8 Appendix B: Some evidence for other claims about existential risk from AI

We systematically reviewed the evidence for claims about misalignment and power-seeking. However, in the course of our research and interviews, we came across some evidence for other relevant claims.

This appendix contains some of the evidence for goal-directedness, situational awareness, and deceptive alignment. It should not be treated as a comprehensive review of the state of the evidence on these topics.

8.1 Some evidence for goal-directedness

Roughly, goal-directedness refers to a property of AI systems to persistently pursue a goal. [34] Goal-directedness has not been well-defined so far, and so reviewing the evidence for goal-directedness is hampered by unclarity about the concept. [35]

That said, it seems plausible that goal-directedness is a direct precondition for goal misgeneralization and for power-seeking,[36] so it is an important claim to assess.

Coherence theorems offer one kind of conceptual evidence for goal-directedness, but the extent to which they apply to future AI systems is contested (Bales, 2023; EJT, 2023; AI Impacts, 2021).[37]

There is limited empirical evidence of goal-directedness in systems so far.[38] One of the researchers we interviewed noted that language models may be particularly unsuited to goal-directedness.[39]

However, individual researchers we interviewed believe that:

- To the extent that language models can simulate humans, they will have the ability to simulate goal-directedness.[40]
- There is a clear trend towards systems acting more autonomously.[41]

[34] In Carlsmith (2022), goal-directedness is referred to as "agentic planning", where AI systems "make and execute plans, in pursuit of objectives, on the basis of models of the world."

[35] "Right now it's really hard to distinguish between real goal-directedness and learned heuristics… I think part of the problem with goal-directedness is we don't really understand the phenomenon that well." [44:00] (AI Impacts, 2023c)

[36] "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) "What I'm expecting is happening here is that current systems are not goal-directed enough to show real power-seeking. And so the power-seeking threat model becomes more reliant on these kind of extrapolations of when there are systems which are more capable, they'll probably be at least somewhat more goal-directed and then once we have goal-directedness, we can more convincingly argue that power-seeking is going to be a thing because we have theory and so on, but there's a lot of uncertainty about it because we don't know how much systems will become more goal-directed." [54:35] (AI Impacts, 2023c)

[37] "Some of the theoretical arguments make the case that goal-directedness is an attractor. I think that's something that's more debatable, less clear to me. There have been various discussions on LessWrong and elsewhere about to what extent do coherence arguments imply goal-directedness. And I think the jury is still out on that one." [42:36] (AI Impacts, 2023c)

[38] "I think the evidence so far at least for language models, there isn't really convincing evidence of goal-directedness." [44:00] (AI Impacts, 2023c)

[39] "It's also possible goal-directedness is kind of hard. And especially, maybe language models are just a kind of system where goal-directedness comes less naturally than other systems like reinforcement learning systems or even with humans or whatever." [40:26] (AI Impacts, 2023c)

[40] "I think generally the kind of risk scenarios that we are most worried about would involve the system acting intentionally and deliberately towards some objectives but I would expect that intent and goal-directedness comes in degrees and if we see examples of increasing degrees of that then I think that does constitute evidence of that being possible. Although it's not clear whether it will go all the way to really deliberate systems, but I think especially to the extent that these systems can simulate humans… they have the ability to simulate deliberate intentional action and planning because that's something that humans can do." [20:20] (AI Impacts, 2023c)

[41] "We are already capable of getting AI systems to do simple things relatively autonomously. I don't think it's a threshold where now it's autonomous, now it's not… I think it's a spectrum and it's just very clearly ramping up. We already have things that have a little autonomy but not very much. I think it's just a pretty straightforward trend at this point." [24:39] (AI Impacts, 2023b)