Answer: Our work on RepE shows that we now have traction on deceptive alignment, which has historically been the most intractable (specific) rogue AI failure mode. We could also use this to identify when an AI acted recklessly, based on its own internal probability and harm estimates. This could also help us ensure that we do not build sentient AIs or AIs that are moral patients.

4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: This directly reduces the existential risks posed by rogue AIs (Carlsmith, 2023), in particular those that are deceptively aligned.
5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? □
6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? □
7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? □
8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility? □

E.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

9. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: This work mainly improves transparency and control. The underlying model is fixed and has its behavior nudged, so it is not improving general capabilities in any broad way.
10. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
    Answer: A diffuse effect is that people may become less concerned about deceptive alignment, which may encourage AI developers or countries to race more intensely and exacerbate competitive pressures.
11. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? □
12. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities? □
13. **Correlation with General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? □
14. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? □

E.3 Elaborations and Other Considerations

15. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
    Answer: Hendrycks et al. (2023) provide four AI risk categories: intentional, accidental, internal, and environmental. This work makes internal risks—risks from rogue AIs—less likely.

    In the past, people were concerned that AIs could not understand human values, as they are "complex and fragile." For example, given the instruction "cure cancer," an AI might give many humans cancer to have more experimental subjects, so as to find a cure more quickly. AIs would pursue some goals, but do so without capturing all the relevant nuances of human values. This misalignment would mean some values would be trampled. But in our ETHICS paper (Hendrycks et al., 2021a), we showed that AIs did indeed have a reasonable understanding of many morally salient concepts. In follow-up work, we showed we can control AI agents to behave more ethically (Hendrycks et al., 2021c; Pan et al., 2023). Now that we can control models to pursue human values to some extent, we need to work to make them their understanding of human values reliable and adversarially robust to proxy gaming.

Since it became possible to instruct AIs to fulfill requests with some fidelity, for many conceptual AI risk researchers, the goal post shifted. It shifted from "outer alignment" to "inner alignment." Demonstrations of safe behavior were deemed insufficient for safety, because an AI could just be pretending to be good, and then later turn on humanity. Deceptive alignment and treacherous turns from rogue AIs became a more popular concern, and the opacity of machine learning models made these scenarios more difficult to rule out. Just as we were the first to comprehensively demonstrate traction on "outer alignment" in our previous work, in this work we demonstrate traction on "inner alignment" and in particular deceptive alignment, as we can influence whether or not an AI lies. Although we have traction on "outer" and "inner" alignment, we should continue working on these directions. At the same time, we should increase our attention for risks that we do not have as much traction on, such as risks from competitive pressures and collective action problems (Hendrycks, 2023).