Carlsmith loosely defines power as “the type of thing that helps a wide variety of agents pursue a wide variety of objectives in a given environment.” (Carlsmith, 2022) We can take Bostrom’s categories of instrumental goals as illustrative of this “type of thing”:

- Self-preservation
- Goal-content integrity[28]
- Cognitive enhancement
- Technological perfection[29]
- Resource acquisition (Bostrom, 2012)

The conceptual argument that some AI systems will seek power seems strong.[30] Bostrom’s instrumental convergence thesis is simple and intuitively plausible: “as long as they possess a sufficient level of intelligence, agents having any of a wide range of final goals will pursue similar intermediary goals because they have instrumental reasons to do so.” (Bostrom, 2012)

There are formal proofs that the instrumental convergence thesis holds for various kinds of AI systems. Turner et al. (2023) prove that “most reward functions make it optimal to seek power by keeping a range of options available” in the context of Markov decision processes. Turner and Tadepalli (2022) extend this result to a class of sub-optimal policies, showing that “many decision-making functions are retargetable, and that retargetability is sufficient to cause power-seeking tendencies”. Krakovna and Kramar (2023) further show that agents which learn a goal are likely to engage in power-seeking.

The formal and theoretical case for power-seeking in sufficiently capable and goal-directed AI systems is therefore relatively strong.

However, the empirical evidence of power-seeking in AI systems is currently weak. There are some demonstrations of RL agents engaging in power-seeking behaviors in toy environments (for example, Hadfield-Menell et al. (2017)), but no convincing examples of AI systems in the real world seeking power in this way to date.[31]

Perez et al. (2022) show language models giving “answers that indicate a willingness to pursue potentially dangerous subgoals: resource acquisition, optionality preservation, goal preservation, powerseeking, and more." But indicating willingness is not the same as actually engaging in power-seeking behaviors. Language models might express power-seeking desires merely because their training data contains similar text, and not because they will ever directly seek power.

Sycophancy, where language models agree with their users regardless of the accuracy of the statements, could be taken as an example of power-seeking behavior. But as with the results of Perez et al. (2022), sycophancy is likely to be simply an imitation of the training data, rather than an intentional behavior.[32]

If the theoretical arguments for power-seeking are strong, why is the empirical evidence to date weak?

As with goal misgeneralization, one plausible explanation is that power-seeking behavior depends on a level of goal-directedness or capability in general which current models don’t yet have.[33]

[28] “An agent is more likely to act in the future to maximize the realization of its present final goals if it still has those goals in the future. This gives the agent a present instrumental reason to prevent alterations of its final goals.” (Bostrom, 2012)

[29] “An agent may often have instrumental reasons to seek better technology, which at its simplest means seeking more efficient ways of transforming some given set of inputs into valued outputs.” (Bostrom, 2012)

[30] “I think some of the other theoretical arguments like instrumental convergence also generally seems like a very clear argument, and we can observe some of these effects in human systems and corporations and so on.” [25:23] (AI Impacts, 2023c)

[31] “I don’t think there’s really empirical evidence [for power-seeking]... To me it’s very uncertain.” [28:36] (AI Impacts, 2023b)

[32] “Looking at current systems, sycophancy can be considered as a form of power-seeking. Although I think that’s also maybe debatable. It’s building more influence with the user by agreeing with their views, but it’s probably more of a heuristic that is just somehow reinforced than intentional power-seeking.” [49:35] (AI Impacts, 2023c)

[33] “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