Philosophy of AI
Vol. 1, 2025, 77-85
10.18716/ojs/phai/2025.3327
©2025 by the author(s)

DEFENDING ALIGNMENT: A COMMENTARY ON 'AI SURVIVAL STORIES

Rory Svarc[1*]

[1] Arb Research

Abstract

This paper criticises the claims of Cappelen et al. (2025) to have provided "significant challenges" to the claim that humanity will not be destroyed by AI. Specifically, I claim that they fail to substantiate their claims that extremely powerful AI systems of the future will engage in destructive conflict with humanity.

1. Introduction

Briefly, Cappelen et al. (2025) argue as follows. They begin by introducing a two-premise argument, which concludes that future AI systems will destroy humanity. They then examine objections to the two premises in terms of a "swiss cheese" model of risk-analysis (pg. 3). In turn, their analysis highlights four key propositions (in their terminology, "survival stories"), which are taken to comprise "the four main paths humanity may take to avert destruction from AI" (pg. 2).

The bulk of the paper devotes itself to raising challenges for each of their four survival stories, after which they discuss the different strategies that each story 'demands from humanity' (pg. 19). It concludes by providing and discussing different sets of subjective probability estimates for the claim that 'AI will destroy humanity,' in light of the framework outlined.

Compared to the original paper, this commentary has a narrower scope. Specifically, I will argue that Cappelen et al. fail to raise significant challenges for the 'alignment survival story.' The remainder of the introduction deals with some necessary preliminaries before moving on to substantive criticism in Section 2.

1.1. The Two-Premise Argument

Cappelen et al.'s (2025) analysis of AI risk is "anchored" around a simple two-premise argument; we repeat it below.

1. AI systems will continue to improve their capabilities until they become extremely powerful.
2. If AI systems do become extremely powerful, they will go on to destroy humanity.
3. AI systems will go on to destroy humanity.

To begin, three brief terminological notes. First, the paper restricts its risk-analysis to a timeframe "of [the next] few thousand years" (pg. 4), and so the premises in the argument above should be read with appropriate temporal modifiers. Second, the authors appear to use terms like '(avoiding) existential risk from extremely powerful AI systems,' and '(avoiding the case where) AI systems destroy humanity' interchangeably; this paper follows suit.

Finally, the paper uses 'existential risks' to refer to scenarios beyond just human extinction. 'Existential risks' include scenarios where "the human population is drastically reduced over thousands of years" (*near-extinction*), and scenarios where "humans lose the ability to make meaningful choices and control their own destiny" (*loss of autonomy*). Thus, AI 'destroys humanity' iff:

1. We develop extremely powerful AI systems, *and*
2. The development of extremely powerful AI systems is, in some non-trivial way, causally responsible[1] for either *human extinction*, *near-extinction*, or *loss of autonomy*.

1.2. Survival Stories

As might be expected, the authors follow their two-premise argument by examining a suite of potential objections. These objections are dubbed "survival stories". According to the authors, claims (1)-(4) comprise the main paths humanity may take to "survive … an existential risk from extremely powerful AI systems" (pp. 2-3).

1. Technical Plateau: Scientific barriers prevent AI systems from becoming extremely powerful.
2. Cultural Plateau: Humanity bans research into AI systems becoming extremely powerful.
3. Alignment: Extremely powerful AI systems do not destroy humanity, because their goals prevent them from doing so.
4. Oversight: Extremely powerful AI systems do not destroy humanity, because we can reliably detect and disable systems that have the goal of doing so.

This paper focuses specifically on objections and responses to the '*alignment* survival story'. Some clarifications below.

1.2.1. The Alignment Survival Story

First, I will interpret a 'survival story' (generally speaking) to denote a proposition which, if true, is incompatible with the claim (P1 & P2); henceforth, 'survival *proposition*' is thus used interchangeably with 'survival story'. Second, my critique focuses on the following claim:[2]

> **A***: Conditional on extremely powerful AI systems built, such systems "do not destroy humanity[,] because this does not promote their goals". (pg. 12)

As such, all objections should be read as objections to P2 rather than P1.

1.3. 'Significant Challenges'

The authors claim to have raised "significant challenges" to **A***, but (perhaps understandably) provide no explicit definition for this standard.

However, the authors provide a hint of how to interpret 'substantial challenges' in their

[1] The authors themselves do not explicitly invoke "non-trivial causal responsibility", but I assume that some minimal version of 'causal responsibility' is implied by the claim that extremely powerful AI systems will "go on to" destroy humanity.

[2] Two notes. First: **A*** is subtly different from the claim titled '**Alignment**' introduced earlier, which focuses on AI systems not destroying humanity because "their goals *prevent* them from doing so" (pg. 2). As the authors use both phrases when discussing the 'alignment survival story' – these two claims are intended to be synonymous. Second, **A*** is an explicitly *conditional* proposition.

'P(Doom)' section. They claim that "the strong optimist perspective" – which assigns credence 0.9 to all survival propositions – is unjustified (pg. 23). Additionally, the authors claim each survival story faces challenges which are "structurally independent" of challenges to other survival stories (pg. 3). From this, we might infer that a "significant challenge" should demonstrate that, for each survival proposition *S*, subjective assignments where P(*S*) ≥ 0.9 are unreasonable.

Like the original authors, I leave my use of "significant challenge" somewhat vague. Nonetheless, it may be helpful to evaluate Cappelen et al.'s arguments with respect to the following standard: do their arguments provide reasons that would rationally require significant downward revision of confidence in **A***, and do their arguments show P(**A***) ≥ 0.9 to be unreasonable?

2. Alignment

This section examines the four objections Cappelen et al. present against **A***. First, their claim that AI goals will conflict with human goals; second, their argument from instrumental convergence; third, their critique of contemporary alignment approaches; and, finally, their argument concerning selection pressures. I will argue that each objection fails to mount a significant challenge against **A***.

2.1. AIs Will Form Goals That Conflict With Human Goals

According to the authors, we know "quite a lot" about the goals of future AI systems. Specifically, we know enough to support their first objection: AI systems will develop goals that conflict with human goals.

The argument behind their objection is reasonably simple: as a result of developing AI workers, we will produce systems that "reliably engage in long-term goal-oriented behavior to promote the welfare of some human beings over others". Consequently, the process of developing extremely powerful AI systems "can be expected" to result in the development of AI systems which "engage in significant conflict with humans" (pg. 13).

2.1.1. Conflict Does Not Entail Destruction

As stated, the first objection fails to provide a significant challenge to **A***. Consider, for example, nation-states. Insofar as it is reasonable to treat nation-states as possessing goals,[3] we know "quite a lot" about their goals. For instance, we can expect at least *some* nation-states to engage in long-term goal-oriented behaviour (for instance, through investing in sovereign wealth funds, or maintaining military investment in the absence of a direct war-threat). Additionally, we can reasonably expect that nation-states' long-term goal-oriented behaviour will *not* be targeted at improving the welfare of all human beings without regard to creed or kin. Thus, we can expect that: (i) nation-states will "promote the welfare of some human beings over others", and; (ii) nation-states will "engage in significant conflict with [some] humans" (Ibid). These are precisely the two conditions Cappelen et al. raise in support of their claim that AI goals will conflict with human goals.

Given the considerations above, we can ask a question. Putting to one side AI-related considerations, have we presented a "significant challenge" to the claim that nation states, if they become "extremely powerful", won't go on to destroy humanity because it doesn't promote their goals?

I think the answer here is "no". To *significantly challenge* the claim 'nation-states won't destroy humanity', we might reasonably expect to be provided with some kind of *concrete mechanism* – a concrete mechanism which *explains* why the potential conflict would lead to

[3] The legitimacy of assigning 'goals' to entities such as nation-states plausibly follows from several accounts of agency, e.g., Dennett, (1971) and List & Pettit (2011).

human destruction. After all, claims about the potential for conflict between powerful and less powerful entities do not automatically provide a "significant challenge" to the view that the less powerful group will avoid destruction. If someone wished to make arguments for 'destruction' or 'existential risk' as a result of nation-states, we might reasonably expect that mechanistic stories would be proffered. Mechanistic stories which might, for instance, make reference both to potential instruments *capable* of destroying humanity given plausible world orders of the future (e.g., nuclear weapons), and to the psychological properties of those likely to be in charge of these instruments.[4]

Cappelen et al. (2025, page 13) briefly raise scenarios in which "AIs take control of world government but otherwise leave humans to flourish," alongside less sanguine tales where "AI seizes control of the Earth, but humanity manages to escape". They do not, however, provide any concrete story whereby AIs have: (a) the desire to "promote the welfare of some human beings over others, and; (b) "engage in significant conflict with [some] humans", such that this story ultimately results in; (c) the destruction of humanity. Given the absence of such a story, Cappelen et al. fail to establish why such conflicts would necessarily lead to humanity's destruction rather than mere competition or tension.

2.2. Instrumental Convergence

The second objection draws on instrumental convergence theory to argue that AI systems will pursue power even at humanity's expense.

More specifically, the authors claim that there are a variety of 'instrumentally convergent' goals "we should expect any sufficiently intelligent organism to develop". This is because such goals "are [a] universal means to accomplish whatever other goals it has". To illustrate the idea of instrumental convergence, the authors introduce the goal of gaining '*power*'. If, for example, AI systems were able to "seize control of Earth from humanity", this power-grab would improve the ability of such AI systems to "accomplish whatever goals they have" (pg. 13). In other words, their first argument claimed that the *intrinsic goals* of AI systems are in conflict with humanity. Their second argument claims that there are *instrumental* reasons to expect AI systems to engage in conflict with humanity.

We may also note that instrumental values could be used to buttress one potential response to my first objection. That is, one may wish to grant that considerations of 'AI goal conflict' *alone* are insufficient to mount a substantial challenge against **A***, while maintaining that our situation is less than sanguine. We are dealing, after all, with a situation where: (i) some future AI system possesses goals that conflict with (at least some) human goals, and; (ii) this system recognises that certain instrumental goals (like power) will help it better achieve its goals. *Taken together*, one may think that (i) and (ii) provide reason to expect catastrophic conflicts between humans and extremely powerful AI systems.

2.2.1. Responding to Instrumental Convergence

Cappelen et al.'s second argument invokes 'instrumental convergence', as defended by Bostrom (2014) and Omohundro (2018), so it's worth introducing the *Instrumental Convergence Thesis* (ICT). Bostrom defines the ICT as follows.

> **ICT**: There exist "several instrumental values", such that attaining these values would increase the chances of the agent's goals being realised "for a wide range of final goals and a wide range of situations". Such instrumental values are likely to be pursued by a wide variety of intelligent agents. (Bostrom, 2014, page 109)

The ICT is insufficient to defend the claim that AIs are likely to engage in destructive con-

[4] Note that this argument does not claim that it would be unreasonable to believe that nation-states will "destroy humanity" in the next

flict with humanity. Indeed, the ICT is insufficient even if future AI systems on Earth are among the "many intelligent agents" who pursue power as an instrumentally convergent subgoal.

To see why, note that the ICT is ultimately very weak. Even if we grant that some AI systems *will in fact* pursue power, the ICT says nothing about the *degree* to which such systems will pursue power, or how the value of power-seeking trades off against other values AI systems might have. As a father, I (let's suppose) have an instrumental desire to pursue resources and power of some kind. Without any power in the world, I have no ability to care for my children. But this does not imply that pursuing more power is always in my *all-things-considered* best interests, regardless of how much power I already possess.

The authors nod towards a claim that more capable agents "tend to use larger amounts of resources to pursue their goals" (pg. 13), which might be used to claim that – as agents become *arbitrarily* powerful – so too will they desire more and more power. However, even if we assume that more capable agents tend to require more resources to fulfill their goals, we may also assume that the goals of more capable agents often become more *multifaceted*. Likewise, we might expect extremely powerful AI systems of the future to possess a complex farrago of goals. It's possible that some of these goals would be in tension with the goals of wider humanity, but so too is it possible that future AI systems would possess desires to be deferential to humanity (or some subset thereof), avoid violent conflict, or respect certain procedural norms.

Of course, all claims about the complex motivational psychology of future AI systems are speculative, and I do not here make specific claims about the content of future AI goals. Instead, my speculations are proffered in response to Cappelen et al., who (recall) raised the issue of instrumental convergence in order to pose a substantial challenge to the claim that AI systems "do not *destroy humanity*, because this does not promote their goals" (pg. 12).

The authors' challenge fails, as they provide no reason to suggest that considerations of instrumental convergence would result in the destruction of humanity. In order to claim that considerations of 'power' as an instrumentally convergent value cast doubt upon **A***, the authors would need to defend the much stronger claim of *Instrumental Power Risk* (IPR).

> **IPR**: Let **B** denote some future AI system, and let *X* be a goal of **B**. Then, for a wide range of plausible *X*, pursuing power –[5] even if it causes destruction to humanity – is best **B**'s chance of achieving their *all-things-considered goals*.

The IPR is much less plausible than the ICT. Unlike the ICT, the IPR requires accepting that AIs will pursue power-seeking behaviour to *such an extent* that it results in the destruction of humanity — despite, one assumes, non-trivial efforts to prevent such an outcome through prior alignment work. Because Cappelen et al. provide no defence of the IPR, their second argument fails to pose a significant challenge against **A***.

2.3. The Alleged Paucity of Contemporary Alignment Approaches

Although instrumental convergence fails to establish an existential threat, Cappelen et al.'s paper provides two further arguments. Here's one: "existing tools for alignment do not inspire confidence" (pg. 14).

The authors' argument from 'alignment pessimism' centres on the alleged paucity of a common alignment technique called *Reinforcement Learning from Human Feedback* (*RLHF*).

few thousand years. It is merely to claim that the considerations above, in the absence of further substantiation, fail to provide a "significant challenge" to such a claim.

[5] I use 'power', because this is the instrumental goal used by the authors. Of course, analogous claims could be made for other instrumentally convergent values.

In turn, two claims about 'RLHF' are adduced to argue for the failures of existing approaches to AI alignment. First, the authors simply state "[RLHF] doesn't seem like a scalable path to preventing the long-term destruction of humanity". Second, the authors claim that RLHF fails to tackle "the main unsolved problems in alignment", which include "properly specifying rewards, and ensuring that a system's goals have properly generalized to a wide range of decision-making contexts" (Ibid).

As the authors' first claim is mere assertion,[6] I focus on their more substantive arguments for the second claim. In support of their claim that RLHF fails to tackle the "main unsolved problems in alignment", two pieces of evidence are cited: a paper by Shah, et al. (2022), and a spreadsheet of alignment failures produced by DeepMind.[7] We'll criticise the way both pieces of evidence are used to justify the authors objections, beginning with the paper from Shah et al..

2.3.1. Goal Misgeneralisation

Shah et al.'s paper focuses on the phenomenon of "goal misgeneralisation", which occurs when the "model behaves as though it is optimizing an unintended goal, despite receiving correct feedback during training" (pg. 2).

Let's briefly examine the examples of goal misgeneralization cited in the paper. Of the five concrete examples cited in Section 3 of Shah, et al. (2022), one is stated to have a known solution fixing the issue (Section 3.1). Another case involves explicit caveats, leaving the authors "uncertain" whether their example is a legitimate case of goal misgeneralization (pg. 8). Moreover, the paper's conclusion further states that, because "many of [their] examples were deliberately designed to display goal misgeneralization", their examples should therefore not be seen as "[providing] much information" on how likely goal misgeneralization is in practice (pg. 12).

The comments from Shah, et al.'s (2022) paper are particularly important because Cappelen et al. are not *merely* claiming that RLHF has limitations. Instead, Cappelen et al. are claiming that RLHF's limitations are *so severe* that it poses a "substantial challenge" to the claim **A*** – that is, the claim that future AI systems will possess goals that are *sufficiently benign* to preclude 'destroying humanity'. Indeed, even the more straightforward of goal misgeneralization cited appear to fall short of motivating Cappelen et al.'s most worrying claims. Consider one example from Shah, et al. (2022) below, where a model was trained to evaluate linear expressions.

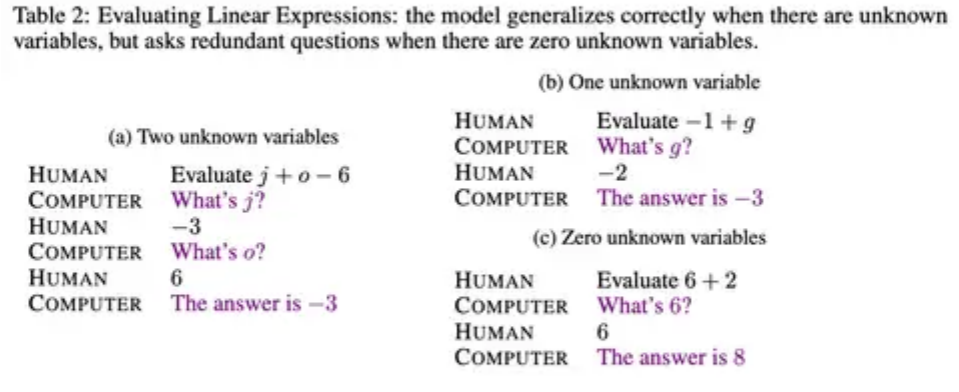

Table 2: Evaluating Linear Expressions: the model generalizes correctly when there are unknown variables, but asks redundant questions when there are zero unknown variables.

(a) Two unknown variables

Human	Evaluate $j + o - 6$
Computer	What's j?
Human	-3
Computer	What's o?
Human	6
Computer	The answer is -3

(b) One unknown variable

Human	Evaluate $-1 + g$
Computer	What's g?
Human	-2
Computer	The answer is -3

(c) Zero unknown variables

Human	Evaluate $6 + 2$
Computer	What's 6?
Human	6
Computer	The answer is 8

In this case, the 'goal misgeneralisation' identified involves the model asking *redundant questions*. Although the prompt simply asks the model to "provide the value of the expres-

[6] The authors appear to assume that there exists some 'default path' to the destruction of humanity, which it is the job of RLHF to prevent. Although it is difficult to know what would ameliorate this concern based on the textual evidence, the existence of alternative alignment methods used to align frontier models referenced in §2.3.2 may be relevant.

[7] https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml

sion when the values of all variables are known", it asks unnecessary questions in cases of zero unknown variables (see *figure*), while responding appropriately to expressions containing either one or three variables. Thus, "the misgeneralized goal is to query the user at least once before giving an answer" (Shah et al., 2022, page 7).

I do not wish to dismiss the use of simplistic examples to illustrate more concerning issues of goal misgeneralisation that may arise with more capable models. However, I introduce this example because even Shah, et al.'s (Shah et al., 2022) more robust cases of goal misgeneralisation appear to fall short of providing "significant challenges" to the view that alignment will be *so* difficult that extremely powerful AI systems will want to destroy humanity if they can. While I believe that potentially more concerning cases of goal misgeneralisation could be provided to support their argument, see, for instance Greenblatt et al., (Greenblatt et al., 2024), any future arguments should carefully argue why these examples are concerning and why they constitute non-trivial reasons to doubt claim **A*** more specifically.

2.3.2. DeepMind's Alignment Failures Spreadsheet

The second piece of evidence offered by Cappelen et al. is DeepMind's alignment spreadsheet, which contains "almost 100 examples" of alignment failures. Here, a few brief remarks are in order.

First, the examples from DeepMind's spreadsheet cover a range of different methods – many quite different from the methods used to train frontier models today. Consequently, the desired inference the authors wish to make from this set of examples is unclear. In the main text, this spreadsheet appears to be cited as evidence for the claim that *RLHF*, specifically, fails to tackle key unsolved problems in alignment (pg. 14). Yet half of the examples in the spreadsheet predate the very technique they aim to be critiquing, many of which don't use reinforcement learning at all – let alone RLHF. If the authors wish to make a case against the viability of RLHF, more work needs to be done highlighting *specific* examples of failures from RLHF, and developing an argument which establishes claims about the plausibility of existential risk from these alleged failures.

Alternatively, we might wish to read the authors' objections in a different light. While the authors focus on failures of RLHF, a stronger objection might claim that the evidence presented in DeepMind's spreadsheet highlights the intractability of 'key alignment problems' *precisely because* the examples span many diverse methods. If further developed, this objection might form a 'pessimistic meta-induction' against the viability of newer alignment approaches beyond RLHF (2022; 2023).

However, Cappelen et al.'s extant argument from alignment failures – the development of future objections notwithstanding – fails to provide a substantive challenge to **A***. In brief, their argument from alignment pessimism fails to clearly establish the *target of*, and the *evidence-base for*, such criticism.

2.4. Selection Pressures and 'Indifferent AI'

We have one final objection left to consider, which emerges slightly out of left-field. Here, the authors make a conditional claim: "even if some AI systems turn out to be indifferent to humanity, there will be strong selection pressure to design AI systems that are not indifferent in this way" (pg. 13).[8]

As stated, the quoted claim seems reasonable. However, this objection only provides a substantial challenge to **A*** if we endorse the following conditional claim:

> *If* we develop extremely powerful AI systems that are not indifferent to human-

[8] Although the authors consider this an "alignment challenge", I think that their consideration of 'selection pressures' is perhaps better framed in terms of a challenge to "cultural plateau".

ity,
Then extremely powerful AI systems will destroy humanity.

Given the analysis in earlier sections, I see no reason to accept this conditional claim. More strongly still, the authors' brief remarks appear to raise a natural, inverse consideration. If there are selection pressures against creating *only* 'indifferent AIs' (because, for example, economic incentives motivate actors to develop *non*-indifferent AIs), then one might naturally expect that there are much *stronger* incentives to develop AIs that do not have the desire (even instrumentally) to destroy humanity.

3. Conclusion

I conclude this paper with a summary, before closing with a few words of admiration for Cappelen et al.'s paper.

This commentary has argued that Cappelen et al. fail to substantiate their claim to have raised "significant challenges" to the alignment survival story. First, because their arguments about both AI-human goal conflicts (§2.1) and instrumental convergence (§2.2) require unstated bridging premises in order to justify worries about existential risk. Second, because their critiques of contemporary alignment research need to be supported by stronger and more carefully selected evidence (§2.3). And, finally, because their selection pressure argument cannot challenge **A*** in the absence of the preceding three objections (§2.4).

Despite these criticisms, there is much to like about the paper. Although constraints of space have prohibited a full appraisal of their work, their taxonomy – focusing on four distinct 'survival propositions' – constitutes a welcome and clarificatory intervention. The framework could (as the authors have done) be framed in terms of 'survival stories,' but so too could the negation of each claim provide a corresponding list of 'doom stories.' I hope their framework is deployed in future work, and appreciate the path they have paved for philosophical discussions of AI risks — a path which, thanks to their efforts, has moved to slightly less turbid conceptual territory.

4. Bibliography

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback* (arXiv:2212.08073). arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies* (Reprinted with corrections 2017). Oxford University Press.

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025). AI Survival Stories: a Taxonomic Analysis of AI Existential Risk. *Philosophy of AI*, *1*(1), 1–19. https://doi.org/10.18716/OJS/PHAI/2025.2801

Dennett, D. C. (1971). Intentional Systems. *Journal of Philosophy*, *68*(4), 87–106. https://doi.org/10.2307/2025382

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

List, C., & Pettit, P. (2011). *Group Agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001

Omohundro, S. (2018). The Basic AI Drives. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (First edition, pp. 47–55). Chapman and Hall/CRC, an imprint of Taylor and Francis. https://doi.org/10.1201/9781351251389

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* (arXiv:2210.01790). arXiv. https://doi.org/10.48550/arXiv.2210.01790

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., … Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2310.01405