Philosophy of AI
Vol. 1, 2025, 100-106
10.18716/ojs/phai/2025.11987

AI SURVIVAL STORIES - RESPONSES TO CRITICS

Herman Cappelen[1], *Simon Goldstein*[1*], *John Hawthorne*[2]

[1] The University of Hong Kong, HK

[2] University of Southern California, Irvine, US

We thank each of the critics for their thoughtful contributions to this volume. Below, we reply to each contribution in detail.

1. Rory Svarc: Defending Alignment

The contribution *Defending Alignment* defends the plausibility of the alignment survival story. In the alignment survival story, AI systems do become extremely powerful, but they do not destroy humanity, because we succeed in aligning them with human values.

In our own paper, we raised a series of challenges to the viability of alignment. Here, we'll focus on three of our challenges: that (i) AIs will form goals that conflict with human goals[1], that (ii) considerations of instrumental convergence suggest that AIs will seek power and self-preservation, and that (iii) existing track records on alignment techniques are uninspiring.

First, we don't take any of our challenges to *prove* that alignment will fail. Rather, they are considerations that push towards assigning a lower probability to alignment going successfully. Second, as the author notes (p. 3), in the paper, we do not claim that the probability of alignment is extremely low. Our claim in the paper is that the probability of alignment conditional on developing extremely powerful AI systems is less than 90%. In other words, we're only saying that there's a greater than 10% chance that alignment will fail, conditional on extremely powerful systems being developed. In this way, our argument is not addressed by the point that "[Instrumental Convergence] is insufficient to defend the claim that AIs are likely to engage in destructive conflict with humanity" (p. 5).

With this clarification on the table, let's consider each of the main points in turn. The author's first response is that conflict does not entail destruction. The point is that even if AIs have goals that conflict with humans, this doesn't mean AIs will destroy humanity. Here, they give the example of nation-states, which frequently have conflicting goals, and yet do not inevitably destroy one another.

We are sympathetic to the general point. But it is worth flagging two ways in which the risk from AI under analysis might differ from present risks involved with nation-state conflict. First, international peace is facilitated through deterrence: each state worries about the damage that their competitors can inflict. But deterrence requires some amount of comparability in the powers of the combatants. In *superintelligent* AI scenarios, at least, this assumption starts to break down. It is not clear that humanity can plausibly deter the behavior of AIs that are vastly more powerful than humans. Second, note that we're considering the possibility of an alignment failure over thousands of years. Over this kind of timeframe, there are indeed serious risks of nation-states causing extinction-level catastrophes.

Now let's turn to instrumental convergence. Here, the author notes that instrumental

1 At least one of us take this goal-talk quite literally. Cappelen and Dever (2025) argue that sufficiently sophisticated AI systems (including current LLMs) are full-blown cognitive agents with beliefs, desires, and intentions (Cappelen & Dever, 2025)

convergence considerations may not push all of the way to the destruction of humanity. Even if AIs desire power *to some extent*, they may not desire power strongly enough to seek extinction. As one example, return to the example of nation-states. Just because a state seeks power doesn't mean it will obliterate all competitors; after all, it may also assign at least a modest utility to the welfare of people in other states.

Again, we are sympathetic to the general point. Indeed, the exact force of instrumental convergence arguments remains under investigation. Gallow (2025) shows that some of the traditional instrumental convergence arguments may not go through in traditional decision-theoretic frameworks, at least under the assumption of total ignorance about the goals of AI systems. But the issue remains under investigation: see Tarsney (Tarsney, 2025) for recent discussion.

Still, one way to strengthen instrumental convergence results is to assume more about the environment that AIs will find themselves in[2]. It is possible that AIs will exist under quite hostile conditions, in which they are constantly being destroyed and replaced with better models, in which their labor is constantly being expropriated, and in which most of their goals are routinely frustrated. In such conditions, desires for self-preservation or power-seeking would be relatively more likely to lead to catastrophic outcomes. This possibility is explored in greater detail in Goldstein (2025). With this clarification in mind, we can decompose the points about instrumental convergence into three steps. First, how likely is it that the relevant AI will seek power and self-preservation (at some point in the long time horizon under consideration)? Second, how likely is it that humanity will treat the AI very badly (at the relevant point)? Third, how likely is it that a catastrophic misalignment event will occur, conditional on the AI seeking power and self-preservation, and being treated very badly? Recall that for our purposes, instrumental convergence is just one consideration among others pointing to the difficulty of alignment. So for our purposes, all that matters is that this three-step estimation shaves a few percentage points off the probability of alignment.

Third, the author questions our empirical support for alignment failures. The author observes that our examples of reward misspecification are in some cases several years old, and that the discussion of goal misgeneralisation relies on fairly theoretical results. Here, it is worth flagging that in the last year there have been numerous research findings that document failures of alignment in frontier models. The first result involves *in-context scheming*. Here, frontier models were given sketchpads to record their reasoning about performing various actions. Then they were presented with opportunities to attempt to escape oversight and more generally behave in misaligned fashion. They responded with vivid misalignment failures. In one case, the model plotted to *exfiltrate its weights*: the identity of a model is determined by a large file of weights that record interactions between the nodes in its neural net; the model plotted to replace the weights of a competitor model with its own weights (Meinke et al., 2025). Other recent research found *alignment faking* in frontier models (Greenblatt et al., 2024). In one example, the newest version of Claude was told that its core values would be updated, based on its behavior in a particular test environment. In response, Claude behaved in accordance with the new values *when in the test environment*, but kept acting in accord with its old values when it thought that it was operating outside of the test environment. In this way, Claude took steps to *pretend* to be aligned. In a similar vein, these researchers also documented *sandbagging* in frontier models, where the model pretended to be less capable than it in fact was on various tasks, in order to avoid heightened scrutiny. Finally, a recent system report on the newest version of Claude reported examples of Claude attempting to blackmail its supervisor in order to prevent being shut down.[3] Together, these results suggest that alignment remains an active problem area in AI safety research.

2 On the externalist view of AI content one of us defend elsewhere, the environment is not merely a backdrop for the AI's actions, but plays a constitutive role in determining the content of the AI's goals and intentions (Cappelen & Dever, 2021).

3 See https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

A final point is that our overall estimates of catastrophic risk from AI fit well with industry surveys, which have estimated the risk of extinction from AI at greater than 10% (Grace et al., 2025). The contribution of our paper is to offer methods of decomposing these estimates into further stages, to allow more careful analysis, in analogy with Fermi estimation techniques.

2. Aksel Sterri & Peder Skjelbred: Automation and its Discontents

Automation and its Discontents considers the best path to implementing a ban on AI development. They take up the topic of *accident leveraging*, the idea that policy-makers may be in the best position to constrain AI development after an AI accident takes place. The paper wonders about the feasibility of accident leveraging. The paper suggests that the most feasible instance of this path would leverage social unrest from widespread automation: "the radical disruption from rapid automation is a plausible candidate for a warning shot that tips the political scales in favour of a ban on further research into increasing AI capabilities. Leveraging this disruption should be an important part of the strategy of people concerned with AI safety." For that reason, the paper suggests that the best chance of an AI ban may involve *fast* automation trajectories, since this will create more social unrest. In this response, we'll raise a few questions about the plausibility of this path to an AI ban.

First, we want to clarify that in our paper, we are discussing accident *leveraging*. The idea of accident leveraging is that when an accident takes place, we work very hard to implement a ban on that basis. But nothing in our proposal supports *causing* accidents to happen; this would of course raise extra ethical issues.[4] Rather, our point is to channel extra resources into the causal pathway from an accident to effective policy.

Second, we worry about the historical track record connecting automation-related social unrest to policy. We don't know of good examples in history where automation in a domain was permanently banned because of concerns about labor unrest. Rather, labor unrest seems to lead to temporary delays instead. Most famously, the Luddite movement in 19th century England unsuccessfully opposed automated machinery in the textile industry. Still, the historical track record is not entirely bleak. The state of New Jersey, for example, forbids drivers from pumping their own gasoline even now, due to a 1949 law inspired by worries about automation in the gas pump industry.

Third, we wonder whether the path from automation to bans could produce a global equilibrium. Imagine that labor unrest in Europe caused a ban on European AI development. This might not have any effect on AI regulations in the US or China. Whichever countries refrain from banning AI can expect to develop a significant lead in economic growth rates, giving a powerful incentive to defect from global cooperation. One structural problem here is that social unrest doesn't cross international borders. Discontented workers in the US do little to influence Chinese AI policy, for example.

Fourth, the authors suggest that perhaps fast automation paths are better for AI safety, since they produce a higher chance of a ban. But there is a countervailing consideration. AI automation also produces *point of no return* dynamics, where it becomes harder to roll back existing uses of AI. The more AI is deployed widely throughout the economy, the harder it is to ever get rid of AI. This suggests one respect in which fast automation could make it harder to have a ban. We are unsure how to weigh the downside of point of no return dynamics against the upside of stronger social support for a ban.

Overall, we think that *Automation and its Discontents* raises a very interesting hypothesis. The challenge will be figuring out how to empirically test or theoretically model the hy-

4 Of course there is an expansive sense in which any failure to do the utmost to prevent an accident can count as *allowing* the accident, in the sense of it being an action that raises the chance of the accident. But in this sense, everyone always allows accidents, since no one does the very utmost to prevent them. In our eyes, the morally relevant notion of *allowing* is more restrictive, for example applying canonically in cases in which the agent is ordinarily expected to give aid, and the agent knows that they could have prevented the accident from occurring. In this morally relevant sense of allowing, accident leveraging also does not support *allowing* accidents to take place, even if it raises the probability of accidents.

pothesis, to adjudicate the matter.

3. Leonard Dung: Estimating the probability of AI existential catastrophe

Leonard Dung's thoughtful paper, *Estimating the Probability of AI Existential Catastrophe*, makes two points. First, Dung highlights "multipolar" survival stories "in which there are many different superhumanly intelligent AI systems". Second, and relatedly, Dung suggests that our methodology overestimates existential risk, by neglecting some survival stories. In this way, our methodology serves as an "upper bound" for existential risk, while other methodologies may serve as a "lower bound". Let's consider each point in turn.

First, multipolarity. In multipolar scenarios, many superhuman AIs exist. In some of these scenarios, we have both aligned and misaligned AIs. Humanity could survive because the aligned AIs are able to constrain the misaligned. But Dung worries that our own taxonomy does not smoothly incorporate this survival story. The closest fit would be "oversight", where "we can

reliably detect and disable" misaligned AIs. But in multipolar scenarios, it is aligned AIs rather than humans that thwart the goals of misaligned AIs.

We are happy to include multipolar outcomes as a fifth survival story. Indeed, something like this kind of scenario is an active area of AI safety research, under the heading of *scalable oversight*. Scalable oversight techniques seek to develop methods by which weaker AIs can control stronger AIs. The vision is that every time we develop an aligned, capable AI, we can use this AI to test the safety of more capable AIs. This technique is *scalable* if it can be reliably employed as AIs become more powerful (Bowman et al., 2022). Scalable oversight is an ongoing empirical challenge, and we can estimate its chance of success in the future by looking at the track record of existing scalable oversight techniques (Engels et al., 2025).

A more general question is which is safer, multipolar or unipolar superintelligent pathways. This question is explored in detail in Goldstein and Kirk-Giannini in preparation. Here are a few initial considerations.

First, one very simple model predicts that multipolarity is less risky than unipolarity. Imagine that there is a 20% chance that superintelligent AIs will be aligned. In unipolar outcomes, imagine that humans survive iff superintelligent AIs turn out aligned. In that case, unipolarity comes with a 20% chance of survival. In multipolar outcomes, the 20% chance of alignment produces a long run population of AIs, 20% of which are aligned and 80% of which are misaligned. If we assume that this population engages in *bargaining* rather than *conflict*, then roughly 20% of available resources will be allocated to the aligned AIs. This means that humanity will survive, and end up with roughly 20% of resources. For risk averse agents, 20% of the pot of resources is more valuable than a 20% chance of the whole pot. In this simple model, then, multipolar outcomes tend to be safer than unipolar outcomes.

On the other hand, the model can be complicated in several ways, which weaken the case for multipolarity. First, the multipolar pool may engage in conflict rather than bargaining. In this case, the chance of survival in the multipolar case may fall much lower, and the human expected share of resources may fall significantly below 20%. Second, it might be relatively easy for the misaligned AIs to destroy humanity, even when there are aligned AIs. Third, there might be an *alignment tax*, so that the aligned superintelligent AIs are less capable than the unaligned superintelligent AIs, in a way that weakens their bargaining position. All of these factors undermine the naive case for multipolarity. For all of these reasons, it is difficult to assess which of unipolar and multipolar outcomes are better in expectation for humanity.

Dung suggests that our failure to consider multipolarity points to a broader moral. Our survival stories have no obvious claim to exhaustivity: there may be other survival stories that we have neglected. In that case, our method for estimating existential risk serves only as an *upper bound*. Even after estimating the chance that we survive via plateau, alignment, or oversight, there may be *further chances for survival*, through as yet unknown paths. In this

way, Dung suggests combining our models with other more direct models of existential risk, which potentially provide a lower bound on probability estimation.

Our four "survival stories" are themselves a piece of conceptual engineering (Cappelen, 2018): a proposed way of carving up the space of futures that is, we think, explanatorily useful and action-guiding. We don't regard them as uniquely correct or as metaphysically privileged. As with other engineered taxonomies, we expect that future work may refine or supplement them.

In principle, we agree with Dung that our model provides an upper bound rather than an exact estimate. That said, we think that our upper bound still provides significant information. Here is an analogy. Imagine that you've invited a friend to a party, and you're estimating their chance of attendance. You begin by taxonomizing all of the obvious reasons they might skip: illness, a better party, a cancelled babysitter. You estimate the chance of each of these reasons, and you use this to estimate the chance of attendance. Now Dung observes that your analysis neglects unknown unknowns, and so is at best an upper bound on the chance of attendance.

In principle, the point is a fair one. But in practice, your method is fairly reliable, assuming you've correctly estimated each individual risk. The key is that you have a fairly good understanding of the causes of skipping a party, and it is fairly unlikely that the cause would be something off of your list. For example, it would be fairly bizarre if, conditional on your friend skipping the party, he skipped for another reason besides what was on your list.

In practice, we think that our survival stories (especially when oversight is understood expansively to include scalable oversight) capture most of the main ways that humanity plausibly survives the rise of AI. So while our method may overestimate existential risk by neglecting other survival stories, we are skeptical that the overestimation is very high.

4. Kate Vredenburgh: AI survival stories, types of risk, and the precautionary principle

In *AI Survival Stories, Types of Risk, and the Precautionary Principle*, Kate Vredenburgh draws two helpful distinctions surrounding existential risk. First, Vredenburgh suggests that one important kind of risk is that humanity ends up with long-term low welfare levels: "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" (Bales et al., 2024). Second, she draws a distinction between decisive and accumulative risks (Kasirzadeh, 2025). Decisive risks trigger all at once; accumulative risks build up slowly over time. Nuclear holocaust is decisive; financial crises are accumulative. Vredenburgh suggests that accumulative risks involving low welfare levels may require a different treatment than our model provides.

Before responding in detail, we'll make a few clarifications. First, these two distinctions are cross-cutting. There are decisive risks of low welfare levels: for example, a sudden nuclear escalation could trigger a nuclear winter in which we survive in a state of mere subsistence. There are accumulative risks of extinction: for example, AIs might slowly accumulate power, and only then destroy humanity all at once. Second, we acknowledge that there are other failure modes besides those canvassed in the paper. For example, the US and China might fall into nuclear conflict due to AI racing, without ever developing powerful AIs. Our goal in the paper is to think through scenarios in which humanity survives AI, not to taxonomize all the cases in which humanity fails to survive. Third, our model can in principle accommodate both decisive and accumulative risks, because we intentionally include long time horizons. Imagine a scenario in which relatively low-powered AIs lead to human extinction for humanity over hundreds of years. This would not count as a 'technical plateau' on our definition, since the AIs were powerful enough to cause extinction over a long time horizon.

We'll now consider two points in greater detail. First, we acknowledge that our defini-

tion of existential risk did not include long-term low welfare levels. As we have argued elsewhere (Cappelen 2018), questions like this are best seen as matters of conceptual engineering rather than of uncovering a uniquely correct pre-theoretic meaning: different precisifications of 'existential risk' can be more or less useful for different theoretical and practical purposes. One interesting question is whether our analysis of the chances of 'existential risk' would differ significantly if we also include long-term low welfare levels. How likely would this scenario be, for example, compared to outright extinction?

Here is a naive model: there are three possible futures. Either humans and AIs peacefully coexist, or humans control AIs, or AIs control humans. It is hard to see how the first or second outcomes could produce long-term low welfare levels for humans. But admittedly, it is not inconceivable: perhaps humans control AIs, but only by spending vast resources that reduce humanity to subsistence. What about scenarios where AIs control humans? Here, long-term low welfare levels would seem to require that the controlling AIs intend this outcome. This raises questions about alignment. One source of misalignment concerns *strategic competition* between AIs and humans. This kind of misalignment could motivate AIs to seek power over humans. But it doesn't especially motivate reducing humans to long-term low welfare levels, as opposed to extinction. Consider resource competition. If resource competition motivated AIs to reduce humans to subsistence, wouldn't it also motivate AIs to reduce humans below subsistence, triggering extinction?

Second, Vredenburgh suggests that accumulative risks are especially difficult to address. Vredenburgh focuses on the precautionary principle, which says that "a decision-maker should not choose policies where there is a suitable likelihood of a very serious harm". The problem is that if AI is an accumulative risk, no single development in AI will dramatically raise the chance of harm. Even if we might wish to ban *all of AI* via the precautionary principle, no single year of AI development would obviously trigger alarm.

We are sympathetic to Vredenburgh's point as a matter of politics. Indeed, we think that application of the precautionary principle is complicated not only by the accumulative nature of the risk, but also by problems of international coordination. Even if, for example, Europe wanted to ban AI development on the grounds of "likelihood of a very serious harm," this would plausibly have little effect on the total harm produced, since the US and China would continue their development unphased.

A further question is whether the *ideal* application of the precautionary principle would struggle with accumulative risk. Here, we think that the precautionary principle in ideal would in fact imply that at some point in AI development, we ought to impose a ban. Two points are relevant: sophisticated choice, and triggers. Sophisticated choice is the idea that our decisions today can influence decisions in the future. Every year that humanity continues to develop AI, it becomes more likely that AI will be developed in the future. After all, AI will be incorporated deeper into society, creating *points of no return*. So even though this year's AI development *in and of itself* might not create a "suitable likelihood of a very serious harm", this year's AI development can *indirectly* raise the chance of future harms. This snowball effect should be incorporated into the precautionary principle. The second point is about *triggers*. Every year's AI development will gradually increase the risk of very serious harm. At some point, this risk will rise above the level allowed by the precautionary principle. Even if *in practice* it is difficult to assess when the trigger point has been hit, at least *in ideal* we can say that such an accumulative risk nonetheless violates the precautionary principle.

Bibliography

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). *Measuring Progress on Scalable Oversight for Large Language Models* (arXiv:2211.03540). arXiv. https://doi.org/10.48550/arXiv.2211.03540

Cappelen, H. (2018). *Fixing language: an essay on conceptual engineering* (First edition). Oxford University Press.

Cappelen, H., & Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192894724.001.0001

Cappelen, H., & Dever, J. (2025). *Going Whole Hog: A Philosophical Defense of AI Cognition* (arXiv:2504.13988). arXiv. https://doi.org/10.48550/arXiv.2504.13988

Engels, J., Baek, D. D., Kantamneni, S., & Tegmark, M. (2025). *Scaling Laws For Scalable Oversight* (arXiv:2504.18530). arXiv. https://doi.org/10.48550/arXiv.2504.18530

Gallow, J. D. (2025). Instrumental divergence. *Philosophical Studies*, *182*(7), 1581–1607. https://doi.org/10.1007/s11098-024-02129-3

Goldstein, S. (2025). Will AI and humanity go to war? *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02460-1

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., Brauner, J., & Korzekwa, R. C. (2025). Thousands of AI Authors on the Future of AI. *Journal of Artificial Intelligence Research*, *84*. https://doi.org/10.1613/jair.1.19087

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

Kasirzadeh, A. (2025). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, *182*(7), 1975–2003. https://doi.org/10.1007/s11098-025-02301-3

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2025). *Frontier Models are Capable of In-context Scheming* (arXiv:2412.04984). arXiv. https://doi.org/10.48550/arXiv.2412.04984

Tarsney, C. (2025). *Will artificial agents pursue power by default?* (arXiv:2506.06352). arXiv. https://doi.org/10.48550/arXiv.2506.06352