only possible to observe the behavior of the system in question, not its inner workings, so we cannot know what goal (if any) a system has learned. Examples to date only conclusively show behavioral or functional goal misgeneralization.[22]

Furthermore, it's often hard to distinguish goal misgeneralization from capability misgeneralization, where the system's capabilities also fail to generalize.[23] In the abstract, goal misgeneralization is distinct from capability misgeneralization: "a system's capabilities generalize but its goal does not generalize as desired. When this happens, the system competently pursues the wrong goal." (Shah et al., 2023) But in real-world settings, the wrong goal may often lead to capability failure. A system which learns to competently predict that tumors with rulers are malignant based on its training data will fail to competently predict actual malignancy when tested on more diverse data (Narla et al., 2018). Insofar as goal misgeneralization comes with capability misgeneralization, AI systems which learn very misgeneralized goals are unlikely to be deployed.

There are several possible explanations of the weakness of evidence on goal misgeneralization so far.

Goal misgeneralization might require a level of goal-directedness which current systems don't yet have,[24] or an ability to reason at higher levels of abstraction.[25] Reliably identifying goal misgeneralization might also require more advanced interpretability techniques.[26] Alternatively, the distinction between behavioral and 'actual' goal misgeneralization may be misplaced: if sufficiently capable systems engage in behaviors which look like goal misgeneralization, then functionally they are misaligned whether or not their internal representations match our description of goal misgeneralization.

So there are some reasons to expect the current evidence of goal misgeneralization to be weak, even if the phenomenon eventually arises strongly. Nevertheless, so far the evidence for goal misgeneralization remains reasonably speculative.[27]

3.3 The evidence for power-seeking

The presence of misaligned goals in and of itself need not pose an existential risk. But if AI systems with misaligned goals successfully and systematically seek power, the result could be existential.

In Carlsmith (2022), power-seeking is defined as "active efforts by an AI system to gain and maintain power in ways that designers didn't intend, arising from problems with that system's objectives."

[22] "I think right now the examples we have are more like behavioral goal misgeneralization where you just have different behaviors that are all the same in training but then they become decoupled in the new setting but we don't know how the behavior is going to generalize. We call it goal misgeneralization maybe more as a shorthand. The behavior has different ways of generalizing that are kind of coherent. We can present it as the system learned the wrong goal, but we can't actually say that it has learned a goal. Maybe it's just following the wrong heuristic or something. I think the current examples are a demonstration of the more obvious kind of effect where the training data doesn't distinguish between all the ways that the behavior could generalize." [37:11] (AI Impacts, 2023c)

[23] "I think it's a less well understood phenomenon... it can be hard to distinguish capability misgeneralization from goal misgeneralization." [33:16] (AI Impacts, 2023c)

[24] "Specifying something as goal misgeneralization also requires some assumption that the system is goal-directed to some degree and that can also be debatable." [33:16] (AI Impacts, 2023c)

[25] "The story of you train an AI to fetch a coffee and then it realizes that the only way it can do that is to take over the world is a story about misgeneralization. And it's happening at a very high level of abstraction. You're using this incredibly intelligent system which is reasoning at a very high level about things and it's making the error at that high level... And I think the state of the evidence is... we've never observed a misgeneralization failure at such a high level of abstraction, but that's what we would expect because we don't have AIs that can even reason at that kind of level of abstraction." [28:36] (AI Impacts, 2023b)

[26] "The mechanism is a lot less well understood. I think to really properly diagnose goal misgeneralization we would need better interpretability tools." [36:30] (AI Impacts, 2023c)

[27] "I think [the evidence for goal misgeneralization] is not as strong [as for specification gaming]." [33:16] (AI Impacts, 2023c) "These generalization failures at new levels of abstraction are notoriously hard to predict. You have to try and intuit what an extremely large scale neural net will learn from the training data and in which ways it will generalize... I'm relatively persuaded that misgeneralization will continue to happen at higher levels of abstraction, but whether that actually is well described by some of the typical power-seeking stories I'm much less confident and it's definitely going to be a judgment call." [28:36] (AI Impacts, 2023b)