interviewees pointed out, it is a matter of judgment whether this is best interpreted as specification gaming or as a simple capability failure.[20]

The evidence is strong that AI systems will be subject to specification gaming to some degree. It remains unclear whether specification gaming will be sufficiently serious to pose an existential risk. In order to cause large-scale harms, misspecified goals would need to be subtle enough that systems were still deployed in high-stakes settings, but diverge extremely from intended goals in deployment. To date, no examples of specification gaming in AI systems have been catastrophic, so there is no direct evidence of this degree of harm from specification gaming.

There are some tentative signs that specification might become a more serious problem as models become more capable. In initial experiments, larger language models and language models with more RLHF are more prone to sycophantic answers, and to expressing a desire to seek power and avoid shutdown (Perez et al., 2022). Insofar as these behaviors are indeed caused by specification gaming,[21] this is cause for concern. Another study has found that when goals are misspecified, more capable RL agents will diverge more from intended goals than less capable agents, suggesting that specification gaming may worsen as capabilities improve. The same study also found that the divergence between intended and misspecified goals was sometimes very sudden, which might make it hard to anticipate and prevent such problems arising in deployment (Pan et al., 2022).

Overall, the evidence for specification gaming is strong, though it remains unclear whether the scale of the problem will be sufficient to pose an existential risk.

3.2.2 The evidence for goal misgeneralization

Another route to AI systems developing misaligned goals is goal misgeneralization, where systems develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.

"Goal misgeneralization is a specific form of robustness failure for learning algorithms in which the learned program competently pursues an undesired goal that leads to good performance in training situations but bad performance in novel test situations." (Shah et al., 2022b)

The underlying mechanism behind goal misgeneralization is distributional shift, where there are systematic differences between the training distribution and the test distribution. Distributional shift is a very widely documented phenomenon in AI systems (Leike et al., 2017; Quinonero-Candela et al., 2022), and out-of-distribution robustness remains unsolved (Hendrycks et al., 2021; Liu et al., 2023). This provides a reason to expect goal misgeneralization to arise.

However, the empirical evidence for goal misgeneralization is currently weak, in spite of the prevalence of distributional shift.

There are examples of goal misgeneralization in AI systems (DeGrave et al., 2021; Langosco et al., 2023; Shah et al., 2022b). However, these examples do not conclusively show that goal misgeneralization will arise in a harmful way.

Firstly, all of the examples of goal misgeneralization we have found take place in demonstration, rather than in deployed systems. Sometimes these demonstrations involve very obvious and crude differences between the training data and the test data. For instance, Langosco et al. (2023) train a CoinRun agent exclusively on mazes where the cheese is always in the upper right hand corner, and show in testing that the agent learns to navigate to the upper right rather than to the cheese. This shows that goal misgeneralization can occur when the training data is very different to the test data - but doesn't provide evidence for goal misgeneralization in more realistic settings. We have not found any evidence of real-world harm from goal misgeneralization so far.

Secondly, it is currently not possible to demonstrate conclusively that examples of goal misgeneralization actually involve systems learning a goal which is correlated in training but not deployment. It is

[20] "With some of the language model examples, I think you can ask the question, is this really specification gaming, or is it capability failure, or something like that? I think sometimes there's a bit of a judgment call there." [29:45] (AI Impacts, 2023c)

[21] That is, the systems are following the specified goal of generating text which receives high positive feedback from humans, but this comes apart from the goal of generating helpful, honest and harmless text. See also Krakovna (2020).