkies may vainly wish to be free from their inadvertently acquired addictions. This is only because the lows of illegal, dangerous and often self-defeating drug-taking ultimately outweigh the ephemeral highs of ill-chosen chemical euphoria. When, on the other hand, one opts once-and-for-all for an architecture of body-and-soul orgasmic sublimity, then one opts as well for a lifetime's freedom from second thoughts.

12: "I sometimes like being sad; it's an experience I wouldn't wish to lose."

An agreeable, wistful melancholy, a haunting lullaby nostalgically recalled from childhood, or perhaps the bitter-sweet memory of a long-lost love, are certainly preferable to the hell of unmitigated depression. Yet all too many types of experiences are unambiguously dreadful. They have no redeeming features at all. They don't issue in great works of art, literature and scholarship etc. They would be far better abolished. All the positive aspects of the more complex and ambivalent states one may undergo can in future be magnified and sharpened; nothing enjoyable need be lost. But the negative undercurrents which still diminish the value and enjoyment of more perceptibly composite states can be chemically subtracted out.

13: "Without suffering, there can be no personal development; unearned happiness leads to stasis."

Suffering is often just coarsening and brutalising. If one is sunk in hopeless despair, or even caught in the grip of an ill-defined malaise, it is as difficult to care about one's inner

growth as it is to care about other people. Personal growth is more likely to unfold if one's appetite for life gets steadily keener. This will occur if one's experiences get progressively richer and more rewarding. Odysseys of self-exploration across the hedonic landscape can offer scope for ever-deepening self-discovery and idealised self-reinvention. Odysseys of pain and misfortune are as likely to desensitise or crush one's spirit as develop it.

Under the grisly genetic status quo, cultivating a sense of personal development is a comforting form of rationalisation, e.g: if I hadn't lost my legs in the accident 20 years ago I would never have become a great artist. So it proved a blessing in disguise after all! Prospectively, however, if one were told 20 years of suffering lay ahead if one sacrificed one's legs, but boundless self-development would follow in consequence, then one still wouldn't opt for it; and quite rightly too. As long as suffering is biologically inevitable, fitfully at least, then its optimal rationalisation is important solace for its victims. Thus reading this manifesto may cause more distress than joy to inveterate rationalisers; I just trust any unease will be mild and temporary. Yet when the biochemistry of suffering becomes only an optional neural add-on, the solace that rationalisation provides will impede the abolition of the miseries that demand it.

14: "Why bother with this intentional flotsam and jetsam at all if happiness itself is supposedly the overriding goal? In the context of the biological program, aren't intentional objects really free-floating and inessential frills to be varied or discarded at will? Isn't invoking "sublimity", "beauty", "love", etc, intellectually dishonest? Aren't they just rhetorical camouflage to win over those whose ideal pleasures tend to the respectably cerebral and the ethereal rather than the orgiastic?"

Our emotions have been pretty thoroughly encephalised by evolution. So it is certainly easier to give some hint of the nature of the paradise that awaits us by evoking, one may hope, the feelings one's audience associates with their own most cherished fantasies and objects of desire. Advocating happiness bereft of any nominal focus, on the other hand, entails working with a lifeless and unpersuasive abstraction. Advocating "hedonism" in the abstract is even worse. The term evokes something shallow, one-dimensional and amoral. Unfortunately, that's the price of sacrificing an underlying seriousness of moral purpose for the sake of a snappy manifesto title.

Naturally, what we think and say we're happy "about" is likely to change as the transition to paradise-engineering unfolds. Many highly-charged intentional objects of contemporary desires will seem historical curiosities even a few decades hence. In common with the particular time- and culture-bound conceptions of heaven and the good life in, say, different eras of the Christian and Islamic traditions, today's favourite intentional objects may indeed be only of derivative value. The mesolimbic dopamine system is doing most of the real causal work. But if the lure of such idols can motivate us to act on the promise of the biological program, then they will have more than served their purpose.

There are, however, substantive reasons why non-arbitrary intentional objects, and indeed an ever-greater scientific understanding of the world, should remain accessible into the indefinite future. The pragmatic advantages of the intentionalist strategy compared to wirehead bliss have already been cited. Sometimes it's useful to be able to look after oneself. There are powerful ethical reasons for keeping intentionalism as well. For ethically it is imperative that the sort of unspeakable suffering characteristic of the last few hundred million years on Earth should never recur elsewhere. If such horror might exist anywhere else in the cosmos, presumably in the absence of practical

intelligence sufficiently evolved to eliminate its distal roots, then this suffering too must be systematically sought out. It needs to be extirpated just as hell-states will have been on Earth. Such inter-stellar rescue missions won't be possible if post-humans have all become wedded to the functional equivalent of wirehead-style pleasure-frenzies. This is because planning, executing and then stewarding ethically-run ecosystems of primordial extra-terrestrial life will require ultra-high technology, wide-ranging research, and a very long time. Subject to a number of assumptions about the origin of information-bearing self-replicators, any primordial life-forms - as distinct from some of their possible artificial successors - will be carbon-based. If multi-cellular evolution occurs, such alien life-forms will quite plausibly run on the same pleasure-pain axis as we do. Of course, this is all hugely speculative. And if trying to save the world is ambitious, then trying to save the universe smacks of hubris; so this avenue won't be pursued further here.

A negative utilitarian will still think that the striving for ever greater extremes and varieties of pleasurable experience while there remains any suffering whatsoever in this universe is a frivolous distraction from what morally matters. S/he may be right. Certain contrived scenarios aside, however, the direct genetic and intra-cranial routes to paradise may serve the different flavours of utilitarianism equally well.

15: "Many of the greatest scientific and artistic achievements of humanity were born of tremendous struggles against adversity. Abolishing the biological substrates of suffering would mean there could be no fruitful inner struggle or creative tension, and hence no more Newtons, Picassos or Beethovens. Scientific and artistic genius demands a capacity for fierce criticism, both of one's own work and the ideas of others. Even if inducing a state of perpetual

euphoria is consistent with bodily self-survival, the lack of critical self-insight such states entail would bring intellectual progress to a halt for ever."

It is worth distinguishing between the destiny of the humanities and the sciences after heaven has been biologically implemented. For a start, the exquisite aesthetic experiences on offer to our genetically enriched descendants may inspire an unprecedented flowering rather than a withering of the fine arts. Our current enjoyment of, say, Van Gogh's "Sunflowers" or Leonardo's "The Last Supper" will seem distracting tickles in comparison. Those who would deny that beauty is in the eye of the beholder might, or might not, be impressed by the disposition of paint on canvass which inspires these rhapsodies. Yet any reservations will last only so long as they remain trapped in the neurochemical orthodoxy of the past. At present, cultivating a fastidious unresponsiveness to certain forms of artistic production is taken as a badge of sophistication and discernment; but then that is our loss.

One blessing of the transcendent beauty awaiting discovery is that it will not depend on the vagaries of artistic genius for its production. The mind/brain lacks "beauty centres" of the same relatively well-defined architecture as its meso-limbic pleasure-system. Yet once the neurochemical signature of aesthetic appreciation is pieced together, its varieties can then be enhanced and selectively amplified. It should be recalled that perennial happiness can as easily lead to more being done in one's life rather than less. Intense episodes of creative energy today are often indistinguishable from mild euphoric hypomania. Some temperamentally laid-back lotus-eaters in the era ahead may indeed ultimately opt for meditative bliss and serenity. On the other hand, post-Transition society will probably be shaped by hypomanic "high-achievers" of formidable dynamism and productivity. Today's thrusting, can-do go-getters will seem lackadaisical in comparison.

The modes of well-being optimal for doing first-rate science and mathematics are obviously different from those best for practising first-rate art, poetry or sex. There is no reason why they should be less intense and rewarding. As to any lack of critical insight, there are also intellectual advantages to be derived from states of invincible well-being. Criticism of one's ideas in modern academia, for instance, is commonly taken as a full-frontal assault on the ego. In the future, critical scrutiny may be actively solicited and ecstatically welcomed. This might prove conducive to markedly better scholarship.

16: "The proposals of HI are too fanciful ever to gain credence, or even deserve serious critical consideration. They make a mockery of all our current values, aspirations and life-projects. A program so abhorrent to one's common-sense and moral intuitions belongs to the realm of vulgar science fiction rather than serious applied science or ethical debate."

Science has comprehensively confounded "common-sense" in all empirical matters. Our traditional ethical intuitions, when wrapped in secular guise, are less susceptible to experimental challenge. It would be a piece of singular good fortune if the least testable aspects of common-sense folk-wisdom just happened to be the ones that could most be relied on. At the very least, intellectual honesty demands that radically counter-intuitive challenges to received value-systems should receive close critical appraisal. The "values, aspirations and life-projects" typical of, say, classical antiquity or the Indian sub-continent may easily seem ridiculous to the jaundiced contemporary eye. Likewise, the disparate intentional objects with which our own well-being now seems inseparably bound may eventually be seen as no less superstitiously revered. They objectively matter, but only because they objectively matter to us. So on the assumption that ethics amounts to something more than truth-valueless word-spinning, then it is worth at least

considering the merits of ethical standpoints no less repugnant to common sense than, say, the theories of contemporary physics.

Appearances to the contrary, there is in any case a sense in which this paper, however superficially outlandish its substance, does not demand any revolutionary transformation of the core values of our secular culture. Its thrust stems from taking a quite conventional principle with the utmost seriousness it deserves. Only a minority of contemporary philosophers or laypeople are expressly utilitarians. Yet a diffuse and unsystematic utilitarianism is extremely widespread in society. It permeates the outlook of many people who never use the term. More interestingly, perhaps, an extraordinarily large proportion of non-, or even professedly anti-, utilitarian positions are argued on, or are underlain by, grounds which on examination prove subtly utilitarian.

Paradoxically, for utilitarian reasons it is nonetheless probably all to the good, this side of paradise at least, that at least some expressly non-utilitarian values are still held. This is because traditional folk-verities offset the acute discomfort many people still feel at the full implications of an exclusively utilitarian ethic.

Of course, one does not have to be a utilitarian to endorse the proposals of this manifesto. To those who are broadly sympathetic to the ethical utilitarian approach, however, then the biological program amounts, figuratively at least, to a gift from the gods.

17: "Being trapped in a chemical paradise would leave one wholly at the mercy of the ruling elites. The authorities could then treat people as puppets to be manipulated at will for their own ends."

The image that provokes this anxiety is presumably that of a drug-pacified class of helots. Perhaps a chemically enslaved underclass will work sweatshop hours for their masters simply to get their next chemical hit. In this fanciful scenario, it is in fact debatable who, if anyone, would really be exploiting whom. Also, certain sanctions are effective only if threatened rather than applied. No group is more ungovernably rebellious towards law and authority than addicts deprived of their fix. Moreover, in our society, the idea of the ruling elites engaging in a conspiracy to keep their population happy while they stoically shoulder the burdens of office tends to overtax the imagination; this is one conspiracy theory too far.

In any case, the conventional equation of happiness and docility owes more to distant memories of Huxley's *Brave New World* than to any deep reflection on the genetic, sociobiological and social-scientific literature. Prozac-style serotonin-enhancing mood-boosters, for instance, dramatically and consistently increase the status in the social pecking-order of the animals to whom they're administered. Such drugs may even lead them to reject a subordinate role altogether. It is revealing, too, that the manifestations of euphoric mania and melancholic depression also serve as descriptions of people occupying alpha and omega status-roles respectively. Mania, unlike most mental disorders, is most common in the upper social and economic classes. It typically involves an exaggeration of behaviour associated with achieving dominant status. By contrast, depression is most common among the poor. Even in today's society, the persistence of depressive states and behaviour fosters stable hierarchies of social dominance. From the perspective of evolutionary psychology, the typical depressive syndrome is part of an adaptive coping-process. "Endogenous" depression involves the passive submission to a prolonged or uncontrollable stress. The elevated levels of cortisol and pain-relieving beta-endorphin characteristic of official clinical depression are also those which promote

physiological adaptation to prolonged stressors. In the ancestral environment, depressive behaviour reduced the risk of physical damage by its tendency to reduce fighting within the group. In the post-Darwinian world, by contrast, depression simply won't exist.

So the "Brave New World" objection needs to be turned on its head. Given the correlation between depressed mood and low social status, the project of radically enriching the mood and motivation of the bulk of the population will probably leave people much less, not more, vulnerable to exploitation by a power-elite. In *Brave New World*, members of the populace were effectively the opiated and tranquillised dupes of the ruling authorities. Soma was a pacifying agent of social control. The consequences of genetically pre-programming happiness, however, will be very different. This is because everyday mental super-health will undermine the biological underpinnings of the dominance- and submission-relationships characteristic of our evolutionary past. More specifically, boosting the efficiency of tyrosine hydroxylase, for instance, won't just act to elevate mood. The consequently enhanced noradrenaline function in the locus coeruleus will tend to diminish subordinate behaviour. These simplistic "one neurochemical, one behaviour" stories are, of course, travesties of the truth, justified only on grounds of expository convenience. This doesn't challenge the essential point.

This point is that happiness, and an enhanced responsiveness to a wider range of rewards, is potentially hugely empowering. We're eternally slaves to the pleasure-pain axis; but a biologically enriched apparatus of pleasure and value-creation will help people assume a greater sense of control of their own lives. As noted, an all-action lifestyle fuelled by dopamine-driven well-being contrasts with the "learned helplessness" and "behavioural despair" characteristic of fatalists convinced that suffering is simply The Human Predicament. Either way, we shouldn't simple-mindedly project the power-and-submission relationships typical of early humans on the African savannah into the

indefinite future. For the genetic basis of our core repertoire of social behaviour will first be tweaked and then drastically recoded. Too many sci-fi romances rely on extrapolating primate dominance-rituals into the indefinite future. That's what makes sci-fi soap operas set in one million years time so curiously (and so spuriously) intelligible. Whereas over the next few millennia and beyond, we'll have the chance to leave endless re-enactments of the ritual power-plays of the ancestral environment ever further behind.

18: "I'd rather stay in touch with Reality than live in an escapist fantasy world."

Some people enjoy the lucky conviction they have more intimate relations with reality than the rest of us. A robust sense of intimacy is of course all the easier if one holds an agreeably commonsensical direct realist view of perception. Unfortunately, common sense is ill-named and at variance with the neuropsychological and quantum mechanical facts. Yet even a virtual worlder, for whom an awake mind/brain can aspire only to real-time data-driven simulations, may be sensitive to the charge of wanting to live in a fool's paradise, blissed out of his head, come-what-may. Better, surely, to live like a sad but wise Socrates than as a happy pig.

Happy pigs should not be despised, but Socratic intellectual heavyweights can be happy too. In a magically transfigured environment in which all one's fellow creatures were fabulously well, it is not clear at all why occupying an affectively neutral or pensive state should promote greater realism and representational fidelity. Perhaps the only way to grasp the actual nature of the unexplored celestial chemistry that beckons is to try becoming blissfully happy as well; and this is surely as good a reason as any for seeking maximal comprehension.

19: "Any creature which enjoyed perpetual bliss would no longer be me. I'm defined as much by my sorrows as my joys."

Winning £20 million on the national lottery, say, would wreak quite radical changes on most people's consciousness and sense of self-identity. It may nonetheless be suspected that the millions of punters who indulge their gambling streak are untroubled by the thought that their picking the lucky number will allow "somebody else" to enjoy the proceeds.

Philosophically, the notions of an enduring metaphysical ego, or for that matter of so-called "relative" identity, are indeed problematic if not incoherent. So in that sense the anxiety noted above is well-founded. Yet in such case any anxiety over personal (non-)identity applies no less to the psychochemical Dark Ages than to the post-Transfiguration era. One's namesake elsewhere in space-time who fell asleep last night is neither token nor even type-identical with the different configuration of matter and energy which bears one's name right now. Fortunately, even if personal identity is formally disavowed, one can normally muster the degree of altruism necessary to promote the future well-being of one's multiple namesakes, and likewise the namesakes and successors of one's family and friends. If contemporary notions of personal identity are ever culturally displaced by a different metaphysic, it may be hoped that our successors can muster the necessary degree of altruism too.

20: "When much of the world is still mired in poverty, hunger and disease, it is at best a flippant irrelevance to dream up hedonistic utopias. Their practice, if not aim, will be the cocooning of an already over-privileged planetary elite. We should instead concentrate on putting all our efforts into ensuring that

everyone in the Third World has enough to eat, clean water supplies, a decent education and medical care and a civilised standard of living."

By most objective indices of well-being (the rates of marital breakdown, crime, suicide, clinical depression and other forms of psychiatric illness etc), the urban-industrial Western elite scores poorly compared to the materially underprivileged masses of the Third World. So the relative good fortune of the inhabitants of liberal capitalist democracies is easily overstated.

An "us and them" approach to life has its limitations. Within the next few hundred years, the invidious distinctions of class, nationality and race which poison the contemporary world will become redundant. On all but the most optimistic projections, the great majority of the world's population aren't going to achieve First World lifestyles for the foreseeable future; but we most assuredly do have the resources to enable the whole planetary population to be magnificently happy. If, for a start, a minute fraction of the resources currently poured into zero-sum status-goods and consumer fripperies were diverted to researching the development of safe, cheap, effective mood brighteners, delayed-action designer euphoriants, and genetically pre-programmed mental super-health, then we would all be far better off. This is no less true of the jaded plutocrat than the impoverished Third World peasant.

21: "The idea of spending one's entire life consumed by whole-body-orgasmic states of hyper-crack-like intensity and euphoria is simply grotesque. It is an affront to human dignity."

Unbridled sensual bliss will be merely one of the flavours of pleasure on the psychochemical menu, though not one that should cause us any embarrassment. In our

own time, the dignified nature of such natural and short-lived routes to pleasure as sex is not always readily apparent to the untutored eye either. The more conspicuous pursuit of money, power and status characteristic of selfish DNA-driven civilisation tends to compromise human dignity in subtler but much more insidious ways. Champions of human dignity do not on the whole forswear such lifestyle choices, and understandably so; (in)dignity is very much in the eye of the beholder. Being made to suffer, however, is arguably the greatest indignity of all.

22: "The track-record of utopianism, whether romantic or allegedly scientific, is uniformly disastrous. Appalling crimes are committed on the assumption that the end justifies the means. A dystopian result is far more likely."

A "dystopia" where everyone is superlatively happy and fulfilled is surely the ultimate misnomer. Perhaps if one's concept of perennial happiness still evokes images of bland and sterile monotony, then the charge may seem reasonable. In fact, the worst coercive excesses one can imagine, albeit somewhat implausibly, from a notional regime of state-sponsored hedonism might stem from the imposed penal sanction of compulsory biological euphoria - perhaps objectionable, but scarcely a cruel (though certainly an unusual) punishment.

23: "Genetically pre-programmed euphoria would undermine the basis of all human relationships. All this fancy verbal window-dressing about combining perpetual ecstasy with love, empathy, beauty etc is only superficial. Say, for example, some terrible physical misfortune overtakes a friend; after all, accidents can happen in even the best-run utopias. One will still be ecstatically

happy: love for one's friend may indeed feel intense; but it is completely shallow if one can't grieve for a tragedy that befalls her."

By hypothesis, one's friend will be incapable of suffering; however badly mangled his or her body. Indeed s/he will still be happy, albeit, we shall assume here, less intensely than before. Perhaps some of her favourite pleasure-cells are damaged. Let us also assume, in this scenario, that the molecular substrates of volition have long since been identified and toned up. One has chosen to blend the biochemical substrates of pleasure with those of dopaminergic "incentive" motivation rather than blissed-out satiety. If this is the case, then one will strive with all one's prodigiously augmented will-power to find means to restore one's friend to a state of maximal well-being. One will try far harder in dopaminergic overdrive than would be psychophysiologically possible if one were stuck in one's current comparatively weak-willed and ineffectual state. Thus a life of unremitting happiness doesn't entail that friendship is shallow or inauthentic; on the contrary, one will have the motivational resources to express depth of personal commitment all the more.

This is not to say that relationships won't change in many different ways after the Transition occurs. At present, for example, friendship often consists of offering mutual support in times of hardship and despair. In future, it may consist of a shared celebration of life.

24: "One big risk posed by the global species-project of The Hedonistic Imperative is that (post)humanity will get "stuck" in a better, but perhaps still severely sub-optimal, state. Evolutionary progress, if one may be allowed to use

such a term, would thereby come to an end. This is too high a price to be paid, or to run the risk of paying."

This worry shouldn't be lightly dismissed. But perhaps three points are worth making here.

First, natural selection has promoted such an abundance of dreadful states that even a severely sub-optimal (by whose criteria? - presumably not the sublimely fulfilled super-beings themselves) result would ethically be far preferable to today's status quo; and indeed preferable to any of our often hellish world's environmentally-tweaked successors.

Second, the danger of getting irreversibly stuck is still present even if genetic engineering and psychopharmacology are renounced in favour of time-honoured "peripheralist" approaches to making the world a better place. In fact, for what it's worth, psychoactive drugs potentially offer a form of "simulated annealing" [in artificial neural network-speak], enabling us to escape entrapment in local minima - though sometimes the jolt may be too uncontrollably violent and even dangerous to be commonly useful e.g. taking psychedelic agents such as lysergic acid diethylamide (LSD), ketamine or DMT.

Third, the idea that the paradise-engineering project sketched in HI would more readily lead to us getting "stuck" stems, I think, from its conflation with one or both of its two immediate intellectual antecedents of which I'm aware. These are opiated-style quiescence à la Brave New World; and the endless, uncontrollably orgasmic lever-pressing frenzy of a rat-/human-driven pleasure-machine. Both stereotypes are deceptive. One consequence of enhancing dopamine function in the manner stressed in this manifesto is that not merely is overall motivation deepened, but also the range of different activities one finds rewarding is increased (*cf.* the recent excitement over

finding the D4 "novelty-loving" gene). Consequently, the likelihood of an organism or a species getting stuck in rut is diminished, though certainly not eliminated, by a strategy which incorporates boosting key receptor sub-types of dopamine-mediated process. It's worth noting that there is an experimentally demonstrable tendency of anti-dopaminergic mood-darkeners and flatteners, notably the D2-blocking major tranquillisers, to reduce incentive-motivation and novelty-seeking behaviour. They are "rut-inducers". Analogously, most of us Dark Age humans, stuck on a hedonic treadmill way down in the historical abyss, don't realise just how trapped we are.

On the other hand, there's a sense in which getting generically "stuck" in paradise is precisely what some of us are after.

25: "The eradication of suffering via genetic engineering and nanotechnology is an admirable goal. So why the disproportionate and perhaps (since so easily misinterpreted) irresponsible emphasis on mood-elevating drugs?"

Advanced genetic engineering and nanotechnological paradise-construction may yield states of conscious existence so wonderful and god-like that the notion of chemically fine-tuning them will seem absurd. What transhuman super-being would wish to contaminate the natural beauty of his or her soul-stuff with alien dirt? Yet some boring level-headedness about prospective time-scales is in order. It is true that the human genome of three-billion-odd base-pairs has now been decoded. A far greater problem for intelligently encephalised paradise-production is the combinatorial explosion issue. This arises, quite inevitably, from a genotype's differential expression in differing environments. Airily invoking "genetic algorithms" and "quantum computation", for

instance, is not wrong; but it tends to gloss over the formidable technical difficulties first to be overcome.

In the meantime, many people alive today will want biologically underwritten fulfilment for themselves and their loved ones. Born tantalisingly just prior to the Transitional era, they will have only the suspect stop-gap of enhancements to contemporary psychopharmacology to fall back on. Their access to cheap-and-cheerful paradises born of quick-and-dirty chemical fixes will, no doubt, seem dreadfully makeshift by the exalted lights of our more distant posterity. This doesn't mean that next century's pharmacotherapies should be damned with the knee-jerk invocation of "drugs" conjured up by our own era's ill-judged recreational excesses. For one of the paradoxical effects, for instance, of a mind-healing strategy using even present-day selective serotonin reuptake blockers can be an enhanced sense of undrugged "normality" in the user. Such a sense can coincide with a biographically abnormal brightening of mood. Unacknowledged everyday states of derealisation, depersonalisation, and indeed other modes of depressive weirdness more typically associated with "bad trips" and "bad drugs", are in fact disturbingly common. Low-grade forms are frequent even in the absence of any exogenous agent to precipitate them. Moreover it's worth recalling that a subjective sense of humdrum, drug-naïve normality is itself just a chemically-induced adaptation. Neither we nor our blissful descendants need feel at all "drugged"; even if, in a sense, that's what we are; and always have been. But if we want to glimpse, rather than talk about, the naturalistic implementation of Paradise, then our generation(s) at least will need to use psychoactive tools-of-the-trade to get there.

In any case, given that so much of our very essence is comprised by the chemical ingredients of our recent meals, it's not as though one's ontological integrity as a pure spirit-being, or whatever, will be under threat from alien soul-pollutants. The difference

between a drug and a nutrient, after all, reflects little more than the accidents of evolutionary history.

26: "The whole manifesto presupposes a Benthamite utilitarian ethic. If we don't accept its utilitarian presuppositions, then the abolitionist project collapses."

The abolitionist project isn't hostage to a single contested family of ethical theories. For it's not only utilitarians who abhor cruelty and suffering. Admittedly, the utilitarian may find it a matter of moral indifference whether our potentially ecstatic descendants opt to become wireheads, blissed-out junkies, or emotionally enriched post-Darwinian superminds. On the hypothetical felicific calculus, it's the sustainable intensity of our well-being (or the minimisation of malaise) that counts, not its peculiar flavours. But utilitarianism is a highly controversial ethic. So this manifesto, at least, lays stress on the quite extraordinary diversity of options for paradise-engineering. These options embrace a spectrum of intellectual, psychedelic, aesthetic, empathetic and even spiritual modes of well-being far richer than anything accessible today. There's no obvious moral imperative driving us to unrefined pleasure-maximisation culminating in a perpetual cosmic orgasm.

Nevertheless, many contemporary thinkers will balk at any form of scientific utopianism. It's not that non-utilitarian ethicists typically argue that the texture ("what it feels like") of unpleasantness is inherently valuable. Instead, most non-utilitarians believe that a capacity for mental distress as well as physical pain serves an important functional role in life itself - and it always will. The many faces of suffering have been harnessed by natural selection [or more traditionally, Divine Providence] to promote the plurality of values that

non-utilitarians uphold. Individual happiness is only one of those values. Much of what we care about isn't reducible to a unidimensional pleasure-pain axis.

Yet bioscience and nanotechnology promise more than the abolition of suffering and the enrichment of our emotional well-being. Critically, the new technologies allow us potentially to create the functional analogues of aversive states - analogue states that can play similar or even enhanced functional roles in the informational economy of an upgraded organism, but without the "raw feels" of suffering as we know it. Genetically constrained gradients of immense well-being - or smart neurochips with the right functional architecture - can be harnessed to animate our lives and promote what non-utilitarians typically value, but without the texture of subjective nastiness. If this prediction is borne out by the implementation of the new neurotechnologies, then the core of the secular anti-abolitionist case collapses. For only the most misanthropic nihilist would contend that despair, agony and malaise are inherently good. Suffering that serves no instrumental purpose at all, not even the interests of the genes whose inclusive fitness it once served, can be phased out without loss.

Of course, functionalist philosophy of mind may turn out to be wrong. As the functionalist alleges, minds may indeed implement the same computation/function in different ways and in different substrates, but perhaps effective nociception, say, must always have an unpleasant textural essence. Functionalism fails to explain the "hard problem" of consciousness; and our ignorance of why sentience (or anything at all) exists may infect everything else - including plans to get rid of suffering. It would seem very odd to claim that the texture of experience is functionally irrelevant or incidental to the role played by its biological substrates. For it's the sheer nastiness of suffering that ostensibly drives the abolitionist project in the first place. Yet we know we can build programmable silicon robots and embedded artificial neural networks to emulate the functional architecture of

organic life-forms: we already engineer robotic sensory capacities, basic "appetitive" states, and the behavioural capacity to avoid noxious stimuli in ways that mimic feats of conscious human agency but without the merest whiff of sentience. On the other hand, today's robots are still primitive in their capabilities; and bionic implants are barely in their infancy. We can't simply extrapolate present-day technical successes into the indefinite future. Perhaps, contra functionalism as understood today, a subjective texture of unpleasantness will prove functionally indispensable for, say, certain critical acts of judgement or discernment, or introspective self-examination. If these capacities are accorded a value potentially greater than the abolition of suffering, and if their subjective nastiness is functionally essential to the role they perform, then the abolitionist project may prove to have a more restricted appeal than the wider consensus canvassed here. If so, then seemingly abstruse debates about functionalist philosophy of mind would have an ethical significance beyond their technical merits.

Whatever the truth of functionalism, many non-utilitarian ethical positions are inconsistent with an abolitionist agenda; all the world's major religions for a start, with the ambiguous exception of Buddhism. Ethical systems that mandate the infliction of misery on other sentient beings against their will can't be reconciled with any form of paradise-engineering. But on the whole, religious and secular ethicists alike aren't so much hostile to abolitionism as simply oblivious to its very possibility. Jesus, Mohammed and Buddha didn't have anything to say on molecular genetics and nanotechnology. Indeed, it's only in the past few decades that the abolitionist project could be contemplated as technically feasible on Earth. Now that its blueprint can at least be formulated, all utilitarians should be abolitionists. But there's no need to turn utilitarian to endorse abolitionism: what's indispensable is an absence of malice.

27: "There will never be a Post-Darwinian Transition. There will always be selection pressure."

So long as there is ageing and death - i.e. for many centuries and perhaps millennia - there will indeed be selection pressure. But in the new reproductive era, the nature of that selection pressure will be different. In the old Darwinian era, "natural" selection is based on random genetic variations, i.e. genetic mutations that are random with respect to what is favoured by natural selection; and it is blind. Nature has no foresight. By contrast, post-Darwinian, "unnatural" selection will be neither blind nor random nor socially unregulated. For reproductive decisions will be taken by informed actors in anticipation of the likely neuropsychological effects of suites of alleles that are purposely pre-selected or designed. Genes predisposing to vicious traits that were adaptive in our Darwinian past will be at a selective disadvantage when we choose the attributes of our offspring, not through a cruel genetic lottery as at present, but by rational design.

The imminent arrival of cloning and designer babies brings profound ethical dilemmas of its own - not least because the new reproductive technologies will precede any post-abolitionist era of mature paradise-engineering. As life-span increases, and the ageing process is progressively defeated, will reproductive decisions remain the prerogative of individuals as now? Or will reproductive decisions be taken societally? All one's libertarian instincts will be alarmed at this prospect. But the carrying capacity of Earth won't allow more than 50 to 100 billion people at most. Either way, there will be selection pressure in the sense that some genes and behavioural dispositions will lose out, at least until we become quasi-immortals and reproduction effectively ceases.

Of course, this heralded post-Darwinian Transition might not be to a civilisation based on paradise-engineering. Post-Darwinian society may be based on something else altogether. Yet because the texture of suffering isn't adaptive per se, whatever its

current role in our legacy wetware, we can predict that the unsavoury genetic coalitions that manufacture its substrates will pass into evolutionary history.

28: "Paradise-engineering is impossible. It would not be evolutionarily stable. Game-theoretic modelling demonstrates that selfishness is always the most profitable strategy possible for replicating units - whether genes or "memes" - susceptible to invasion by "defectors". Invincibly happy life-forms are inherently more vulnerable than their discontented, anxious and malaise-driven counterparts. A society of genetically pre-programmed ecstatics could not arise, let alone endure. It would be an environment open to invasion by mean-spirited defector mutants who would replace the hardwired sweethearts. Unpleasant states of consciousness will last forever."

This objection conflates two issues. Could it ever be an evolutionarily stable strategy for our descendants to be 1) innately happy? 2) innately unselfish?

The answer to the first question depends on the sort of happiness hardwired. Are we modelling a civilisation of, say, quasi-immortal superminds animated by gradients of genetically programmed well-being? Or wireheads and their genetic equivalents - a "blissed out" rather than cerebral hedonism? Clearly, the option of global wireheading [or lifelong immersive virtual realities etc] isn't an evolutionarily stable strategy, at least until the ageing process is conquered. This is because wireheads have no inclination to breed and certainly not to raise children. By contrast, fitness-enhancing gradients of well-being - and traditionally, ill-being - or their functional analogues can serve to motivate, protect and preserve us. Such gradients are adaptive when they are "encephalised" by evolution - and ultimately, shaped by rational design. Uniform

euphoria [or chronic depression] and its insentient robotic analogues isn't adaptive. For this sort of functional architecture doesn't impel its subjects to do anything, learn anything - or nurture children. Either way, genetic fitness isn't inseparably tied to a particular texture of experience, but to the way we behave and reproduce.

The controversial answer to the second question - namely that it is today's hardwired quasi-sociopathy that will prove evolutionarily unstable - sounds woolly-minded and naive, not to say biologically illiterate. Surely a civilisation founded on blissful altruists can't amount to a viable strategy? "Hardwired sweetheart" scenarios aren't pivotal to the abolitionist project. They are also hugely more speculative. So why is blissful altruism an option for paradise-engineering worth exploring? Surely selfishness always wins?

Fortunately not. The (technical) genetic and metaphorical, behavioral and psychological senses of "selfish" are easy to confuse. This is because today they overlap so closely. Paradise-engineering can never be based on genetic unselfishness. But a genetic predisposition to altruism - in the metaphorical, behavioral and psychological senses of "altruistic" - can be evolutionarily stable against so-called defectors if and when it is also genetically selfish, i.e. Darwinian fitness-enhancing. This is how our capacity for kindness, compassion and empathy - however meagre - arose in the first place. Even today, a genetic predisposition to individual "saintliness" isn't always a losing strategy; recall the self-sacrificing holy man who attracts devoted female admirers and becomes the proverbial father of his nation. But on the whole, a capacity to cheat, to compete and to lie has proved adaptive; humans evolved as Machiavellian apes. Thus the proposal that unnatural selection pressure could ever cause "saintliness" to spread in a society of (non-clonal, genetically diverse) ecstatics looks implausible in practice. Surely alleles which promote competitiveness could never be outcompeted? Won't our descendants be, at best, happier egotists?

Now this may of course be the case. Yet decoding the human genome puts us on the brink of a major discontinuity in the mode of selection of self-replicating DNA - an evolutionary transition as profound as any in the history of life on earth. The long-term consequences of our capacity to rewrite our own code for the nature of adaptive - and maladaptive - traits may be very different from what we imagine. In the Darwinian era of "natural" selection, a regime of blind, random genetic variation typically promotes an indifference to the fate of most of our fellow genetic vehicles. In the environment of evolutionary adaptation, this predisposition enhanced the inclusive fitness of our DNA. We have a "theory of mind", but our minimal capacity for empathy is limited mostly to kith and kin. So callousness has flourished. "Nice guys" get eaten or outbred. Darwin himself spoke of "the clumsy, wasteful, blundering low and horridly cruel works of nature." By contrast, the impending post-Darwinian era of "unnatural" selection portends genotypes that will be pre-selected/designed in anticipation of their desired effects. So genetic variation will no longer be random and undirected. Its consequences will be collectively planned - imperfectly at first, then eventually perhaps via simulation and game-theoretic modelling with quantum supercomputers.

So questions of how we actually take the reproductive decisions, and on what criteria, are going to be crucial. What sort of traits do we want our offspring to have? Modelling post-Darwinian societies is immensely complex: post-humans may well rewrite their own individual genotypes ["genetic bootstrapping"] as well as the germ-line; and cloning will be trivially easy in the technical sense. Forms of "group selection" that simply weren't viable in the Darwinian Era become workable when reproductive decisions are collectivised; the "tragedy of the commons" can be forestalled. In a post-ageing world, reproduction may well be rare - and become progressively rarer as the carrying capacity of Earth [and ultimately the galaxy?] is reached. But taking a (very) crude genes' eye-

view, in the era of designer babies a variant allele coding for, say, enhanced love-and-nurturance-inducing oxytocin expression, or a sub-type of serotonin receptor etc predisposing to unselfishness in the metaphorical, behavioral and psychological senses, may be differentially pre-selected and customised in preference to alleles promoting, say, sexual jealousy, aggressiveness or sociopathic behaviour. Genetically influenced "altruistic" traits that carry a higher payoff in the technical selfish genetic sense aren't susceptible to invasion by mean-spirited "defector" mutants - even if genetic variation were to remain random rather than directed. Thus in generations to come, the genetic and non-genetic senses of the word "selfish" may diverge. Indeed, as the abolition of suffering becomes first technically feasible, and later trivially easy, then the language and institutions of traditional morality may become archaic relics from a vanished age. What sort of values will replace them is hard to say. But as our descendants rewrite the vertebrate genome, and redesign the global ecosystem via nanotechnology, harsh "unnatural" selection pressure may penalise the very sorts of nasty traits that were genetically adaptive in the Darwinian Era. On this analysis, post-Darwinian superminds will be extraordinarily benevolent; but paradoxically, the science of paradise-engineering will have its origins in genetic selfishness.

Perhaps. Let's take a more pessimistic scenario. Assume that (post)humans continue to be selfish in every sense. After all, just because allegedly we all (obliquely) seek happiness, this doesn't mean we seek happiness for everybody. Just because successful and intelligent life-forms will be able to underwrite their own happiness, why assume that they'll care about others? Let's further assume, contrary to the optimistic functionalist arguments above, that the textures of invincible happiness do inevitably make any coalition of alleles that promotes them potentially genetically vulnerable. After all,

invincible well-being wasn't a viable strategy on the African savannah; why should it triumph in an era of artificial selection?

Does this pessimistic set of assumptions predict the persistence of a legacy architecture of misery and malaise? Will unpleasant states of consciousness really last for ever?

No, not necessarily, not even then. The more vulnerable that enhanced well-being allegedly makes us, the more our self-interest will lie in ensuring that all others are happy and well-disposed too; and in ensuring that any novel life-forms we create in the new reproductive era are constitutionally happy and benevolent. If the discontent of others potentially threatens our own well-being, then genetically underwriting their empathetic bliss serves our self-interest. If mutant psychopaths pose a potential danger [though in fact strict sociopathy tends to diminish inclusive fitness even in the primordial Darwinian era], then self-interest dictates using prophylactic germ-line therapy against genes promoting sociopathy and its sub-syndromal variants; this is one state-space of genetic options whose full exploration we can live without. In the past, natural selection ensured that selfishness, in every sense of the word, frequently paid off. This entailed "winners" causing often severe suffering to losers. According to rank theory, the far greater incidence of the internalised correlate of the losing [behavioral] sub-routine, depression, compared to the winning sub-routine, euphoric (hypo)mania, attests to the terrible price that social animals have paid for the advantages of group living. Until now, blind genetic competition has ensured overt individual competitiveness among reproductive vehicles. There has been a sometimes physically violent struggle for the best mates and scarce resources. Winners and losers alike have been trapped on the same hedonic/dolorous treadmill. But when unlimited emotional well-being is possible for everyone at no cost to the well-being of others - and an unlimited diversity of good

experiences is accessible to all via immersive VR - then only sustained malevolence, not mere egoism, will suffice to perpetuate the cruelties of the old order.

None of this proves that our descendants will really be smarter, nicer and happier - the magic trinity predicted and endorsed here. This is scenario-spinning, not true game-theoretic modelling. There are suppressed premises and controversial assumptions in all the above arguments for paradise-engineering. Which strategies will really prove stable remains to be seen. The nature of the ultimate winning strategy is open. Certainly a transformation of human nature isn't going to arise through a world-wide spiritual awakening, an innovative package of socio-economic reforms, or a spontaneous desire to be nice to each other. But it's quite possible that, in the long run, the Darwinian genetic program based on suffering and quasi-sociopathy will lose out. Misery is not a stable strategy because by its nature rational agents seek to escape it; and soon a society of intelligent agents will have the collective capacity to do so.

29: "There is a contradiction at the heart of the abolitionist project. On the one hand, it is argued that suffering will be eradicated by biotechnology. On the other hand, it is claimed that no one will be forced to be happy: our freedom will allegedly be enhanced, not restricted, by the option of unlimited bliss. But perversely or otherwise, some people will always choose to be miserable - or at least to retain the traditional biological capacity to be so. Thus abolitionism can't be reconciled with an absence of compulsion."

Prescription and prediction are easily muddled. It is advocated that all involuntary suffering should be abolished. It is predicted that all suffering will be abolished. On this perspective, our descendants are no more likely to submit themselves to emotional pain

and malaise than we would today opt to undergo a major surgical operation without an anaesthetic.

In practice, an ethic of absolute personal freedom is probably untenable. Even the devout libertarian will sanction, say, the administration of a foul-tasting medicine to an unwilling sick youngster, or the forcible injection of an anaesthetic into a struggling animal before veterinary surgery. We sometimes override the choices and desires of simple minds. It would be cruel to do otherwise. Non-human animals, the severely mentally disabled and very young children don't know their own interests; mature adult humans are presumed different. The problem here is that super-intelligent extraterrestrials - or our own advanced descendants - may perceive us, primitive *Homo sapiens*, as comparatively no less mentally simple than are toddlers or pets in our eyes today. Any advanced intelligence may discern the analogous way that Darwinian minds are locked in dysfunctional cycles of self-abuse - unaware of our own interests. If so, then should we/small children be allowed to keep on hurting ourselves so badly?

As libertarians, we must presumably answer yes. This stance would seem hard to reconcile with a utilitarian ethic. For what are a few minutes of unpleasantness compared to an eternity of bliss? Yet even to moot the involuntary treatment of malcontents, let alone advocate its practice, is a dangerous line of argument for the abolitionist to pursue. For the misconception that anyone is going to coerce us into being happy is one of the biggest ideological obstacles to the future abolition of suffering. Fortunately, it is a mistake to believe that even a utilitarian ethicist is committed to mandatory therapy for the emotionally sick. This is because even the hint of compulsion causes distress to most people - thereby sabotaging the abolitionist project and defeating the utilitarian's own ends.

So the spectre of dissident emotional primitives being dragged kicking and screaming into the pleasure chambers must not become the defining image of abolitionist ideology. Conjuring up such a travesty of paradise-engineering doesn't show that a utilitarian ethic is mistaken. Instead it illustrates that the advocacy of compulsion is not a truly utilitarian policy at all. Like so many arguments against a utilitarian ethic, it relies on misconceived policy prescriptions wrongly derived from the sovereign utility-maximising principle.

In reality, abolitionists may call themselves fanatical libertarians on solid utilitarian grounds. For the freedom to transcend our Darwinian past and to choose our own homeostatic level of well-being is one of the most persuasive arguments for the abolitionist case.

30: "Why invoke nanotechnology? Surely genetic engineering alone can abolish suffering?"

If the abolitionist project is to be complete, then it must embrace the rest of the living world. In terrestrial ecosystems, the higher vertebrates can be genetically redesigned using foreseeable extensions of existing technologies. But pain and suffering will still fester in less accessible parts of the animal kingdom, e.g. in the oceans. Fortunately, within a few centuries, our descendants will have the capacity to use self-replicating nanobots armed with supercomputing power to redesign the marine ecosystem. Today, needless to say, this sounds like the wildest science fantasy. But even if we rely only on extrapolation, not revolutionary conceptual and technical breakthroughs, then the implementation of the abolitionist program is still grounded in relatively well-understood science. The reason that the prospect of molecular hedonic engineering hasn't yet been

explored by nanotechnology theorists is not that the technology involved is uniquely challenging. It's because tough-minded technocrats have different ends in mind.

In the present era, of course, it is hard to feel deeply exercised by the plight of marine invertebrates. We may feel that we have worries enough nearer home. But it is not pleasant to be eaten alive, even if one is a small mollusc. In paradise, it won't happen.

31: "Suppose that biotechnology really does give birth to an entirely new reproductive era. Suppose that humanity really is destined, as claimed in HI, for an era of ubiquitous designer babies - the so-called post-Darwinian transition. This transition may not be to an era of paradise-engineering. The biological basis of suffering may never be abolished. For if prospective parents are free to choose the attributes of their children, their typical priority will not be the creation of offspring who are innately happy. Instead, innumerable "pushy" parents will continue to seek children who are smarter, better-looking, competitively driven, more "successful" - and choose genotypes to match. Such parental bias can be explained, ultimately, by evolutionary psychology. At present, of course, prospective parents can't directly select allelic combinations of genes that promote such traits. In tomorrow's genetic supermarket, they may be granted an opportunity to do so. But if so, then selection pressure - albeit artificial or "unnatural" selection pressure - will favour exaggerated versions of traits that were adaptive in the old Darwinian era of natural selection. The outcome of the imminent reproductive revolution won't be a civilisation founded on genetically pre-programmed bliss."

Assume, plausibly, that within a few decades prospective parents will be able to choose the genetic dial settings for their kids' emotional well-being - the average "set-point" on our emotional thermostat around which well-being (or ill-being) tends to fluctuate. Grant too the key premise of the objection: many parents do indeed care far more about the worldly "success" of their children than they do about their children's personal (un)happiness. This doesn't entail that the substrates of suffering will be re-created indefinitely. Even parents for whom the emotional well-being of their offspring is trivial - of no more significance than, say, choice of eye colour - are still likely to opt for higher rather than lower dial settings on the hedonic treadmill i.e. alleles and allelic combinations that predispose their children to flourish. For most parents do prefer, on balance, their children to be temperamentally happy rather than miserable, even if happiness is only one desired attribute among many - perhaps not the most important - and in some instances perhaps only a minor or incidental trait. "I don't care what they do when they grow up, so long as they're happy" expresses, not a revolutionary sentiment, but a clichéd platitude of Western liberal society. This preference is explicable in part because happiness, and the spectrum of behavior associated with the "winning sub-routine", is positively correlated with social dominance and reproductive success. Ambitious parents certainly don't want to produce "losers". Depressive or anxiety-ridden kids can't compete effectively against their peers. A tendency to low mood, and the spectrum of subordinate behaviour with which depression is associated, may have been genetically adaptive for low-status tribal weaklings on the African savannah. For depressive behaviour, contingently activated, can be a viable fallback strategy for stressed low-status tribal animals in an adverse social environment. This may explain why depressive disorders are so common. But a genetic predisposition to low spirits, or at least anything like unipolar depression as distinct from bipolarity, is not part of an

optimal reproductive strategy for potential "winners". If intelligently engineered, a genetically enhanced sense of well-being is empowering. Its behavioural phenotypes are potentially far more adaptive than the predisposition to learned helplessness and behavioural despair characteristic of the depressive spectrum. So in the new reproductive era, pushy parents in particular are likely to shun depressive genotypes. What guise their children's well-being may take is another question. True emotional enrichment transcends the simple-minded recipes discussed here - mere modulations of the old Darwinian repertoire of sadness, happiness, disgust, fear, jealousy, anger and loneliness. Indeed the enriched emotional palette of our descendants may assume textures conceptually unimaginable to primordial Darwinian lifeforms. Our post-human successors may be rapturously happy about things we've never dreamed of, in ways we can't imagine, and in a conceptual scheme that hasn't yet been invented. But in today's terms, parents who are ambitious in a conventional sense for their family may seek an egoistic rather than empathetic kind of well-being for their children. Such parents may also favour (genotypes predisposing to) hypomanic exuberance rather than serene happiness. Backward-looking parents may even opt to endow their children with functional analogues of older Darwinian traits, but set against a much higher emotional baseline. None of this suggests that parents will opt, in the long run, for allelic combinations whose expression induces suffering or even unpleasantness in their carriers - even if medical ethics committees were to license their (re-)creation. Aside from anything else, children who are genetically predisposed to be depressive, sour-tempered or brattish are less rewarding to raise than children who are abundantly joyful and loving. Pre-selecting one of the nastier Darwinian genotypes for one's progeny would be self-defeating. In an era of artificial selection, the partially heritable bundle of traits we call "lovability" promises to be highly adaptive for (post)humans and their household pets alike.

The above account inevitably falls short on detail. Empirical cross-cultural studies of the (partially) heritable characters most favoured by contemporary parents for their offspring may serve as a better guide to the nature of tomorrow's designer babies. However, such a yardstick implausibly assumes an absence of state regulation and control over parental genetic choices. Likewise, the question of the future intensity settings of genetically pre-programmed happiness is here left open. Oversimplifying hugely, and treating happiness on a crude one-dimensional scale, will successive generations of genetically enriched (post)humans tend to be a bit happier, or blissfully happy, or orders of magnitude happier than their Darwinian ancestors, as predicted in HI? Most parents today, if pressed, might express a preference for their children to be very happy rather than happy; but only a minority of early adopters would opt for superkids who were constitutionally sublimely happy. Thus in the near future, the dial settings on enhanced kids' emotional thermostats will probably encode lives animated by (homeostatic gradients of) modest well-being rather than (homeostatic gradients of) sublime bliss. Analogously today, parents are typically most comfortable with the idea of rearing clever children rather than a family of geniuses. Yet as our conception of psychological health is enriched, so presumably will its socially acceptable norms. Ambitious parents usually aspire to a higher quality of life for their offspring than their own. This generalisation holds even though a comparative poverty of ambition may initially induce many parents to settle for comfortable mediocrity for their kids rather than mental superhealth. Perhaps this pleasure-deficit will be remedied in our lifetime by somatic gene therapy and genetically personalised mood-enrichers; perhaps not. But ultimately our descendants are no more likely to pre-select genotypes coding for inherently nasty states of mind than they are likely to pre-select genotypes coding for neuropathic pain. The historical record notwithstanding, human perversity has its limits.

32: "There is a flaw, possibly a fatal flaw, in HI. Yes, there probably will be a reproductive revolution. True, over time, prospective parents are unlikely to choose "nasty" genotypes for their children. Yes, this reproductive shift may even represent a major evolutionary transition in life on Earth. But, critically, a large percentage of the population will presumably continue to have children by "natural" means - whether out of bioconservative ideology, religious conviction, or just normal teenage fecklessness. Among this percentage of natural reproducers, a large and unknown number of couples will themselves be the offspring of natural methods of reproduction. Therefore a lot of the nastier code in our old Darwinian genome will be retained, together with the propensity to suffering it entails. Perhaps the natural reproducers will eventually interbreed with mature designer babies of more distant posterity. Who knows what will be the long-term consequences of mixing rational re-design and a legacy genome? But either way, unless the ideology of abolitionism is universally adopted as a value system - or ruthlessly enforced by a coercive state apparatus of unprecedented intrusiveness into the female body - then the global abolition of suffering will be postponed indefinitely. HI is a nice idea. But it's hard to see how it could work."

The key premise of the Objection is probably correct. So long as any pure-bred Darwinians continue to procreate by natural means, then suffering in some form or other will persist. The persistence of suffering is inevitable if archaic humans also reject as "unnatural" (etc) the other two core technologies of mood-enhancement, i.e. wireheading and sustainable pleasure drugs. So what grounds are there for believing that natural reproduction as practised today will ever cease? This is quite a radical prediction. And

even if the abolition of natural reproduction is technically feasible, isn't its disappearance too high a price to pay for mental superhealth and a cruelty-free world?

The reason for predicting that within a few centuries all human reproduction will be rigorously controlled, both in its timing and in its nature, stems from a second momentous technological revolution in prospect, namely the conquest of ageing. Whether you estimate that curing senescence will take another 100 years or 500 years, this genetic-cum-nanotechnological revolution is destined to sweep away the plague of human mortality. First on the horizon are interventions to prevent age-associated diseases (Alzheimer's, osteoporosis, cardiovascular disease, age-related memory decline, etc). Such primitive gene therapies are only the harbinger of a massive repair-and-renovation job on the human genome. This mega-project will tackle the fundamental biology of ageing itself. Replacing the biology of ageing is much more ambitious. Since rational design of the genome from scratch is impossible, we can only "bootstrap" our way to millennial lifespans - a formidable genetic challenge. But as the era of eternal youth unfolds, our descendants are not going to pre-select genotypes predisposing to ("for") age-associated diseases or senescence for their future offspring. Nor, realistically, are members of the older generation likely to shun rejuvenating somatic gene therapies for themselves. In consequence, the current slowdown in global population growth will reverse. The planet will fill up and approach the limits of its carrying capacity.

This physical constraint on our ability to multiply will recede but will stay intact even if you think we are destined to colonise the galaxy, or even if (fancifully and implausibly) you think we are going to "upload" ourselves onto computers, or even if you think the sky's the limit and intelligent life is limited in its expansion potential only by our world's Bekenstein bound. Even if individual mobility and resource consumption weren't an issue either, since we'll all be plugged into immersive VR or an analogue of the Matrix (etc),

then this physical constraint still holds: if we phase out ageing and become quasi-immortals, then we'll quite literally run out of Lebensraum in the absence of strict reproductive controls. The libertarian will find these words as uncomfortable to read as they are to write.

HI ducks the question of the specific social and biomedical mechanisms regulating reproduction in a post-ageing society. This omission is deliberate: control of human reproduction, whether sexual or clonal, will be a generic feature of any post-ageing civilisation. The need for social mechanisms of reproductive control on pain of Malthusian catastrophe isn't a specific peculiarity of the abolitionist project. If (post)humans aren't going to grow old and die, as we do today, then we can't go on having children at will indefinitely. A regime based on genetic Russian roulette will be replaced by an ethically responsible policy of planned parenthood.

At what cost? Other things being equal, state-regulated birth-control might be expected to cause widespread and profound personal distress. Only a small minority of people in human society are happy to remain childless. Infertility causes much heartache. For most people, having children is to a greater or lesser degree our raison d'être. For evolutionary reasons, it would be astonishing if this were other than the case. We may fear death and growing old; but typically what makes life meaningful - and our death bearable - is the lives of our children and grandchildren. Thus as we're constituted at present, the spectre of restrictions on our right to procreate is a disturbing idea. An intimate realm of our lives that has hitherto been essentially private could be in danger of intrusion by the state. Even a Chinese-style one-child campaign strikes the Western mind as a draconian curb on personal freedom.

So how will this dilemma be resolved? At present, we may try and persuade ourselves that we wouldn't want to stay eternally youthful. But if the option of eternal youth or

even its semblance were there, then it would be naïve to think most people wouldn't discard a lifetime of rationalisations and seize it. This bold statement might seem to imply a rather facile biotechnological determinism. For it is being assumed without argument that just because 1) we don't really want to grow old; and 2) technically it will be feasible to live indefinitely, we will therefore opt to do so - barring traumatic wetware accidents of course, though even here the use of prudent automated off-site self-backup policies should allow restores from last working copy. But for all its pitfalls, some sort of biotechnological determinism here is well-founded. Our fear of ageing, death and dying is simply too deeply rooted in the Darwinian psyche for us to perpetuate the senile holocaust into the era of mature genomic medicine. Renouncing the option of quasi-immortality may be conceivable in theory. Yet who'll opt to live (and die) as a disposable Darwinian "crumbly" if one can live and look like a Greek god?

The solution to the psychological dislocations such sustainable youth may entail is more likely to be biological than sociological. Just as biotechnology can potentially allow us to become better, more loving parents (e.g. by use of agents that induce oxytocin receptor gene overexpression, etc), so conversely biotech can curb the craving to have children when reproduction is infeasible. These techniques may be pharmacological or genetic or both. Godlike lifespans needn't have any adverse effects on our mental health; quite the reverse. Genetically enriched humans can feel utterly divine, not just look it. For lifelong well-being can potentially take many guises; and most forms of emotional enrichment won't entail living vicariously through the lives of our immediate biological descendants - natural as this habit of mind still seems in our late Darwinian world.

Switching on or off some of our deepest human desires sounds more like a dystopian nightmare than a recipe for paradise-engineering. Who is to orchestrate the switching; and how? No such hard choices are thrust upon us today. We just reproduce, decline into

our dotage and then die. Yet re-engineering the human mind and body alike can still strike even secular minds as almost sacrilegious. We admire excellence in the design of inorganic technology even as we abhor its prospect in ourselves. But whatever the mechanisms, if we cure ageing and don't intervene to regulate other primordial human traits as well, then intolerable psychological stress and social conflict are presumably inevitable. All sorts of ugly scenarios can be envisaged if life-extension technologies are pursued in isolation from mental health research and therapeutic interventions to match.

Nothing in this analysis of a post-ageing world proves that the control of (post)human reproduction also entails the design of psychologically superwell (post)humans. In overcoming ageing, it is possible, if sociologically unlikely, that we will opt to leave our repertoire of hunter-gatherer emotions unchanged - just as, conversely, it is technically possible we will conquer suffering without scrapping death and ageing. The response set out here aims rather to show why haphazard sexual reproduction isn't an inevitable fixture of tomorrow's post-Darwinian society; and how in future the creation of pain-ridden humans will demand an implausible measure of premeditation. So too, one day, may the creation of perishable human beings destined to grow old and die.

Yet just how likely in practice are our descendants to be eternally youthful, superintelligent, superempathetic - and to live happily ever after? A reality-check might seem in order. The post-ageing era is still far enough away to make any predictions hazardous. Those of us still in thrall to our Darwinian gut-instincts will find these scenarios all smack of wish-fulfilment and idle fantasy - mere fairy tales masquerading as science. HI certainly glosses over some very grim late Darwinian nastiness looming in the decades ahead: nuclear warfare, bioterrorism, global pandemics - and the usual soul-destroying tragedies of Darwinian-style personal life. Certainly, any futurology based on radical discontinuities rather than extrapolation rarely rings true at the time. But the

(potential) beauty of genetic engineering, quantum supercomputing and utopian nanotech is the way these technologies can be used to convert wishful thinking into sublime reality. What it means to be "realistic" will shortly be redefined. One reason for researching the prospects of a post-Darwinian civilisation is that paradise-engineering can deliver a practical solution to everything that's wrong with the world today.

33: "If (1) HI is correct, And if (2) HI should apply to all sentient beings, not just those on earth, Then (3) We have a moral obligation to spread throughout the universe as quickly as is practical, eliminating aversive experience and maximizing pleasure gradients everywhere.

Furthermore, if also (4) There are a very large number (let's say at least millions) of intelligent life forms elsewhere in the universe, Then (5) It's a virtual certainty that at least some of them (and more likely, most of them) are substantially more intelligent than us, And (6) It's a virtual certainty that at least some of them are at least equally driven to their goals, at least some subset of which are likely to apply to the entire universe.

We can subdivide the life forms mentioned in (6) into three categories: Category A consists of those life forms which have the same goals and choose the same means as HI. This sounds unlikely but might not be. Consider: If (7) morality is absolute rather than relative (i.e. there is some correct way to behave), and if (8) morality has attractors (i.e. most or all sufficiently intelligent life forms will discover the right way to behave and at least some of them will choose to behave that way), and if (1) then (9) at least some other life forms will find HI persuasive and will work toward it.

If (9) and (4), and if (10) the most advanced life forms are best equipped to determine and then carry out HI to maximize the chances of success, then (11) it's probably the case that there is no need for humans to get involved in HI. This logic isn't airtight, however. For example, if (12) all life forms reason this way, then none would act, assuming that some other life form would take care of HI (unless one or more life forms thought or knew that they were the most advanced). In addition, it might be the case that (13) the best implementation approach involves several life forms, not just the most advanced one (perhaps to accomplish the goals of HI more quickly). Nevertheless, it seems fairly clear that if (9) and (4), then it's highly unlikely that humanity is in the best position to implement universe-wide HI.

Category B consists of those life forms which have the same goals but choose different means than us. Some of the points in Category A would apply, but an additional conclusion given (5) seems to be that we should trust their judgement. This appears to be true even those life forms felt that the best approach included elimination of earthly life (and other similar life forms elsewhere).

Category C consists of those life forms which have different goals. If (6), then I believe that it is a virtual certainty that Category C is not empty; i.e., at least some life forms will have different goals than HI. If this is the case, and if (5), then it doesn't seem to matter much what we do, as the outcome will almost certainly be the goal of whichever life form is most advanced. This doesn't imply that (14) working toward earth-level HI goals is entirely pointless, but it does seem to substantially restrict the value of such efforts, making them local and temporary." [with thanks to Tom Murcko]"

Most people believe that the complete abolition of suffering in *Homo sapiens* is impossible. Extending the circle of compassion to other animals via ecosystem redesign and genetic engineering seems even more far-fetched. So the prospect of some kind of cosmic rescue mission to promote paradise engineering throughout the universe has a distinct air of science fiction. This may, of course, be the case. The timescales are certainly daunting even for a single galaxy of 400 billion stars some 100,000 light years across - on the order of millions or perhaps tens of millions of years. The level of intellectual, political and sociological cohesion over time required to mount such a project eclipses anything human society could organise today. Moreover, recent evidence from distant type Ia supernovae suggests that the expansion of the universe isn't slowing as hitherto supposed, but accelerating owing to poorly understood "dark energy". In consequence, perhaps only our local galactic supercluster will ever be accessible to our descendants.

Viewed purely as a technical challenge, however, the use of self-reproducing, autonomous robots - "von Neumann probes" - to explore and/or colonize our galaxy is both feasible and well-researched. The difference is that their purpose hasn't normally been conceived as a mercy mission for pain-ridden ecosystems that may have evolved elsewhere. [Ironically, notional "berserker probes" that sterilise all life have been discussed in science fiction, albeit not with a negative utilitarian ethic in mind.] Plausibility aside, it is ethically obligatory for utilitarians anywhere to maximise the well-being of all accessible sentience if it's technically feasible to do so - in the absence of any countervailing argument like the Objection above. Less clearly, an obligation to promote the substrates of well-being throughout the cosmos is arguably a disguised implication of various ethical systems that deplore merely "unnecessary" suffering. What "necessary suffering" might mean here is critical but ambiguous.

The most problematic premise in the Objection is perhaps number 4, i.e. the hypothetical existence of millions of other intelligent lifeforms. This assumption relies on the Drake equation or one of its variants in estimating the number of extraterrestrial civilizations with which we might come in contact. Any such assumption must overcome the Fermi paradox: "Where are they?" No discernible sign of extraterrestrial life exists - whether its artefacts, physical presence or signals. There may indeed be an indefinitely large number of technologically advanced civilisations in the Multiverse as a whole, or in other domains, or in other branes on "braneworld" scenarios, or even in our domain outside the "Hubble Bubble" [according to the chaotic inflationary universe scenario pioneered by physicist Andre Linde, quantum fluctuations divide the inflationary universe into a vast multitude of exponentially large domains or "mini-universes" where the laws of low-energy physics may be different]. Counterintuitively, as Max Tegmark points out, one popular cosmological model apparently predicts that each of us has an effectively identical twin in a galaxy typically around $10^{10^{28}}$ metres away. These distance scales are quite dizzying.

The point in this context is that even if we are unique to the known universe, we need not be "special" - which would entail a rejection of the normal Copernican assumption. If inaccessible civilisations do exist beyond our cosmic event horizon, their superintelligent inhabitants may well have transcended their evolutionary origins just as we are poised to do too. If such superbeings are benevolent, they will presumably [given "moral attractors"] rescue others physically accessible to being saved within their light-cone ("Category A"). It would be nice to think that cross-species deliverance from suffering was a universal law; the Objection raises the disturbing possibility ("Category C") that it isn't. The existence of hypothetical advanced lifeforms with the same goals as us but who choose different means ("Category B") might indeed shift the onus of responsibility away

from the junior civilization. Yet how common is the multiple independent origin of technologically advanced civilizations within a cosmically narrow (space)time-frame?

This is all extremely speculative. Extensive scanning of the electromagnetic spectrum discloses no evidence that technologically sophisticated life exists in our galaxy, or anywhere else in the observable universe. This absence of evidence extends to what Russian astrophysicist Nikolai Kardashev described as "Type III civilizations" - supercivilizations that would employ the energy resources of an entire galaxy. Their electromagnetic signature could in principle be detected by SETI (Search for ExtraTerrestrial Intelligence) researchers as well. Nothing has been found. The search continues.

Many explanations of "The Great Silence" have been mooted. Why assume, for instance, that intelligent extraterrestrials will manifest anything resembling the motives, values, conceptual framework or colonial expansionism of contemporary *Homo sapiens*? Is our conception of intelligent life and its signature too impoverished for us to have even located the relevant search-space to investigate? But (very) tentatively, the conservative explanation of why an immense ecological niche remains unfilled is that the silence is just what it seems. No technologically advanced, spacefaring civilisations exist within our few billion odd light years neighbourhood. It's up to us.

This conclusion doesn't mean we are locally alone. The Objection is right to take the status of sentient beings in other worlds extremely seriously. If we could really be confident that Earth-based organisms were the only lifeforms in the accessible universe, or if only minimally sentient microbial life exists in other worlds, then eliminating suffering on our planet would effectively discharge our ethical responsibilities. Once our world was cruelty-free, we could retreat into our own private nirvanas - or perhaps build heaven-on-earth and terraform it beyond. Yet it's also possible that complex life and

suffering - perhaps intense suffering - exists in alien ecosystems within our cosmic event horizon; and such lifeforms are impotent to do anything about their plight, i.e. they are as helpless as are all but one species on contemporary Earth. The presence of such malaise-ridden lifeforms would be undetectable to us with current technology. We have no empirical evidence of their existence one way or the other.

So how likely is such a scenario on theoretical grounds? Life's origins apparently lie early in Earth's 4.6 billion-year history. Deceptively perhaps, its rapid emergence suggests that the process may be relatively "easy" - and thus spontaneously repeated on a massive scale on Earth-like planets across the cosmos. Yet we still can't explain how the primeval "RNA world" preceding our DNA regime came into being. Nor can we yet synthesise life in vitro, or computationally simulate its genesis on Earth. So it's quite possible that only a freakish chain of circumstances allowed life to get started in the first instance. Piling improbable event on improbable event, another chain of contingent circumstances over several billion years allowed multicellular eukaryotic life to evolve. Eventually, life arose with the capacity to rewrite its own source code. It's unknown how many significantly different developmental pathways exist leading to organisms capable of scientific technology, or where the biggest evolutionary bottlenecks lie.

There is another imponderable here too. How likely is it that any primordial alien life will undergo suffering, or even be sentient, if its substrate differs from our familiar organic wetware? We know that our silicon (etc) robots can be programmed to exhibit the quasi-functional analogues of "mental" and "physical" pain and pleasure, and display a repertoire of "emotional" behaviour without any relevant "raw feels". Will putative extraterrestrials likewise be akin to zombie automata - "intelligent" or otherwise? [If so, would their fate matter?] Or more plausibly, will extraterrestrial life be sentient like us (or perhaps hypersentient)?

Here at least we can rationally speculate: the answer is probably the latter, though these modes of sentience may be very different. For there are powerful reasons for thinking that all primordial information-bearing self-replicators must be carbon-based, owing to the functionally unique valence properties of the carbon atom. Likewise, primordial life-supporting chemistries probably require liquid water. [If and when organic life becomes technologically advanced enough to build silicon robots, create "post-biological" digital life, design self-replicating nanobots, run "simulations" in quantum computers, etc, all bets are off.] If such primordial organic life ever reaches a multicellular stage, then the binary coding system of a pleasure-pain axis embedded in a nervous system is an informationally efficient solution to the challenges of the inner and outer environment, albeit brutishly cruel. So if hypothetical early alien life stumbled upon the molecular mechanisms underlying the pleasure-pain axis, then the information-processing role of its gradients will plausibly have been harnessed by natural selection to boost the inclusive fitness of self-propelled organisms - as it has on Earth. No "programmer" or designer is needed. Moreover, given the comparatively narrow range of habitats in the physical universe that could sustain primordial multicellular life, the phenomenon of convergent evolution may mean that all such life, wherever it evolves, isn't going to be quite so exotic as astrobiologists sometimes suppose. [By contrast, advanced life and consciousness could be unimaginably exotic.] If so, then the same abolitionist blueprint for ecosystem redesign and genomic rewrites should be applicable to other planetary biospheres - if we decide to intervene in Darwinian worlds rather than retain their ecological status quo.

That's a lot of ifs. Right now, it's difficult to care deeply about the plight of creatures who may not even exist, or who may be accessible only to our distant post-human descendants. Ecological charity, one feels, begins at home. Yet such indifference may be

a reflection of our limited psychology, not a moral argument for inertia. Naturally, we may all be mistaken in ways that exceed our conceptual resources to imagine or describe. Alternatively, something on the lines of the Objection may be correct. Certainly we rarely, if ever, understand the full ramifications of what we are doing. It's hard enough to plan ahead for the next five years, let alone envisage interstellar travel for the next five million. [This is one good reason not to get trapped in a rut of wirehead hedonism or its chemical counterparts rather than strive for superintelligent well-being.] Yet to opt for a deliberate policy of non-interference - whether in the lives of our suffering fellow humans, non-human animals, or primordial extraterrestrials - is no less morally fraught than paternalistic intervention. The argument that we should do nothing until we fully understand its implications cuts little ice in an emergency - and the horrors of a living world where babies get eaten alive by predators, creatures die of hunger, thirst, and cold, etc, must count as morally urgent on all but the most Disneyfied conception of Mother Nature. Analogously, it would be morally reckless for us to shun the use of, say, anaesthetics, pain-killers, veterinary interventions and similar "unnatural" novelties on the grounds that their use poses unknown risks - even though these risks surely exist and should be researched with all possible scientific rigour.

There are indeed ethical pitfalls in "playing God". These pitfalls would be even greater if [as the Objection assumes] there exist god-like extraterrestrial lifeforms better equipped than us to do so. Yet on both a domestic and cosmological scale, moral hazards exist for absentee landlords as well as for hands-on managers. Inaction can be culpable too. Here on Earth, there might seem a moral imperative to intervene and rescue, say, a drowning toddler on (almost) any ethical system at all. But what if that child grows up to be Hitler's grandfather (etc)? We can't know this, since we don't yet carry pocket felicific calculators. Yet the risk is presumably worth taking: we don't let the child drown.

Likewise, if your hand is in the fire, you withdraw it. If you are benevolent, then you do the same to rescue a small child or animal companion who is suffering similar agony - whether you are formally a utilitarian ethical theorist or not. The moral sceptic might argue that all value judgements are truth-valueless; but s/he can't argue consistently that we ought to believe this - or behave in one way rather than another. Taking the abolitionist project to the rest of the galaxy and beyond sounds crazy today; but it's the application of technology to a very homely moral precept writ large, not the outgrowth of a revolutionary new ethical theory. So long as sentient beings suffer extraordinary unpleasantness - whether on Earth or perhaps elsewhere - there is a presumptive case to eradicate such suffering wherever it is found.

34: "Why does HI lay such stress on gradients of well-being? From an ethical perspective, wouldn't a permanent maximum of bliss be better?"

A motivational system based entirely on heritable gradients of well-being is a less radical prospect than the abolition of motivation altogether. This is because hardwiring constant maximum bliss entails discarding the information-signalling role of the pleasure-pain axis completely - not just recalibrating its scale. Barring some extraordinarily advanced technology, uniformly happy beings will be out-reproduced. So for the foreseeable future, at any rate, encoding a physiological maximum of lifelong bliss is simply not an evolutionarily stable strategy. Then there's ideology to consider. If maximising gross cosmic happiness depends on (post)humans embracing a classical utilitarian value system, it's presumably an unlikely scenario on that score too. Pluralist or perhaps quasi-utilitarian value systems are more sociologically plausible. Yet HI's (tentative) forecast that a motivational regime of gradients of bliss will be conserved indefinitely is itself no more than a conjecture. One counterargument is that choosing less fulfilling states of

mind runs counter to the hedonic roots of our decision-making psychology itself. When mature technologies of emotional self-mastery become ubiquitous, it's uncertain who - if anyone - will really settle for what subjectively feels like an inferior option. What dial-settings will rational agents choose for their own mood-range when freed from the old Darwinian roulette? In practice, informed preference utilitarianism and classical utilitarianism tend to converge. Just possibly, the cumulative outcome of our choices may be the transcendence of traditional decision-making. As a slogan, "freedom to control one's emotions" invites readier assent than "freedom to enjoy limitless bliss". What's unclear is whether the ultimate cosmic outcome will be substantially different - or ethically, whether it ought to be so. Obviously, care should be taken here to separate normative judgement from positive prediction. Certainly, billions of years of pan-galactic hedonism isn't quite what Jeremy Bentham had in mind when first enunciating the greatest happiness principle. A lawyer by training, Bentham had in mind institutional and legislative reform. Yet harnessing biotechnology to a classical utilitarian ethic dictates saturating the cosmos with blissful euphoria/positive value and then computationally sustaining this theoretical maximum indefinitely - whether in the form of discrete superminds or perhaps a Borg-like collective mind. The logic of "hedonistic" utilitarianism is inexorable, even if its premises can be challenged.

The issue of whether we should encode hedonic gradients or constant happiness should be distinguished from the related question of so-called "higher" versus "lower" pleasures, i.e. the notional value of whatever we may be happy "about". Gradients of cerebral well-being (or ill-being) can certainly facilitate critical discernment, rational decision-making and motivated behaviour. Yet as our rapidly evolving computer software attests, neither qualia nor an organic substrate are essential to this functional role. So as our integration with intelligent software increases, the "texture" of subjective dips of bliss may turn out

to be functionally unnecessary for sentient organic life too. Tomorrow's technologies of fine-grained emotional control may enable early post-humans, for instance, to amplify their most treasured second-order desires for, say, cultural excellence, intellectual acumen and moral integrity while banishing the baser carnal passions. But after exploring the richest hedonic backdrop to whatever it is one most values - whether highbrow or lowbrow by today's lights - will anyone revert to hedonically impoverished states on discovering what they've been missing? Does our contemporary revulsion from crude wireheading, for instance, lie in the unvarying bliss that it yields - or merely its unedifying focus? Thus it's conceivable, as the Objection implies, that our distant descendants will enjoy some kind of ceaseless rapture - perhaps contemplating unimaginably sublime beauty or love or elegant mathematical equations. Or, less portentously, hilariously funny jokes. Naturally, these examples are purely illustrative, since post-humans may be imbued with kinds of blissful experience whose categories *Homo sapiens* can't name or conceive. Perhaps post-humans will be temperamentally meditative; perhaps dynamic. Perhaps they'll live in augmented organic virtual reality; or perhaps they'll live in designer VR paradises run on different bylaws from our presumptive basement. Perhaps they'll inherit a recognisable descendant of ordinary waking primate consciousness; or perhaps they'll live in unknown realms of utopian psychedelia. Unfortunately, our ignorance of the potential varieties of blissful experience contributes to the misconception that such well-being will necessarily be "thin" or unidimensional rather than diverse. But whatever the scenario, there's indeed no guarantee that a rational superintelligence will tolerate any decrements of well-being, information-signalling or otherwise.

The Objector's vision of unvarying bliss doesn't appeal to the dominant Western ethos. For the most part, modern capitalist societies prize innovation, creativity and change. So

the prospect of a civilisation based (merely) on gradients of extreme well-being may be less unsettling than a future of constant bliss - though either condition is alien to Darwinian life. We associate permanence with stagnation; and passivity with low motivation and malaise. So any "static" vision fails to inspire. From a broader evolutionary perspective, self-propelled bodies exhibiting goal-directed behaviour arose early in the history of multicellular life on earth. This architecture has been strongly conserved over hundreds of millions of years. Looking ahead to an era when intelligent life has conquered raw suffering, and to an era when we can modulate our core emotions at will, enhanced hedonic gradients and/or their functional analogues may lead our post-human descendants, and/or our intelligent robots/cyborgs, to radiate and colonize every niche of the accessible multiverse within our light cone/galactic supercluster and intelligently re-engineer it. But what then? The (hypothetical) discipline of secular eschatology won't always be the idle fancy it seems at present. After we can effectively ring the changes within the finite state-space of matter and energy in our cosmic neighbourhood, which kinds of supersentience will be judged worth instantiating? To use a lame analogy, will we opt endlessly to replay mediocre games of chess or painting-by-numbers? Or confine ourselves to the state-space of perfection? Is status quo bias as irrational in post-Darwinian paradise as it is in Darwinian purgatory? On the Objector's "constant bliss" scenario, everything formerly unpleasant or mediocre - from avoidance of noxious stimuli to the mundane maintenance of the infrastructure of civilisation - will presumably have been computationally "offloaded" onto our intelligent machines/prostheses. Critically, selection pressure will no longer operate since post-humans will have occupied every possible niche and engineered themselves to have become effectively immortal. The old era of frenetic "action" - the sound and fury of imperfect lives played out against a backdrop of restless discontent and scarcity

economics - will belong to our animalistic ancestry. Even the transitional era defined by gradients of cerebral euphoria will have been left behind. Quite possibly the molecular signature of all valuable experience will have been identified; and its substrates amplified to the fullest. Indeed, given the pleasure principle plus advanced technology, an evolutionary trajectory to the presumed attractor of ideal states of sentience may be inescapable. Once the transition to grown-up consciousness is complete, the theoretical possibility of venturing outside this state-space may be even less likely than, say, our now deciding to revisit the lives of savages in caves. If and when intelligent life reaches cosmic superheaven, perhaps the baroque scaffolding that got us there will be kicked away. Eternal bliss needn't be orgasmic in the sense of lacking all intentional objects beyond itself; but presumably even this must be an open question. Either way, "timeless" bliss doesn't have to feel static. Mastery of the neurochemistry of time perception may allow each here-and-now to have a vast temporal depth, rich internal dynamics, and subjectively to last an eternity. But perhaps speculations about the far future of cosmic consciousness are best avoided.

It should be stressed that all such wild post-Darwinian scenarios are remote - and vastly more speculative than the abolition of suffering or radical motivational enrichment. Hitherto in history, fitness-enhancing gradients of discontent have been the motor of progress - intellectually, socially, aesthetically, morally, personally. Most of the discontent endemic to the living world has indeed been unproductive; but not all of it. So harnessing the information-bearing role of its functional analogues - i.e. dips or anticipated dips of subjective well-being that still feel wonderful, but not sublime - is a more practical stopgap than encoding constant bliss. After all, we're barely on the eve of the reproductive revolution of designer babies, let alone an era of advanced paradise-engineering. In the near-to-medium term, recalibrating the genetic dial-settings that

regulate hedonic tone is a less challenging bioengineering task than offloading everything to smart machines and replacing the old motivational and affective homeostatic control mechanisms of organic life completely. Gradient-surfing is also more ideologically realistic. Moreover even on the more conservative gradients-of-bliss scenario, any subjective "cost" of hedonically sub-optimal states, i.e. information-signalling dips in well-being - is presumably acceptable to all but the most ardent utilitarian ideologues. Thus in future our hedonic baseline of mental health can still be richer than today's peak experiences. Assuming that the information-signalling role of gradients in well-being is indeed retained, any functional decrements of bliss can still be small. Even if the gradients are exceedingly subtle, there is no risk of a "Buridan's ass" scenario. [Buridan's ass was a mythical mediaeval equine which starved to death from indecision after being presented with the option of two equally appetising stacks of hay]. It's depressives who are prone to procrastinate; by contrast, happy people are typically decisive, extremely happy people more so. Indeed, HI predicts that our immediate descendants at least will not be "passively" uniformly happy, but hypermotivated, albeit on a much higher plateau of well-being than our current neural architecture can support. Enriching the reward centres of contemporary organic life will tend to heighten both its sense of purpose and purposeful behaviour - though to what end we don't know. Admittedly, this association of enhanced motivation with enhanced well-being may only be a contingent fact of our neural architecture - an accident of evolutionary history. The mesolimbic dopamine ("wanting") and mu opioid ("liking") neurotransmitter systems have co-evolved; their functional roles can in principle be disentangled. But again, a separation is scarcely imminent. (Post)human agency still has a long future.

Depending on the strength of our bioconservative prejudice, gradients of adaptive well-being needn't be heritable. In principle, designer drugs, neurochip implants, nanobots, or

autosomal gene therapy could achieve the same result - even within the constraints of a contemporary genome. But if our existing motivational system is defective, then it would seem cruel not to cure the pathology rather than transmit it to future generations. We wouldn't now consider it ethical deliberately to pass on genes for, say, a chronic pain syndrome on the grounds that our future pain-wracked offspring should be "free to choose" whether they wanted to be pain-free or not. Ethically, are our more pervasive syndromes of psychological malaise any different? Why shouldn't mental superhealth be heritable too?

How about the very long-term future? Normative judgements aside, will motivation in the traditional sense endure as long as sentient life itself? Could a future informational economy of mind based on gradients of bliss culminate in some sort of timeless cosmic paradise? Early in the 21st century, at any rate, this sort of question is probably too difficult to answer.

35: "Why the headlong rush to paradise engineering? Why not wait until we have the wisdom to understand the implications of what we're doing? Let's get it right."

We are faced with a "bootstrap" problem. Human beings may only ever be wise enough to understand the ramifications of what we're doing *after* we have enhanced ourselves sufficiently to be able to do so. Perhaps La Rochefoucauld was wiser than he knew: "No man is clever enough to know all the evil he does." Our species may take pains to avoid building a fools' paradise or some sort of Brave New World. But when, and by what means, will we ever be intelligent enough to be sure of succeeding? When will we be wise enough to avoid making mistakes that we haven't even conceived? As the reproductive,

infotech and nanotech revolutions unfold, (post)humans are bound to seek ways to make ourselves incrementally smarter. Does it really make sense to postpone a parallel emotional enrichment - assuming, naïvely, that emotional and cerebral intelligence could be so cleanly divorced? After all, narrowly-conceived intelligence-amplification carries risks of its own; greater wisdom may depend on emotional enrichment rather than being a prerequisite for it. For example, it transpires that genetically engineered "Doogie mice", endowed with an extra copy of the NR2B subtype of NMDA receptor, have not merely superior memories, but a chronically enhanced sensitivity to pain. Imagine if, prior to clinical trials, ambitious prospective human parents had rashly arranged to insert multiple copies of the gene in their designer babies to give them a future competitive advantage in education. The outcome might be pain-ridden child prodigies. Vastly more subtle and complex pitfalls doubtless lie ahead that make any steps towards a post-human civilisation problematic, not just paradise-engineering. If the risk-reward ratio of a proposed intervention is unfavourable, then clearly a potentially life-enriching drug, gene therapy (etc) shouldn't be rushed. But sometimes the risk-reward ratio is unclear. A more intractable problem is that some risks may be unknown, or inadequately quantified, or both.

So is the Objection essentially correct? Should we opt to conserve the genetic status quo of Darwinian life? Or at best defer the prospect of distinctively emotional enrichment to the presumed wisdom of our distant descendants?

Delay would be morally reckless for the following reason: ethically, even a non-negative utilitarian can agree that it's critical to distinguish between the relief of present suffering and the refinement of future bliss - between the moral urgency of the abolitionist project and the moral luxury of a (hypothetical) full-blown paradise-engineering. The risk-reward ratio of proposed interventions will shift as life on Earth gets progressively better - both

for an individual and for civilisation as a whole. We demand a far higher level of proven safety from an improved version of aspirin, for example, than from a potentially life-saving anti-AIDS drug. By parity of reasoning, the same yardstick should apply to their affective counterparts, the different forms of psychological distress. If, fancifully, we were already living in some kind of heaven-on-earth, or even just in a civilised, pain-free society, then it would indeed be foolish to put our well-being at risk by hazardous and premature enhancements designed to make life even better. Bioconservativism might be a wise policy. The Objection might then be tenable. Manifestly, we don't dwell anywhere of the sort.

Compare the introduction of pain-free surgery. In the pre-anaesthetic era, a surgical operation could be tantamount to torture. Patients frequently died. Survivors were often psychologically as well as physically scarred for life. Then a wholly unexpected breakthrough occurred. Within a year of William Morton's demonstration of general anaesthesia at Massachusetts General Hospital in 1846, ether and chloroform anaesthesia were being adopted in operating theatres across the world - in Europe, Asia and Australasia. Instead of embracing this utopian dream-come-true, would it have been wise to wait 30 years while conducting well-controlled trials to see if agents used as general anaesthetics caused delayed-onset brain damage, for instance? Ideally, yes. Should prospective studies have first been undertaken comparing the safety of ether versus chloroform? Again, yes - ideally. Rigorous longitudinal studies would have been more prudent. In the mid-19th Century, there were no professional anaesthesiologists, no balanced anaesthesia, no patient monitoring apparatus, muscle relaxants or endotracheal intubation. The mechanisms of anaesthesia in the central nervous system weren't understood at all. Nor, initially, were the principles of antiseptic surgery: only the combination of anaesthesia plus antisepsis could ever make surgery comparatively safe.

If the use of anaesthetics had led to delayed-onset long-term brain damage (etc), then the medical doubters might now be hailed as uncommonly prescient - instead of enduring the "enormous condescension of posterity", relegated to a footnote in our incorrigibly Whiggish potted histories of medicine.

Despite these caveats, the world-wide introduction of general anaesthesia in surgery is, by common consent, one of the greatest triumphs of medical history. Why the precipitate haste of its adoption? In essence, anaesthetic use spread rapidly across the world because the horrors of extreme physical pain entailed by surgery without anaesthesia were judged by most (but not all) physicians and their patients to outweigh the potential risks - even though the risks weren't properly known or adequately quantified. Surgeons, too, were able thereafter to attempt ambitious life-saving interventions that were effectively impossible before. By our lights, early anaesthesia was appallingly crude, just as narcotic analgesia remains to this day. But the moral urgency of getting rid of suffering - whether its guise is "physical" or "mental" or both - is obscure only to those not caught in its grip. This is why almost everyone will "break" under torture; and why, globally, hundreds of thousands of depressed people take their own lives each year: in fact "mental" pain effectively kills more people than its nominally physical counterpart. If one is looking for historical role-models, then perhaps Dr John Snow - "the man who made anaesthesia a science" - may serve as an exemplar. As the use of surgical anaesthesia spread like wildfire in the late 1840s, Snow didn't advocate the "safe", bioconservative option of abstinence or delay. That would have been callous. But unlike some of his more gung-ho medical colleagues, Snow was mindful of the potential risks of the seemingly miraculous discovery. His introduction of standardised dosing through efficient inhalers and careful patient monitoring saved many lives. Moral urgency is not a license for recklessness.

Like most analogies, this one is far from exact. Currently millions of sentient creatures, human and non-human, are indeed stricken by suffering no less grievous than patients in the pre-anaesthetic, pre-opioid analgesic era; and likewise, exciting but largely unproven technologies exist to remedy their plight. So to that extent, the historical parallel holds. But statistically, most people are not in the throes of extreme psychological distress. Thus if one is currently relatively satisfied with one's life, and if one's dependants are relatively satisfied too, then there are strong grounds for caution over experimenting with ill-tested interventions that promise to enhance one's existing well-being. Thus the advent of a putative sustainable mood-enricher to reset one's emotional thermostat, a novel intellect-sparing serenic to banish unwanted anxiety, an illuminating new psychedelic, a super-empathogen, a genius-pill (or whatever) might represent a tantalizing prospect. Yet they should presumably undergo rigorous prior testing before general public licensing - however dazzling the anticipated benefits. It might seem that delay is the only responsible option; there can be wisdom in inaction.

The pitfall to this "safety-first" approach lies in the extreme risk of moral complacency it breeds. Hundreds of millions of human beings, and billions of non-human animals, are not in such a fortunate position. On a universalist utilitarian ethic, or simply a Buddhist-style ethic of compassion, we should systematically apply the same level of urgency to relieving their suffering as one would be justified in exercising if one were oneself tormented by intense pain or suicidal despair. Extreme suffering is the plight of billions of sentient beings alive today, whether in our factory-farms, in a Darwinian state of nature, or in a depressed neighbour. Desperate straits mandate taking risks one would otherwise shun.

On the face of it, if one aims to lead a cruelty-free lifestyle, one may disclaim personal complicity in such suffering. But this moral opt-out clause may be delusive. Simply by

deciding to have genetically unenriched children, for instance, one perpetuates the biology of suffering by bringing more code for its substrates into world. A healthy caution toward untested novelties should not collapse into status quo bias.

Any plea, then, for institutionalized risk-assessment, beefed-up bioethics panels, academic review bodies, worse-case scenario planning, more intensive computer simulations, systematic long-term planning and the institutionalized study of existential risks is admirable. But so is urgent action to combat the global pandemic of suffering. "The easiest pain to bear is someone else's".

36: "HI claims that once the biological substrates of suffering have been abolished, it is 'inconceivable' that suffering will ever be recreated. But this isn't so. According to the Simulation Argument, there is a significant likelihood that we ourselves are living in an ancestor-simulation run by our advanced descendants. If this is the case, then our simulated status entails that posthumans will not eradicate suffering. The Simulation Argument implies that our descendants will re-introduce suffering via their ancestor-simulations, or they never opted to abolish suffering in the first instance."

[http://www.simulation-argument]

The Simulation Argument (SA) is perhaps the first interesting argument for the existence of a Creator in 2000 years. It is worth noting that SA is distinct from the traditional sceptical challenge of how one can ever know that one's senses aren't being manipulated by an evil Cartesian demon, or be sure that one isn't just a brain in a nefarious neurosurgeon's vat, and so forth. SA is also distinct from the controversial but non-sceptical inferential realist theory of perception: inferential realists believe that each of us

lives in egocentric simulations of the natural world run by a real organic computer i.e. the mind-brain. Instead, SA claims that given exponential growth in computing processing power and storage capacity, the entire universe as commonly understood could be a simulation run on an ultrapowerful computer built by our distant descendants. We may really be living in one of posterity's versions of The Matrix. SA's important subtlety - the subtlety that catapults SA from idle philosophical fancy to serious scientific metaphysics - is that if multiple ancestor-simulations are destined to be created whose inhabitants are subjectively indistinguishable from ourselves, then statistically it is much more likely that we are living with the great majority in one of these indistinguishable simulations rather than with the minority in pre-simulation Reality. Or rather, SA concludes that at least one of the following three propositions must be true: 1. Almost all civilisations at our level of development become extinct before becoming technologically mature; 2. The fraction of technologically mature civilizations that are interested in creating ancestor-simulations is almost zero; 3. You are almost certainly living in a computer simulation. Actually, SA's proposed trilemma may shortly be simplified. The first of SA's three disjuncts, the extinction scenario, can be effectively excluded within a century or two - an exclusion that ostensibly increases the likelihood one is living in a cosmic mega-simulation. For humans are poised to colonise worlds beyond the home planet, thereby rendering global thermonuclear war, giant asteroid impacts, a nanotech "grey goo" incident, superlethal viral pandemics and other Earth-ravaging catastrophes impotent to extinguish intelligent life. Even on the most apocalyptic end-of-the-world prophecies, intelligent life will presumably survive in at least low-density branches of the universal wave function. In the far future, superintelligent posthumans may at some stage mass-produce ancestor-simulations. If so, these computer simulations of ancestral life may include billions of

human primates whose inner lives, the simulation hypothesis suggests, may be subjectively indistinguishable from our own.

What should we make of this? First, a familiar sociological point. The dominant technology of an age typically supplies its root-metaphor of mind - and often its root-metaphor of Life, The Universe and Everything. Currently our dominant technology is the digital computer. We may have finally struck lucky. Yet what digital computers have to tell us about the ultimate mysteries of consciousness and existence remains elusive. At any rate, no attempt will be made here exhaustively to discuss SA except insofar as its conclusion impacts on the abolition of suffering. But it's first worth raising a few doubts about the technical feasibility of any kind of simulation hypothesis. These doubts will then be set aside to consider the likelihood that a notional superintelligence that did have the computing technology to run full-blown ancestor-simulations would ever choose to do so.

One problem with SA is that it rests on a philosophical premise for which there is no evidence, namely the substrate-independence of qualia - the introspectively accessible "raw feels" of our mental lives. This premise is probably best rephrased as the substrate-neutrality or substrate-invariance of qualia: SA functionalism doesn't claim that the colours, sounds, smells, emotions, etc, of subjective first-person consciousness can be free-floating, merely that any substrate that can "implement" the computations performed by our neural networks will conserve the textures of human experience. The substrate-neutrality assumption is intended to rule out a [seemingly] arbitrary "carbon chauvinism": take care of the computations, so to speak, and the qualia will take care of themselves. SA aims to quantify the likelihood of our living in an ancestor-simulation with a principle of indifference: the probability that we are living in a simulated universe rather than primordial Reality is equal to the fraction of all people that are actually simulated people. Critically for the argument, SA assumes the subjective

indistinguishability of "real" from hypothetical post-biological "simulated" experiences. SA proposes that the power of posthuman supercomputers may allow vastly more simulated copies of people to exist than ever walked the Earth in the ancestral population. This is because once a single "master program" is written, copying its ancestor-files is trivially easy if storage space is available. Hence SA's claim that if posthumans ever run ancestor-simulations, then we are almost certainly in one of them. But here is the rub. The prior probability to be assigned to our living in a simulated universe depends on the probability one assigns to the existence of superadvanced civilisations that are both able and willing to create multitudes of sentience-supporting ancestor-simulations. And there is simply no evidence that such computationally simulated virtual "people", if they ever exist, will be endowed with phenomenal consciousness - any more than computationally simulated hurricanes feel wet. SA postulates that consciousness will supervene or "result" from supercomputer programs emulating organic mind/brains with the right causal-functional organization at some suitably fine-grained level of detail. The physical substrates of the putative supercomputer used to simulate sentient creatures like us will supposedly influence our kinds of consciousness only via their influence on computational activities. But it's worth noting that silicon etc robots/computers can already emulate and exceed human performance in many domain-specific fields of expertise without any hint of consciousness. It's unclear how or why generalising or extending this performance-gap will switch on inorganic sentience - short of the physical "bionization" of our robots/computers via organic implants. Without qualia, we ourselves would just be brainy zombies; yet qualia are neither necessary nor sufficient for the manifestation of behavioural intelligence. Thus some very stupid organic creatures suffer horribly. Some very smart silicon systems and digital sims aren't sentient; they can defeat the human

world-champion at chess. We're clearly missing something: but where are we going wrong?

For SA to work in the absence of a scientific explanation of consciousness, some kind of cross-substrate qualia conservation postulate must be assumed on faith. Yet if phenomenal consciousness is really feasible in other substrates or virtual machines, does this synthetic consciousness have the same generic texture as ours - or might not synthetic consciousness be as different as is waking from dreaming (or LSD-like) consciousness? Assuming conscious minds can be "implemented", "uploaded" or "emulated" in other substrates, what grounds are there for supposing that the uploads/simulated minds retain all, or any, particular qualia at every virtual level - assuming their specific textures are as computationally incidental to the mind as are the specific compositions of the pieces in a game of chess? Granted that biological minds can be scanned, digitized and uploaded to/simulated in another medium, will the hypothetical sentience generated be sub-atomic, nano-, micro-, (or pan-galactic?) in scale? Can abstract virtual machines really generate spatio-temporally located modes of consciousness? Are multiple layers of qualia supposed to be generated by virtual beings in a nested hierarchy of simulations? Are the stacked qualia supposed to be epiphenomenal, i.e. without causal effect; if so, what causes subjects like us to refer to their existence? By what mechanism? If ancestor-simulations are being run, then what grounds exist for assuming the conservation of type-identical qualia across multiple layers of abstraction? Are these layers of computational abstraction supposed to be strict or, more realistically, "leaky"? SA undercuts the [ontological] unity of science by treating Reality as though it literally has levels. Yet there is no evidence that virtual machines could have the causal power to generate real qualia; and the existence of "virtual" qualia would be a contradiction-in-terms.

None of the above considerations entail that phenomenal consciousness or unitary conscious minds are substrate-specific. Perhaps the problem is that there are microfunctional differences between organic and silicon etc computers/robots - microfunctional differences that our putative Simulators might emulate on their supercomputers with software that captures the fine-grained functionality which coarser-gained simulations omit. After all, it's question-begging to describe carbon merely as a "substrate". The carbon atom has functionally unique valence properties and a unique chemistry. The only primordial information-bearing self-replicators in the natural world are organic precisely in virtue of carbon's functional uniqueness. Perhaps the functional uniqueness of organic macromolecules extends to biological sentience. These microfunctional differences may be computationally irrelevant or inessential to a game of chess; but not in other realms. Suppose, for example, that the binding problem [i.e. how the unity of conscious perception is generated by the distributed activities of the brain] and the unitary experiential manifolds of waking/dreaming experience can be explained only by invoking quantum-coherent states in organic mind-brains. Admittedly, this hypothesis resolves the Hard Problem of consciousness only if one grants a monistic idealism/panpsychism that most scientists would find too high a price to swallow. But on this account, the fundamental difference between conscious biological minds and silicon etc computers is that conscious minds are quantum-coherent entities, whereas silicon etc computers (and brains in a dreamless sleep, etc) are effectively mere classical aggregates of microqualia. Counterintuitively, a naturalistic panpsychism actually entails that silicon etc robots are zombies.

A proponent of the simulation hypothesis might respond: So what? A functionally unique organic neurochemistry needn't pose an insurmountable problem for a Simulator. After all, there is no reason to suppose that a classical computer can't formally calculate

anything computable on a quantum computer, since (complications aside) a quantum computer is computationally equivalent to a Turing machine, albeit hugely faster. So if silicon etc supercomputers could simulate biological mind-brains with their putative quantum-coherence as well, then qualia might still "emerge" at this layer of abstraction. The technicalities of SA's original, classical formulation aren't essential to the validity of its argument. SA still works if it's recast and the organic mind/brain is a quantum computer. The snag is that this defence of SA conflates the simulation of extrinsic and intrinsic properties: formal input-output relationships and the felt textures of experience. Computational activity that takes milliseconds will not feel the same as computational activity that takes millennia - quite aside from any substrate-specific differences in texture or absence thereof. If quantum coherence is the signature of conscious mind, then conscious biological minds are implicated in the fundamental hardware of the universe itself - the computationally expensive, program-resistant stuff of the world. As David Deutsch has stressed, the computations of a quantum computer must be done somewhere. If our minds by their very nature tap into the quantum substrate of basement reality, then this dependence undercuts the grounds for believing that we are statistically likely to inhabit an ancestor-simulation - though it doesn't exclude traditional brain-in-a-vat style scepticism.

Of course, none of the above reasoning is decisive. We simply don't understand consciousness. Many scientists and philosophers would dispute that quantum theory is even relevant to the problem. Or perhaps we are simulated quantum mind/brains running on a post-silicon quantum supercomputer. Or perhaps the laws of quantum mechanics itself are an artefact of our simulation in some kind of posthuman "computronium". Who knows. Here we are veering into more radical forms of scepticism. But if insentient simulations of humans (etc) are feasible, then one may reasonably

doubt all three disjuncts of SA. Maybe neither the premises nor the conclusions of SA are true. Intelligent life is not headed for extinction. Some of our descendants may conceivably run multiple ancestor-simulations in low-density branches of the universal wave function. It is exceedingly unlikely that we are participants in one of them.

However, let's set aside technical doubts about computationally simulated sentience. Assume that posthumans have solved the Hard Problem of consciousness. The explanatory gap has been closed without unravelling our entire conceptual scheme in the process. Or perhaps qualia can themselves be digitally encoded and computationally re-created at will. Assume too that some analogue of Moore's Law of computer power is not just a temporary empirical generalisation: computer power continues to increase indefinitely until superintelligence has to grapple with the Bekenstein bound - unless this limit on the entropy or information that can be contained within a three-dimensional volume is itself supposed to disclose the granularity of our simulation. Assume further that a supercivilisation reaches a stage of development where it has the technical capacity to run an abundance of ancestor-simulations and simulate [a fragment of] the multiverse disclosed by contemporary physical science - though computationally simulating the infinite-dimensional Hilbert space of quantum-mechanics is no task for the faint-hearted. Finally, if the ancestor-simulations running are supposed to be cheap simulacra rather than faithful replications, let's assume like SA that the computational savings in taking "reality-shortcuts" outweigh the computational cost of the supervisory software - although in practice the computational price of intervening when ancestor-simulants get too close to discovering their ersatz status could make skimping on our Matrix a false computational economy. Granted all the above, then consider the scenario proposed in SA. Of all the immense range of alternative activities that future Superbeings might undertake - most presumably inconceivable to us - running ancestor-simulations is

one theoretical possibility in a vast state-space of options. On the one hand, posthumans could opt to run paradises for the artificial lifeforms they evolve or create. Presumably they can engineer such heavenly magic for themselves. But for SA purposes, we must imagine that (some of) our successors elect to run malware: to program and replay all the errors, horrors and follies of their distant evolutionary past - possibly in all its classically inequivalent histories, assuming universal QM and maximally faithful ancestor-simulations: there is no unique classical ancestral history in QM. But why would posthumans decide to do this? Are our Simulators supposed to be ignorant of the implications of what they are doing - like dysfunctional children who can't look after their pets? Even the superficial plausibility of "running an ancestor-simulation" depends on the description under which the choice is posed. This plausibility evaporates when the option is rephrased. Compare the referentially equivalent question: are our posthuman descendants likely to recreate/emulate Auschwitz? AIDS? Ageing? Torture? Slavery? Child-abuse? Rape? Witch-burning? Genocide? Today a sociopath who announced he planned to stage a terrorist attack in the guise of "running an ancestor-simulation" would be locked up, not given a research grant. SA invites us to consider the possibility that the Holocaust and daily small-scale horrors will be recreated in future, at least on our local chronology - a grotesque echo of Nietzschean "eternal recurrence" in digital guise. Worse, since such simulations are so computationally cheap, even the most bestial acts may be re-enacted an untold multitude of times by premeditated posthuman design. It is this hypothetical abundance of computational copies that lends SA's proposal that one may be living in a simulation its argumentative bite. At least the traditional Judeo-Christian deity was supposed to be benevolent, albeit in defiance of the empirical evidence and discrepancies in the Biblical text. But any Creator/Simulator who opts to run prerecorded ancestor-simulations presumably knows of the deceit practised on the

sentient beings it simulates. If the Simulators have indeed deceived us on this score, then what can we be expected to know of unsimulated Reality that transcends our simulation? What trans-simulation linguistic apparatus of meaning and reference can we devise to speak of what our Deceiver(s) are purportedly up to? Intuitively, one might suppose posthumans may be running copies of us because they find ancestral Darwinian life interesting in some way. After all, we experiment on "inferior" non-human animals and untermenschen with whom we share a common ancestry. Might not intellectual curiosity entitle superintelligent beings to treat us in like manner? Or perhaps observing our antics somehow amuses our Simulators - if the homely dramaturgical metaphor really makes any sense. Or perhaps they just enjoy running snuff movies. Yet this whole approach seems misconceived. It treats posthumans as though they were akin to classical Greek gods - just larger-than-life versions of ourselves. Even if advanced beings were to behave in such a manner, would they really choose to create simulated beings that suffered - as distinct from formally simulating their ancestral behaviour in the way we computationally simulate the weather?

Unfortunately, this line of thought is long on rhetorical questions and short on definitive proof. A counterargument might be that most humans strongly value life, despite the world's tragedies and its everyday woes. So wouldn't a "like-minded" Superbeing be justified in computationally replaying as many sentient ancestral lives as possible, including Darwinian worlds like our own? Even Darwinian life is sometimes fun, even beautiful. Might not our Simulators regard the episodic nastiness of such worlds as a price worth paying for their blessings - a judgement shared by most non-depressive humans here on Earth? Yet this scenario is problematic even on its own terms. Unless the computing resources accessible to our Simulators were literally infinite, a claim of dubious physical meaning, every simulation has an opportunity-cost in terms of

simulated worlds forgone. If one were going to set about creating sentient-life-supporting worlds in a supercomputer, then why not program and run the greatest number of maximally valuable paradises - rather than mediocre or malignant worlds like ours? Presumably posthumans will have mastered the technologies of building super-paradises for themselves, whether physically or via immersive VR. They'll presumably appreciate how sublimely wonderful life can be at its best. So why recreate the ugliness from which they emerged - a perverse descent from posthuman Heaven into Darwinian purgatory? Our own conviction that existing life is worthwhile is itself less a product of disinterested reflection than a (partially) heritable expression of status quo bias. If prompted, we don't believe the world's worst scourges, past or present, should be proliferated if the technical opportunity ever arises. Thus we aim to cure and/or care for the brain-damaged, the mentally ill and victims of genetic diseases; but we don't set out to create more brain-damaged, mentally ill and terminally sick children. Even moral primitives like contemporary Darwinian humans would find abhorrent the notion of resurrecting the nastier cruelties of the past. One wouldn't choose to recreate one's last toothache, let alone replay the world's sufferings to date. How likely are posthumans ever to be more backward-looking, in some sense, than us?

Of course, predictions of "progress" in anything but the most amoral, technocratic sense can sound naïve. Extrapolating an exponential growth in computing power, weapons technology or the like sounds reasonable. Extrapolating an expanding circle of compassion to embrace all sentient life sounds fuzzy-minded and utopian. Certainly, given the historical record, envisaging dystopian possibilities is a great deal more plausible than a transition to paradise-engineering. However, a reflex cynicism is itself one of the pathologies of the Darwinian mind. As our descendants rewrite their own code and become progressively smarter, their conception of intelligence will be enriched too.

Not least, enriched intelligence will presumably include an enhanced capacity for empathy: a deeper understanding of what it is like to be others - beyond the self-centred perspective of Darwinian minds evolved under pressure of natural selection. An enhanced capacity for empathetic understanding doesn't feature in conventional measures of intelligence. Yet this deficit reflects the inadequacy of our Asperger-ish "IQ tests", not the cognitive unimportance of smarter mind-reading and posthuman supersentience. Failure to appreciate the experience of others, whether human or nonhuman, is not just a moral limitation: it is a profound intellectual limitation too; and collective transcendence of humanity's intellectual limitations is an indispensable part of becoming posthuman. If our descendants have any inkling of what it is like to be, say, burned alive as a witch, or to spend all one's life in a veal crate, or simply to be a mouse tormented by a cat, etc, then it seems inconceivable they would set out to (re-)create such terrible states in computer "simulations", ancestral or otherwise. Achieving a God's-eye view that impartially encompasses all sentience may be impossible, even for our most godlike descendants. But posthuman cognitive capacities will presumably transcend the anthropocentric biases of human life. HI argues that posthuman benevolence will extend to the well-being of all sentience; this is technically feasible but speculative.

However, there is a counter to such reassuring arguments. It runs roughly as follows. We can have no insight into the nature of a hypothetical posthuman civilisation that might be capable of running subjectively realistic ancestor-simulations in their supercomputers. Therefore we have no insight into the motivational structure of our Simulators and why they might do this to us. Or perhaps we are merely incidental to their simulation(s) - which exist for a Higher Purpose that we lack the concepts even to express. For instance, perhaps advanced posthumans can command the Planck-scale energies needed hypothetically to create a "universe-in-the-laboratory". For inscrutable reasons, such

posthumans might decide to spin off a plethora of baby multiverses, making it statistically more likely that we are living in one of them rather than in the primordial multiverse. If so, we are emulating/simulating our ancestors in another multiverse that spawned us; and we are destined in turn to emulate/simulate our descendants in baby multiverses to come. This scenario contrasts with messy "interventionist" or conspiratorial simulations where posthuman supercomputers are supposed to be constantly rearranging stuff in our simulated world to keep us in ignorance of our artificial status. The point here is that we can't rule out any of such scenarios because we know absolutely nothing of posthuman ethics - or posthuman values of any kind. Posthuman psychology may simply be unfathomable to *Homo sapiens*, as are our purposes to lesser primates - or to beetles. Or maybe an explanation of our simulated status may be inaccessible to us simply in virtue of our being the ancestor-simulations of real historical people. Our ignorance could be written into the script.

We can't be sure this argument is false. There is nonetheless a problem with the unfathomability response. The prospect of using supercomputers to run ancestor-simulations belongs to the conceptual framework of early 21st Century human primates. The idea resonates with at least a small sub-set of social primates because running ancestor-simulations seems - pre-reflectively, at any rate - the kind of interesting activity that more advanced versions of ourselves might like to pursue. Yet if we have no insight into truly posthuman motivations or purposes, or indeed whether such anthropomorphic folk-psychological terms can bear posthuman meaning, then it's hard to assign any significant probability to our successors opting to run sentient ancestor-simulations. In fact, given the immense state-space of potential options, and the intrinsic squalor of so much Darwinian life, the prior probability we should assign to their doing so might seem vanishingly small - even if the technological obstacles could be overcome.

Contrary to the Objection, then, the existence of a world full of suffering is not evidence that our advanced descendants will never abolish its substrates. The existence of suffering is strong presumptive evidence that our descendants will never run sentience-supporting ancestor-simulations.

Appendix II: Q & A

The questions and responses below were originally posted on quora.com.

How do I believe that the humans around me actually possess consciousness?

The ancient sceptical Problem Of Other Minds is usually reckoned insoluble. Worse, mainstream scientific materialism offers no grounds for believing that one is not surrounded by p-zombies.

(*cf*. https://en.wikipedia.org/wiki/Philosophical_zombie)

However, the conjecture that one is surrounded by sentient beings rather than p-zombies may instead be treated as an experimentally testable hypothesis.

Consider the Hogan sisters (*cf*. "Could Conjoined Twins Share A Mind?":

http://www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-share-a-mind.html

http://www.youtube.com/watch?v=WKwT1Ol3nY0)

Developing technologies of reversible thalamic bridges promise a future of "mind-melding" with other humans and sentient beings from other species. Such utopian technologies should finally lay to rest the philosophical Problem Of Other Minds.

Mind-melding technologies may lead, not just to a Copernican moral revolution, but also a revolution in our conception of decision-theoretic rationality. Naturally, the proposal that mature posthuman ethics and decision-theoretic rationality might converge sounds too good to be true. But once sentient beings can "mind-meld", behaving "selfishly" may come to seem not just immoral but also irrational - akin to harming oneself. Perhaps compare the orthodox metaphysical individualism presupposed by the otherwise excellent Less Wrong Decision Theory FAQ:

http://lesswrong.com/lw/gu1/decision_theory_faq/

Are we quantum computers?

Conventional answer: no. The brain is too "warm, wet and noisy." Approximate decoherence timescales for neuronal superpositions can be calculated.

(*cf*. Max Tegmark: https://www.physicalism.com/quantum-computer.pdf: "Why the brain is probably not a quantum computer")

Intuitively, sub-femtosecond timescales are orders of magnitude too rapid to be harnessed by natural selection. Intuitively again, consciousness "emerges" on a dynamical timescale of milliseconds via patterns of neuronal firings.

Unconventional answer: yes. Our minds have been quantum computers for the past 540 million years. If neurons were the discrete, decohered classical objects of textbook neuroscience, then phenomenal binding of distributed neuronal feature-processors into perceptual objects would be impossible. Without such classically impossible phenomenal binding, the quasi-classical world-simulations of our everyday experience would be impossible too. If your waking or dreaming brain were a classical computer, then you'd at most be what philosophers call a "micro-experiential zombie", i.e. a mere aggregate of Jamesian mind-dust.

Who is right?

Mercifully, experiment rather than philosophising should decide.

Any quantum mind theory that does

http://www.sciencedirect.com/science/article/pii/S1571064513001188

("Consciousness in the universe: A review of the 'Orch-OR' theory")

or doesn't

https://www.physicalism.com/#6 ("an experimentally testable conjecture")

propose modifying or supplementing the unitary Schrödinger dynamics makes empirical predictions that can be experimentally falsified (or confirmed) by molecular matter-wave interferometry.

For some background reading on the phenomenal binding/combination problem, see David Chalmers:

http://consc.net/papers/combination.pdf

How much do our pain thresholds differ?

Pain-sensitivity varies hugely. Many genes are implicated. Here let's focus on the sodium-channel SCN9A gene. The SCN9A gene encodes the voltage-gated sodium-channel type IX a subunit known as Nav1.7. Nonsense mutations of the SCN9A gene abolish the capacity to feel physical pain. Other alleles of SCN9A are associated with unusually high or unusually low pain thresholds. (*cf*. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2096434/)

In principle, humanity could massively reduce the burden of suffering in the world by offering all prospective parents routine access to preimplantation genetic screening for benign "low pain" genes. "Low pain" alleles could also easily be bred in domestic nonhuman animals - and rapidly extended across the rest of the living world via CRISPR-based "gene drives": http://blogs.scientificamerican.com/guest-blog/gene-drives-and-

crispr-could-revolutionize-ecosystem-management/ ("'Gene Drives'" And CRISPR Could Revolutionize Ecosystem Management.")

When a friend of the American composer John Cage asked "Don't you think there's too much suffering in the world?", Cage answered, "No, I think there's just the right amount". Many victims would disagree. Humanity will shortly be able to decide the optimal level of suffering both for members of our own species - and eventually for life itself.

Should we eliminate the human ability to feel pain?

Are our perceptions physically existing somewhere?

Talk of "perceptions" can be misleading. Whether one is dreaming or awake, the mind-brain runs a spatio-temporally located world-simulation. The simulation is entirely internal to the skull: immersive, cross-modally matched organic VR. Thanks to natural selection, when you are awake your world-simulation tends to track - and causally co-vary with - gross, fitness-relevant patterns in the mind-independent world.

The world-simulation metaphor of our minds is ably defended by e.g. cognitive neuroscientist and philosopher of mind Antti Revonsuo (*cf*. Inner Presence: Consciousness as a Biological Phenomenon and contested by e.g. philosopher Daniel Dennett. (*cf*. https://en.wikipedia.org/wiki/Cartesian_theater).

Why does 'anything' exist?

Intuitively, there shouldn't be anything to explain. Bizarrely, this doesn't seem to be the case. One clue to the answer may be our difficulty in rigorously specifying a default state of "nothingness" from which any departure stands in need of an explanation. A

dimensionless point? A timeless void? A quantum vacuum? All attempts to specify an alternative reified "nothingness" - an absence of laws, properties, objects, and events - just end up smuggling in something else instead. Specifying anything at all, including the truth-conditions for our sense of "nothingness", requires information. Information is fundamental in physics. Information is physical. Information, physics tells us, cannot be created or destroyed. Thus wave functions in quantum mechanics don't really collapse to yield single definite classical outcomes. (*cf*. Wigner's friend: https://en.wikipedia.org/wiki/Wigner%27s_friend). Decoherence - the scrambling of phase angles between the components of a quantum superposition - doesn't literally destroy superpositions. Not even black holes really destroy information. (*cf*. https://en.wikipedia.org/wiki/Black_hole_information_paradox)

So naturally we may ask: where did information come from in the first place?

Perhaps the answer is that it didn't. The total information content of reality is necessarily zero: the superposition principle of QM formalises *in*existence.

On this story, one timeless logico-physical principle explains everything, including itself. The superposition principle of quantum mechanics formalises an informationless zero ontology - the default condition from which any notional departure would need to be explained. In 2002, *Physics World* readers voted Young's double-slit experiment with single electrons as the "most beautiful experiment in physics". (*cf*. http://physicsworld.com/cws/article/news/2013/mar/14/feynmans-double-slit-experiment-gets-a-makeover) Richard Feynman liked to remark that all of quantum mechanics can be understood by carefully thinking through the implications of the double-slit experiment. Quite so; only maybe Feynman could have gone further. *If* Everettian QM (*cf*. http://plato.stanford.edu/entries/qm-everett/) is correct, reality consists of a single vast quantum-coherent superposition. Each element in the

superposition, each orthogonal relative state, each "world", is equally real. (*cf*. https://itunes.apple.com/us/app/universe-splitter/id329233299?mt=8 - "Universe splitter" app.) Most recently, the decoherence program in post-Everett quantum mechanics explains the emergence of quasi-classical branches ("worlds") like ours from the underlying quantum field-theoretic formalism. (*cf*. Wojciech Zurek: http://arxiv.org/pdf/0903.5082v1.pdf) The universal validity of the superposition principle in post-Everett QM suggests that the mystery of our existence has a scientific rather than theological explanation.

What does it mean to say that the information content of reality may turn out to be zero? Informally, perhaps consider the (classical) Library of Babel. (*cf*. https://en.wikipedia.org/wiki/The_Library_of_Babel)

The Library of Babel contains all possible books with all possible words and letters in all possible combinations. The Library of Babel has zero information content. Yet somewhere amid the nonsense lies the complete works of Shakespeare - and you and me. However, the Library of Babel is classical. Withdrawing a book from the Library of Babel yields a single definite classical outcome - thereby creating information. Withdrawing more books creates more information. If we sum two ordinary non-zero probabilities, then we always get a bigger probability. All analogies break down somewhere. Evidently, we aren't literally living in Borges' Library of Babel. So instead of the classical Library of Babel, let us tighten the analogy. Imagine the quantum Library of Babel. Just as in standard probability theory, if there are two ways in QM that something can happen, then we get the total amplitude for something by summing the amplitudes for each of the two ways. If we sum two ordinary non-zero probabilities, then we always get a bigger probability. Yet because amplitudes in QM are complex numbers, summing two amplitudes can yield zero. Having two ways to do something in quantum mechanics can make it not happen.

Recall again the double-slit experiment. Adding a slit to the apparatus can make particles less likely to arrive somewhere despite there being more ways to get there. Now scale up the double-slit experiment to the whole of reality. The information content of the universal state vector is zero. (*cf*. Jan-Markus Schwindt, "Nothing happens in the Universe of the Everett Interpretation": http://arxiv.org/pdf/1210.8447v1.pdf) The quantum Library of Babel has no information.

Caveats? Loose ends? The superposition principle has been experimentally tested only up to the level of fullerenes, though more ambitious experiments are planned (*cf*. http://www.nature.com/news/2009/090910/full/news.2009.903.html). Some scientists still expect the unitary Schrödinger dynamics will need to be supplemented or modified for larger systems - violating the information-less zero ontology that we're exploring here.

Consciousness? Does the superposition principle break down in our minds? After all, we see live or dead cats, not live-and-dead-cat superpositions. Yet this assumption of classical outcomes - even non-unique classical outcomes - presupposes that we have direct perceptual access to the mind-independent world. Controversially (*cf*. https://www.physicalism.com/quantum-computer.pdf), perhaps the existence of our phenomenally bound classical world-simulations itself depends on ultra-rapid quantum-coherent neuronal superpositions in the CNS. For if the superposition principle really broke down in the mind-brain, as classical neuroscience assumes, then we'd at most be so-called "micro-experiential zombies" - just patterns of discrete, decohered Jamesian neuronal "mind-dust" incapable of phenomenally simulating a live or a dead classical cat. (*cf*. https://www.physicalism.com/#6)

This solution to the phenomenal binding problem awaits experimental falsification - or

implausible vindication! - with tomorrow's tools of molecular matter-wave interferometry. (*cf*. Non-materialist Physicalism)

What about the countless different values of consciousness? How can an informationless zero ontology possibly explain the teeming diversity of our experience? Well, just as the conserved constants in physics cancel out to zero, and just as all of mathematics can in principle be derived from the properties of the empty set, perhaps the solutions to the field-theoretic equations of QFT mathematically encode the textures of consciousness. If we had a cosmic analogue of the Rosetta stone, then we'd see that these values inescapably "cancel out" to zero too. Unfortunately, it's hard to think of any experimental tests for this speculative conjecture.

"A theory that explains everything explains nothing", protests the critic of Everettian QM. To which we may reply, rather tentatively: yes, precisely.

Is everything made of consciousness?

It's an open question. Formally, the world is exhaustively described by the equations of physics and their solutions. Physics – or rather tomorrow's physics beyond the Standard Model - is causally closed and complete. But physics is silent on the intrinsic nature of the physical: the mysterious "fire" in the equations.

An intuitively plausible philosophical assumption is that this "fire" - the essence of the physical – is *non*-experiential. Thus the equations of quantum field theory describe the behaviour of fields and their excited quanta of *in*sentience. Such an assumption is hard to test experimentally. Moreover, the assumption that the intrinsic nature of the physical is non-experiential would seem inconsistent with the only part of the "fire" in the equations to which one enjoys direct access, namely one's own conscious mind. If the "fire" in the

equations really is non-experiential, we need to explain how consciousness "emerges" (how? where? when? why?) from insentient fields. In addition, we must derive the values and interdependencies of the diverse textures of experience from the underlying formalism of QFT. We must also explain how such emergent consciousness has the causal capacity to allow us to discuss its existence *without* violating the causal closure and completeness of physics.

By contrast, if non-materialist physicalism (*cf*. https://www.physicalism.com) is true, then the world is exhaustively described by the equations of physics; and the solutions to the field-theoretic equations yield the values of consciousness. Traditionally, physicalism is treated as a cousin of materialism. Yet non-materialist physicalism is better viewed as the scientifically literate form of monistic idealism.

Do Holocaust survivors feel empathy for slaughtered animals?

Is it a coincidence that Israel may become the first vegan nation: http://www.israel21c.org/culture/israel-goes-vegan/

Many Holocaust survivors - and their children and grandchildren - have made the connection. When a Nobel laureate like Isaac Bashevis Singer describes the fate of nonhuman animals as "an eternal Treblinka", this is not a parallel a Jewish writer draws lightly.

In later life, even death-camp commandant Franz Stangl recognised the parallel. In Brazil, Stangl gave up eating tinned meat after his train stopped one day next to a slaughterhouse ("Into That Darkness: from Mercy Killing to Mass Murder, a study of Franz Stangl, the commandant of Treblinka" (1974, second edition 1995)).

Of course, all analogies break down somewhere. Thus the Nazis sincerely (and

psychotically) believed in a mythical international Jewish conspiracy against the Aryan race. By contrast, the standard moral argument in favour of meat eating runs "But I like the taste!"

Not merely animal advocates have come to believe that humans are doing something ethically monstrous. In "Sapiens" (2014), Israeli historian Prof. Yuval Noah Harari observes: "Tens of billions of them [nonhuman animals] have been subjected over the last two centuries to a regime of industrial exploitation, whose cruelty has no precedent in the annals of planet Earth. If we accept a mere tenth of what animal-rights activists are claiming, then modern industrial agriculture might well be the greatest crime in history."

What is David Pearce's position on meta-ethics?

For reasons we don't understand, the pain-pleasure axis discloses the world's inbuilt metric of (dis)value. Full-spectrum superintelligence in command of all the first-person and third-person facts will act accordingly. For evolutionary reasons, humans lack such an impartial God's-eye view. The egocentric illusion is immensely adaptive. Hence our epistemological limitations are genetically hardwired.

The psychopath – or rogue zombie AI – is unimpressed.

"Sure", says the psychopath or the sophisticated digital zombie, *"I can see that you're in agony. No doubt your first-person experience of agony has a 'normative aspect' for you. For you, doubtless it's not an 'open question' whether agony is bad or not. I know it's disvaluable for you. But the point is, it's not disvaluable for me! As Hume says, it is* 'not contrary to reason to prefer the destruction of the whole world to the scratching of my

finger'. *Hume's guillotine can't be cheated. I'm not being irrational or immoral in ignoring your desperate cries for help."*

Is today's psychopath or tomorrow's psychopathic zombie AI correct?

No, in my view – simply ignorant.

Perhaps imagine a Borg-like civilisation, or a world of ubiquitous "mind-melding". In such an advanced civilisation, first-person experience is shared more intimately than by mirror-touch synaesthestes or the Hogan sisters (cf. Would it be theoretically possible to experience the conscious experience of another being?) today – including the normative aspect of experience disclosed by the pleasure-pain axis. [If you don't believe that experience can have a normative aspect even for the subject, then perhaps plunge your hand in iced water and hold it there indefinitely until you agree. Language can't dispense with semantic primitives altogether: like redness, (dis)value is a semantic primitive whether one believes in meta-ethical antirealism or not.] The Borg *knows* something that skull-bound humans trapped in our solipsistic island universes cannot grasp. If humans had God-like omniscience, then just as you withdraw your hand from the fire, then humanity would aim to perform the God-like cosmological equivalent – computationally non-trivial as that equivalent may be.

I don't want to downplay the mystery of first-person consciousness and the nature of (dis)value, or the challenge posed by value realism for rational policy-making insofar as one aspires to be an effective altruist. Yet unless modern science is hopelessly mistaken, then – in defiance of all appearances – I'm not really special, and neither are you. If agony and despair are bad for me – and they are! – then they are objectively bad for anyone, anywhere. One's own epistemological limitations don't deserve elevation into a metaphysical principle of Nature. First-person experience can't be relegated to second-

rate ontological status. First-person experience is as objectively real as it gets. In my view, ethics will be computable by full-spectrum superintelligence. The challenge now is to build it.

Effective altruism – Wikipedia

You Are Them by Magnus Vinding

For a contrary view, see:

J. L. Mackie - Ethics: Inventing Right and Wrong (0140135588, 1991).

And for a conception of rationality predicated on traditional metaphysical individualism, perhaps see the Less Wrong Decision Theory FAQ and the Orthogonality thesis – though belief that full-spectrum superintelligence is inherently sentience-friendly isn't an argument for complacency about the risks of AI (*cf*. the "No true Scotsman" fallacy).

What evidence is there for quantum computation in the brain?

Perhaps the strongest empirical evidence that the mind-brain is a quantum computer lies under one's virtual nose, so to speak, in the guise of phenomenally bound perceptual objects ("local" binding), and the unitary subject who apprehends them ("global" binding). However, independent experimental confirmation of this conjecture will depend on next-generation molecular matter-wave interferometry.

If neurons were discrete, decohered classical objects, as we might naively suppose, then organic minds could at most be patterns of membrane-bound "mind-dust": so-called micro-experiential zombies. Individual neuronal edge-detectors, motion-detectors, neurons mediating colour, and so forth could not generate phenomenally bound perceptual objects, nor a quasi-classical world-simulation for those phenomenally bound

dynamical objects to populate, nor a fleetingly unitary phenomenal self who could pose such questions. By way of analogy, compare interconnected but skull-bound American minds communicating over the Internet. Whatever computations these interconnected skull-bound minds might experimentally execute, the collective outcome of the computations is not a pan-continental subject of experience - no continental sunsets or symphonies or migraines, just an information-processing micro-experiential zombie. Neuroscience needs to understand how a waking or dreaming "pack of neurons" is different.

Clues? There is no theoretical or experimental evidence that the unitary Schrödinger dynamics breaks down in the CNS. So let us provisionally assume that unmodified and unsupplemented quantum theory is correct. If so, then neuronal superpositions ("Schrödinger's cat states") of edge-detectors, motion-detectors, colour-detectors must occur: you instantiate such neuronal superpositions right now. Naively, sub-femtosecond quantum superpositions in the warm wet CNS are computationally and phenomenally too short-lived to underpin our minds – ludicrously prolonged by twenty-five orders of magnitude or so compared to Planck-scale physics, but still orders of magnitude shorter than the normal millisecond dynamical time-frames over which everyday common-sense says that consciousness "emerges".

Thankfully, scientific experiment rather than philosophical speculation should resolve the issue. Thermally-induced decoherence in living subjects is too strong for the tell-tale non-classical interference effects diagnostic of neuronal superpositions to be readily detected in the laboratory. However, trained-up *in vitro* neuronal networks (*cf*. https://www.physicalism.com/#6) should suffice to confirm or experimentally falsify the conjecture to the satisfaction of proponents and critics alike.

For some background reading:

http://www.bostonneuropsa.net/PDF%20Files/Mashour/quantumbinding.pdf ("The Cognitive Binding Problem: From Kant to Quantum Neurodynamics")

http://arxiv.org/pdf/0909.1469v3.pdf ("Toward Quantum Superposition of Living Organisms")

https://www.physicalism.com/quantum-computer.pdf ("Why the brain is probably not a quantum computer")

What is reality made of?

Formally, a gigantic wavefunction. Yet what "breathes fire into the equations and makes a universe for them to describe" is unknown. Intuitively, the intrinsic nature of the physical is non-experiential. However, the only part of the "fire" in the equations to which one enjoys direct access, namely one's own conscious mind, discloses properties wholly at variance with materialist metaphysics.

The Penrose Orch-OR theory, like all stories invoking observer-induced state vector reduction, entails modifying or supplementing the unitary dynamics. But in Penrose's approach, quantum state reduction is a gravitational phenomenon. However, no departure from the unitary Schrodinger dynamics has ever been experimentally detected. A large minority of theorists now believe that the superposition principle is universally valid: the state vector of the universe evolves deterministically in accordance with the Schrodinger equation. Classicality is an emergent phenomenon. Wojciech Zurek offers a good overview of the decoherence programme e.g. here:

http://arxiv.org/pdf/1412.5206v1.pdf

What would you do if someone attempted to rescue a prey from its predator?

Applaud and assist. Should we prioritise the interests of human and nonhuman predators or their victims? Do we want to promote a living world where sentient beings harm each other or not?

Until recently, the problem of predation was academic. But the CRISPR genome-editing revolution and the promise of synthetic gene drives mean the entire biosphere will shortly be programmable.

So what is the optimal level of suffering in the living world? Should we aim for conservation biology or compassionate biology? Suppose we encounter an advanced civilisation that has abolished population control by starvation, disease and predation in favour of cross-species immunocontraception. Should we urge this peaceable civilisation to restore ancestral horrors - death by asphyxiation, disembowelment or being eaten alive? Or should all sentient beings be allowed to flourish unmolested?

I'm 17 and just realized that the universe is indifferent to our suffering. The universe still expands. Life goes on. What is the point?

"*As I looked out into the night sky, across all those infinite stars, it made me realize how insignificant they are.*"

(Peter Cook)

Biological minds like ours are part of the universe. For sure, most of the universe is indifferent to suffering. But not all of it. Critically, one species on Earth has mastered its genetic source code. The entire biosphere will soon be programmable. Intelligent moral agents will shortly be able to choose how much suffering and malaise we want to exist in

the living world (cf. https://www.abolitionist.com). In principle, biotechnology can abolish the biology of unpleasant experience in all sentient life.

What is the point of it all?

Well, recall a lot of the suffering in the world isn't raw physical distress. The common experience of not-seeing-the-point-of-it-all is itself part of the problem of suffering. Low mood is associated with feelings of emptiness, hopelessness and futility. Life seems meaningless. Conversely, good mood is associated with a sense of purpose and significance. Compare how boosting mesolimbic dopamine function gives life urgency: a sense of things-to-be-done. Biological interventions can enhance your mood and motivation. Ultimately, the feeling of "pointlessness" can itself be abolished via CRISPR-based genome editing. What's its use?

Right now we're on the brink of a major evolutionary transition in the development of life. Transhumanists believe we should all have the opportunity to feel "better than well" - ideally, a "Triple S" civilisation based on superintelligence, superlongevity and superhappiness.

Yet intuitively, technology can't solve everything. What about the meaning of life? What's it all about?

Cracking that one is indeed a challenge. However, let's leave "meaning" in some transcendent sense to theologians and metaphysicians. Empirically, for reasons we simply don't understand, life based on gradients of intelligent bliss will feel significant beyond the bounds of normal human experience. Even today, no one says, "I feel blissfully happy but my life feels pointless". Take care of happiness and the meaning of life will take care of itself.

Is it immoral to kill an ant?

Like a minority of humans, some ants fail the mirror test (cf. "Are Ants (Hymenoptera, Formicidae) capable of self-recognition?" http://www.journalofscience.net/File_Folder/521-532%28jos%29.pdf). Yet like humans, ants are sentient beings with a pleasure-pain axis (cf. "Morphine addiction in ants") and a capacity to suffer. Insofar as it's immoral to harm any sentient being, regardless of race or species, then yes, it's immoral gratuitously to harm an ant. In the long run, intelligent moral agents may practise high-tech Jainism (cf. High-tech Jainism).

Of course, like most people I think mankind has more important issues to worry about than the well-being of an individual ant. So is one really morally bound to step aside when some humble invertebrate crosses one's path? Get real!

It's a powerful intuition. However, let's bear in mind that compared to posthuman superintelligence, humans will probably be as sentient and sapient as ants. Is superintelligence morally bound to respect the interests of cognitively humble beings like us?

Fortunately for *Homo sapiens*, full-spectrum superintelligence will presumably enjoy a superhuman capacity for perspective-taking and empathetic understanding. IMO it's a capacity humans should aspire to emulate.

What did Hitler think about the Jews that were crying during the holocaust??

Few people dared personally to confront Hitler about the suffering of Jewish people. One exception was Henriette von Schirach, wife to Baldur von Schirach, Gauleiter of Vienna.

When visiting Holland in 1943, Henriette was woken in her hotel by the screams and crying of Jewish women and children outside who were being deported. A sympathetic German soldier explained what was happening. Henriette promised to take the matter up with Hitler. She broke off her visit to the Netherlands. Hitler's secretary Christa Schroeder recalls the row that followed at the Berghof on Good Friday.

"'Be silent, Frau von Schirach, you understand nothing about it. You are sentimental. What does it matter to you what happens to female Jews? Every day tens of thousands of my most valuable men fall while the inferior survive. In that way the balance in Europe is being undermined,' and here he moved his cupped hands up and down like a pair of scales.

'And what will become of Europe in one hundred, in one thousand years?' In a tone which made it evident that he considered the matter closed, he declared: 'I am committed by duty to my people alone, to nobody else!'"

("He Was My Chief: The Memoirs of Adolf Hitler's Secretary" by Christa Schroeder, Frontline Books, 2009)

Henriette and her husband were never invited to the Berghof again.

For the most part, Hitler seems to be have been hard-hearted rather than sadistic. Hitler didn't want to dwell on the suffering he caused any more than, say, factory-farm owners or consumers of meat products want to dwell on the suffering of their victims today.

After an irreversible transition to a blissful existence with boundless cognitive, physical and transcendental euphoria, what would you do?

A chrysalis has limited insight into the nature of life as a butterfly. The metamorphosis you propose is more profound. Even so, intelligent bliss differs from being "blissed out". Therefore let's assume that life based on information-sensitive gradients of bliss also enhances our motivation to act and our sense of social responsibility.

What next?

If there still exists the slightest distress in even the humblest marine invertebrate, then intelligent moral agents aren't entitled to rest. Even after we've reprogrammed the biosphere to eliminate experience below "hedonic zero", we mustn't risk abandoning ourselves prematurely to escapism, i.e. "hedonism " in the baser sense. Ethically speaking, mankind needs to discover the theoretical upper bounds to intelligent moral agency in the cosmos. What are our ultimate cosmological responsibilities? Perhaps the "thermodynamic miracle" (Eric Drexler) of life's genesis means that cosmic rescue missions are impossible or redundant. We may well be alone in our Hubble volume. If so, we don't yet know this.

However, let us assume that all our cosmological duties have been discharged. Nothing exists in our forward light-cone beyond life animated by gradients of intelligent bliss.

What would I do personally?

1) I'd explore psychedelia.

Mapping out the boundaries of one's personal ignorance of the varieties of conscious experience is dauntingly difficult. Compare how even lucid dreamers have only limited insight into the nature of dreaming consciousness – of what it means to be "asleep", let alone to be "awake". Likewise, each of us while awake has only limited insight "from the inside" into what we're lacking *and* into the nature of ordinary waking consciousness itself. What humans naively call ordinary waking consciousness is just one small state-

space of experience among billions of state-spaces. A Mendeleev table for state-spaces of qualia is a distant prospect. In what God-like state of mind could it ever be surveyed? Until then, we're as knowledgeable as earthworms - to a good approximation at any rate.

The remedy for such ignorance might seem self-evident. Use the experimental method! Sadly, most dark Darwinian minds are not robust enough to explore the wilder shores of psychedelia, let alone cope with the alien state-spaces of experience opened up by tomorrow's CRISPR genome-editing. Heaven knows what outlandish state-spaces of psychedelia can be generated with novel genes, alleles and exotic gene-expression profiles. Such "unknown unknowns" needn't scare us. Granted the biology of invincible well-being that you propose, we could all safely become psychonauts. Mastery of our reward circuitry can make "bad trips" on novel designer drugs not just physiologically impossible but also literally inconceivable.

Lest all this sound too breathless, IMO we shouldn't imagine that taking psychedelics is the route to instant wisdom – even when it's safe for us all to become psychedelic investigators. By analogy, imagine a primitive savage who stumbles across a TV with hundreds of different channels. Alas, the TV set is faulty. The channels display only "noise". Likewise, most physically possible state-spaces of experience have never been recruited by natural selection for any information-signalling purpose in living organisms - let alone shared in common by language-users to allow intelligent communication about their properties. Taking psychedelics today typically leads to psychosis or "enlightenment" rather than far-reaching discoveries that stand the test of time. By analogy again, a congenitally blind child who is surgically given the gift of sight is "enlightened". Wow! S/he is also bewildered. Mature visual intelligence takes years, if not decades, to acquire. The same is true of navigating alien state-spaces of consciousness.

Despite these caveats, I think life based on gradients of genetically preprogrammed bliss will lead to a true cognitive revolution - a post-Galilean science of consciousness.

2) I'd aim higher.

Darwinian consciousness is polluted by misery and malaise. By contrast, the biology of lifelong well-being you propose seems almost magical. Yet why stop there? Strip away the considerations of prudence and morality that constrain our personal exploration of pleasure today ("*Pleasure is the greatest incentive to evil.*" - Plato). Artificial intelligence and genome-editing promise to make such practical problems soluble. Empirically, for reasons we don't understand, there is an intimate link between pleasure and value. The experience of lifelong superhuman pleasure will yield the experience of lifelong superhuman value too. Biotech can make everyday life sublime.

The following example may seem homely. I hope it nonetheless makes the point. If like me you star your music collection from 1 to 5 for excellence, then a music collection that yielded a star-rating of 6 to 10 would induce tingles down your spine all day. What if our reward circuitry could be redesigned to yield a default hedonic range of 95 to 100? Critical discernment could be retained. Yet our musical pleasure and capacity for musical appreciation would be out of this world. Today we don't know what we're missing. The same holds for art, beauty, sexuality, introspection, spirituality – and personal relationships.

Trapped in the squalor of Darwinian life, most of us find the prospect of such an elevated hedonic range is fantastical at best. Yet neuroscientists are already homing in on the molecular signature of pure bliss in our twin "hedonic hotspots" in the CNS (*cf*. "Building a neuroscience of pleasure and well-being": http://psywb.springeropen.com/articles/10.1186/2211-1522-1-3). In principle, we can

amplify subjective well-being by orders of magnitude beyond today's "peak experiences". Artificial intelligence researchers sometimes speculate on a future of recursively self-improving software-based AI that bootstraps itself to full-spectrum superintelligence (*cf*. Intelligence explosion). Why not create recursively self-improving happiness too? Rational value-maximisers, at least, should aim for an analogue of Moore's law that embraces recursively self-improving subjective well-being.

Right now, yes, the molecular biology of such hedonic enrichment seems a utopian pipedream. I think our overriding ethical focus should be on mitigating, preventing and eventually abolishing outright the biology of suffering. Human civilisation is based on the exploitation and abuse of sentient beings. Talk of creating a living world based on gradients of superhuman well-being rings hollow. But coming into existence needn't be harmful indefinitely. Mastery of the molecular machinery of bliss promises an exponential growth in intelligent well-being - a major evolutionary transition in the development of life.

Transhumanists believe we should be working for a "triple S" civilisation of superintelligence, superlongevity, and superhappiness.

The welcome gift of personal bliss wouldn't (I hope) change this goal.

Are radical eliminativists about consciousness p-zombies? Or do they misinterpret the nature of their own consciousness?

A good rule of thumb is to try to set out a position with which you disagree more powerfully than the advocacy its smartest proponents and then critique it. As a consciousness realist, I find radical eliminativism almost incomprehensible. This makes devil's advocacy rather difficult. Trying to imagine what it's like to suppose one is a

zombie (e.g. Daniel Dennett, "From Bacteria to Bach and Back: The Evolution of Minds": https://www.amazon.com/Bacteria-Bach-Back-Evolution-Minds/dp/0393242072, p. 363) feels more alien than imagining one has Cotard's syndrome (cf. https://www.washingtonpost.com/national/health-science/zombie-disease-makes-people-think-they-have-died/2015/10/30/ca8ab52c-532f-11e5-933e-7d06c647a395_story.html), or what it's like to be a bat. For the only thing I've ever known, except by inference, has been my own conscious mind. Both the scientific world-picture and the principle of mediocrity suggest I'm in no way special.

However, here goes...

Radical eliminativists regard natural science as our best story of the world. Ultimately, all science derives from physics. Physics is causally closed and complete. The Standard Model is extraordinarily accurate and well-tested. The field-theoretic ontology of physics has no place for first-person experience. Therefore consciousness can't exist.

Radical eliminativists tend to be:

1. drug-naive ("What does a fish know of the water in which he swims?"). Compare researchers who experiment with consciousness rather than just philosophise. e.g. https://erowid.org/experiences/

2. high IQ/AQ (cf. https://www.wired.com/2001/12/aqtest/). People high on the AQ spectrum don't just read other minds differently from neurotypicals. High-AQ folk understand their own minds differently too. The human faculty of introspection is more variable than exteroception. (cf. "The Unreliability of Naive Introspection": http://www.faculty.ucr.edu/~eschwitz/SchwitzPapers/Naive1.pdf) High-AQ eliminativists don't have an introspectively accessible phenomenology of thoughts and feelings in the

same way as do consciousness realists. Perhaps compare Dennett's "heterophenomenology": https://en.m.wikipedia.org/wiki/Heterophenomenology)

3. perceptual naive realists. Direct realists about perception believe they are directly acquainted with the physical properties of medium-sized macroscopic objects as described by an approximation of classical physics. Compare a world-simulation model of perceptual experience in which sunsets and symphonies are as much features of conscious mind as the subtle, thin and elusive cognitive phenomenology of our thought-episodes. (cf. https://mitpress.mit.edu/books/inner-presence)

And

4. don't lucid dream (cf. https://en.m.wikipedia.org/wiki/Lucid_dream), or even remember their dreams. If one is having a lucid dream, then one's entire world-simulation is manipulable at will - and manifestly consciousness-dependent.

And yet...

Before major surgery, the eliminativist materialist insists on general anaesthesia, rather than mere muscle-paralysing agents like curare (cf. "Awareness during Anaesthesia": http://www.anesthesiaweb.org/awareness.php), just like ordinary patients. Why, exactly? This isn't a rhetorical question. Like consciousness realists, radical eliminativists take analgesics for pain-relief - although their pain thresholds may be higher than neurotypicals (cf. the "extreme male brain" theory of ASD. Testosterone has both an anti-introspective and painkilling action.) Here I really do struggle to make sense of eliminativism. My guess is that a radical eliminativist would respond that pain is real, but consciousness realists radically misunderstand its nature: we should reject Sellars' "Myth of the Given" (cf. https://sites.google.com/site/drtimthornton/courses/epistemology/sellars-and-the-myth-

of-the-given). All experience is contaminated by theory. What consciousness realists call the "raw feels" of agony, e.g. the subjective first-person experience of a nasty migraine, should be instead be reinterpreted as a purely physical phenomenon.

If so, then I'd agree - in a sense. Only physical properties are real. First-person facts are real. Yet if subjective pain and pleasure are really physical properties, then the ontology of physics - ultimately the mysterious "fire" in the equations of QFT - is radically different from our naive materialist intuitions about the intrinsic nature of the physical. Here we enter very different territory indeed: https://www.physicalism.com/abstract.html

Is genetic engineering (crispr, gene drive, etc) advanced enough to kill or save billions of people?

Millions of gamers across the world enjoy playing Plague Inc: Evolved (PC). The object of the game is to eradicate the human species by evolving pathogens via a complex set of variables to simulate the severity and spread of the plague. Tomorrow's CRISPR-based "gene drives" (cf. Gene Drive FAQ - Sculpting Evolution) have the capacity to kill billions of sentient beings or make the world a radically better place.

First the scary stuff. "Weaponised" gene drives may democratise weapons of mass destruction (cf. "This could be the next weapon of mass destruction"). Newspaper stories like "New ISIS weapon: 'Supercharged' killer mosquitoes" are sensationalist and (to the best of my knowledge) still unduly alarmist; but the threat of bioterrorism is real (cf. "Why FBI and the Pentagon are afraid of gene drives"). Using cheap molecular tools and laboratory equipment readily available on eBay, an ecologically literate garage biohacker could take out entire ecosystems by targeting one or more “keystone” species (cf. Keystone species). In principle, even a single gene-drive-engineered organism released

in the wild - whether accidentally or deliberately - could crash an entire ecosystem. The novel capacity of synthetic biology to let you "upload" genetic code to your PC, then edit and manipulate the code, and next download the code into revised living organisms heralds the era of computer-designed sentient beings - and computer-designed weaponised organisms that "hijack" evolution and transcend the old constraints of Mendelian inheritance. Using weaponised gene drives, tomorrow's bioterrorists could suppress pollinators in order to destroy a country's agricultural production; modify the host range, transmissibility and virulence of pathogens; make vaccines ineffective and confer resistance to antibiotics, antifungals and antiviral agents; and modify currently innocuous insects to transmit diseases such as malaria, dengue, filariasis - and worse. Depending on their level of sophistication, biohackers - or rogue state actors - could sabotage biosurveillance efforts, circumvent existing diagnostic and detection tools; and defeat potential "reversal drives" designed to overwrite changes introduced by their primary drives.

Worryingly, the deliberate release of gene-drive-engineered organisms into the wild is also potentially anonymous. Effective deterrence, international regulation and enforcement mechanisms, and democratic accountability are all woefully lacking.

If all goes well, CRISPR/Cas9-based gene drives will imminently be used to wipe out the scourge of insect-borne disease. Malaria has killed an estimated half the humans who ever lived (cf. "Portrait of a serial killer"); the disease still kills or sickens millions of human and nonhuman animals each year. However, mosquitoes and other insect vectors can just as readily be weaponised to deliver lethal bacterial toxins to entire human populations. Mercifully, Unit 731 (cf. "Operation Cherry Blossoms at Night") didn’t have access to CRISPR-based gene drives because if they did, the outcome of WW2 might have been very different. By levelling the playing-field for weapons of mass destruction,

weaponised gene drives are likely dramatically to shift the balance of international power. Simultaneous release of multiple independently-targeted gene drives makes biodefense extremely difficult. IMO some of the nastier non-obvious possibilities shouldn't be publicly speculated on even in outline; but the optimal level of self-censorship is unclear. Does the study of global catastrophic and existential risk increase or diminish its likelihood? How do bio-laboratories and academic research institutes protect themselves - and us - against "deep entryism"? Evidently, CRISPR/Cas9-mediated gene drives can't distinguish between Christians, Jews and Muslims; but CRISPR-based gene-drive-engineered organisms could be used as so-called "ethnic bioweapons" (cf. Ethnic bioweapon). Genotype-specific bioweapons can either be finely targeted (cf. "Hacking the President's DNA") or appallingly indiscriminate. We may hope that tomorrow's genetic jihadis will worry about "collateral damage". Unfortunately, some religious fundamentalists think more like Arnaud Amalric than secular bioethicists. [Arnaud Amalric was a Cistercian abbot who played a prominent role in the Albigensian Crusade. When asked by a Crusader how to distinguish the Cathars from the Catholics, Amalric supposedly responded, "Caedite eos. Novit enim Dominus qui sunt eius." Loosely: "Kill them all. God will know His own." cf. Massacre at Béziers]

Religious extremists won't be the only groups tempted to modify the biosphere with rogue drives. Blackmailers, extortionists, and organised crime are already taking an interest in synthetic biology. However, highly motivated idealists and ideologues are at least as worrying as amoral criminals. For example, sooner or later animal rights extremists may decide to tweak e.g. the Lone Star tick (cf. "This bug's bite could turn you into a vegetarian") with a clever gene drive. The way that humans treat nonhumans is indeed monstrous; but such an initiative is not going to help win the battle for hearts and minds.

[The concept of using bioweapons to promote dietary modification isn't entirely new. "Operation Vegetarian" (cf. Operation Vegetarian) isn't the name of a clandestine animal rights plot to turn humans into obligate herbivores, but rather a plan hatched by British Intelligence in WW2 to drop cattle-cakes laced with anthrax spores on Germany. Grazing cattle would then eat the cakes and infect meat-eating German consumers - although not Hitler, who was a vegetarian.]

And then there are Deep Greens who publicly or privately agree with Professor Erik Pianka, who reportedly favours elimination of 90 percent of Earth's human population by airborne Ebola or its equivalent (cf. "Group of scientists gave standing ovation for plan to kill 90 percent of human population with airborne Ebola"). The idea of using gene drives to cull an ecologically damaging invasive species opens up possibilities its originators may not have intended. In addition, some Deep Greens have a depth of ecological knowledge of keystone species needed to bring about a planetary cataclysm that is still (probably) lacking in Islamic fanatics.

Again, depending on the sophistication and motivations of the actors in question, a "Doomsday device" could theoretically be engineered either to eradicate or interfere with the metabolism of keystone species of phytoplankton in the oceans. Phytoplankton contribute between 50 to 85 of the oxygen in Earth's atmosphere. For evolutionary reasons, status quo bias is endemic in human society; but it's far from universally shared (cf. Better Never to Have Been Quotes).

On a brighter note...

Used responsibly and under United Nations auspices, CRISPR-based gene drives will eradicate vector-borne infectious diseases ranging from Zika to malaria. Most ambitiously, gene drives could be used to help create a happy biosphere (cf. gene-

drives . com: "genetically designing a happy biosphere"). Synthetic biology allows intelligent moral agents to "reprogram" Nature. Life on Earth can potentially be wonderful - and perhaps even sublime. “May all that hath life be delivered from suffering”, said Gautama Buddha; and this outcome will shortly be technically feasible - one way or another.

What is the Quantum Mind?

"*There is nothing so absurd that some philosopher has not already said it*."

(Cicero)

All minds are quantum minds. The classical-looking world-simulation you're experiencing now is what a quantum mind feels like from the inside. The same selection mechanism ("quantum Darwinism": https://arxiv.org/pdf/0903.5082.pdf) that explains the emergence of classicality in the mind-independent world also acts on quadrillions of coherent neuronal superpositions ("cat states") in the CNS. This insanely powerful, unremitting Darwinian selection mechanism sculpts what would otherwise be fleeting psychotic noise – i.e. individual sub-femtosecond superpositions of distributed neuronal feature-processors – into a phenomenally bound world-simulation described from within by an approximation of classical physics. Only a quantum mind can phenomenally simulate a classical world. Decohered classical neurons would just be "mind-dust”, as you are in a dreamless sleep.

Investigators working on the foundations of quantum mechanics wonder why experiments ever have definite outcomes at all (*cf*. http://faculty.up.edu/schlosshau...). Why do we never observe smeared-out pointer-readings or live-and-dead cats? Why are superpositions never experienced, only inferred? (*cf*. Double-slit experiment - Wikipedia)

Unanswerable questions usually turn out to be ill-posed.

Alternatively, *only* superpositions are ever experienced. Your experience of determinate experimental outcomes (and live *or* dead cats) consists of coherent neuronal superpositions. It's precisely the fact that the superposition principle of QM never breaks down that allows you phenomenally to simulate a well-behaved classical world where it does. The vehicle of simulation is quantum-coherent; the experiential content of the simulation is robustly classical. Perhaps think of Schrödinger's neurons, not Schrödinger's cat. The classical world-simulations run by our minds have been throwaway quantum computers for the last *c*. 540 million years.

Note this is a conservative story. Its background assumptions involve no new principle of physics, no inexplicable violation of unitarity, no observer-induced "collapse of the wavefunction", just the bare formalism of the unitary Schrödinger dynamics (*cf*. Schrödinger equation - Wikipedia).

Dualist philosophers of mind like David Chalmers disagree. Neither classical or quantum physics can explain phenomenal binding *even if* some form of panpsychism or non-materialist physicalism is true (*cf*. The Combination Problem for Panpsychism - Bibliography - PhilPapers). The "structural mismatch" between the formalism of physics and our phenomenally bound classical world-simulations can't be bridged.

Maybe Chalmers is right.

Yet to prove his case, it's not enough for the dualist to demonstrate a structural mismatch between our minds and some cheesy wet lump of neural porridge occupying the four-dimensional space-time of classical physics. The dualist must demonstrate a structural mismatch between the bound phenomenology of our minds and the

fundamental high-dimensional space required by the dynamics of the wavefunction (*cf*. https://www.physicalism.com/hilb...).

Whether such a structural match does or doesn't exist isn't a "philosophical" opinion.

It's an empirical question to be settled by tomorrow's molecular matter wave-interferometry.

What will the non-classical interference signature reveal?

(*cf*. an experimentally testable conjecture)

As a non-materialist physicalist, I predict – tentatively – that interferometry will yield a perfect structural match, and the Hard Problem of consciousness will be solved.

Perhaps Cicero had a point.

What will life be like in the year 3000?

The history of futurology to date makes sobering reading. Prophecies tend to reveal more about the emotional and intellectual limitations of the author than the future. The optimistic prognosis set out below omits the aftermath of twenty-first century thermonuclear war and other horrors from the early millennium.
But here goes…

Year 3000

1) Superhuman bliss.

Mastery of our reward circuitry promises a future of superhuman bliss – gradients of genetically engineered well-being orders of magnitude richer than today's "peak experiences".

Superhappiness?

Building a neuroscience of pleasure and well-being

2) Eternal youth.

More strictly, indefinitely extended youth and effectively unlimited lifespans. Transhumans, humans and their nonhuman animal companions don't grow old and perish. Automated off-world backups allow restoration and "respawning" in case of catastrophic accidents. "Aging" exists only in the medical archives. SENS Research Foundation - Wikipedia

3) Full-spectrum superintelligences.

A flourishing ecology of sentient nonbiological quantum computers, hyperintelligent digital zombies and full-spectrum transhuman "cyborgs" has radiated across the Solar System. Neurochipping makes superintelligence all-pervasive. The universe seems inherently friendly: ubiquitous AI underpins the illusion that reality conspires to help us.

https://en.wikipedia.org/wiki/Superintelligence:_Paths,_Dangers,_Strategies

https://intelligence.org/

http://www.kurzweilai.net/

https://www.biointelligence-explosion.com/parable.html

4) Immersive VR.

"Magic" rules. "Augmented reality" of earlier centuries has been largely superseded by hyperreal virtual worlds with laws, dimensions, avatars and narrative structures wildly different from ancestral consensus

reality. Selection pressure in the basement makes complete escape into virtual paradises infeasible. For the most part, infrastructure maintenance in basement reality has been delegated to zombie AI.

https://en.wikipedia.org/wiki/Augmented_reality

https://en.wikipedia.org/wiki/Virtual_reality

5) Transhuman psychedelia/novel state spaces of consciousness.

Analogues of cognition, volition and emotion as conceived by humans have been selectively retained, though with a richer phenomenology than our thin logico-linguistic thought. Other fundamental categories of mind have been discovered via genetic tinkering and pharmacological experiment. Such novel faculties are intelligently harnessed in the transhuman CNS. However, the ordinary waking consciousness of Darwinian life has been replaced by state-spaces of mind physiologically inconceivable to *Homo sapiens*. Gene-editing tools have opened up modes of consciousness that make the weirdest human DMT trip akin to watching paint dry. These disparate states-spaces of consciousness do share one property: they are generically blissful. "Bad trips" as undergone by human psychonauts are physically impossible because in the year 3000 the molecular signature of experience below "hedonic zero" is missing.

http://www.shulginresearch.org/home/

https://qualiacomputing.com/

6) Supersentience/ultra-high intensity experience.

The intensity of everyday experience surpasses today's human imagination. Size doesn't matter to digital data-processing, but bigger brains with reprogrammed, net-enabled neurons and richer synaptic connectivity can exceed the maximum sentience of small, simple, solipsistic mind-brains shackled by the constraints of the human birth-canal. The theoretical upper limits to phenomenally bound mega-minds, and the ultimate intensity of experience, remain unclear. Intuitively, humans have a dimmer-switch model of consciousness – with e.g. ants and worms subsisting with minimal consciousness and humans at the pinnacle of the Great Chain of Being. Yet Darwinian humans may resemble sleepwalkers compared to our fourth-millennium successors. Today we say we're "awake", but mankind doesn't understand what "posthuman intensity of experience" really means.

What earthly animal comes closest to human levels of sentience?

7) Reversible mind-melding.

Early in the twenty-first century, perhaps the only people who know what it's like even partially to share a mind are the conjoined Hogan sisters. Tatiana and Krista Hogan share a thalamic bridge. Even mirror-touch synaesthetes can't literally experience the pains and pleasures of other sentient beings. But in the year 3000, cross-species mind-melding technologies – for instance, sophisticated analogues of reversible thalamic bridges – and digital analogs of telepathy have led to a revolution in both ethics and decision-theoretic rationality.

http://www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-

share-a-mind.html

https://en.wikipedia.org/wiki/Mirror-touch_synesthesia

8) The Anti-Speciesist Revolution/worldwide veganism/*invitro*tarianism.

Factory-farms, slaughterhouses and other Darwinian crimes against sentience have passed into the dustbin of history. Omnipresent AI cares for the vulnerable via "high-tech Jainism". The Anti-Speciesist Revolution has made arbitrary prejudice against other sentient beings on grounds of species membership as perversely unthinkable as discrimination on grounds of ethnic group. Sentience is valued more than sapience, the prerogative of classical digital zombies ("robots").

What is high-tech Jainism?

The Anti-Speciesist Revolution

'Speciesism: Why It Is Wrong and the Implications of Rejecting It'

9) Programmable biospheres.

Sentient beings help rather than harm each other. The successors of today's primitive CRISPR genome-editing and synthetic gene drive technologies have reworked the global ecosystem. Darwinian life was nasty, brutish and short. Extreme violence and useless suffering were endemic. In the year 3000, fertility regulation via cross-species immunocontraception has replaced predation, starvation and disease to regulate ecologically sustainable population sizes in utopian "wildlife parks". The free-living descendants of "charismatic mega-fauna" graze happily with neo-dinosaurs, self-replicating nanobots, and newly minted

exotica in surreal garden of edens. Every cubic metre of the biosphere is accessible to benign supervision – "nanny AI" for humble minds who haven't been neurochipped for superintelligence. Other idyllic biospheres in the Solar System have been programmed from scratch.

https://en.wikipedia.org/wiki/CRISPR

https://www.gene-drives.com/

http://www.nybooks.com/articles/2007/07/19/our-biotech-future/

10) The formalism of the TOE is known.

(details omitted: does Quora support LaTeX?)

Dirac recognised the superposition principle as *the* fundamental principle of quantum mechanics. Wavefunction monists believe the superposition principle holds the key to reality itself. However – barring the epoch-making discovery of a cosmic Rosetta stone – the implications of some of the more interesting solutions of the master equation for subjective experience are still unknown.

https://en.wikipedia.org/wiki/Theory_of_everything

https://en.wikipedia.org/wiki/M-theory

https://www.quora.com/Why-does-the-universe-exist-Why-is-there-something-rather-than-nothing

https://www.amazon.com/Wave-Function-Metaphysics-Quantum-Mechanics/dp/019979054X

11) The Hard Problem of consciousness is solved.

The Hard Problem of consciousness was long reckoned insoluble. The Standard Model in physics from which (almost) all else springs was a bit

of a mess but stunningly empirically successful at sub-Planckian energy regimes. How could physicalism and the ontological unity of science be reconciled with the existence, classically impossible binding, causal-functional efficacy and diverse palette of phenomenal experience? Mankind's best theory of the world was inconsistent with one's own existence, a significant shortcoming. However, all classical- and quantum-mind conjectures with predictive power had been empirically falsified by 3000 – with one exception.

https://en.wikipedia.org/wiki/Standard_Model

https://en.wikipedia.org/wiki/Hard_problem_of_consciousness

https://en.wikipedia.org/wiki/Quantum_mind

[Which theory is most promising? As with the TOE, you'll forgive me for skipping the details. In any case, my ideas are probably too idiosyncratic to be of wider interest, but for anyone curious: What is Quantum Mind?]

12) The Meaning of Life resolved.

Everyday life is charged with a profound sense of meaning and significance. Everyone feels valuable and valued. Contrast the way twenty-first century depressives typically found life empty, absurd or meaningless; and how even "healthy" normals were sometimes racked by existential angst. Or conversely, compare how people with bipolar disorder experienced megalomania and messianic delusions when uncontrollably manic. Hyperthymic civilization in the year 3000 records no such pathologies of mind or deficits in meaning. Genetically preprogrammed gradients of invincible bliss ensure that all sentient beings find life self-intimatingly valuable. Transhumans love themselves,

love life, and love each other.

Transhumanism

13) Beautiful new emotions.

Nasty human emotions have been retired – with or without the recruitment of functional analogs to play their former computational role. Novel emotions have been biologically synthesised and their "raw feels" encephalised and integrated into the CNS. All emotion is beautiful. The pleasure axis has replaced the pleasure-pain axis as the engine of civilised life.

An information-theoretic perspective on life in Heaven

14) Effectively unlimited material abundance/molecular nanotechnology.

Status goods long persisted in basement reality, as did relics of the cash nexus on the blockchain. Yet in a world where both computational resources and the substrates of pure bliss aren't rationed, such ugly evolutionary hangovers first withered, then died.

http://metamodern.com/about-the-author/

https://en.wikipedia.org/wiki/Blockchain

15) Posthuman aesthetics/superhuman beauty.

The molecular signatures of aesthetic experience have been identified, purified and over-expressed. Life is saturated with superhuman beauty. What passed for "Great Art" in the Darwinian era is no more impressive than year 2000 humans might judge, say, a child's painting by numbers or Paleolithic daubings and early caveporn. Nonetheless, critical

discernment is retained. Transhumans are blissful but not "blissed out" – or not all of them at any rate.

https://en.wikipedia.org/wiki/Art

http://www.sciencemag.org/news/2009/05/earliest-pornography

16) Gender transformation.

Like gills or a tail, "gender" in the human sense is a thing of the past. We might call some transhuman minds hyper-masculine (the "ultrahigh AQ" hyper-systematisers), others hyperfeminine ("ultralow AQ" hyper-empathisers), but transhuman cognitive styles transcend such crude dichotomies, and can be shifted almost at will via embedded AI. Many transhumans are asexual, others pan-sexual, a few hypersexual, others just sexually inquisitive. "*The degree and kind of a man's sexuality reach up into the ultimate pinnacle of his spirit"*, said Nietzsche – which leads to**(17)**.

https://www.livescience.com/2094-homosexuality-turned-fruit-flies.html

https://en.wikipedia.org/wiki/Object_sexuality

https://en.wikipedia.org/wiki/Empathizing

%E2%80%93systemizing_theory

https://www.wired.com/2001/12/aqtest/

17) Physical superhealth.

In 3000, everyone feels physically and psychologically "better than well". Darwinian pathologies of the flesh such as fatigue, the "leaden paralysis" of chronic depressives, and bodily malaise of any kind are inconceivable. The (comparatively) benign "low pain" alleles of the SCN9A gene that replaced their nastier ancestral cousins have been superseded by AI-

based nociception with optional manual overrides. Multi-sensory bodily "superpowers" are the norm. Everyone loves their body-images in virtual and basement reality alike. Morphological freedom is effectively unbounded. Awesome robolovers, nights of superhuman sensual passion, 48-hour whole-body orgasms, and sexual practices that might raise eyebrows among prudish Darwinians have multiplied. Yet life isn't a perpetual orgy. Academic subcultures pursue analogues of Mill's "higher pleasures". Paradise engineering has become a rigorous discipline. That said, a lot of transhumans are hedonists who essentially want to have superhuman fun. And why not?

https://www.wired.com/2017/04/the-cure-for-pain/

http://io9.gizmodo.com/5946914/should-we-eliminate-the-human-ability-to-feel-pain

http://www.bbc.com/future/story/20140321-orgasms-at-the-push-of-a-button

18) World government.

Routine policy decisions in basement reality have been offloaded to ultra-intelligent zombie AI. The quasi-psychopathic relationships of Darwinian life – not least the zero-sum primate status-games of the African savannah – are ancient history. Some conflict-resolution procedures previously off-loaded to AI have been superseded by diplomatic "mind-melds". In the words of Henry Wadsworth Longfellow, "*If we could read the secret history of our enemies, we should find in each man's life sorrow and suffering enough to disarm all hostility."* Our descendants have windows into each other's souls, so to speak.

19) Historical amnesia.

The world's last experience below "hedonic zero" marked a major evolutionary transition in the evolutionary development of life. In 3000, the nature of sub-zero states below Sidgwick's "natural watershed" isn't understood except by analogy: some kind of phase transition in consciousness below life's lowest hedonic floor – a hedonic floor that is being genetically ratcheted upwards as life becomes ever more wonderful. Transhumans are hyper-empathetic. They get off on each other's joys. Yet paradoxically, transhuman mental superhealth depends on biological immunity to true comprehension of the nasty stuff elsewhere in the universal wavefunction that even mature superintelligence is impotent to change. Maybe the nature of e.g. Darwinian life, and the minds of malaise-ridden primitives in inaccessible Everett branches, doesn't seem any more interesting than we find books on the Dark Ages. Negative utilitarianism, if it were conceivable, might be viewed as a depressive psychosis. "Life is suffering", said Gautama Buddha, but fourth millennials feel in the roots of their being that Life is bliss.

Invincible ignorance? Perhaps.

https://en.wikipedia.org/wiki/Negative_utilitarianism

20) Super-spirituality.

A tough one to predict. But neuroscience can soon identify the molecular signatures of spiritual experience, refine them, and massively amplify their molecular substrates. Perhaps some fourth millennials enjoy lifelong spiritual ecstasies beyond the mystical epiphanies of temporal-

lobe epileptics. Secular rationalists don't know what we're missing.

https://www.newscientist.com/article/mg22129531-000-ecstatic-epilepsy-how-seizures-can-be-bliss/

21) The Reproductive Revolution.

Reproduction is uncommon in a post-aging society. Most transhumans originate as extra-uterine "designer babies". The reckless genetic experimentation of sexual reproduction had long seemed irresponsible. Old habits still died hard. By year 3000, the genetic crapshoot of Darwinian life has finally been replaced by precision-engineered sentience. Early critics of "eugenics" and a "Brave New World" have discovered by experience that a "triple S" civilisation of superhappiness, superlongevity and superintelligence isn't as bad as they supposed.

The Reproductive Revolution

Brave New World

22) Globish ("English Plus").

Automated real-time translation has been superseded by a common tongue - Globish – spoken, written or "telepathically" communicated. Partial translation manuals for mutually alien state-spaces of consciousness exist, but – as twentieth century Kuhnians would have put it – such state-spaces tend to be incommensurable and their concepts state-specific. Compare how poorly lucid dreamers can communicate with "awake" humans. Many Darwinian terms and concepts are effectively obsolete. In their place, active transhumanist vocabularies of millions of words are common. "Basic Globish" is used for

communication with humble minds, i.e. human and nonhuman animals who haven't been fully uplifted.

https://plato.stanford.edu/entries/incommensurability/

https://en.wikipedia.org/wiki/Uplift_(science_fiction)

23) Plans for Galactic colonisation.

Terraforming and 3D-bioprinting of post-Darwinian life on nearby solar systems is proceeding apace. Vacant ecological niches tend to get filled. In earlier centuries, a synthesis of cryonics, crude reward pathway enhancements and immersive VR software, combined with revolutionary breakthroughs in rocket propulsion, led to the launch of primitive manned starships. Several are still starbound. Some transhuman utilitarian ethicists and policy-makers favour creating a utilitronium shockwave beyond the pale of civilisation to convert matter and energy into pure pleasure. Year 3000 bioconservatives focus on promoting life animated by gradients of superintelligent bliss. Yet no one objects to pure "hedonium" replacing unprogrammed matter.

https://en.wikipedia.org/wiki/Interstellar_travel

https://en.wikipedia.org/wiki/Utilitarianism

24) The momentous "unknown unknown".

If you read a text and the author's last words are "and then I woke up", everything you've read must be interpreted in a new light – semantic holism with a vengeance. By the year 3000, some earth-shattering revelation may have changed everything – some fundamental background assumption of earlier centuries has been overturned that

might not have been explicitly represented in our conceptual scheme. If it exists, then I've no inkling what this "unknown unknown" might be, unless it lies hidden in the untapped subjective properties of matter and energy. Christian readers might interject "The Second Coming". Learning the Simulation Hypothesis were true would be a secular example of such a revelation. Some believers in an AI "Intelligence Explosion" speak delphically of "The Singularity". Whatever – Shakespeare made the point more poetically, "*There are more things in heaven and earth, Horatio, Than are dreamt of in your philosophy"*.

As it stands, yes, (24) is almost vacuous. Yet compare how the philosophers of classical antiquity who came closest to recognising their predicament weren't intellectual titans like Plato or Aristotle, but instead the radical sceptics. The sceptics guessed they were ignorant in ways that transcended the capacity of their conceptual scheme to articulate. By the lights of the fourth millennium, what I'm writing, and what you're reading, may be stultified by something that humans don't know and can't express.

https://plato.stanford.edu/entries/skepticism-ancient/

OK, twenty-four predictions! Successful prophets tend to locate salvation or doom within the credible lifetime of their intended audience. The questioner asks about life in the year 3000 rather than, say, a Kurzweilian 2045. In my view, everyone reading this text will grow old and die before the predictions of this answer are realised or confounded – with one possible complication.

Opt-out cryonics and opt-in cryothanasia are feasible long before the conquest of aging. Visiting grandpa in the cryonics facility can turn death into an event in life.

> I'm not convinced that posthuman superintelligence will reckon that Darwinian malware should be revived in any shape or form. Yet if you want to wake up one morning in posthuman paradise – and I do see the appeal – then options exist: Alcor

Since the Hedonistic Imperative now seems technically feasible, what are the largest sociological barriers stopping its realization?

"Whatever is, is right."

(Alexander Pope, Epistle 1 of an Essay on Man. 1733–1734)

Should we conserve the biology of suffering?

Or genetically engineer a civilisation based on gradients of intelligent bliss?

HI was written in 1995. Talk of e.g. "Genetically Engineering Almost Anything" (cf. http://www.pbs.org/wgbh/nova/next/evolution/crispr-gene-drives/) could be dismissed as utopian sci-fi. But as you say, from an engineering perspective, HI is feasible - a transhuman "Triple S" civilisation based on superintelligence, superlongevity and superhappiness. Life on Earth could be wonderful and perhaps even sublime. So why isn't a transhumanist agenda yet mainstream?

Perhaps the single greatest obstacle to abolishing the horrors of Darwinian life isn't religious, ethical and ideological opposition. It's status quo bias. Consider physical pain. Words don't do justice to how unbelievably nasty the experience of raw pain can be. Even "mild" uncontrolled chronic pain can lead to clinical or sub-clinical depression. We now have the technology (cf. https://en.wikipedia.org/wiki/Preimplantation_genetic_diagnosis) to ensure that all

children born into the world are blessed with an extremely high pain tolerance - the kind of pain-threshold of today's genetic outliers who insist, "Pain is just a useful signalling mechanism." Eventually, even "mild" physical pain can be eliminated in favour of pain-free nociception. (cf. "Should we eliminate the human ability to feel pain?" http://io9.gizmodo.com/5946914/should-we-eliminate-the-human-ability-to-feel-pain) In the meantime, no holy religious text proclaims, "Thou shalt not use preimplantation genetic screening to ensure your future children are born with benign 'low-pain' alleles of the SCN9A gene." (cf. https://www.wired.com/2017/04/the-cure-for-pain/ - "How a Single Gene Could Become a Volume Knob for Pain") Yet most religious and secular people continue to have children via the time-honoured genetic crapshoot, trusting that Providence or Mother Nature will lead to a happy outcome. (cf. https://en.wikipedia.org/wiki/Appeal_to_nature)

Or consider the suffering of nonhuman animals, both domestic and free-living ("wild"). Closing and outlawing factory-farms and slaughterhouses would entail minimal personal inconvenience to consumers. No need to wait until cheap gourmet in vitro meat products reach the supermarket shelves. If more people can be induced to explore e.g. plant-based veggieburgers, then meat-eaters would realise that switching to a cruelty-free lifestyle would have a negligible impact on their own quality of life. Once again, the dead weight of tradition hangs heavy. Recognising there is something deeply morally wrong (cf. Speciesism: Why It Is Wrong and the Implications of Rejecting It by Magnus Vinding) with what ordinary, "decent" people have done all their lives doesn't come naturally to most of us. (cf. https://aeon.co/essays/what-will-our-descendants-judge-as-our-greatest-sin - "What will our descendants judge as our greatest sin?")

More ambitiously, the entire biosphere is now programmable via synthetic CRISPR-based gene drives. Vector-borne disease is eliminable – to the benefit of human and nonhuman

animals alike. Our reward circuitry too is reprogrammable, not just in humans, but across the tree of life. Intelligent moral agents will shortly be in a position to choose the optimal level of suffering in the living world in defiance of the "laws" of Mendelian inheritance. Unlike giving up meat, this challenge is computationally non-trivial. But status quo bias means that most people reflexively support "conservation biology", or even reactionary proposals like "re-wilding", without giving the terrible suffering of nonhumans a second thought.

Of course, the problem isn't "just" status quo bias, or even ethical-ideological rationalisation of our daily woes. We shouldn't gloss over well-reasoned objections to any grandiose megaproject to eliminate suffering. Who's going to be in charge? The UN? The World Health Organization? Who will pay? The risks of genome-editing are real. Any critic who pleads for exhaustive prior research before we start editing germ-lines should be respected. The technical obstacles to getting rid of all experience below "hedonic zero" aren't insuperable, at least to the best of our knowledge (cf. The Church-Turing Thesis); but they are still huge. "Mental" distress is complex. The scope for unanticipated side-effects and "unknown unknowns" from biological interventions is indisputably far-reaching. Genes and culture co-evolved. The high genetic loading of hedonic set-points doesn't make socio-economic reform any less urgent. There's also the question of sociologically and technically realistic timescales. Not least, cheating the negative feedback mechanisms of the hedonic treadmill, and genetically raising hedonic set-points so we all feel "better than well", isn't nearly as easy as genetically reducing the burden of physical pain.

Maybe the best way to tease apart principled objections to global biohappiness from mere status quo bias is to pose a thought-experiment. Variants of this thought-experiment can also be devised for any other item on the transhumanist agenda. Ask the

critic to imagine we encounter an advanced civilisation that has rewritten its genetic source-code. Its members are animated entirely by information-sensitive gradients of well-being - a default hedonic state far richer than human “peak experiences”. Let’s assume that the genetically-tweaked descendants of ancestral "wildlife" graze blissfully in their conservation parks. Population sizes are regulated by cross-species immunocontraception rather than starvation, disease and predation. The extra-terrestrials are hyper-intelligent, i.e. they aren’t “blissed out” (cf. gradients.com - "An information-theoretic perspective on life in Heaven"). Yet most of their sensual, intellectual and psychonautic delights are alien to us. ("The limits of pleasures are as yet neither known nor fixed, and we have no idea what degree of bodily bliss we are capable of attaining" - Jean Anthelme Brillat-Savarin). Now for the crux. What credible arguments might human bioconservative critics use to persuade this advanced civilisation to re-introduce the biology of involuntary suffering and malaise - and all the other nasty states of mind that were fitness-enhancing in their ancestral environment? Depending on the degree of convergent evolution, perhaps their ancestors too once experienced jealousy, resentment, envy, spite, depression, status-anxiety, existential angst – all the ghastly stuff we call “part of what it means to be human”. What exactly are their superhappy minds missing? Should they practise “re-wilding” and bring it back?

Quite possibly they'd view human primitives as in the grip of a depressive psychosis.

Would they be right?