Philosophy of AI
Vol. 1, 2025, 96-99
10.18716/ojs/phai/2025.11986
©2025 by the author(s)

AI SURVIVAL STORIES, TYPES OF RISK, AND THE PRECAUTIONARY PRINCIPLE

Kate Vredenburgh[1*]

[1] London School of Economics

Cappelen, Goldstein, and Hawthorne's article offers a refreshing new perspective on questions of AI safety. Debates over AI and existential risk standardly start from a background assumption of humanity's continued survival, and then reason about the probability of catastrophic outcomes. This paper, by contrast, flips that mode of reasoning on its head: a decision-maker should start from the assumption that powerful AI systems will destroy humanity, and then reason about the different outcomes in which humanity is saved from such an existential threat. Doing so enables us to better partition the space of possible events in which saving occurs, and so to come to more justified probabilities of humanity being saved. The article also illustrates the value of this approach by defending particular claims about promising paths to avert existential risk. To do so, it puts forth a model that different parties to the debate can use to calculate the probability that humanity iwill bedestroyed. The authors also defend a number of claims about the most likely mechanisms for realizing each of the four survival stories.

There's a lot to agree with in the authors' contribution to the debate over existential risk. However, I will focus on two ways in which I take the model to be incomplete, and propose two significant amendments. They are: an expanded account of existential risk; and, revised model of the causal pathways by which each of the survival stories - technical, cultural, alignment, and perfect safety - might be realized. I'll then close by proposing an additional reason in support of the authors' claim that accidents are a more promising pathway to the cultural plateau than sophisticated reflection, and reflect on the likelihood of such an accident that is both morally permissible and causally effective.

There is often some fuzziness in what people mean when they claim that AI is an existential threat to humanity. Here, the authors are admirably clear: they cash out what it is to be an existential threat as a threat that poses a risk of the extinction or partial extinction of humanity, or as seriously threatening the autonomy of living people. Notably absent from this characterization, however, are reductions in wellbeing. A world that is populated by superintelligent or supernumerous AI agents could be one in which the human population is composed of people who have lives that are so low in average individual wellbeing that this state of affairs is on a par with human extinction or serious and widespread autonomy loss. Cappelen, Goldstein, and Hawthorne take it that there is at least a good chance that AI will compete with humans for resources in order to, say, accomplish its goals (see also Bales, D'Alessandro, and Kirk-Giannini 2024). Such competition could lead to a resource allocation between AI and human populations, such that each human life has, on average, an unacceptably low level of wellbeing. Such a scenario deserves to be classified as an existential threat, alongside widespread loss of human autonomy.

On its own, this omission is not particularly significant, and can easily be incorporated into the model. However, the omission has, I hazard, deeper roots in how many participants in debates around AI safety conceive of the causal mechanisms by which an existential threat is activated. And, expanding the space of possible mechanisms will require complicating the model defended by Cappelen, Goldstein, and Hawthoren.

To make this point, I will draw heavily on a recent article by Kasirzadeh (2025), in

which she distinguishes between decisive and accumulative AI risks. Decisive risks are probabilities of an "abrupt large-scale events that lead to humanity's extinction or cause an unrecoverable decline in its potential" (Kasirzadeh 2025: 5). Decisive risks are caused by agents, or a group of agents intentionally acting together, who have centralized control over mechanisms by which they are able to cause humanity's (near) extinction or loss of autonomy. The threat of nuclear war is such a decisive risk: there are some agents who have centralized control over nuclear weapons, and they could intentionally trigger a series of events that would lead to human extinction.

Furthermore, the assumption of decisive risks also influences the model that the authors develop to calculate the probability of human extinction. In particular, it casts doubt on the accuracy of the model's claim that reaching one of the four survival stories - technical, cultural, alignment, and perfect safety - will prevent humanity from facing a further existential risk. According to the model, if there are technical limitations on the development of supercapable artificial agents (term taken from Bales, D'Alessandro, and Kirk-Giannini 2024), or on the supernumerosity of suitably intelligent agents, then humanity will not face cultural, alignment, or imperfect safety risks. However, if the risk from supercapable or supernumerous AI systems is *aggregative,* rather than decisive, then reaching a technical plateau does not rule out cultural risk. Aggregative risks are those that weaken different systems over time, making them vulnerable to a triggering event that initiates the (near) extinction, or significant autonomy or welfare loss, of humanity (Kasirzadeh 2025). Importantly, these systems are made vulnerable due to the pervasive use of AI across political, economic, military, and other domains, where these different systems become interconnected but remain fairly decentralized. In other words, in cases of aggregative risk, AI creates an existential threat because of the pervasive and interconnected deployment of AI in organizations and institutions that affect the autonomy, wellbeing, or existence of the human population, without suitable safeguards. Above, nuclear war was presented as an example of a decisive risk; here, we can instead consider financial crises as a paradigm aggregative risk. Financial crises are not created by a single, powerful bank able to manipulate the world economy. Instead, they come about in a highly interconnected banking ecosystem in response to a triggering event that exploits banks' interconnected vulnerabilities. And, importantly, they come about because there was a gradual loss of resilience within the financial system that was neither noticed nor corrected (Battiston et al. 2016)

Taking aggregative risk into account is important because it leads to different implications for how different survival stories may come about. As the authors note, one of the advantages of a focus on survival stories is that it shifts the burden of proof onto those who are confident of humanity's survival in the face of increasingly powerful AI. What are they assuming about how institutions and individual actors will cope with AI, alongside assumptions about AI agents' capabilities and goals? Goldstein, Cappelen, and Hawthorne's contribution helpfully pushes us to consider both facets of human survival: not only how will AI agents develop, but how will our institutions need to change in order to cope with them?

Considering aggregative risk pushes us to widen our view on the likely pathways to human survival, as reducing aggregative risk will require different mechanisms than reducing decisive risk (Kasirzadeh 2025). As the authors discuss, mustering public opinion around an AI ban may require a significant adverse event due to AI technologies, analogous to accidents in nuclear power stations. However, while an accident-triggered cultural plateau is compelling for decisive risk, it is less so for at least some types of accumulative risk. Some accumulative risks will have the characteristic that the interconnected system in which they arise are robust against external shocks until the system is weakened such that a tipping point is reached. Again, by example of the financial system: decades of deregulation led to banks reducing the amount of cash they held in reserve, but, until a certain tipping point of a suitably low level of cash reserves was reached, the banking system was still robust to external shocks. Before the tipping point, however, banks could borrow from other banks if they were over-stretched in their cash reserves; and so, failures would not trigger stricter regulation. Thus, because aggregative risk is often created by a gradual increase in the

fragility of an interconnected and complex system, monitoring and testing the resilience of the overall system are important ways to reduce aggregative risk. For example, in the wake of the 2007 financial crisis, stress testing has become an important means for central banks to understand the impact of sudden shocks or changes in important financial variables on a portfolio, and so to gauge vulnerability to such events based on banks' ability to meet capital and liquidity requirements (Foglia 2009). More work needs to be done here to develop such strategies for measuring the vulnerability of social systems to AI risk.

How likely, however, is humanity to reach the cultural plateau by instituting such a monitoring system and banning AI, or doing so in the wake of a serious accident? Here, I want to close by opening up the question of whether humanity is likely to reach a cultural plateau by morally permissible means, and whether rational reflection on risk is likely to lead us there. Here, I want to begin with the latter question, and offer a further argument in support of the author's favoring of an accident-driven cultural plateau, rather than sophisticated reflection, to avert decisive risk. It is often tempting to look to European regulation and its use of the *precautionary principle* in the hopes that sophisticated reflection on the part of politicians will lead to a ban on powerful AI systems (Steele 2006). The precautionary principle is often posed as an alternative to standard expected utility or cost-benefit analysis. Very roughly, the precautionary principle states that a decision-maker should not choose policies where there is a suitable likelihood of a very serious harm. It has some *prima facie* appeal in exactly these sorts of cases in which there is an activity that produces great benefits, but where there is a small chance of a very serious harm. And, it might seem to counsel banning the development of powerful AI technologies that pose an existential risk, as long as the probability of human extinction were suitably high.

However, it is not clear that the precautionary principle will require that decision-makers ban risky AI technologies, or that it will even permit them to do so. In order for the precautionary principle to be action-guiding, a decision-maker has to choose the scope of choices over which it applies (Thoma 2022). But, for many new technologies, the probability or the severity of human-level extinction will only be significant enough to trigger the precautionary principle if the decision-maker considers a set of choices (Thoma (2022) terms these *cumulative risk cases*). However, for any individual choice, the probability or severity will not be high enough to trigger precaution. And, as Thoma motivates, none of the individual actions are enough to push the case over the precautionary principle's threshold and trigger the principle: even if the agent can identify that they are close to the threshold, the threshold is likely vague, and one more act of developing or deploying AI does not seem to make a difference.

We can illustrate this point using the example of aggregative risk, which is more obviously a cumulative risk case. In many cases of aggregative risk, the severity of potential harm accumulates as more and more AI systems are embedded across different interconnected domains. But, the severity of harm is only significant when one considers all those systems together. However, cases of cumulative risk are likely to be common among the set of AI existential risks. For example, a case of decisive risk can also be a cumulative risk case: each individual choice to develop or deploy another slightly more advanced AI system adds to the total probability of human extinction, when one considers the set of all those choices. But, again, each choice individually adds very little risk, and does not put one over the threshold of the precautionary principle to trigger caution.

The problem is that the precautionary principle does not require a particular scope of activity to evaluate either the probability or the severity of harm. And, on some scopes of activity, caution will not be permissible, much less not required. If a decision-maker considers a single choice about AI technologies, the precautionary principle is likely to require developing and implementing the technology, since the risk of harm from that choice is very small, but the value created is significant. There is a push to develop advanced AI systems for economic gains, as the authors discuss, but also a push to develop AI to better deliver public services, for medical use cases, or to help manage scarce resources. But, if a decision-maker considers the set of all decisions to develop and implement advanced AI systems, the precautionary principle would likely recommend caution. And, because the dif-

ferent scopes of decisions do not agree on the required course of action, we cannot use some kind of robustness principle to decide what to do. The matter of scope leaves us in an uncomfortable place. Decision-makers will not just disagree about the probabilities of certain events, but also about the scope at which we ought to evaluate choices about AI. And so, choices that will keep humanity on one of the four plateaus will be determined by cultural or other factors that influence the time horizon that decision-makers choose. Given humanity's near term thinking in the face of the threat from climate change, there is good reason to be pessimistic that politicians will adopt a scope for the precautionary principle that leads to the banning of powerful AI systems.

We thus seem to be left with the "deploy and hope for accidents" strategy, which Cappelen, Goldstein, and Hawthorne take to be more promising than sophisticated reflection. A final worry, however, is that this strategy cannot be implemented in a reasonably effective and morally permissible way. One moral concern centers around whether the people who are being harmed by an accident caused by AI are being used as a mere means to benefit other current and future people. Say that a regulator is concerned about the existential threat posed by AI systems under development, but knows that no party will have the political clout to pass an AI ban. The regulator believes that an accident is the most likely means to rally political support for an AI ban. So, they do not interpret AI safety regulation as requiring significant safety measures. As they intend that the harm from an accident produces certain good effects for others, it seems as if those harmed are being used as a mere means. And so, according to certain background moral commitments, some causal pathways to an accident-triggered cultural plateau will not be morally permissible. By contrast, those harmed wouldn't be used as a mere means if the regulator had not intended for there to be an accident, but had instead intended to promote economic growth through loose safety regulation (at least, if one accepts the doctrine of double effect). However, in that case, the "deploy and hope for accidents" strategy does not look as promising, as it requires an accident severe enough to trigger a ban, but not so severe that it poses an existential threat. We are thus left in a bind: if the relevant decision-makers abide by those moral commitments, humanity's survival will be largely left to luck.

References

Bales, Adam, William D'Alessandro, and Cameron Domenico Kirk-Giannini. 2024. "Artificial Intelligence: Arguments for Catastrophic Risk." Philosophy Compass 19(2): e12964. doi:10.1111/phc3.12964.

Battiston, Stefano, J. Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G. Haldane, Hans Heesterbeek, Cars Hommes, et al. 2016. "Complexity Theory and Financial Regulation." Science 351(6275): 818–19. doi:10.1126/science.aad0299.

Foglia, Antonella. 2009. "Stress Testing Credit Risk: A Survey of Authorities' Approaches." International Journal of Central Banking (18).

Kasirzadeh, Atoosa. 2025. "Two Types of AI Existential Risk: Decisive and Accumulative." Philosophical Studies. doi:10.1007/s11098-025-02301-3.

Steele, K. 2006. "The Precautionary Principle: A New Approach to Public Decision-Making?" Law, Probability and Risk 5(1): 19–31. doi:10.1093/lpr/mgl010.

Thoma, Johanna. 2022. "Time for Caution." Philosophy & Public Affairs 50(1): 50–89. doi:10.1111/papa.12204.