7 Appendix A: Carlsmith's argument for existential risk via power-seeking AI

The following table maps between the premises of Carlsmith (2022)'s argument, and the claims used in this report (see Table 1). Claims within the scope of this report are bolded.

Note that the claims used in this report are not identical to Carlsmith's premises, though they are closely related.

Carlsmith		**Claims used in this report**
By 2070:		(*Preconditions: Timelines*) The relevant AI systems will be developed in the not-too-distant future.
1. It will become possible and financially feasible to build AI systems with the following properties:	*Advanced capability*: they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).	(*Preconditions: Capabilities*) Some AI systems will be highly capable, in the sense that they are able to perform many important tasks at or above human level
	Agentic planning: they make and execute plans, in pursuit of objectives, on the basis of models of the world.	(*Preconditions: Goal-directedness*) Some AI systems will be goal-directed, in that they pursue goals consistently over long time periods.
	Strategic awareness: the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.	(*Preconditions: Situational awareness*) Some AI systems will be aware that they are AI systems, and whether they are in training or deployment.
	(Call these "APS"—Advanced, Planning, Strategically aware—systems.)	
2. There will be strong incentives to build and deploy APS systems.		
3. It will be much harder to build APS systems that would not seek to gain and maintain power in unintended ways (because of problems with their objectives) on any of the inputs they'd encounter if deployed, than to build APS systems that would do this, but which are at least superficially attractive to deploy anyway.		(*Misalignment*) Some capable AI systems will develop goals which are misaligned with human goals. (*Misalignment: Specification gaming*) Some capable AI systems will learn designer-specified goals which diverge from intended goals in unforeseen ways. (*Misalignment: Goal misgeneralization*) Some capable AI systems will develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.
4. Some deployed APS systems will be exposed to inputs where they seek power in unintended and high-impact ways (say, collectively causing >$1 trillion dollars of damage), because of problems with their objectives.		(*Power-seeking*) Some capable, misaligned AI systems will seek power in order to achieve their goals.