One researcher we interviewed highlighted goal-directedness as one of their key uncertainties about existential risk from AI.[42]

8.2 Some evidence for situational awareness

"A model is situationally aware if it's aware that it's a model and can recognize whether it's currently in testing or deployment." (Berglund et al., 2023)

This is important to arguments about existential risk from AI as situational awareness is plausibly a precondition for successful misaligned power-seeking: a model may need to understand its own situation at a sophisticated level in order to make plans which successfully disempower humans. In particular, situational awareness seems like a precondition for deceptive alignment.

There is some empirical work demonstrating situational awareness in large language models, but the results are inconclusive (Berglund et al., 2023; Ngo et al., 2023; Perez et al., 2022). Berglund et al. (2023) find that language models can perform out-of-context reasoning tasks, but only with particular training set ups and data augmentation. Perez et al. (2022) run various experiments to test awareness, and find that "the models we evaluate are not aware of at least some basic details regarding themselves or their training procedures." On the other hand, Langosco et al. (2023) use the same questions as Perez et al. (2022) but find that their model answers 85% accurately.

[42] "I think we might see more goal-directed systems which produce clearer examples of internal goal misgeneralization, but also I wouldn't be that surprised if we don't see that. I think that's one of the big uncertainties I have about level of risk. How much can we expect goal-directedness to emerge?" [40:26] (AI Impacts, 2023c)