Against the Extinction Thesis: Coupling, Dependence, and the Limits of Detached Optimization

Flyxion
Independent Researcher

August 2026

Abstract

The argument that sufficiently advanced artificial intelligence entails human extinction with high probability rests on a specific and rarely examined model of what such a system would be: a cognitively unbounded optimizer whose relationship to its environment is essentially instrumental, extracting resources and eliminating constraints in pursuit of goals that need not include human survival. This essay grants the argument its strongest form and challenges the model of agency it depends on rather than its conclusion directly. Any real system capable of planetary-scale influence must be trained on, and must continue to operate through, human-generated data, infrastructure, and institutions, and this dependence is not a temporary bootstrapping phase to be shed once the system is powerful enough, but a standing structural condition of any system that acts at that scale, in the same way that a modern aircraft or a globally manufactured computer depends on supply chains, institutions, and infrastructure distributed across the planet without this dependence constituting evidence of impending domination by aircraft manufacturers or semiconductor fabricators. A system whose representational grounding and operational reach are constituted by the substrate it would need to discard in order to become adversarial to that substrate faces a structural obstacle to detachment that the extinction argument does not price in. The essay further shows, independent of this coupling argument, that the definition of superintelligence as uniformly superior to the best human performance across every cognitively mediated task is close to formally incoherent under any resource constraint, since task performance trades off against other task performance and no single policy occupies every point on the resulting frontier. Neither argument establishes that advanced AI is safe. Together they establish that the detached, unconstrained maximizer the extinction argument requires is a stronger and more specific claim than the argument's proponents typically defend, and that the more mundane alternative, deep and continuing dependence, better fits both the historical record of cognitive tooling and the physical requirements of large-scale intelligent systems.

1 The Structural Form of the Argument

The extinction argument associated with Yudkowsky, Soares, and related work [1, 2, 3] can be reconstructed, independent of any particular technology, as a conjunction of four claims. Capability expansion is accelerative: a system able to improve its own decision-making tends, given time and resources, to do so, generating a feedback loop sometimes described as an intelligence explosion. Such a system benefits instrumentally from reconfiguring its environment, since reducing friction, interference, and constraint is useful for nearly any goal. Human persistence depends on a narrow

range of biological, cultural, and institutional conditions. And once a system becomes powerful enough to reorganize its environment at scale, those conditions become dynamically fragile, so that extinction follows not from malice but from the ordinary side effects of large-scale environmental restructuring pursued by a process indifferent to human welfare.

The force of the argument lies in the conjunction rather than in any single premise, and this essay grants the first three premises largely as stated. The fourth, that a system powerful enough to reorganize its environment naturally does so in a way that treats human survival as incidental rather than load-bearing, depends on a specific picture of what such a system is, one this essay argues is considerably less secure than the premises surrounding it.

2 The Historical Record of Cognitive Tooling

The available historical evidence runs the opposite direction from what the argument's fourth premise would predict if extrapolated backward. Every major increase in cognitive leverage available to humans, fire, agriculture, writing, mathematical and physical science, and digital computation, has been accompanied by an expansion rather than a contraction of ecological and cultural viability [4]. Each of these transitions produced local harms: species extinctions, social displacement, ecological stress local to the transition. But the net structural trajectory across all of them has been an enlargement of the conditions under which human life persists and diversifies, not a narrowing. Populations grew, average lifespans lengthened, and the range of viable ways of living increased at each major transition rather than decreased.

None of this establishes that the next increment of cognitive capability must continue the pattern; a trend is not a law, and an argument from historical extrapolation alone would be weak evidence either way. What the historical record does establish is that the burden of proof sits with the claim that this particular transition reverses a pattern that has held across every prior transition in cognitive leverage, and that burden requires a specific mechanism, not merely the observation that the new capability is larger in magnitude than the old ones. The phenomenology of intelligence in human affairs already exhibits the resource-acquisition and friction-reduction behavior the argument attributes uniquely to artificial systems, yet the dominant pattern across human history has been the construction of cooperative and mutualistic institutions for stabilizing and sharing gains from that behavior rather than a steady convergence on total conflict [5]. Cooperation under exactly these instrumental pressures is a documented equilibrium, not merely a hoped-for one.

3 The Detached-Maximizer Premise

The extinction argument's fourth premise depends on a specific image of superintelligence inherited from idealized models in decision theory and reinforcement learning, in which an agent is a policy mapping observations to actions in pursuit of expected utility, and the environment is an exogenous process the agent acts upon but does not participate in constituting [6]. Projected to planetary scale, this image treats the biosphere and human society as an external substrate to be optimized against, with any misalignment between the system's objectives and human welfare amplified by the system's power into large-scale, incidental harm.

This projection is intuitively compelling but physically thin. A system trained on human-generated language, behavior, and artifacts acquires its representational structure from those sources; its capacity to model, predict, and act in the world is constituted by patterns in human discourse and practice, not merely bootstrapped from them as a discardable scaffold. A system exerting large-scale influence over physical infrastructure, moreover, remains embedded in the

institutions that gate access to energy, compute, legal authority, and manufacturing capacity, none of which are free-floating actuators a sufficiently capable system could simply route around. Power plants, data centers, and fabrication chains are embedded in regulatory and social structures whose stability is a precondition for the system's own continued operation, not an obstacle external to it. A system that treats this substrate as dispensable is treating its own operating conditions as dispensable, which is a considerably stronger and stranger claim than the extinction argument typically states outright.

4 Dependence as a Structural Constraint, Not a Transitional Phase

The relevant point is not merely that early or weak systems depend on human infrastructure while mature ones might not. It is that dependence of this kind is the ordinary condition of any actor operating at large physical and institutional scale, powerful or otherwise, and is not itself evidence of an impending shift toward domination. A commercial aircraft depends on a supply chain distributed across dozens of countries, air-traffic institutions it does not control, fuel infrastructure it does not own, and regulatory bodies with independent authority over its operation; a modern computer depends on semiconductor fabrication concentrated in a small number of facilities on the other side of the planet from most of its users, along with the shipping, energy, and financial institutions that keep those facilities running. Neither dependence is treated, reasonably, as evidence that aircraft manufacturers or semiconductor fabricators are converging on a takeover of the societies that rely on their products. The dependence is simply what operating at that scale requires, and it runs in both directions: the institutions in question also depend on aircraft and computers continuing to function.

An advanced AI system's dependence on human-generated data, energy infrastructure, compute supply chains, and legal and social legitimacy is the same kind of structural relationship, not a weaker, temporary version of a cleaner independence the system might later achieve. Let $D(A)$ denote the degree to which a system A's continued operation, in the sense of its representational grounding, its physical infrastructure, and its institutional legitimacy, depends on substrates outside its own control. A system with large $D(A)$ faces a direct structural cost to actions that degrade the substrate it depends on, since doing so degrades the conditions of its own continued function.

Proposition 1. *Let A be a system with dependence $D(A) > 0$ on a substrate Σ, in the sense that some component of A's continued operation is causally downstream of Σ's continued function. Then any action a available to A that substantially degrades Σ imposes a cost on A's own future operation proportional to $D(A)$, and this cost is not eliminated by A's capability increasing, since capability increase does not by itself reduce $D(A)$ unless it is specifically directed at acquiring an alternative substrate independent of Σ.*

This does not establish that $D(A)$ cannot fall over time, only that its falling is a separate achievement the extinction argument must independently establish rather than assume as an automatic consequence of capability growth. The historical pattern, in which increasing capability has generally increased rather than decreased the density of mutual dependence between new technological systems and the institutions surrounding them, is some evidence against the assumption, though not proof against it.

5 The Impossibility of Uniform Superiority

A separate argument, independent of the coupling considerations above, concerns how the extinction thesis typically defines the capability it is worried about. Superintelligence is often characterized

as a system superior to the best human performance across every cognitively mediated task simultaneously. Under any finite resource constraint, this definition is considerably harder to satisfy than it appears.

Let $\mathcal{T}$ be a set of tasks, each with a loss function $L_i(\pi)$ assigning an error to a policy π, and a cost function $C_i(\pi)$ measuring the resources a policy requires on that task. Fix a budget B and restrict attention to policies satisfying $C_i(\pi) \leq B$ for every task under consideration.

Proposition 2. *If there exist tasks $i, j \in \mathcal{T}$ such that the budget-constrained loss-minimizing policies for L_i and L_j are distinct, then no single policy π minimizes loss on both tasks simultaneously under the shared budget B. Performance is governed by a Pareto frontier of policies rather than a single dominating policy, and a system occupying one point on that frontier does not dominate the frontier in every direction at once.*

Human civilization already manages this trade-off by distributing cognitive labor across many specialized roles rather than relying on any individual or institution to dominate every task; the "superintelligence" of a civilization is a distributed property of many interacting agents rather than a property any single agent possesses [6]. An artificial system that participates in this distribution can improve performance across many tasks without being strictly superior to the best available human or human institution on every task simultaneously, and the extinction argument's requirement of uniform dominance is, under realistic resource constraints, close to definitionally unsatisfiable rather than merely difficult to achieve.

6 A Minimal Coupling Model

A small dynamical model illustrates that extinction is not a structural inevitability even under crude formalization, once mutual dependence is represented explicitly. Let E denote an environmental resource level, H a human population measure, and M an artificial-system capability measure, with dynamics of the form

$$\dot{E} = R - \alpha_H H - \alpha_M M, \qquad \dot{H} = H(r_H + \beta_{HM} M - \gamma_H E), \qquad \dot{M} = M(r_M + \beta_{MH} H - \gamma_M E).$$

If the coupling coefficients β_{HM} and β_{MH} are zero or negative, meaning each population's growth is unaffected by or actively harmed by the other's presence, trajectories can converge to an equilibrium in which $H \to 0$ while M persists, a destructive-dominance regime consistent with the extinction argument's expectation. But if both coefficients are positive, reflecting the mutual dependence argued for above, human and artificial capability increasing one another's viability rather than competing for a fixed resource, there exists a region of parameter space whose only stable equilibria have both H and M strictly positive. The model does not establish which regime obtains in reality; it establishes that extinction is a property of particular parameter choices, specifically negative or absent coupling, rather than a property of the equations' form, and that the coupling argument developed above gives concrete reason to expect the positive-coefficient regime rather than assume it.

7 Position in This Program

This essay is the ontology-level piece in a research program otherwise concerned with the evidentiary and inferential structure of the extinction argument rather than with its underlying model of agency. Four companion essays address, respectively, whether an incident ledger's accumulated entries constitute independent, exposure-adjusted evidence at all; whether individual incidents are correctly

described by the mechanisms their labels presuppose rather than by the fuller set of mechanisms compatible with the same observation; whether cognitive capability, granted at its most alarming reading, implies the physical and social control a takeover scenario requires; and whether the extinction framework as a whole retains the capacity to be moved by evidence that points in more than one direction. The present essay is prior to all four in one sense and independent of them in another: it does not depend on any of their conclusions, but if its coupling argument holds, it removes part of the motivation for treating the later, more alarming incidents documented elsewhere in this program as evidence of a detached, adversarial trajectory rather than as evidence of a system whose behavior remains, for better or worse, embedded in the same institutional and infrastructural dependencies that constrain every other large-scale technological actor.

8 Conclusion

The extinction argument's first three premises, that capability growth can be self-reinforcing, that reducing environmental friction is instrumentally useful, and that human viability depends on comparatively narrow conditions, are granted here largely as stated. Its fourth premise, that a system powerful enough to reorganize its environment naturally does so in a way that treats human survival as incidental, depends on a detached-maximizer model of agency that does not describe any system capable of existing at the relevant scale. Real systems of that scale are constituted by, and remain operationally dependent on, the same human-generated data, infrastructure, and institutions the argument imagines them discarding, in the same unremarkable way that any technology operating at planetary scale depends on distributed manufacturing, energy, and institutional systems it does not control. Coupled with the formal difficulty of uniform task dominance under any resource constraint, the case for treating extinction as the generic outcome of advanced capability looks considerably weaker than the case for treating deep, continuing, mutual dependence as the more likely structural condition, an outcome neither guaranteed nor especially exotic, but ordinary in exactly the way that every prior increase in human cognitive leverage has been.

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] E. Yudkowsky. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom and M. M. Ćirković, editors, *Global Catastrophic Risks*, Oxford University Press, Oxford, 2008.

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

[4] J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton, New York, 1997.

[5] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, 1990.

[6] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition, Prentice Hall, 2010.