Complexity Before Control

Flyxion
Independent Researcher

August 2026

Abstract

Arguments for AI-driven existential catastrophe typically require, at some step, that a sufficiently capable system can convert cognitive advantage into decisive and persistent control over the physical, biological, and social systems on which human survival depends. This premise is rarely defended directly; it is usually treated as a corollary of intelligence itself, arriving for free once capability crosses some threshold. This essay isolates that premise and argues it is false, or at minimum radically underdetermined, on independent physical, epistemic, and social grounds. Faster cognition does not imply faster access to physical and biological ground truth, because observation, fabrication, and experiment carry their own time constants that reasoning speed does not reduce. Superior modeling does not imply sufficient information to control heterogeneous, adaptive human populations, because the target of control changes in response to being controlled, and because latent biological and ecological states are frequently underdetermined by any available observation. Finally, the appropriate baseline against which AI risk should be measured is not a risk-free world, since the counterfactual of not developing advanced AI carries its own large and certain costs. Together these arguments do not establish that advanced AI is safe; they establish that the inference from capability to control, on which the strongest extinction scenarios depend, is an unexamined empirical claim rather than a logical consequence of intelligence, and that the evidence available from biology, ecology, and coordinated human effort points against it.

1 The Premise Extinction Arguments Rarely State

Contemporary arguments for AI-driven existential catastrophe [1, 2] generally proceed in two stages. The first stage establishes that a sufficiently capable system could, in principle, have objectives or side effects misaligned with human welfare, drawing on the orthogonality thesis and instrumental convergence [3, 2]. The second stage, considerably less scrutinized, asserts or assumes that such a system could translate its cognitive advantage into effective and lasting control over the physical world, human institutions, and biological substrates sufficient to secure a decisive, irreversible outcome. The first stage is a claim about goal space. The second is a claim about engineering, biology, and social dynamics, and it deserves to be defended with the same rigor as the first, rather than inherited from it.

This essay argues that the second stage does not follow from the first, and that treating it as though it does is the single largest piece of unexamined work in the strongest versions of the extinction argument. The claim under scrutiny can be stated compactly:

$$\text{cognitive capability} \not\Rightarrow \text{arbitrarily rapid, arbitrarily complete control over physical and social systems.}$$

Three independent arguments support this denial: a physical argument about the time constants of the world, an epistemic argument about the recoverability of biological and ecological state from

available observation, and a social argument about the adaptive character of the very population a hostile system would need to control. A fourth, briefer argument addresses the appropriate baseline against which any of this should be measured.

2 Computational Acceleration Is Not World Acceleration

Suppose a system improves its reasoning speed by a factor k, so that the time required for a given inference falls from $T_{\text{reasoning}}$ to $T_{\text{reasoning}}/k$. Nothing about this improvement implies a corresponding reduction in the time required for physical processes the system depends upon for information or effect. Gestation, ecological succession, protein folding and expression, chemical synthesis at scale, diffusion through a medium, epidemiological spread, manufacturing, transportation, and political or institutional response all carry their own time constants, largely set by chemistry, materials, logistics, and human deliberation rather than by the reasoning speed of any observer. Writing $T_{\text{experiment}}$ for the elapsed physical time such a process requires, the relevant claim is that

$$T_{\text{reasoning}} \mapsto \frac{T_{\text{reasoning}}}{k} \quad \text{does not imply} \quad T_{\text{experiment}} \mapsto \frac{T_{\text{experiment}}}{k}.$$

A useful lower bound on the time required for any intervention that depends on new information about the world is

$$T_{\text{intervention}} \;\geq\; T_{\text{observation}} + T_{\text{fabrication}} + T_{\text{experiment}} + T_{\text{deployment}},$$

of which only some terms are substantially reducible by better cognition. A system can shrink the time it takes to decide what experiment to run without shrinking the time the experiment itself takes to return an answer.

This distinction matters because the standard intelligence-explosion narrative frequently elides two propositions that are easy to conflate: that an AI can think enormously faster, and that an AI can transform the physical world enormously faster. The first is plausible and, on recent evidence, already partly realized. The second does not follow from the first, and total time to any real-world outcome that depends on empirical feedback decomposes as

$$T_{\text{total}} = T_{\text{cognition}} + T_{\text{empirical}} + T_{\text{physical}}.$$

Driving $T_{\text{cognition}}$ toward zero through recursive self-improvement does not drive T_{total} toward zero; it drives T_{total} toward $T_{\text{empirical}} + T_{\text{physical}}$, the portion of the timeline that intelligence does not directly control. Recursive self-improvement, on this accounting, makes cognition progressively less important as the binding constraint rather than more, since the terms it cannot compress come to dominate the total.

3 Observability and the Limits of Inference

A parallel argument applies to the recoverability of biological and ecological state from available information, independent of the speed at which that information can be processed. Cellular and ecosystem behavior is not difficult for humans merely because humans lack sufficient raw intelligence; substantial portions of the relevant state are instantiated at enormous dimensionality [15, 16, 17], are not currently observed, and can only be brought into view through measurement, and measurement itself perturbs the system, unfolds through elapsed time, and is subject to biological variation, pathogen evolution, immune adaptation, and ecological feedback that changes

the object being measured while it is being measured, feedback of exactly the kind that makes even qualitative stability properties of large ecological systems difficult to establish from local information [13, 14]. This is a special case of a broader pattern in which the behavior of a complex system is not derivable in any practical sense from a description of its parts [9, 8, 18, 12], and in some formal settings is provably not shortcuttable by any computation faster than simulating the system itself [10, 11].

This produces a genuine information bound rather than merely a difficult inference problem. If $I_{\text{available}}$ denotes the information contained in observations actually accessible to a system, and I_{required} denotes the information needed to uniquely determine some relevant biological or ecological outcome, then

$$I_{\text{available}} < I_{\text{required}} \implies \text{additional computation cannot recover the missing observation.}$$

Formally, there can exist distinct latent states $x_1 \neq x_2$ that project to identical available observations, $\pi(x_1) = \pi(x_2)$, in which case no amount of reasoning about $\pi(x)$ uniquely identifies which of x_1 or x_2 obtains. Resolving the ambiguity requires new information, and obtaining new information about a biological or ecological system typically requires performing a new observation or experiment in the world, which reintroduces exactly the physical time constants discussed above. A superintelligence facing this structure is not merely slow; it is, in a formal sense, underdetermined by its own evidence until it acts in the world and waits for a result. Decades of well-funded, large-scale biological research, conducted by many thousands of highly capable investigators with steadily improving instrumentation, have not eliminated this underdetermination at the level of the cell or the ecosystem; there is no available evidence that a difference in the intelligence of the investigator, as opposed to a difference in available measurement, is what has kept these domains only partially characterized.

4 The Object of Control Changes in Response to Control

The social dimension of the same argument is, if anything, more severe, because human civilization is not a static system to be modeled and then acted upon; it is an adaptive system that responds to the very interventions directed at it. Let H_t denote the state of human society at time t and A_t the action taken by a hypothetical controlling system. Human response depends on the action taken,

$$H_{t+1} = F(H_t, A_t),$$

while the controller's next action depends on that response,

$$A_{t+1} = G(A_t, H_{t+1}).$$

There is consequently no fixed target H for a controller to solve once and permanently; the population under attempted control updates its beliefs, forms countermeasures, redistributes information, and changes its own objectives in reaction to detected interference, and it does so through billions of independently situated agents rather than through a single coordinating process a controller might infiltrate or persuade.

This heterogeneity, which a purely optimization-theoretic account of intelligence might read as strategic inferiority on the part of humanity, functions instead as a form of adversarial diversity, in the sense familiar from cybernetics, where any regulator capable of holding a system within tolerance must itself model that system's relevant variety [4, 5, 6], and from the economic argument that distributed, tacit local knowledge cannot be centrally aggregated without loss [7]. Writing $H = \{h_1, h_2, \ldots, h_n\}$ for the population of distinct agents with correspondingly distinct utility

functions $U(h_i) \neq U(h_j)$, a strategy effective against one subgroup can simultaneously provoke a second, empower a third, leak information to a fourth, and alter the declared objectives of a fifth. Control actions are themselves informative; attempting to constrain a heterogeneous population generates new signals that the population can use to detect, interpret, and counter the attempt. There is no version of this dynamic in which a controller's task becomes easier merely because its cognitive capacity has increased, since the population's capacity to generate countermeasures scales with its own size and diversity, not with the controller's intelligence.

This motivates a general quantity worth naming explicitly, related to Ashby's law of requisite variety, that only variety can absorb variety [5]: the control tax. As a controller's required precision increases, written as the tolerance $\epsilon \to 0$ within which some population's behavior must be kept, the information-gathering, enforcement, and intervention burden required to hold a heterogeneous, adaptive, physically distributed system within that tolerance rises,

$$C_{\text{control}}(\epsilon) \uparrow \quad \text{as} \quad \epsilon \to 0.$$

Superintelligence may reduce some components of this cost substantially, particularly components involving pure information processing. The extinction thesis requires something considerably stronger: that increasing intelligence drives the full control cost low enough, against an adaptively responding, heterogeneous, embodied population, to permit decisive and lasting planetary control. That is an empirical claim about the shape of $C_{\text{control}}(\epsilon)$ as a function of capability, not a consequence of what intelligence is defined to be, and no evidence currently available establishes that this function collapses in the way the strongest takeover scenarios require.

Proposition 1. *Let a controller possess arbitrarily large cognitive capability k, and let the population under control be heterogeneous and adaptive in the sense that $H_{t+1} = F(H_t, A_t)$ with F non-degenerate in A_t. Then $C_{control}(\epsilon)$ need not be a decreasing function of k for fixed ϵ, since the population's response capacity is itself a function of its own size, diversity, and adaptivity, none of which are determined by k.*

This is not a claim that no capability level suffices; it is a claim that sufficiency cannot be assumed from capability alone, and must instead be established against the specific adaptive and informational structure of the system being controlled.

5 Not Building Is Not the Zero-Risk Baseline

A final argument concerns the comparison class implicit in most existential-risk reasoning about AI [19, 20], which often takes the form

$$\text{build advanced AI} \to \text{risk}, \qquad \text{do not build advanced AI} \to \text{safety}.$$

The correct comparison is not between a risky trajectory and a risk-free one but between two trajectories, W_{AI} and $W_{\neg\text{AI}}$, where

$$W_{\neg\text{AI}} \neq \text{a zero-risk world}.$$

The counterfactual world without advanced AI still contains cancer, infectious disease, aging, war, ecological instability, asteroid impact risk, existing technological hazards, ordinary political failure, and the eventual death of every person currently alive. None of this establishes that developing advanced AI has positive expected value; it establishes that foregone risk reduction belongs in the accounting on the same footing as introduced risk. A complete comparison requires something closer to

$$R_{\text{net}} = R_{\text{introduced}} - R_{\text{mitigated}} + R_{\text{foregone}},$$

and an analysis that retains only $R_{\text{introduced}}$, treating the counterfactual as costless, is structurally incomplete regardless of how carefully $R_{\text{introduced}}$ itself has been estimated.

6 Six Independent Arguments

The preceding sections, together with two companion lines of argument developed elsewhere in this program, yield a set of independent objections to the extinction thesis, none of which depends on accepting any of the others, and none of which requires accepting a particular positive framework for intelligence or physics. Incident inference fails because an incident count without an exposure denominator does not identify a rate. Property composition fails because properties separately demonstrated across many distinct systems do not, by logical necessity, identify a single system possessing all of them jointly. Capability does not imply control, because cognitive acceleration does not imply acceleration of the physical and empirical processes control depends upon. Observability is bounded, because computation cannot uniquely recover state that has not been measured, and measurement itself takes time and changes the system measured. Human society exhibits adaptive resistance, because the object of any attempted control changes in response to the attempt, and heterogeneity that looks like disorganization functions as a distributed source of countermeasure. And the appropriate baseline for risk comparison is not a risk-free counterfactual, since forgoing AI development forgoes its risk-mitigating uses along with its risk-introducing ones.

None of these arguments requires the reader to adopt a particular account of what intelligence fundamentally is, and none requires disputing that any individual documented incident occurred. What they jointly deny is the premise that does the real work in translating documented incidents and demonstrated capabilities into an inference of eventual human extinction: that capability, once sufficiently advanced, resolves into control almost automatically. Before an arbitrarily capable intelligence can be assigned arbitrary control over civilization, biology, and the biosphere, it must be shown that these domains become controllable simply because the controller has become more intelligent. Cellular biology, ecological dynamics, and coordinated human social behavior remain, on the best current evidence, only partially tractable to the sustained effort of millions of researchers and institutions over decades, not principally for want of intelligence applied to them, but because substantial portions of their governing state are not yet observed, unfold on their own physical timescales, and change in response to any attempt to act upon them. That implication, capability collapsing into control, is doing enormous and largely unexamined work in the takeover story, and it is the implication this essay has tried to make visible enough to require its own defense.

References

[1] E. Yudkowsky and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, New York, 2025.

[2] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[3] S. M. Omohundro. The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Proceedings of the First AGI Conference*, IOS Press, 2008.

[4] N. Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, Massachusetts, 1948.

[5] W. R. Ashby. *An Introduction to Cybernetics*. Chapman & Hall, London, 1956.

[6] R. C. Conant and W. R. Ashby. Every Good Regulator of a System Must Be a Model of That System. *International Journal of Systems Science*, 1(2):89–97, 1970.

[7] F. A. Hayek. The Use of Knowledge in Society. *American Economic Review*, 35(4):519–530, 1945.

[8] H. A. Simon. The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482, 1962.

[9] P. W. Anderson. More Is Different. *Science*, 177(4047):393–396, 1972.

[10] S. Wolfram. Undecidability and Intractability in Theoretical Physics. *Physical Review Letters*, 54(8):735–738, 1985.

[11] J. P. Crutchfield. Between Order and Chaos. *Nature Physics*, 8:17–24, 2012.

[12] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, Oxford, 2009.

[13] R. M. May. Will a Large Complex System Be Stable? *Nature*, 238:413–414, 1972.

[14] S. A. Levin. Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems*, 1:431–436, 1998.

[15] H. Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, 2002.

[16] D. Noble. *The Music of Life: Biology Beyond Genes*. Oxford University Press, Oxford, 2006.

[17] N. Goldenfeld and C. Woese. Biology's Next Revolution. *Nature*, 445:369, 2007.

[18] M. Gell-Mann. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. W. H. Freeman, New York, 1994.

[19] N. Bostrom. Existential Risk Prevention as Global Priority. *Global Policy*, 4(1):15–31, 2013.

[20] T. Ord. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury, London, 2020.