- **Goal misgeneralization**,[16] where some capable AI systems develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.

3.2.1 The evidence for specification gaming

One route to AI systems developing misaligned goals is specification gaming, where AI systems learn the goals which they are given, but these goals are misspecified and come apart from intended goals.

"Specification gaming is a behavior that satisfies the literal specification of an objective without achieving the intended outcome." (Krakovna et al., 2020) If sufficiently powerful AI systems were to be deployed in high-stakes settings, then the difference between the literal specification and the intended outcome could become extreme, leading to catastrophic outcomes (Bostrom, 2014; Pueyo, 2018).

Specification gaming is a well-established phenomenon, both in general and in the context of AI systems.

In non-AI contexts, there are numerous examples of variants of specification gaming,[17] in economics (Braganza, 2022; Chrystal and Mizen, 2003; Goodhart, 1984; Kelly and Snower, 2021; Lucas, 1976), education (Berliner and Nichols, 2005; Campbell, 1979; Elton, 2004; Fire and Guestrin, 2019; Koretz, 2008; Strathern, 1997; Stroebe, 2016), healthcare (O'Mahony, 2017; Poku, 2016) and other areas.[18] It is clear that at least in human and social systems, such dynamics are widespread.

In the context of AI systems, there are both theoretical demonstrations of specification gaming given certain model assumptions (Beale et al., 2020; Hennessy and Goodhart, 2023; Manheim and Garrabrant, 2019; Zhuang and Hadfield-Menell, 2021), and many empirical examples of specification gaming in AI systems, both in toy environments and in deployment (Krakovna et al., 2020).[19]

For example, OpenAI trained an agent to play the game CoastRunners. The agent was rewarded for hitting targets along the course of a boat race. But instead of racing to the finish line, the agent discovered a loophole where it could race in a circle, repeatedly crashing and setting itself on fire, to earn maximum points (Jack Clark, 2016).

While a majority of clear examples of specification gaming in AI systems arise in toy environments like CoastRunners (Krakovna, 2020), there are already some examples of deployed AI systems engaging in specification gaming, and of this behavior leading to harm, particularly in the areas of bias and misinformation.

For example, a healthcare screening system deployed in 2019 was trained to predict health care costs. As less is spent on Black patients' care because of unequal access to healthcare, the algorithm rated Black patients as less sick than White patients even where Black patients had more underlying chronic illnesses (Obermeyer et al., 2019).

Falsehoods generated by large language models can also be viewed as the result of specification gaming, though here the case is less clear. Language models trained to accurately predict the next token frequently generate false content (Collective, 2023a,b; Heaven, 2022), but as one of our

hacking (Amodei et al., 2016; Skalse et al., 2022), reward misspecification (Ngo et al., 2023), and Goodhart's law (Hennessy and Goodhart, 2023; Manheim and Garrabrant, 2019; Thomas and Uminsky, 2022).

[16] "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). Goal misgeneralization is related to goal drift (Hendrycks et al., 2023) and distributional shift (Amodei et al., 2016; Leike et al., 2017).

[17] For discussions about a cluster of related concepts including Goodhart's Law and proxy failure, see Amodei et al. (2016); John et al. (2023); Manheim and Garrabrant (2019); Thomas and Uminsky (2022).

[18] See Table 1 in John et al. (2023) for a collection of examples.

[19] The database linked to from this post contains over 70 examples of specification gaming. See also Hadshar (2023).

"One form of the problem has also been studied in the context of feedback loops in machine learning systems (particularly ad placement), based on counterfactual learning and contextual bandits. The proliferation of reward hacking instances across so many different domains suggests that reward hacking may be a deep and general problem, and one that we believe is likely to become more common as agents and environments increase in complexity." (Amodei et al., 2016). "Reward hacking—where RL agents exploit gaps in misspecified reward functions—has been widely observed" (Pan et al., 2022).