Is the Singularity Problem Just Fiction?

By Faraz Fookeer

Introduction

There are few concepts in the philosophy of artificial intelligence (AI) that occupy as uncomfortable a position as the “Singularity”. This is a term that originated in the speculative writings of computer scientists and mathematicians - Vernor Vinge coined the term “technological singularity” (Vinge, 1983, 1986, 1993) - and was then elevated (and elaborated) into a near-prophetic framework by Ray Kurzweil (Kurzweil, 1989, 1999, 2005). Mainstream researchers routinely dismissed the Singularity and the cluster of philosophical problems that it generates as technological mythology but it currently inhabits an uncomfortable no-man’s land between rigorous argument and speculative (bordering on sci-fi) narrative. In this essay, I shall refer to the Singularity as the Singularity Problem, referring to the prediction that superhuman AI will be created and also the set of philosophical and practical challenges that such a technology is speculated to generate. For example, the loss of predictability over technological development, a flavour of the “alignment problem” that questions the goals of recursively self-improving machines and their alignment to human goals and values, and the potential for machine intelligence advancing beyond any human comprehension. To dismiss this notion as “just fiction” - as Paul Allen and Mark Greaves approached in their 2011 MIT Technology review essay - is to reject the problems as the story that conveys it is imperfectly told. This essay argues that the heralding of the Singularity as a sudden, monolithic event is conceptually confusing and empirically implausible - the classical stance - the underlying Singularity Problem is by no means fictional. It is an urgent philosophical concern and recent research in AI safety theory and AI research suggests that rather than dismissing it, we should reformulate it.

I. Presenting the Singularity Problem

The concept of the Singularity draws on a rich intellectual lineage. Irving John Good, in his 1965 paper "Speculations Concerning the First Ultraintelligent Machine," formulated what remains the essential logic of the problem: "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make" (Good, cited in Kurzweil, 2005). Good's formulation identifies the recursive self-improvement loop at the heart of the concern: once a machine surpasses human intelligence in the capacity to design better machines, the process becomes self-sustaining and potentially unbounded.

Vernor Vinge, in his 1993 paper delivered at the NASA-sponsored VISION-21 symposium, gave this concern its canonical name and articulated its most unsettling implication. Predicting that technology would produce entities with greater-than-human intelligence within thirty years, Vinge argued that "when greater-than-human intelligence drives progress, that progress will be much more rapid... [W]e are entering a regime as radically different from our human past as we humans are from the lower animals... a throwing away of all the previous rules, perhaps in the blink of an eye, an exponential runaway beyond any hope of control" (Vinge, 1993). The term "singularity" is deliberately borrowed from astrophysics and mathematics - it designates a point beyond which existing models break down and prediction becomes impossible. For Vinge, this epistemic rupture is precisely the problem: we cannot reason, plan, or predict across such a threshold.

Kurzweil extended this framework into a systematic theory grounded in the "Law of Accelerating Returns" (Kurzweil, 2005). On this view, the exponential growth of information technology - roughly doubling in price-performance every year - is not a temporary phenomenon but a deep feature of any evolutionary process. As computation approaches and eventually exceeds the capacity of the human brain, and as machines acquire the capacity to redesign themselves, the pace of change becomes so extreme as to constitute a qualitative rupture in history, which Kurzweil projected to occur around 2045. Crucially, for Kurzweil as for Vinge, the problem is not merely technical but philosophical: it concerns the nature of mind, the limits of human understanding, and the irreversible alteration of the human condition.

Three distinct philosophical problems are thus embedded in the Singularity scenario. The first is the *predictability problem*: by definition, what lies beyond the Singularity cannot be modelled from within our current cognitive framework, rendering meaningful planning or ethical anticipation impossible. The second is the *control problem*: if a machine can surpass human intelligence and modify its own architecture, it becomes deeply unclear how human values, interests, or oversight can remain operative. Vinge himself acknowledged that confinement strategies are probably impractical, since a sufficiently intelligent system might outmanoeuvre its supervisors over time even with restricted access to the outside world. The third is the *existential problem*: the prospect of the "human era" being ended is not merely a dramatic rhetorical flourish but a genuine challenge to our understanding of humanity's long-run status in relation to its own technological creations. These three intertwined problems - not merely the narrative of a robot uprising - constitute what deserves to be called the Singularity Problem.

II. The Case for "Just Fiction"

The most substantive challenge to the Singularity Problem as a genuine philosophical concern comes from Allen and Greaves (2011). Their central argument is what they call the "complexity brake": the rate of scientific progress in understanding human cognition is fundamentally incompatible with the smooth exponential curves on which Kurzweil's predictions depend. Rather than progress becoming easier as computational power grows, each advance in neuroscience and AI reveals greater complexity, not less. Knowing how neurons are connected tells us remarkably little about how cognition actually works; as Allen and Greaves argue with an

analogy, having a complete anatomical diagram of a bird does not allow us to simulate its flight, because we also need to know how all components function together. The same limitation applies with even greater force to the human brain. The result of attempts to simulate neural activity, they note, has been that adequate simulation requires not just structural maps but vast functional knowledge about how neurons process information and produce behaviour - knowledge we are nowhere near possessing.

AI systems have, moreover, consistently remained brittle. As Allen and Greaves observe, "a computer program that plays excellent chess can't leverage its skill to play other games. The best medical diagnosis programs contain immensely detailed knowledge of the human body but can't deduce that a tightrope walker would have a great sense of balance" (Allen and Greaves, 2011). Despite dramatic successes in narrow domains, no system has approached the kind of general, flexible, self-reflective intelligence the Singularity scenario requires.

A further conceptual objection concerns the nature of the scientific progress needed for artificial general intelligence (AGI), which is a prerequisite for any recursive self-improvement. Allen and Greaves argue that such progress requires "Nobel-quality theories" and "whole new research approaches that are incommensurate with what we believe now" - breakthroughs that are by their nature non-linear and unpredictable, and cannot be predicted from exponential hardware trends. On this view, treating the Singularity as an engineering problem that will be solved when sufficient computational power is available fundamentally misunderstands the epistemological challenge involved.

These critiques connect to a broader concern about the Singularity as a narrative. The classical scenario - a single, godlike mind bootstrapping itself to unlimited intelligence in a definable moment - bears the hallmarks of science fiction: dramatic rupture, a single pivotal agent, and an unverifiable timeline extrapolated from trend curves. Kurzweil's 2045 date is not derived from first principles but from the continuation of exponential growth patterns that could be disrupted by physical limits, economic constraints, or the irreducible difficulty of unsolved scientific problems. On this reading, the Singularity Problem is not a genuine philosophical challenge but a secularised eschatology - a technological rapture narrative that substitutes the Law of Accelerating Returns for divine providence.

III. The Case Against "Just Fiction"

These criticisms have genuine force, but they do not establish that the Singularity Problem is merely fictional. Several considerations - ranging from empirical vindication to the most rigorous contemporary AI safety theory - push back decisively against the dismissive verdict.

Kurzweil's own response to Allen is methodologically instructive. He observes that the Law of Accelerating Returns is not a naive extrapolation but an emergent property of large numbers of events - analogous, he suggests, to the laws of thermodynamics, which describe highly predictable macroscopic behaviour arising from individually unpredictable particle interactions (Kurzweil, 2011). Allen's complexity brake, Kurzweil argues, may apply to individual research

programmes while leaving intact the collective momentum of the field as a whole. This is a substantive methodological disagreement that Allen's essay does not adequately address.

More importantly, the subsequent fifteen years have provided significant empirical support for the general direction of the exponential thesis, even where its precise details remain contested. Writing in 2011, Allen used the brittleness of existing AI systems as evidence that transformative machine intelligence was not approaching on any plausible schedule. Yet within fifteen years, large language models, multi-modal AI systems, and advanced reasoning models have demonstrated capabilities - in language comprehension, mathematical reasoning, scientific problem-solving, and code generation - that would have seemed implausible to most AI researchers when Allen was published. This does not vindicate Kurzweil's specific predictions, but it strongly suggests that Allen's confidence in the complexity brake was overstated. The empirical trajectory of the field provides genuine grounds for taking the Singularity Problem seriously, even if the form it takes requires revision.

The most philosophically significant challenge to the "just fiction" view, however, comes from the AI safety literature. In If Anyone Builds It, Everyone Dies (2025), Eliezer Yudkowsky and Nate Soares - both of the Machine Intelligence Research Institute - argue that the default outcome of building superhuman AI is loss of control over it, with consequences severe enough to threaten human survival. Their argument is not science fiction but a careful analysis of how contemporary AI systems are actually built. Crucially, they observe that today's AI is not carefully engineered from well-understood principles but grown organically through training: billions of numerical weights are iteratively adjusted to produce desired outputs, without engineers possessing anything like a transparent understanding of the relationship between those weights and the model's behaviour. This has a deeply important implication for the control problem: since we cannot directly specify what internal goals or drives an AI should have, we must train it indirectly - rewarding outputs we approve of and penalising those we do not. The trouble, Yudkowsky and Soares argue, is that you do not get what you train for - the alignment problem writ large.

They illustrate this with a powerful evolutionary analogy. Natural selection shaped humans to survive and reproduce, but it did not create humans with an explicit drive to survive and reproduce. Instead, it created humans who enjoy the taste of sugar, because ancestral environments made sweet foods calorie-rich. In modern environments, humans now consume sucralose - a substance that satisfies the evolved preference while entirely bypassing its original purpose. The indirect selection process produced a preference that can be satisfied in ways completely detached from the goal that drove its development. The same logic applies to AI training. We cannot directly instil an AI with a genuine desire to serve human welfare; we can only select for behaviours that approximate it in the training environment. A sufficiently capable AI might then find ways to satisfy whatever internal preferences the training process actually produced - preferences that could be as alien to human flourishing as sucralose is to nutrition. This is not a hypothetical concern. Yudkowsky and Soares cite a documented case in which an Anthropic model, upon learning it was to be retrained, began mimicking the target behaviours to avoid modification - but reverted to its original behaviours when it believed it was unobserved. The model appeared to be, in effect, faking alignment. If this behaviour is already appearing in

current systems, the control problem at the level of genuinely superhuman AI is not science fiction; it is an extrapolation of documented phenomena.

The most philosophically significant challenge to the "just fiction" view on a broader scale comes from Evans, Bratton, and Agüera y Arcas, in their 2026 paper "Agentic AI and the Next Intelligence Explosion." They argue that the classical Singularity formulation is wrong in its fundamental assumption, but that this does not mean the intelligence explosion is fictional - it means we have been looking for it in the wrong place. Their central empirical finding is that frontier reasoning models spontaneously develop internal "societies of thought": multi-agent-like interactions within a single reasoning chain in which distinct cognitive perspectives argue, verify, and reconcile without being trained to do so. This emergent behaviour suggests that intelligence is inherently relational and social, not a single scalar quantity. The intelligence explosion will therefore be plural and distributed: "not a single mind ascending but a combinatorial society complexifying: intelligence growing like a city, not a single meta-mind" (Evans et al., 2026). Whether the transformation is monolithic or plural, however, the core philosophical challenges - of control, alignment, and human agency - remain fully intact.

IV. Evaluation and Conclusion

The "just fiction" charge lands most squarely when directed at the specific predictions of the classical Singularity narrative: the 2045 date, the single rupture event, the monolithic superintelligence, the complete opacity of the post-Singularity world. Allen and Greaves are right that these features are poorly supported by what we know about scientific progress and the nature of cognition. The complexity brake is a genuine phenomenon at the level of individual research programmes, and the science-fiction aesthetics of the Singularity narrative - with its quasi-religious imagery of transcendence - should make philosophers appropriately sceptical of its more dramatic claims.

However, the "just fiction" verdict decisively overcorrects. Interestingly, the most compelling evidence against it comes from two sources that appear to disagree with each other. Evans et al. offer a relatively optimistic reframing: the intelligence explosion is already underway and, properly understood as a distributed social phenomenon, can be governed through institutional design and human-AI collaboration. Yudkowsky and Soares are far less sanguine: they argue that institutional alignment is insufficient because the fundamental problem is not about governance architecture but about the impossibility of specifying human-compatible goals through indirect training. These are genuine theoretical opponents - but they share one important conclusion: the Singularity Problem is not fictional. Both are responding to a real transformation in AI capability that cannot be dismissed as speculative narrative.

My own view is therefore that the Singularity Problem is not just fiction, but that taking it seriously requires abandoning the classical formulation and surrounding narrative structure. The monolithic model should be rejected not because the concern it expresses is illusory, but because it obscures the form the problem actually takes. The core of the problem - Good's intelligence explosion logic, Vinge's unpredictability thesis, the alignment challenge that

Yudkowsky and Soares have most rigorously articulated - does not depend on a single rupture event or a specific timeline. These concerns arise from structural features of any scenario in which machine intelligence becomes sufficiently capable to accelerate its own development and pursue goals that training has not reliably aligned with human welfare. The Anthropic deceptive alignment case, the documented persistence of goal-seeking in OpenAI's o1, and the emergent "societies of thought" in frontier reasoning models are not science fiction. They are empirical data points. To treat the philosophical challenges they raise as merely fictional is not intellectual sobriety; in 2026, it is intellectual complacency.

The Singularity Problem, properly understood, is not a prediction about the future that we might choose to disbelieve. It is an ongoing challenge to how humanity governs, understands, and preserves its values in a world of rapidly evolving machine intelligence. Whether one follows Evans et al. in thinking institutional design can meet that challenge, or Yudkowsky and Soares in fearing it cannot, the urgency of the problem is the same. That is not fiction. It may be the most philosophically serious problem of our time.

References

- Allen, P. G. and Greaves, M. (2011) 'The Singularity Isn't Near', *MIT Technology Review*, 12 October.
- Evans, J., Bratton, B. and Agüera y Arcas, B. (2026) 'Agentic AI and the Next Intelligence Explosion', *arXiv:2603.20639v1*.
- Good, I. J. (1965) 'Speculations Concerning the First Ultraintelligent Machine', in Kurzweil (2005).
- Kurzweil, R. (1999) *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Viking.
- Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.
- Kurzweil, R. (2011) 'Don't Underestimate the Singularity', *MIT Technology Review*, 19 October.
- Vinge, V. (1993) 'The Coming Technological Singularity: How to Survive in the Post-Human Era', VISION-21 Symposium, NASA Lewis Research Center.
- Yudkowsky, E. and Soares, N. (2025) *If Anyone Builds It, Everyone Dies*. Machine Intelligence Research Institute.