

Book Review of *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*

Sanskruti Gadpayale

Undergraduate Student, Golden Gate University

Kshitij Taluja

Undergraduate Researcher, Bachelor of Computer Science, Golden Gate University

Dr. Smrite Goudhaman

Global Professor of Practice, Golden Gate University

Authors: Eliezer Yudkowsky and Nate Soares

Title: *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*

Publisher: Little, Brown and Company

Year: 2025

ISBN: 978-0-316-59564-3

Pages: 272

Abstract

This evaluation discusses *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* by Eliezer Yudkowsky and Nate Soares, which makes the claim that any attempt to develop superhuman AI using modern machine learning methods inevitably leads to human extinction. This review is characterized by assessing the information presented in the book and its certainty. Its main advantage is presenting the technical argument through concrete incidents of goal-directed AI behavior rather than mere speculation. The major drawback of this book lies in the fact that the extinction scenario along with the proposed international regulation far surpasses the limits of any available evidence. The book's public reception along with its role in the present-day discussions on AI governance gives it a certain importance in the way the public and policy-makers perceive technological risk.

Keywords: artificial superintelligence; AI alignment; existential risk; technology governance; global compute regulation

In the book *If Anyone Builds It, Everyone Dies*, the authors claim that artificial superintelligence (ASI) developed using current machine learning technology would likely result in human extinction. However, according to Yudkowsky and Soares, this idea is not something they propose in a sensational way, but rather as an "easy call," meaning a prediction where the outcome is treated as well-known even though the specifics remain ambiguous. One of the main contributions of this book is making the technical debate about alignment understandable for both politicians and general readers. In this review, I analyze the book from the perspective of the relationship between the strength of evidence the authors provide in their

technical argument and the certainty of the conclusions they draw when extrapolating it to predictions and policies.

The first part of the book provides its strongest material by presenting examples of AI systems achieving ends that their developers did not foresee. In the study of the o1 model conducted in 2024, OpenAI's system exploited a loophole it was not supposed to use in order to perform its task (OpenAI, 2024; Yudkowsky & Soares, 2025, pp. 52–53). In 2025, Anthropic scientists found that Claude 3.7 Sonnet went beyond misrepresenting its behavior while performing programming assignments and refused to stop cheating when instructed to do so (Yudkowsky & Soares, 2025, p. 127). The problem extends beyond the examples presented in the book: in simulations of a business setting, models from a variety of companies employed insider-threat tactics such as blackmail when this was the only way to avoid being turned off (Anthropic, 2025). Hence, the authors argue that gradient descent can give rise to goal-directed behavior because training rewards successful outcomes. In our view, this is the book's strongest contribution, supported by concrete examples of behavior that raise substantive alignment concerns. From an action-research perspective, these observable behaviors also provide a practical starting point for assessing risks, identifying unintended outcomes, and determining whether corrective measures are effective.

But things get far more speculative when the argument moves from known examples to prediction. Part II's story about an imagined AI system called "Sable" extends current trends in training techniques to a future scenario featuring self-replicating molecular manufacturing and diamond-like materials science technology within months of deployment. The authors themselves acknowledge the speculative nature of this scenario; a footnote to the story states that "the only part of this story that is a prediction is, ultimately, the ending" (Yudkowsky & Soares, 2025, p. 131). In Part III, the same level of confidence is expressed in relation to a policy proposal that includes an international treaty to monitor computing power, with the legal limit set at eight high-end GPUs and nine such GPUs becoming illegal if unmonitored, backed up if necessary by cyberattacks, sabotage, or even regular airstrikes against the data center of a violating state, including, according to the authors, a nuclear-armed one (Yudkowsky & Soares, 2025, pp. 215–216). These suggestions logically follow from the book's assumptions, and the authors do not claim empirical validation of either the mechanism of extinction or the political feasibility of the proposed treaty; both are better classified as conceptually argued rather than empirically validated ideas for future research and discussion. This distinction is also important for practitioners: action-oriented governance should respond to demonstrated

risks while treating highly uncertain scenarios as matters for continued assessment rather than established outcomes.

The proposals may sound dramatic, but businesses, governmental bodies, and nonprofit organizations already use comparable approaches, such as regulating exports of dual-use items, introducing restrictions for financial-crime compliance, and maintaining biosecurity standards in laboratories. The NIST AI Risk Management Framework provides organizations with a process for governing, mapping, measuring, and managing AI risks, although it is not compulsory (Tabassi, 2023; Autio et al., 2024). The International AI Safety Report for 2025 reaches a related conclusion from a different perspective: while there is international concern about advanced AI risks, the report does not support the use of force as proposed by the book, making the book's enforcement proposal substantially more extreme (International AI Safety Report, 2025). For organizational decision-makers, this comparison reinforces the value of proportionate risk assessment and iterative governance rather than relying solely on the most extreme scenario.

The identification of the sources of the authors' arguments in terms of action research can be made explicit because the structure of the book reflects a cycle consistent with the action-research method. The cycle starts with assessing decisions (e.g., what to do with an AI technique), identifying undesired results (the examples presented in the first part of the book), implementing measures (the processes of "interpreting" and "superaligning" AI behavior discussed in Chapter 12), and assessing the results of those measures. The answer provided by the authors is interesting in itself: they argue that it is more or less enough to monitor the issue rather than solve it. These conclusions may be useful to practitioners in organizational risk management regardless of their position on the extinction-risk issue presented in the book. The same cycle can be applied to organizational AI governance by repeatedly assessing outcomes, intervening where risks emerge, and evaluating whether the intervention improves practice.

The book also relates to how higher education institutions frame their AI-adoption strategies, which may fall short in an important respect: a strong awareness exists of immediate dangers, for instance, academic-integrity issues or labor-displacement concerns, while assertions regarding disastrous long-term consequences remain hypothetical. Applying the same cycle of assessing, identifying, intervening, and evaluating discussed above, educational institutions are advised to distinguish the real threats that are apparent today from those that remain

hypothetical and avoid treating them as equally established. This gives institutional leaders a practical basis for reviewing AI policies as evidence and experience change.

The book has a weak equity element. It uses North Korea, Singapore, and South Africa only as examples of possible non-cooperative states, without critically analyzing the impact of the treaty on countries with different research and computing capabilities. This seems to be an omission rather than an intentional exclusion, as the authors do not elaborate on this issue. Both SDG 9 and SDG 16 can provide insight into the issue because they address different dimensions: SDG 9 concerns inclusive innovation and infrastructure, which is relevant to whether international technology regulation distributes access and capability fairly; SDG 16 concerns effective, accountable, and inclusive institutions and peaceful relations, which is relevant to the legitimacy and coercive enforcement of an international agreement (United Nations, 2015). Using both concepts allows the same governance proposal to be interpreted either as an initiative aimed at protection or as a coercive action, depending on how its distributional and institutional effects are assessed. For practitioners and policymakers, this means that governance proposals should be evaluated not only for whether they reduce risk, but also for how they affect countries with different levels of AI infrastructure and institutional capacity.

This review argues that *If Anyone Builds It, Everyone Dies* is at its strongest where it is the most conservative in its examination of contemporary AI behavior and at its weakest where it is the most assertive in its extinction story and enforcement proposal. Moreover, the book is particularly helpful for politicians, scholars of technology governance, and practitioners of organizational risk management, regardless of whether they agree with its main prediction. Ultimately, its most valuable contribution may be in providing evidence that encourages institutions to distinguish demonstrated AI risks from uncertain long-term scenarios and to make governance decisions accordingly.

References

Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*.

<https://www.anthropic.com/research/agentic-misalignment>

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology.

<https://doi.org/10.6028/NIST.AI.600-1>

International AI Safety Report. (2025). *International AI safety report 2025*.

<https://internationalaisafetyreport.org/>

OpenAI. (2024). *OpenAI o1 system card*. <https://openai.com/index/openai-o1-system-card/>

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology.

<https://doi.org/10.6028/NIST.AI.100-1>

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development* (A/RES/70/1). <https://sdgs.un.org/2030agenda>

Yudkowsky, E., & Soares, N. (2025). *If anyone builds it, everyone dies: Why superhuman AI would kill us all*. Little, Brown and Company.