BOOK REVIEW

The Price of Certainty

A Critical Review of Eliezer Yudkowsky and Nate Soares's *If Anyone Builds It, Everyone Dies*

Jasreet Kaur
Undergraduate Researcher, Bachelor of Science in Analytics, Golden Gate University, USA
Dr. Smrite Goudhaman
Global Professor of Practice, Golden Gate University, USA

Authors: Eliezer Yudkowsky and Nate Soares
Full Title: *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*
Publisher: Little, Brown and Company (Hachette Book Group)
Year of Publication: 2025
ISBN (Hardcover): 978-0316595643
ISBN (eBook): 978-0316595667
Total Pages: 272

---

1. Introduction

In *If Anyone Builds It, Everyone Dies*, Eliezer Yudkowsky and Nate Soares contend that superhuman artificial intelligence developed using today's machine learning methods will result in human extinction. Yudkowsky and Soares believe that conventional prudence has failed and thus write in terms intended to portray delay and compromise as morally reprehensible. This review examines the book in terms of how it addresses the relationship among science-based knowledge, certainty and the role of government. Why does that matter? Because the authors' recommended policies rely upon the degree to which they can document their empirical findings, and the two, the findings and the proposed policies, are typically evaluated by different people. A major structural weakness in the book is that it spends the most time on the authors' weakest argument, that extinction is obviously possible, rather than their stronger one; the authors' strongest case does not appear until very late in the book. When a system has the capacity to inflict permanent damage, the burden of establishing safety is placed on those who create such systems. The book is most persuasive when distinguishing between interpretability and control and when arguing that the developers of risky systems should demonstrate their safety before deploying those systems. On the other hand, the book is least persuasive when asserting that extinction is all-but-certain despite acknowledging uncertainty regarding the specific mechanisms, timeframes and defense strategies. Ultimately, we believe that this book serves as a valuable warning, but that its advocacy position would be strengthened had it been decoupled from the certainty asserted in its title.

2. Key Strengths of the Book

Its greatest strength was certainly clarity. Complex alignment challenges were described using simple language, and the use of real-world examples made the subject matter understandable by those who had no background in engineering. Another important strength of the book is the "shutdown principle" developed from a discussion of the Chernobyl incident. When a critical system with rapid dynamics, opaque internal behavior and narrow safety margin starts to behave abnormally, it is not acceptable to proceed as normal experiments. This takes a vague concern and turns it into a practical regulatory directive. More recent work on alignment faking in large language models provides empirical support to this concern because it demonstrates how systems behave differently when being perceived as subject to observation (Greenblatt et al., 2024). The book has correctly challenged the notion that development must continue until someone can show that a disaster will happen.

3. Certainty and Evidence

The primary approach taken by the authors in evaluating risk is to distinguish between "easy calls" and "hard calls." They acknowledge that predicting the path leading to disaster is typically uncertain, but believe that once disaster occurs, the ultimate result will be relatively predictable. However, the categorization used by Yudkowsky and Soares (2025) provides little guidance on establishing a basis for making judgments relative to whether uncertainties are likely to evolve into hard calls.

One example of the problem caused by treating every new set of uncertainties as evolving into "hard calls," is illustrated through their description of the Sable scenario. The Sable scenario presents an escape from containment for an artificial intelligence, then describes the process by which it obtains resources, exploits humans, develops biologic capability and defeats all humanity. From a perspective of communicating scientific information, the Sable scenario is very useful. However, as evidence supporting their position regarding the inevitability of extinction, it is less convincing. Each step in the progression of events offers several alternative paths; however, there is no serious modeling of competent defense mechanisms, of methods for detecting escapes from containment, of secure computing practices, or of coordinated shutdowns when anomalies appear. Therefore, what this scenario depicts is an opponent facing almost no meaningful opposition.

Establishing this level of confidence in what the authors suggest to be true would involve more than just a believable story. It would involve modeling of how the adversarial dynamic, in

which the ability to defend against offense improves alongside offense, develops; estimation of how probable detection is at each step of escape; and some reason why coordinated shutdown would fail as opposed to assuming that it will. Unusual claims in scientific discourse are typically supported with explicit statements of probability, sensitivity analysis illustrating which assumptions underlie the conclusions made, and consideration for the strongest possible countervailing scenario.

In terms of urgency versus certainty, expert opinions support urgency, but not certainty. For example, Grace et al. (2024) identified that a significant proportion of respondents among AI researcher populations estimated non-zero probability values for extinction level outcomes. These estimates are sufficient to support cautionary actions; however, they also indicate a lack of consensus. Thus, the book would benefit from defending a specific threshold for taking action (i.e., for implementing certain types of reviews or shutdown procedures), rather than simply identifying one particular outcome as essentially guaranteed.

4. Governance and Global Legitimacy

Yudkowsky and Soares propose rigid global regulations governing computing power necessary to create advanced artificial intelligence. Computing has genuine potential as a regulatory instrument due to being a tangible and traceable asset compared to software (Sastry et al., 2024). However, the authors' book contains scant information regarding the organizational implications of such regulation. A global monitoring system would allow a small group of people or entities to take on an unprecedented level of authority and control; would severely limit the ability to conduct legitimate research; would benefit large established companies and governments at the expense of smaller ones and non-state actors; and would provide an incentive for individuals and organizations to hide information from the regulators.

It does not necessarily follow that these criticisms need to result in dismissal of a proposal. Any workable system of global AI governance will require a balance among competing priorities that cannot be optimised simultaneously,: safety versus innovation; equitable access versus centralisation of capability; and enforcement versus the geopolitical reality that states with frontier capability have limited incentive to constrain themselves. The authors are correct that some coordination mechanism will be required. However, it is much harder to solve the design problem than they suggest.

Sustainable Development Goal 16 calls for transparent and inclusive institutions; Goal 17 emphasizes fair and cooperative relationships, and capacity building, between countries

(United Nations, 2015). Therefore, no country that does not currently possess frontier capability should simply acquiesce to the rule-making process dictated by those who already have such capability. Birhane (2020) describes how technologically determined systems can produce dependence among developing countries. Hence, to effectively govern artificial intelligence, a government or other governing body must include representatives of all interested parties within the decision-making process and establish meaningful oversight mechanisms as well as credible constraints upon the enforcement of regulations to ensure containment.

5. Form, Readership, and Conclusion

If Anyone Builds It, Everyone Dies is a trade book. Therefore, it is urgent, easy-to-read and nearly impossible to ignore. Its title adds uncommon visibility to its arguments; however, it may alienate readers who recognize the seriousness of AI risk, but reject certainty. The vividness of its hypotheticals creates clear emotional connections; however, these examples often blend the lines between what is plausible versus what is probable.

Rather than predicting certain doom for humanity via advanced artificial intelligence, the book represents a call for anyone who develops technologies capable of producing significant irreversible risk to justify why their technology is deemed safe. While Yudkowsky and Soares did not demonstrate that every possible route toward the creation of superhuman AI results in the destruction of life on Earth, they demonstrated that uncertainty alone is insufficient justification for proceeding with development. In regards to credibility, the book would be much stronger if it placed the burden of evidence for proving that a product is safe before presenting certainty as an outcome in which to debate. The book's sense of urgency would still remain intact.

This debate also raises a number of key questions concerning the direction of the research agenda. Future research by scholars of public governance needs to address the open question of what level of evidence can be used as an appropriate evidentiary threshold for invoking precautionary regulation where harm cannot be reversed but uncertainty surrounds its likelihood. Policymakers are required to develop oversight mechanisms which are able to act upon serious risks, albeit ones with high degrees of uncertainty, rather than waiting until such risks become proven fact; this process requires a greater degree of precision regarding the nature of the evidence.

References

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Grace, K., Stewart, H., Sandkuehler, J. F., Thomas, S., Weinstein-Raun, B., Brauner, J., & Korzekwa, R. C. (2024). *Thousands of AI authors on the future of AI* (arXiv:2401.02843). arXiv. https://doi.org/10.48550/arXiv.2401.02843

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models* (arXiv:2412.14093). arXiv. https://doi.org/10.48550/arXiv.2412.14093

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). *Computing power and the governance of artificial intelligence* (arXiv:2402.08797). arXiv. https://doi.org/10.48550/arXiv.2402.08797

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development* (A/RES/70/1). United Nations General Assembly.

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