Philosophy of AI
Vol. 1, 2025, 93-96
10.18716/ojs/phai/2025.2854

ESTIMATING THE PROBABILITY OF AI EXISTENTIAL CATASTROPHE: CONVERGING ON THE ANSWER FROM OPPOSITE ENDS

Leonard Dung[1]

[1] Ruhr Universität Bochum, DE

In their wonderful paper, Cappelen et al. (2025) propose a novel "survival story" approach for estimating the probability of an existential catastrophe caused by AI (short: "AI doom"). With this, they turn the traditional methodology for estimating the probability of AI doom on its head. The traditional approach asks what one needs to assume to conclude that AI will cause an existential catastrophe and then assesses the probability of these assumptions (Carlsmith, 2022; Dung, 2024). By contrast, the survival story method asks what one needs to assume to reach the result that AI will *not* cause an existential catastrophe, i.e., that humanity will "survive".

As the authors indicate, this reframing highlights that the burden of proof does not lie on one side of the debate specifically: While providing compelling arguments for a high probability of AI doom is hard, arguments for AI survival also face difficult challenges. No party can rest content and assume that their view is the default which is assumed true until proven otherwise. Instead, both parties should only be confident once they have provided strong arguments.

In this commentary, I will focus on how the survival story approach advances probabilistic estimation of AI doom. While I highly commend Cappelen et al.'s methodology, I make two points: First, in their paper, they – to some extent – neglect "multipolar" survival stories in which there are many different superhumanly intelligent AI systems. Second, this is an instance of the general issue that their methodology threatens to overestimate the probability of AI doom, by overlooking important survival stories. The survival story methodology can be seen as providing an upper bound for the probability of AI doom. Since the traditional methodology provides a lower bound for the probability of AI doom, both methodologies should be combined.

A survival story is a scenario in which AI doom does not occur. Cappelen et al. propose four kinds of survival stories: 1. Technical Plateau. 2. Cultural Plateau. 3. Alignment. 4. Oversight. These are taken to be exhaustive: If there is no AI doom, one of these survival stories must obtain. Equipped with this assumption, Cappelen et al. show that one can compute the probability of AI doom. In the first step, one needs to estimate four probabilities: the probability that survival story A (say, Technical Plateau) fails (i.e. does not occur), that B (say, Cultural Plateau) fails conditional on A failing, that C fails conditional on A and B failing, and that D fails conditional on A, B, and C failing. In the second step, one multiplies these probabilities. The product is the probability that all survival stories fail jointly. By assumption, this means that AI doom occurs.

I will be concerned with the survival stories approach generally, not just the specific stories Cappelen et al. discuss. I think that the distinctive weakness of this approach is that the survival stories which are considered may not be an exhaustive partition of the space of possible futures without AI doom. If so, then the approach may *overestimate* the probability of AI existential risk.

For example, multipolar scenarios are scenarios in which many AI systems with superhuman capacities come to exist. Some of these scenarios do not fit well within the authors'

taxonomy. One may imagine that these AI systems gradually become more and more powerful, without bound, so that Plateau stories seem to not apply. At the same time, many of these systems may be misaligned with human goals, so that Alignment does not apply. Cappelen et al. view multipolar scenarios like these as instances of Oversight. However, this is (at least) a fairly atypical case of oversight which is defined in the paper as: "**Oversight**: Extremely powerful AI systems do not destroy humanity, because we can reliably detect and disable systems that have the goal of doing so."

One may imagine that some of these superhuman misaligned systems are removed from control by humans and other AI systems completely, so that oversight in any normal sense of the term does not occur. The only reason that AI doom does not occur is that many similarly powerful AI systems – we can imagine some or all of them to be unaligned with human goals – exist. If one would try to take control, others would potentially stop them.

Moreover, the authors' arguments seem to not directly apply to such a survival story. For if there are sufficiently many superhuman systems and they do not coordinate much, it may be that no single system is able to bring about doom. So, in the ideal case, we may be as safe as in a Plateau story. This may be similar to why no human has caused an existential catastrophe yet: every single human lacks the capability, and it is hard for large groups of humans which want to permanently suppress or kill all other humans to coordinate.

A natural reply is that – whether multipolar scenarios can be counted as cases of oversight or not – it is easy to formulate a set of survival stories so that it is logically guaranteed that they exhaust the space of scenarios in which AI doom does not happen. In the trivial case, one may formulate a number of stories N and then add story N+1 which is defined to be the disjunction of all scenarios in which no story N obtains but AI doom does not happen. It is also possible to extend the categories of stories Cappelen et al. discuss widely enough that it is logically guaranteed that any scenario without AI doom belongs into one of them.

However, crucial for assessments of AI existential risk is not whether a scenario is *logically contained* within the kinds of survival stories considered. Instead, it matters whether the scenario has been explicitly considered so that it influences the estimate of the probability that the survival story obtains. If certain kinds of multipolar scenarios have not been considered when estimating the conditional probability of Oversight, then the estimate will be an over-estimate, even if Oversight logically includes these scenarios. Relatedly, one cannot estimate the probability of a disjunction of survival stories if one has no concrete idea which scenarios might be contained in the disjunction (because the scenarios are defined to be distinct from all of the ones that one has explicitly considered). So, the vulnerability of the survival story approach is that it may produce over-estimates of AI existential risk, since relevant survival stories may be missed, either by being a logically distinct alternative to the stories which have been factored into the probabilistic estimate or by not having been explicitly considered, even though they might have significant probability. Adding a disjunctive survival story which, by definition, refers to all stories that have been missed is not a solution.

The solution appears once one realizes that the survival story approach is the opposite of the traditional approach. On the traditional approach, one formulates arguments which contain the conclusion that AI doom will occur. There are different possible arguments for such a conclusion, for instance based on AI misalignment (e.g. Bostrom, 2014; Cotra, 2022; Dung, 2023) or based on intentional misuse of AI through humans (Friederich, 2023; Yum, 2024). So, any particular argument is only a "lower bound" on the probability of AI doom. One reaction is to make these arguments more general, so that they encompass as many concerns relevant to the probability of AI doom as possible and get closer to being point estimates, not just lower bound estimates, of such risks. However, this would make such arguments less concrete and more intricate which plausibly has the consequence that the relevant probabilities are harder to estimate, if it is easier to assess more concrete and specific scenarios than general principles and premises. Embracing multiplicity, one may instead regard different arguments for AI doom as different possible "AI destruction stories". Then,

one could compute the probability of AI doom by summing the probabilities of each individual destruction story (after conditionalizing them analogously how Cappelen et al. do it with survival stories).

No matter which option one chooses: There remains the risk that one misses important destruction stories. This would lead to an *underestimate* of the probability of AI doom. Thus, the survival story and the traditional destruction story approach have opposing weaknesses: The former encourages overestimates of the probability of AI doom, while the latter encourages underestimates. For this reason, the former approach supplies upper-bound estimates of AI doom, while the latter provides lower-bound estimates. This suggests that one should combine both approaches. This way, one can justify judgements about the *probability range* within which AI existential risk is located, without any blindspots, even if important survival or destruction stories are missed.

Moreover, potential occurrences of conflicts between estimates based on both methods – e.g. the lower bound estimate of AI doom being higher than the upper-bound estimate – can be used as an independent test for the correctness of these estimates. Such conflicts would show that (at least) one of the estimates is flawed.

Since the soundness of this approach does not depend on exhaustively partitioning the space of possible outcomes of the development of AI, it may also encourage the exploration of a much larger number of much more concrete scenarios. By considering and examining an increasing number of scenarios, we can make our assessments of the probability of AI existential catastrophe increasingly precise.

References

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press.

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025). Ai Survival Stories: A Taxonomic Analysis of Ai Existential Risk. *Philosophy of Ai*.

Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* (arXiv:2206.13353). arXiv. https://doi.org/10.48550/arXiv.2206.13353

Cotra, A. (2022). *Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover*. Lesswrong. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 138. https://doi.org/10.1007/s11229-023-04367-0

Dung, L. (2024). The argument for near-term human disempowerment through AI. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01930-2

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00268-7

Yum, I. (2024). Language Agents and Malevolent Design. *Philosophy & Technology*, *37*(3), 104. https://doi.org/10.1007/s13347-024-00794-0