Is Alignment Unsafe?

Cameron Domenico Kirk-Giannini
[Draft — Please cite published version when available.]

Abstract: Inchul Yum (2024) argues that the widespread adoption of language agent architectures would likely increase the risk posed by AI by simplifying the process of aligning artificial systems with human values and thereby making it easier for malicious actors to use them to cause a variety of harms. Yum takes this to be an example of a broader phenomenon: progress on the alignment problem is likely to be net safety-negative because it makes artificial systems easier for malicious actors to control. I offer some reasons for skepticism about this surprising and pessimistic conclusion.

With few exceptions, technical and theoretical work in AI safety has proceeded from the assumption that finding a way to align artificial systems with human values would be a net safety-positive intervention. Only if we can control the increasingly powerful AI systems we develop can we prevent them from causing a variety of harms ranging from existential catastrophe to the perpetuation of unjust social structures. Indeed, this presumed connection between control and safety is so established that one major focus of the emerging literature on AI sentience and moral standing is how to reconcile human safety with the possibility that it might be morally impermissible to dominate, constrain, or control artificial systems in certain ways (Goldstein and Kirk-Giannini 2023a, Tubert and Tiehen 2024).

Inchul Yum (2024) articulates a more pessimistic vision of the relationship between control and safety: by making powerful AI systems easier for various human actors to use, alignment strategies increase the probability that such systems will be misused by malicious actors to cause harm. For Yum, it follows that aligning artificial systems is likely to be a net safety-negative intervention. Yum focuses on the "language agent strategy" discussed in recent work of mine (Goldstein and Kirk-Giannini 2023b), but he is explicit that his remarks are meant to generalize: there is "a broader conflict between alignment risk and misuse risk" (2024, 3).

Yum's contribution to the wider conversation about alignment and human safety is valuable, and I share many of his concerns and starting points. The central claim Simon Goldstein and I make about the language agent strategy in our paper is that it reduces the probability of a *misalignment catastrophe* — a catastrophe that "result[s] from humans losing control over an AGI system" (1). We explicitly set aside consideration of catastrophes that result from the activity of malicious actors, though we take this to be a serious issue. Yum adopts the conclusion of our argument about the risk of an alignment catastrophe as a starting point; he argues that it is precisely *because* the language agent strategy makes alignment less difficult that it is likely to be net safety-negative. So I agree with Yum on the likely impact of the language agent strategy when it comes to the alignment problem, as well as on the claim that malicious use is a second serious source of risk from advanced AI technologies and the idea that there may be difficult trade-offs between reducing the risk of a misalignment catastrophe and reducing the risk of malicious use.

I am not, however, convinced of Yum's pessimistic conclusion that the language agent strategy — along with successful alignment strategies more generally — is likely to be net safety-negative.

In assessing the likely safety implications of the language agent strategy, it is important to keep in mind which scenarios we are comparing. To a first approximation, an alignment strategy is likely to have a net positive impact on safety if the probability of a safe outcome conditional on adopting that strategy is higher than the probability of a safe outcome conditional on some baseline. For simplicity, we can assume this baseline is allowing research on AI technology to progress at its normal rate without adopting any alignment strategy.

Keeping this methodological point in mind weakens the considerations Yum uses to motivate the idea that the language agent strategy is likely to be net safety-negative. For example, consider Table 1, in which he lists a number of positive and negative outcomes of "developing powerfully aligned AI systems using language models" (2024, 9). Among the positive outcomes are improved systems for cancer detection and new AI productivity tools that contribute to economic growth. Among the negative outcomes are AI-based cyberattacks and the expansion of surveillance.

Yum is careful to note that the entries in Table 1 are not a complete picture of the possible positive and negative outcomes of adopting the language agent strategy, and also that language agents are not the only AI systems that could lead to the outcomes listed. Yet the form of his subsequent argument does not fully take into account the significance of these qualifications. In particular, he argues that the language agent strategy is likely to be net safety-negative because the negative items in Table 1 outweigh the positive items.

One problem here is that many of the outcomes listed in Table 1 do not plausibly depend on the language agent strategy, or alignment efforts more generally. Systems for surveillance, cancer detection, education, and so forth depend on advances in computer vision and base LLM technology which are foreseeable whether or not the alignment problem is solved — and the same is true of most of the entries in the table. To assess the likely impact of alignment strategies, we should not weigh up the positive and negative entries in a table like Table 1; instead, we should weigh up the *expected difference* that adopting an alignment strategy will make to the likelihood and value or disvalue of each entry compared to the baseline.

Another problem is that a sound assessment of the likely impact of alignment strategies requires a table that is complete. For example, Yum includes the development of advanced bioweapons among the possible negative outcomes of advanced AI. There is indeed a serious concern for alignment research here, since making it easier to develop advanced bioweapons plausibly increases the risk of catastrophic outcomes from malicious use. But Yum does not include the possibility of a misalignment catastrophe among the possible negative outcomes of advanced AI. This is an important omission, since the main case for the language agent strategy is that it reduces the probability of a misalignment catastrophe. What Yum needs in order to establish his pessimistic conclusion about alignment is an argument that alignment strategies are likely to increase the expected disutility from possibilities like bioweapons falling into the hands of malicious actors *more* than they decrease the expected disutility from possibilities like a rogue AI destroying humanity.

The closest Yum comes to offering an argument of this kind is his discussion of the idea that "chaotic results [from misaligned AI systems] are often relatively harmless or benign" (2024, 13). In this connection, he draws on examples including "translation software comically failing to properly render a name, a virtual assistant unhelpfully responding about cheese locations, and image generation creating bizarrely merged photos" (2024, 13). Of course, if misalignment is sure to be benign rather than catastrophic, we should not trade an increased risk of malicious misuse for a

reduced risk of misalignment. But these examples will leave many who worry about a misalignment catastrophe rather cold: their concern is not with simple systems delivering chaotic results, but rather with advanced, agentive systems functioning in ways that are both unintended *and competent* (Carlsmith 2021, Bales et al. 2024).

Works Cited

Bales, A., D'Alessandro, W., and Kirk-Giannini C. D. (2024). Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass* 19(2): e12964.

Carlsmith, J. (2021). Is power-seeking AI an existential risk? arXiv Preprint: <https://arxiv.org/pdf/2206.13353>.

Goldstein, S. and Kirk-Giannini, C. D. (2023a). AI wellbeing. PhilPapers Preprint: <https://philpapers.org/rec/GOLAWE-4>.

Goldstein, S. and Kirk-Giannini, C. D. (2023b). Language agents reduce the risk of existential catastrophe. *AI & Society*. Online First.

Tubert, A. and Tiehen, J. (2024). Existentialist risk and value misalignment. *Philosophical Studies*. Online First.

Yum, I. (2024). Language agents and malevolent design. *Philosophy & Technology*. Online First.