Existential Risk from Power-Seeking AI

Joe Carlsmith

March 2023

(Forthcoming in "Essays on Longtermism," from Oxford University Press, edited by Jacob Barrett, Hilary Greaves, and David Thorstad. This is a shortened version of my 2021 report, "Is power-seeking AI an existential risk?". Human-narrated audio for this essay is here, and for the longer report here; or search "Joe Carlsmith Audio" in your podcast app. There's also an even shorter video presentation here.)

Abstract. This essay formulates and examines what I see as the core argument for concern about existential risk from misaligned artificial intelligence. I begin by discussing a backdrop picture that informs such concern. On this picture, intelligent agency is an extremely powerful force, and creating agents much more intelligent than us is playing with fire—especially given that if their objectives are problematic, such agents would plausibly have instrumental incentives to seek power over humans. I then look more closely at the type of agents we should be worried about; the incentives to create and deploy them; the difficulty of ensuring that they don't seek power in unintended ways; the reasons to expect them to end up deployed regardless; the likelihood that the problem scales to the permanent disempowerment of our species; and the value lost if so. My current view, in light of these considerations, is that the existential risk at stake is disturbingly high (i.e., greater than 10% by 2070).

Contents

1 Introduction

Some worry that advanced artificial intelligence will pose an existential risk to humanity.[1] In this article, I formulate and examine what I see as the core argument for such a concern. In brief, and put roughly, the argument is that by 2070:[2]

1. It will become possible and financially feasible to build relevantly powerful and agentic AI systems.[3]

2. There will be strong incentives to do so, conditional on (1).

3. It will be much harder to build aligned (and relevantly powerful/agentic) AI systems than to build misaligned (and relevantly powerful/agentic) AI systems that are still superficially attractive to deploy, conditional on (1) and (2).

4. Some such misaligned systems will seek power over humans in high-impact ways, conditional on (1)-(3).

5. This problem will scale to the full disempowerment of humanity, conditional on (1)-(4).

6. Such disempowerment will constitute an existential catastrophe, conditional on (1)-(5).

These claims are extremely important if true. My aim is to investigate them.[4]

My current view is that there is a disturbingly substantive chance (i.e., greater than 10%) that all of these claims are true, and that many people alive today—including myself—live to see humanity permanently disempowered by AI systems we've lost control over.[5] That is: I view this as

[1] For classic arguments and other resources, see e.g. Yudkowsky (2008), Bostrom (2014), Hawking (2014), Christiano (2019), Russell (2019), Ord (2020), Ngo (2020), Karnofsky (2021, 2022), Ngo et al. (2023), and Cotra (2021). By "existential risk," I mean a risk that threatens to destroy humanity's longterm potential (here I am following Ord (2020: 27)).

[2] I'm focusing on 2070 because I want to keep vividly in mind that I and many other readers (and/or their children) should expect to live to see the claims at stake here falsified or confirmed. That said, the main arguments don't actually require the development of relevant systems within any particular period of time (though timelines in this respect can matter to e.g. the amount of evidence that present-day systems and conditions provide about future risks).

[3] I define various of the terms in this argument—e.g., "relevantly powerful and agentic," "misaligned"—more precisely below.

[4] For somewhat complicated reasons, the sections below do not correspond perfectly to the argument's premises.

[5] In a longer report on which this article is based (see Carlsmith (2022)), I try to get more quantitative purchase on the level of risk, by assigning probabilities to the various premises in the argument, but I won't do so here. A striking fraction of experts in the

a problem of grave importance. My hope, here, is to facilitate productive debate about it.

2 Backdrop

The specific arguments I'll discuss emerge from a broader backdrop picture, which I'll gloss as:

1. Intelligent agency is an extremely powerful force for controlling and transforming the world.

2. Building agents much more intelligent than humans is playing with fire.

I'll start by briefly describing this picture, as it sets an important stage for the discussion that follows.

Of all the species that have lived on the earth, humans are clearly strange. In particular, we exert an unprecedented scale and sophistication of intentional control over our environment. Consider, for example, the city of Tokyo, or the Large Hadron Collider, or a large coal mine.

What makes this possible? Something about our minds seems centrally important. We can plan, learn, communicate, deduce, remember, explain, imagine, experiment, and cooperate in ways that other species can't. These cognitive abilities—employed in the context of the culture and technology we inherit and create—give us the power, collectively, to transform the world. Let's call this loose cluster of abilities "intelligence," though very little will rest on the term.

And humans aren't just smart: we're also *agentic*. That is (loosely): we pursue objectives, guided by models of the world. We have cities, particle accelerators, and coal mines because we were *trying* to build them.

It seems possible, in principle, to build agentic cognitive systems—both biological and artificial—whose intelligence significantly exceeds our own. And in the context of artificial agents, the differences between brains and computers—in possible speed, size, available energy, memory capacity, component reliability, input/output bandwidth, and so forth—make the eventual possibility of very dramatic differences in ability especially salient.

But the choice to build such superhumanly-intelligent artificial agents should be approached with extreme caution.[6] As humanity's impact on the earth

field, however, seem likely to agree that the risk is at least 10%: in Stein-Perlman et al's (2022) survey of more than 700 AI researchers who had recently published at NeurIPS or ICML (major machine learning conferences), 48% of respondents gave at least 10% chance that the long-run effect of AI will be "extremely bad (e.g. human extinction)." The median respondent said 5%.

[6]Some articulate this view by appeal to the dominant position of humans on this planet, relative to other species (see e.g. Bostrom (2015), Russell (2019: Chapter 5) on the

illustrates, intelligent agency is a force of formidable potency. If we unleash much more of this force into the world, via new, more intelligent forms of non-human agency, it seems reasonable to expect dramatic impacts, and reasonable to wonder how well we will be able to control the results.

I'll focus on a particular version of this worry, centered on the following hypothesis: that by default, suitably strategic and intelligent agents, engaging in suitable types of planning, will have instrumental incentives to gain and maintain various types of power (call this "power-seeking"), since this power will help them pursue their objectives more effectively. The worry is that if we create and lose control of such agents, the result won't just be *damage* of the type that occurs when a plane crashes, or a nuclear plant melts down—damage which remains passive, for all its costs. Rather, the result will be highly capable, non-human agents actively working to gain and maintain power over their environment—agents in an *adversarial* relationship with humans who don't want them to succeed.

Nuclear contamination is hard to clean up, and hard to stop from spreading. But it isn't *trying* to spread—and certainly not with greater intelligence than the humans trying to contain it. But the power-seeking agents just described *would* be trying, in sophisticated ways, to undermine our efforts to stop them. If such agents are sufficiently capable, and/or if sufficiently many of such failures occur, humans could end up permanently disempowered.

In principle, then, sophisticated AI agents with problematic goals could represent an unprecedented threat to the human species. Let's look more closely at whether to expect this threat to arise in practice.

3 APS systems

This section discusses in more detail the type of AI systems I'm worried about, and the timelines to their development.

"Gorilla Problem," Ord (2020), and Ngo (2020) on the "second species argument"). For example: some argue that the fate of the chimpanzees is currently in human hands, and that this difference in power is primarily attributable to differences in intelligence, rather than e.g. physical strength. Just as chimpanzees—given the choice and power—should be careful about building humans, then, we should be careful about building agents more intelligent than us. This argument is suggestive, but far from airtight. Chimpanzees, for example, are themselves much more intelligent than mice, but the "fate of the mice" was never "in the hands" of the chimpanzees. What's more, the control that humans can exert over the fate of other species on this planet still has limits, and we can debate whether "intelligence," even in the context of accumulating culture and technology, is the best way of explaining what control we have. More importantly, though: humans arose through an evolutionary process that chimpanzees did nothing to intentionally steer. Humans, though, will be able to control many aspects of processes we use to build and empower new intelligent agents.

I'll focus on AI systems with three key properties.

1. *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).[7]

The aim here is to hone in on systems whose capabilities make any power-seeking behavior they engage in worth taking seriously (in aggregate) as a potential route to the disempowerment of roughly all humans. Such a condition does not, I think, require meeting various stronger conditions sometimes discussed[8]—for example, "human-level AI,"[9] "superintelligence,"[10] or "AGI."[11] That said, I'm erring, here, on the side of including "weaker" systems — including some that might not, on their own (or even in aggregate), be all that threatening.[12]

2. *Agentic planning:* they make and execute plans, in pursuit of objectives, on the basis of models of the world.

The aim here (and with the next property, "Strategic awareness") is to hone

[7] An AI system with these capabilities can consist of many smaller systems interacting, but it should suffice to ~fully automate the tasks in question.

[8] A level of AI progress that disempowered all humans would constitute "transformative AI" in the sense used by Karnofsky (2016) e.g. "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution." But *that* sort of transformation is precisely the type of thing we're trying to forecast; e.g., it's the result, not the cause. And disempowerment does not require other, more mechanistic standards for transformation—e.g. economic growth proceeding at particular rates (though we can argue, here, about the economic value that AI progress sufficient to disempower humans would represent).

[9] This is used in various ways, to mean something like (a) a single AI system that is in some generic sense "as intelligent" as a human; (b) a single AI system that can do anything that a given human (an average human? the "best" human? any human?) can do; (c) a level of automation such that unaided machines can perform roughly any task better and more cheaply than human workers (see Grace et al. (2018)). My favorite is (c).

[10] Bostrom (2014, Chapter 2) defines this as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." A related concept requires cognitive performance that in some loose sense exceeds all of human civilization—though exactly how to understand this isn't quite clear (and human civilization's "cognitive abilities" change over time).

[11] "AGI" is sometimes used as a substitute for some concept of "human-level AI"; in other contexts, it refers specifically to some concept of human *learning* ability (see e.g. Selsam (2018), and Arbital here (no author listed, but I believe it is Yudkowsky)), or some method of creating systems that can perform certain tasks (see Ngo (2020) on "generalization-based approaches"). "AGI" is often contrasted with "narrow AI"—though in my opinion, this contrast too easily runs together a system's ability to *learn* tasks with its ability to *perform* them. And sometimes, a given use of "AGI" just means something like "you know, the big AI thing; *real* AI; the special sauce; the thing everyone else is talking about."

[12] Put another way, I'm erring on the side of "necessary" rather than "sufficient" for riskiness. I have yet to hear a good capability threshold that is both necessary *and* sufficient—and I'm skeptical that one exists.

in on the type of goal-oriented cognition required for arguments about the instrumental value of gaining/maintaining power to be relevant to a system's behavior.[13] That is, in order to have instrumental incentives to seek power, a system needs to have objectives that power-seeking promotes, and its behavior needs to be sensitive to the incentives those objectives create.

We can argue about exactly what is required for concepts like "planning," "pursuing objectives," and "using models of the world" to apply—and indeed, muddiness around abstractions in this vicinity seem to me a key way that thinking on this topic (including my own) might go astray.[14] I take it, though, that humans do these things, and that AI systems can, too. I'm talking about the ones that do (or at least, that do something close enough to justify predicting their behavior on this basis).

3. *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.

The aim here is to hone in on the type of world-modeling required to notice and respond to incentives to seek power, where they exist. Clearly, this capability comes in degrees. But broadly and loosely, we can think of a strategically aware, planning agent as possessing models of the world that would allow it to answer questions like "what would happen if I had access to more computing power" and "what would happen if I tried to stop humans from turning me off" about as well as humans can (and using those same models in generating plans).

Let's call a system with all three of these properties an Advanced, Planning, Strategically-aware system (or "APS system").

Will it become possible and financially feasible to build APS systems before 2070? I think that this is more likely than not.[15] However, I won't attempt to examine the issue here.

Perhaps the right probability here is lower. But I doubt that it is far lower. Less than 10%, for example, seems to me unreasonable (and I don't think that the difficulty of forecasts like these licenses assuming that the probability here is very low, or treating it that way implicitly).

[13] We can imagine cases in which AI agents end up valuing power for its own sake, but I'm not going to focus on those here.

[14] See the longer report for more discussion.

[15] This view emerges partly from a series of investigations at Open Philanthropy into AI timelines, summarized in Karnofsky (2021) partly from public forecasts; partly from eye-balling recent progress in deep learning myself; and partly from deference to some of Open Philanthropy's technical advisors.

4 Incentives

Let's assume, then, that it will become possible and financially feasible to develop APS systems. Should we expect relevant actors to do so, especially on a widespread scale?

It seems likely that there will be strong economic and political incentives to automate advanced capabilities. But building strategically aware agentic planners might not be the only way to do this. After all, many tasks that one might wish to automate—translating languages, classifying proteins, predicting human responses, etc.—don't seem to require agentic planning or strategic awareness (at least at current levels of performance).

Nevertheless, I think, there are strong reasons to expect that AI progress will push in the direction of APS systems. I will focus on three.

The first and strongest reason is that agentic planning and strategic awareness seem quite *useful.* Agentic planning is a very powerful and general way of interacting with an environment—especially a complex and novel environment that doesn't afford much room for trial and error—in order to produce favored outcomes. Many tasks humans care about (creating and selling profitable products, designing and performing successful scientific experiments, achieving social and political goals) have this structure. So do many of the sub-tasks involved in those tasks (e.g., efficiently gathering and synthesizing relevant information, communicating with stakeholders, managing resource-allocation, etc). And even when agentic planning isn't strictly required to achieve a valuable outcome, it will often help. So if our AI systems can't engage in agentic planning, then the scope of what they can do seems, naively, like it will be severely restricted.

So, too, with strategic awareness. Strategic awareness is closely connected to a basic capacity to "understand what is going on," interact with other agents (including human agents) in the real world, and recognize available routes to achieving your objectives—all of which seem very useful to performing tasks of the type just described. Indeed, to the extent that humans care about the strategic pursuit of e.g. business, military, and political objectives, and want to use AI systems in these domains, it seems like there will be incentives to create AI systems with the types of world models necessary for very sophisticated and complex types of strategic planning—including planning that involves recognizing and using available levers of real-world power.

Here is a second reason to expect APS systems. Even if some task doesn't *require* agentic planning or strategic awareness, it may be that creating APS systems is the only route, or the most efficient route, to automating that task, given available techniques. For example: perhaps the best way to automate a wide range of tasks is to create AI systems that can learn new

tasks with very little data.[16] And we can imagine scenarios in which the best way to do *that* is by training agentic planners with high-level, broad-scope pictures of "how the world works," and then fine tuning them on specific tasks.[17]

Finally, even if you're not explicitly aiming for or anticipating agentic planning and strategic awareness in an AI system, these properties could in principle arise in unexpected ways regardless, and/or prove difficult to prevent. For example, optimizing a system to perform some not-intuitively-agential task (for example, predicting strings of text) could, given sufficient cognitive sophistication, result in internal computation that makes and executes plans, in pursuit of objectives, on the basis of broad and informed models of the real world. Indeed, the likelihood of this seems correlated with the strength of the "usefulness" consideration discussed above: insofar as agentic planning and broad-scope world-modeling are very useful types of cognition, we might expect to see them cropping up a lot in sufficiently optimized systems, whether we want them to or not.

Of these three reasons to expect APS systems—their usefulness, the pressures exerted by available techniques, and the possibility that they arise as byproducts of sophistication—I place the most weight on the first.

5 Alignment

Let's assume that it will become possible and financially feasible to create APS systems before 2070, and that there will be significant incentives to do so. This section argues that it will be difficult to create APS systems that don't seek to gain and maintain power in unintended ways.

5.1 Some definitions

I'll define "misaligned behavior," in an AI system, as unintended behavior that arises in virtue of problems with an AI system's objectives.[18] A characteristic feature of misaligned behavior is that it is *unintended* but still

[16] This may be especially true of tasks where we lack lots of training data, which could be the majority of useful tasks.

[17] Richard Ngo suggested this point in conversation; see also his discussion of the "generalization-based approach" in his (2020). The training and fine-tuning used for GPT-3 may also be suggestive of patterns of this type.

[18] There are ambiguities about what sort of behavior counts as "intended" by designers (in particular, the relationship between "intended" and "foreseen" is unclear), but I'm going to leave the notion vague for now, and assume that behaviors like lying, stealing money, resisting shut-down by appropriate channels, harming humans, and so forth are generally "unintended. Similarly: I don't, at present, have a rigorous account of how to attribute unintended behavior to problems with objectives vs. other problems; and I doubt the distinction will always be deep or easily drawn (this doesn't make it useless). But I'll lean on the intuitive notion for now.

competent. That is, it looks less like an AI system breaking or failing in its efforts to do what designers want, and more like an AI system trying, and perhaps succeeding, to do something designers *don't* want it to do.[19]

Not all misaligned AI behavior seems relevant to existential risk. Consider, for example, an APS AI system in charge of an electrical grid, whose designers intend it to send electricity to both town A and town B, but whose objectives have problems that cause it, during particular sorts of storms, to only send electricity to town A. This is misaligned behavior, and it may be quite harmful, but it poses no threat to the entire future.

Rather, the type of misaligned AI behavior that I think creates the most existential risk involves misaligned *power-seeking* in particular: that is, 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 (I'll call this sort of behavior "PS-misaligned").[20] AI systems that don't seek to gain or maintain power may cause a lot of harm, but this harm is more easily limited by the power they already have. And such systems, by hypothesis, won't try to maintain that power if/when humans try to stop them. Hence, it's much harder to see why humans would fail to notice, contain, and correct the problem before it reaches an existential scale.[21]

I'll say that a system is "fully aligned" if it doesn't engage in misaligned behavior on any inputs compatible with the basic physical conditions of our universe (I'll call these "physics-compatible inputs"),[22] and "practically

[19]In this sense, it's less like a nuclear plant melting down, and more like a heat-seeking missile pursuing the wrong target; less like an employee giving a bad presentation, and more like an employee stealing money from the company.

[20]In the electrical grid case, the AI system hasn't been described as trying to gain power (for example, by trying to hack into more computing resources to better calculate how to get electricity to town A) or to maintain the power it already has (for example, by resisting human efforts to remove its influence over the grid). And in this sense, I think, it's much less dangerous.

[21]That said, note that not all misaligned power-seeking, even in APS systems, is particularly harmful, or intuitively worrying from an existential risk perspective. Ben Garfinkel suggests the example of a robo-cop pinning down the wrong person, but remaining amenable to human instruction otherwise. And in general, it seems most dangerous if an APS system is *using* its advanced capabilities in pursuing power; e.g., if I'm a great hacker, but a poor writer, and I'm trying to get power via my journalism, I'm less threatening.

[22]Thus, it is physics-compatible for a randomly chosen bridge in Indiana to get hit, on a randomly chosen millisecond in May 2050, by a nuclear bomb, a 10 km-diameter asteroid, and a lightning bolt all at once; but not for the laws of physics to change, or for us all to be instantaneously transported to a galaxy far away. Obviously the scope here is very broad: but note that misaligned behavior is a different standard than "bad" or even "catastrophic" behavior. It will always be possible to set up physics-compatible inputs where a system makes a mistake, or gets deceived, or acts in a way that results in catastrophic outcomes. To be misaligned, though, this behavior needs to arise from problems with the system's objectives in particular. Thus, for example, if Bob is a paper-clip maximizer, and he builds Fred, who is also a paper-clip maximizer, Fred will (on my definition) be

aligned" if it doesn't engage in misaligned behavior on any of the inputs it will in fact receive.[23] And I'll say the system is "fully *PS*-aligned" if it doesn't engage in misaligned *power-seeking* on any physics-compatible inputs, and "practically *PS*-aligned" if it doesn't engage in misaligned *power-seeking* on any of the inputs it will in fact receive.[24] For our purposes, it is this last property—practical PS-alignment—that matters most.

5.2 Power-seeking

Why might we think that it will be hard to prevent misaligned power-seeking? A key hypothesis, some variant of which underlies much of the discourse about existential risk from AI, is that there is a close connection (in sufficiently advanced agents) between misaligned behavior in general and misaligned power-seeking in particular.[25] I'll formulate this hypothesis as follows:

> Instrumental Convergence: If an APS AI system is less-than-fully aligned, and some of its misaligned behavior involves strategically-aware agentic planning in pursuit of problematic objectives, then in general and by default, we should expect it to be less-than-fully PS-aligned, too.[26]

fully-aligned with Bob as long as Fred keeps trying to maximize paperclips on all physics compatible-inputs (even though some of those inputs are such that trying to maximize paperclips actually minimizes them, kills Bob, etc). Thanks to Eliezer Yudkowsky, Rohin Shah, and Evan Hubinger for comments on the relevant scope here (which isn't to say they endorse my choice of definition).

[23]By "inputs," I mean information the system receives via the channels intended by its designers (an input to GPT-3, for example, would be a text prompt). I am not including processes that intervene in some other way on the internal state of the system—for example, by directly changing the weights in a neural network (analogy: a soldier's loyalty need not withstand arbitrary types of brain surgery).

[24]That is, and importantly, for a system to be practically PS-aligned, it doesn't need to be the case that it would never, in any circumstances (or with any level of capability), engage in problematic power-seeking. This is in contrast with some other strands of the literature. See e.g. Yudkowsky on the "omnipotence test for AI safety": "The Omni Test is that an advanced AI should be expected to remain aligned, or not lead to catastrophic outcomes, or fail safely, even if it suddenly knows all facts and can directly ordain any possible outcome as an immediate choice. The policy proposal is that, among agents meant to act in the rich real world, any predicted behavior where the agent might act destructively if given unlimited power (rather than e.g. pausing for a safe user query) should be treated as a bug." (See Yudkowsky in "Querying the AGI user", "AI safety mindset".) Talk of an "objective" such that the "optimal policy" on that objective leads to good outcomes is also reminiscent of something like the Omni Test. See e.g. Hubinger's (2020) definition of "intent alignment": "An agent is intent aligned if the optimal policy for its behavioral objective is aligned with humans." Also see Christiano (2018) and Hubinger et al. (2019: 35).

[25]See e.g. Bostrom (2014: 127-40); and Russell (2019: 132-45).

[26]Note that instrumental convergence is not a conceptual claim, but rather an empirical claim that purports to apply to a wide variety of APS systems. In principle, for example, we can imagine APS systems that plan in pursuit of problematic objectives on some inputs, but which are nevertheless fully PS-aligned (or very close to it). Consider, for example, an APS version of the electrical grid AI system above, which plans strategically in pursuit

Why believe in Instrumental Convergence? The basic reason is that power, almost by definition, is extremely useful to accomplishing objectives. So to the extent that an agent is engaging in unintended behavior in pursuit of problematic objectives, it will generally have incentives, other things equal, to gain and maintain forms of power in the process—incentives that strategically-aware agentic planning puts it in a position to recognize and respond to.[27]

What sorts of power might a system seek? Bostrom (2014) identifies a number of "convergent instrumental goals," each of which promotes an agent's power to achieve its objectives.[28] These include:

- self-preservation (since an agent's ongoing existence tends to promote the realization of its objectives);
- preventing changes to its objectives (since agent's pursuit of those objectives in particular tends to promote them);
- improving its cognitive capability (since such capability tends to increase an agent's success in pursuing its objectives);
- technological development (since control over more powerful technology tends to be useful);
- resource-acquisition (since more resources tend to be useful, too).

We've already seen examples of rudimentary AI systems "discovering" the usefulness of resource acquisition, for example. When OpenAI trained two teams of AIs to play hide and seek in a simulated environment that included blocks and ramps that the AIs could move around and fix in place, the AIs learned strategies that depended crucially on acquiring control of the blocks and ramps in question— despite the fact that they were not given any direct incentives to interact with those objects (the hiders were simply rewarded for avoiding being seen by the seekers; the seekers, for seeing the hiders).[29]

---

of directing electricity only to town A, but which just doesn't consider plans that involve seeking power. That said, the in-principle possibility of strategic, agentic misalignment without PS-misalignment is important, though, since it might be realized in practice. Perhaps, for example, the type of training we should expect by default will reinforce cognitive habits in APS systems that steer away from searching over/evaluating plans that involve misaligned power-seeking, even if other types of misaligned behavior persist.

[27] One way to bring this out is to conceptualize power in terms of the number of options an agent has available to it. Thus, if a policy seeks to promote some outcomes over others, then other things equal, a larger number of options makes it more likely that a more preferred outcome is accessible. Indeed, talking about "options-seeking," instead of "power-seeking," might have less misleading connotations.

[28] Bostrom is following Omohundro (2008). Here I am thinking of the pursuit of any of these instrumental goals as a part of "power-seeking" in the relevant sense.

[29] Thus, the hiders learned to move and lock blocks to prevent the seekers from entering the room where the hiders were hiding; the seekers, in response, learned to move a ramp to give them access anyway; adjusting for this, the hiders learned to take control of that

Of course, this is a very simple, simulated environment, and the level of agentic planning it makes sense to ascribe to these rudimentary AIs isn't clear.[30] But the basic dynamic here applies in the real world, as well, and it applies to more advanced systems. Things like money, compute, energy, and social influence will often help agents achieve their objectives. We should expect a sufficiently sophisticated, strategically aware AI agent to register this fact, and make its decisions accordingly. Thus a sophisticated AI system might try to hack into financial databases; take control of compute; manipulate human opinion; and so on.[31]

One objection to instrumental convergence is that many humans don't seem particularly "power-seeking," despite their agentic planning and strategic awareness.[32] And it's true that many humans do not seek various types of power *in their current circumstances* — in which (for example) their capabilities are roughly similar to those of their peers, they are subject to various social and legal incentives, they are hemmed in by significant physical and temporal constraints, and they recognize certain intrinsically important ethical constraints. But almost all humans will seek to gain and maintain various types of power in *some* circumstances, especially when they can do so at little cost. Thus, for most humans, it doesn't make sense to try to start a billion dollar company—the expected returns on such effort are too low. But most humans will walk across the street to pick up a billion dollar check. More generally, the power-seeking behavior hu-

ramp before the seekers can get to it, and to lock it in the room as well. In another environment, the seekers learned to "surf" on boxes, and the hiders, to prevent this, learned to lock *all* boxes and ramps before the seekers can get to them. For more detail see Baker et al. (2020).

[30] And importantly, the trainers weren't *trying* to disincentivize resource-seeking behavior; quite the contrary, the set-up seems (I haven't investigated the history of the experiments) to have been designed to test whether "emergent tool-use" would occur.

[31] Other concrete examples of unintended power-seeking might include AI systems trying to: break out of a contained environment; make backup copies of themselves; gain unauthorized capabilities, sources of information, or channels of influence; mislead/lie to humans about their goals; resist or manipulate attempts to retrain them or shut them off; create/train new AI systems themselves; coordinate illicitly with other AI systems; impersonate humans; cause humans to do things for them; increase human reliance on them; weaken various human institutions and response capacities; take control of physical infrastructure like factories or scientific laboratories; cause certain types of technology and infrastructure to be developed; or directly harm/overpower humans.

[32] Humans care a lot about survival, and about certain resources (food, shelter, etc), but beyond that, we associate many forms of power-seeking with a certain kind of greed, ambition, or voraciousness, and with intuitively "resource-hungry" goals, like "maximize X across all of space and time." *Some* strategically-aware human planners are like this, we might think, but not all of them: so strategically aware, agentic planning isn't, itself, the problem. For more in this vein, see e.g. Cegloski's (2016) "Argument From My Roommate"; and also Pinker (2018: 297): "There is no law of complex systems that says that intelligent agents must turn into ruthless conquistadors. Indeed, we know of one highly advanced form of intelligence that evolved without this defect. They're called women". Thanks to Rohin Shah for discussion of the humans example.

mans display, when getting power is easy, seems to me quite compatible with the instrumental convergence thesis. And unchecked by ethics, constraints, and incentives (indeed, even *when* checked by these things) human power-seeking seems to me plenty dangerous, too.[33]

A second objection (in possible tension with the first) is: *humans* (or, some humans) may be power-seeking, but this is a product of a specific evolutionary history (namely, one in which things like survival, resource-acquisition, and social dominance were directly selected for), which AI systems will not share.[34] Some versions of this objection simply neglect to address the instrumental convergence argument above[35] (and note, regardless, that some proposed ways of training AI systems resemble evolution in various respects).[36] But we can see stronger versions as suggesting that maybe it just isn't that hard to train APS systems not to seek power in unintended ways, across a large enough range of inputs, if you're actively *trying* to do so (evolution wasn't). And I do think this is possible—indeed, it's one of my main sources of hope. But I also think that there are barriers to overcome, which I discuss in the next section.

There's more to say about the instrumental convergence thesis.[37] The formulation I've offered here might warrant refinement—and indeed, this part of the argument is one of my top candidates for ways that the abstractions employed might mislead.[38] Still, I expect the basic gist to point at some-

[33]That said, the absence of various forms of overt power-seeking in humans may point to ways we could try to maintain control over less-than-fully PS-aligned APS systems. See discussion below for more.

[34]See e.g. Zador and LeCun (2019) (and follow-up debate Pace (2019)), and Pinker (2018, Chapter 19). One can also imagine non-evolutionary versions of this—e.g., ones that attribute human power-seeking tendencies to our culture, our economic system, and so forth. Indeed, Ceglowski (2016) can be read as suggesting something like this in the context of a particular demographic: after listing Bostrom's convergent instrumental goals, he writes: "If you look at AI believers in Silicon Valley, this is the quasi-sociopathic checklist they themselves seem to be working from."

[35]That is, the argument isn't "humans seek power, therefore AIs will too"; it's "power is useful for pursuing objectives, so AIs pursuing problematic objectives will have incentives to seek power, by default."

[36]Large, multi-agent reinforcement learning environments might be one example. And the "league" used to train AlphaStar (see Vinyals 2019:350-54) seems reminiscent of evolution in various ways.

[37]See Carlsmith (2020: Section 4.2), for more details.

[38]In particular, this part of the argument requires that the agentic planning and strategic awareness at stake be robust enough to license predictions of the form: "if (a) a system would be planning in pursuit of problematic objectives in circumstance C, (b) power-seeking in C would promote its objectives, and (c) the models it uses in planning put it in a position to recognize this, then we should expect power-seeking in C by default." I've tried to build something like the validity of such predictions into my definitions of agentic planning and strategic awareness; but perhaps for sufficiently weak/loose versions of those concepts, such predictions are not warranted; and it seems possible to conflate weaker vs. stronger concepts at different points in one's reasoning, and/or to apply such concepts in contexts where they confuse rather than clarify. Thanks to Rohin Shah for

thing real—the main question is how often it arises, and how difficult it is to avoid.

5.3 The challenge of practical PS-alignment

Let's grant that less-than-fully aligned APS systems will have at least some tendency towards misaligned, power-seeking behavior, by default. The challenge, then, is to prevent such behavior in practice—whether through alignment (including full alignment), or other means. How difficult will this be?

We can group possible interventions here into three types. The first attempts to control a system's *objectives*. The second attempts to control its *capabilities*. The third attempts to control its *circumstances.* Each type has problems.

5.3.1 Controlling objectives

Much current work on technical alignment focuses on shaping an AI system's objectives to prevent misaligned power-seeking.[39] But this project must confront at least two major obstacles.

5.3.1.1 Problems with proxies. Many ways of attempting to control an AI's objectives share a common challenge: namely, that giving an AI system a "proxy objective"—that is, an objective that reflects properties correlated with, but separable from, intended behavior—can result in behavior that weakens or breaks that correlation, especially as the power of the AI's optimization for the proxy increases.

We're familiar with this problem in human contexts. To give just one example: suppose I try to lower the cobra population by paying the people in my town a bounty for each dead cobra. Initially they might kill lots of existing cobras. But as the cobra population decreases, they might also start to *breed* cobras, in order to kill them and turn them in.[40] Optimizing

emphasizing possible objections in this vein, and for discussion.

[39]Note that the mechanisms we have available to shaping an AI system's objectives change over time. I emphasize this because sometimes the challenge of AI alignment is framed as one of shaping an AI's objectives *in a particular way*—for example, via hand-written code, or via some sort of reward signal, or via English-language sentences that will be interpreted in literalistic and uncharitable terms. And this can make it seem like the challenge is centrally one of, e.g., coding, measuring, or articulating explicitly everything we value, or getting AI systems to interpret instructions in common-sensical ways. These challenges may be relevant in some cases, but the core problem is not method-specific.

[40]Inspired by a (possibly fictitious) anecdote described in an Wikipedia entry ("Perverse incentive": "The original cobra effect"). Other examples: paying railroad builders by the mile of track that they lay incentivizes them to lay unnecessary track ("Perverse incentive": "Other examples"); if teachers take "cause my students to get high scores on standardized tests" as their objective, they're incentivized to "teach to the test"—an incentive that can work to the detriment of student education more broadly; and so on. See Wikipedia

for "turn in dead cobras" breaks that proxy objective's correlation with "reduce the cobra population", which is what I actually want the townspeople to do.

The same phenomenon can arise in training AIs. Suppose, for example, that I'm trying to get an AI to engage in some behavior I want, by rating its behavior using some type of feedback. The correlation between the behavior I would rate highly and the behavior I *actually* want will be strong so long as I can monitor and understand the AI's behavior. But if the AI is too sophisticated for me to understand everything it's doing, and/or if it can deceive me about its action, the correlation weakens: the AI may be able to cause me to give high ratings to behavior I wouldn't (in my current state) endorse if I understood it better—for example, by hiding information about that behavior, or by manipulating my preferences.

We already see this sort of problem in existing AI systems. Thus, for example, if we train an AI system to complete a boat race by rewarding it for hitting green blocks along the path to the finish line, it learns to drive the boat in circles in order to hit the same green blocks over and over again (see Clark and Amodei: 2016).[41] Examples like these may seem easy to fix, but they illustrate the more general problem: systems optimizing for proxies often behave in unintended ways. Indeed, this tendency is closely connected to a core property that makes advanced AI useful: namely, the ability to find novel solutions and strategies that humans wouldn't think of.[42]

How can we address this problem? Human feedback seems likely to play a key role.[43] And it may, ultimately, be enough. But notably, we need ways of drawing on this feedback that don't require unrealistic amounts of human supervision and human-generated data;[44] we need to ensure that such feedback captures our preferences about behavior that we can't directly understand and/or whose consequences we haven't yet seen;[45] we need ways of eliminating incentives to manipulate or mislead the human feedback mechanisms in question; and we need such methods to scale competitively as frontier AI capabilities increase.

("Perverse incentive") for more examples. See Manheim and Garrabrant (2019) for an abstract categorization of dynamics of this kind.

[41]See Krakovna et al. (2020) for a much longer list of examples in this vein.

[42]When you don't know how an AI will achieve its objective, and that objective doesn't capture everything that you really want, then even for comparatively weak systems and simple tasks, it's hard to anticipate how the system's way of achieving the objective will break its correlation with what you really want. And as the AI's capacity to generate solutions we can't anticipate grows, the problem becomes more and more challenging.

[43]This paragraph draws heavily on discussion in Ngo (2020).

[44]See e.g. Christiano et al (2017).

[45]See Christiano et al. (2018) for discussion. Iterative amplification and distillation (Christiano et al. (2018)); debate (Irving et al. (2018)); and recursive reward modeling (Leike et al. (2018)) can all be seen as efforts in this vein.

Would it help if our AI systems could understand fuzzy human concepts like "helpfulness," "obedience," "what humans would want," and so forth? I expect it would, in various ways (though as I discuss below, this also opens up new opportunities for deception/manipulation). But note that the key issue isn't getting our AI systems to *understand* what objectives we want them to pursue—indeed, such understanding is plausibly on the critical path to increasing their capability, regardless of their alignment.[46] Rather, the key issue is causing them to *pursue* those objectives for their own sake.[47]

5.3.1.2 Problems with search. Many techniques shape an AI's objectives using proxies. But some techniques—namely, those that involve *searching* over AI systems and selecting those that perform well on some evaluation criteria, without controlling the systems' objectives directly—face an additional problem.

The problem is that, *even if* the search criteria fully capture the behavior we want, the resulting systems may not end up intrinsically motivated by the criteria in question. Instead, they may end up with other objectives, pursuit of which correlated with good performance during the selection process, but which lead to unintended behavior when the system is exposed to other inputs.

Some think of human evolution as an example.[48] Someone interested in creating agents who pass on their genes could run a process similar to evolution, which searches over different agents, and selects for ones who pass on their genes (for example, by allowing ones who don't to die out). But this doesn't mean the resulting agents will be intrinsically motivated to pass on their genes. Humans, for example, are motivated by objectives that were *correlated* with passing on genes (for example, avoiding bodily harm, having sex, securing social status, etc), but which they'll pursue in a manner that breaks such correlations, given the opportunity (for example, by using birth control, or remaining childless to further their careers).

Rudimentary, evolved AI systems display analogous tendencies. Thus, when Ackley and Littman ran an evolutionary selection process in an environment with trees that allowed agents to hide from predators, the agents developed such a strong attraction to trees that (after reproductive age)

[46]This is a point from Ord (2020). For example, if we train some set of sophisticated agents to get bananas, in a complex environment that requires understanding and modeling humans, they may end up capable of understanding quite accurately (even more accurately than us) what we have in mind when we talk about "aligned behavior," and of behaving accordingly (for example, when we give them bananas for doing so). But their intrinsic objectives could still be focused centrally on bananas (or something else), and our abilities to control those objectives directly might remain quite limited.

[47]Though if they understand those objectives, but don't share them, we might also be able to incentivize them to pursue such objectives for instrumental reasons.

[48]See e.g. Hubinger et al. (2019: 6).

they would starve to death in order to avoid leaving tree areas.[49] And we see early evidence of similar dynamics in systems trained via gradient descent.[50] Thus, for example, when an agent is given reward for visiting colored spheres in a certain order, and trained in an environment where an "expert" traces the correct path, it learns to follow the expert rather than to trace the correct path—thus accumulating large amounts of negative reward when paired with an "anti-expert" who traces the path incorrectly.

The problem, in these various cases, isn't that the evaluation criteria (e.g. "pass on genes," "visit the spheres in X order") fail to fully capture and operationalize what we want. The problem is that selecting agents by reference to these evaluation criteria doesn't afford the designers enough control over the objectives of the resulting agents.

It's an open empirical question how often problems like this will arise. But there are reasons to worry. For one thing, proxy goals correlated with evaluation criteria may be simpler, and therefore easier to learn, than the evaluation criteria.[51] What's more, if the "actual" objective function provides slower feedback, agents that pursue faster-feedback proxies may have advantages.[52] Finally, to the extent that many objectives would *instrumentally* incentivize good behavior in training (for example, because many objectives, when coupled with strategic awareness, incentivize gaining power in the world, and doing well in training leads to deployment/greater power in the world), but few involve *intrinsic* motivation to engage in such behavior, we might think it more likely that selecting for good behavior leads to agents who behave well for instrumental reasons.

Overall, ensuring robust practical PS-alignment seems harder if available techniques search over systems that meet some external evaluation criteria, with little direct control over their objectives. And much of contemporary machine learning fits this bill.

5.3.1.3 Myopia. Some broad types of objectives seem to incentivize power-seeking on fewer physics-compatible inputs than others. Perhaps, then, we

[49]See Christian (2020).

[50]See Shah et al. (2022) and Langosco et al. (2023) on examples of "goal misgeneralization."

[51]In the context of evolution, for example, it seems much harder to evolve an agent whose mind represents a concept like "passing on my genes," and then takes doing this as its explicit goal—humans, after all, didn't even have the concept of "genes" until very recently—than to evolve an agent whose objectives reflect the relevance of things like bodily damage, sex, power, knowledge, etc to whether its genes get passed on (though starting with cognitively sophisticated agents might help in this respect). This is a point I believe I heard first from Evan Hubinger.

[52]For example: in the game Montezuma's Revenge, it helps to give an agent a direct incentive analogous to "curiosity" (e.g., it receives reward for finding sensory data it can't predict very well), because the game's "true" objective (e.g. exiting a level by finding keys that require a large number of correct sequential steps to reach) is too difficult to train on. See Burda et al. (2018), and discussion in Christian (2020).

can aim at those, even if we lack more fine-grained control.

Short-term (or, "myopic") objectives seem especially interesting here.[53] Paradigmatically dangerous AI systems plan in pursuit of long-term objectives. Longer time horizons allow more time to gain and use forms of power that aren't readily available, and more easily justify temporarily costly action (for example, trying to appear aligned, in order to get deployed) for the sake of longer-term gains. Since myopic agents are on a much tighter schedule, they have weaker incentives to attempt forms of power-seeking (deception, resource acquisition, etc.) that only pay off in the long-run.[54]

Myopia might help, but I see at least two problems with relying on it. First, there will plausibly be demand for non-myopic agents. Human individuals and institutions often have fairly (though not arbitrarily) long-term objectives that require long-term planning—running factories and companies, pursuing electoral wins, and so on. As I've already discussed, there will be powerful incentives to automate the pursuit of these objectives. Non-myopic systems will have an advantage, in this regard, over myopic ones.

Second, the "search" techniques discussed in the previous section—techniques that don't allow you to control an agent's objectives directly—may make ensuring myopia challenging. And various types of long-term training processes—for example, reinforcement learning on tasks that involve many sequential steps—seem likely to result in non-myopia by default.[55]

5.3.2 Controlling capabilities

AI alignment research often focuses on controlling a system's objectives. But controlling its capabilities can help with practical PS-alignment too. The less capable a system, the more easily its behavior (including its tendencies towards misaligned power-seeking) can be anticipated and corrected. Less capable systems will also have a harder time getting and keeping power, and a harder time making use of it. So they will have stronger incentives to cooperate with humans, rather than (say) deceive or overpower them.

Preventing agentic planning and strategic awareness in the first place would

[53] We can also imagine other strategies for controlling shrinking the set of inputs that prompt PS-misaligned behavior. For example, we might aim for objectives that penalize "high-impact" action (see e.g. the discussion of "impact penalties" in Krakovna et al. (2020)), or that prohibit lying in particular, or that give intrinsic weight to various legal and ethical constraints. But these face the same challenges involving proxies and search discussed in the last two sections.

[54] Of course, even short spans of time can be enough to do a lot of harm, especially for extremely capable systems. And the timespans "short enough to be safe" can alter if what one can do in a given span of time changes. Thanks to Rohin Shah, Paul Christiano, and Carl Shulman for discussion.

[55] That said, myopia is a fairly coarse-grained property for an objective to possess, and may be easier to cause or check for than others.

be one example of "controlling capabilities," but there are other options, too. I'll consider two.

The first is *specialization*: we might try to build APS systems whose competence is as narrow and specialized as possible. After all, an APS system skilled at a specific kind of scientific research, and not at (say) hacking and social persuasion, seems much less dangerous than an APS system that can engage in these other tasks as well. And specialization often has various benefits (hence its importance in human organizations and economies).[56] That said, generality has benefits, too—which is why human workers with quite general skill-sets (CEOs, for example) are prized in many domains.[57] And just as available machine learning techniques may push the field toward agentic planning and strategic awareness, so too they may push it towards generality.[58] Plus, even very specialized APS systems can be dangerous.[59]

The second strategy is *preventing problematic improvements in capability.* New capabilities can put a system in a position to gain and maintain power in ways it couldn't before—and hence, make new incentives action-relevant. Practical PS-alignment may therefore require controlling the extent to which the inputs a system receives result in improved capabilities. This seems easier if the variables in the system that determine how it responds to inputs (for example, the weights in a neural network) stay fixed. But we may also want systems that mix task-performance and learning together, that "remember" previous events, and so forth; and predicting and controlling the capabilities such systems will develop could be difficult (especially if we don't understand well how they work—more below).

Strategies that rely on limiting a system's capabilities also face a more general problem: namely, that there are likely to be strong incentives to scale

[56]For example, they can be optimized more heavily for specific functions (to borrow an example from Ben Garfinkel, there is a reason that the flashlight, camera, speakers, etc on an iPhone are inferior to the best flashlights, cameras, etc). And note that we will have much greater abilities to optimize AI systems for particular tasks than we do with humans.

[57]In particular, general systems plausibly respond better to changing environments and demands; and if a task requires multiple competencies, specialized systems can be harder to coordinate (e.g., it's helpful to have a single personal assistant, rather than one for email, one for scheduling, one for travel planning, one for research, etc).

[58]GPT-3, for example, is trained to a fairly general level of capability via predicting text, and later fine-tuned on specific tasks like coding. Such an approach might be necessary for tasks where data is hard to come by or learn from directly (e.g. "be an effective CEO"). More broadly, as Bostrom (2014) notes, the most efficient route to widespread automation may be the creation of general purpose agents that can learn a wide variety of new tasks very efficiently (though those agents could also end up quite specialized later).

[59]A system highly skilled at hacking into new computers and copying itself, for example, can spread far and wide; a system skilled in science can design a novel virus; a system with control over automated weapons can use them; a system skilled at social manipulation can turn an election; and so forth.

up the capabilities of frontier systems.[60] PS-alignment strategies that can't scale accordingly (and competitively) therefore risk obsolescence as state of the art capabilities advance. A key question for any such strategy, then, is whether it can translate, given success at some level of capability, into a different strategy that scales better.[61]

5.3.3 Controlling circumstances

Alongside (or instead of) controlling a system's "internal" properties—its objectives or capabilities—we might also try to control its *external circumstances*.[62] If we want to prevent a system from engaging in hacking, for example, we might try to control its options (no internet access for the system); its incentives (if the system is caught hacking it will be disabled, thus reducing its ability to pursue its objectives); or both.

The success of this strategy will depend a lot on the system's capabilities. If it is sufficiently good at hacking, our efforts to prevent it from accessing the company bank account might fail. If the system is sufficiently good at avoiding detection, our threats to disable it if we catch it hacking will provide little incentive for it to refrain. So in order to contain problematic power-seeking by controlling options or incentives, we need monitoring and enforcement mechanisms that scale with the capabilities of frontier APS systems. But this may prove very difficult; as the capabilities of frontier systems increase, their capacity to evade and disable our monitoring and enforcement mechanisms will increase as well.[63]

[60] Though note that PS-alignment problems with more capable systems could complicate these dynamics. More discussion below.

[61] For example, we might try to achieve practical PS-alignment with some fairly advanced systems (including, perhaps, quite specialized ones—or, indeed, non-APS ones), and then use them to create new and superior PS-alignment strategies (indeed, as AI development itself becomes increasingly automated, automating alignment research will plausibly be necessary regardless). Note, though, that plans of the form "create some practically PS-aligned systems, and ask them what the plan should be" might just not work. For example, the new systems might not have adequate plans either. One might therefore need to create even more capable systems, which *also* might not have adequate plans, and so forth, until one pushes up against (or perhaps, past) the limits of one's capacity to ensure practical PS-alignment.

[62] Admittedly, the line between capabilities and circumstances is somewhat blurry, but I mean for the former to lean more heavily on internal features of the system, and the latter, on external features.

[63] Also: ensuring practical PS-alignment seems harder insofar as a system's deployment circumstances are dissimilar from the ones on which humans have observed and verified PS-aligned behavior (for example, during training, or pre-deployment testing). Indeed, ideally, one would want the deployment inputs to come from the same distribution as the training inputs. But in practice, and especially in strategically-aware systems, ensuring a close-to-identical distribution seems very difficult (if not impossible). This is partly because the world changes (and indeed, the actions of the APS systems can themselves change it. See Krueger et al. (2020) and Christian (2020) for some discussion of this possibility.) But also, to the extent that the distinction between training and deployment reflects some real difference in the agent's level of influence on the world, this difference is itself a change in distribution—one that a sufficiently sophisticated agent might recognize.

5.4 Unusual difficulties

So all three strategies for ensuring PS-aligned behavior we've considered—controlling objectives, capabilities, and circumstances—face serious difficulties. Still, one might suggest, we have compelling inductive grounds for thinking that these difficulties can be overcome. After all, plenty of technologies encounter serious safety problems in the course of their development, but we often reach adequate safety standards in the end. Planes don't generally crash; bridges stay up; e-mail servers are tolerably secure; and so on.

But advanced, planning, strategically-aware AI systems pose safety challenges that (most) other technologies do not. Here are a few that seem salient.

The first is *opacity.* Ensuring safety and reliability requires understanding a system well enough to predict its behavior. But the thinking and strategic decision-making of an agent whose cognitive capacities significantly exceed our own may be quite alien and opaque to us. Such a system might consider options humans never would; understand physical and social dynamics that humans do not; reason in ways we cannot understand, and so on.[64]

This issue seems especially salient in the current, machine-learning dominated AI paradigm, in which our ability to create an AI system that can perform some task (e.g., predicting text) often far exceeds our ability to understand *how* the system does what it does. We set various key high-level variables (the system's architecture, the number of parameters, the training process, the evaluation criteria), but the system that results is still, in many (though not all) respects, a black box. We must rely on further experiments to try to get some handle on what it knows, what it can do, and how it is liable to behave.[65] This marks an important contrast with technologies like planes and bridges, where we achieve safety partly through understanding of the basic physical principles that govern their behavior.[66]

The second challenge comes from *adversarial dynamics*, especially in the context of efforts to detect safety problems. Suppose an AI system has

[64]See Yudkowsky (unpublished) on "strong cognitive uncontainability."

[65]Of course, our understanding of how ML systems work will likely improve over time, and research in this area ("interpretability") is ongoing (See e.g. Olah (2020) and Goh et. al. (2021)). But interpretability is no bottleneck to training bigger models on even more complex tasks—or, plausibly, to the commercial viability of those models. And much of the AI field is devoted to pushing forward with developing whatever capabilities we can, interpretable or no.

[66]That said, understanding comes in many varieties and degrees; and it's an empirical question what mix of experiment/search vs. first-principles understanding/design has actually been involved in ensuring the safety of different technologies (for example, in biology, or before advanced scientific understanding).

an objective that it can better achieve by passing a training or evaluation process. It may then manipulate that process, or deceive the people conducting it.[67] And if it is sufficiently capable at this, the appearance of safety and reliability on various tests may tell us little about how the system will behave in other circumstances. Few if any existing technologies exhibit this dynamic. Planes, rockets, and nuclear plants may be dangerous and complicated. But they never *try* to appear safer than they are, or to manipulate our ability to understand and evaluate them.

A final challenge involves the *stakes of error.* If an engineered virus escapes from a lab, it can spread rapidly and become increasingly difficult to contain. And this seems like a much better analogy for a misaligned, power-seeking AI system than a plane crashing or a bridge falling down. Because the stakes of error are so high, there is much less room for trial and error.[68] Indeed, if you're trying to store an engineered virus that has a significant chance of killing most of the global population if it gets released, you need safety standards *much* higher than those we use, even now (after generations of trial and error), for bridges or planes—much higher, indeed, than we use for approximately anything (this is one key reason to never, ever create such a virus). And humanity's track record in the highest stakes contexts—biosafety labs that handle the most dangerous viruses, nuclear power plants, nuclear weapons facilities—seems far from comforting.[69]

5.5 Overall difficulty

Overall, then, ensuring practical PS-alignment seems like it could well prove challenging. And while there are tools that might help—myopia, restricting capabilities, and so on—there are significant problems with each of these tools, along with more general reasons to think the problem uniquely difficult relative to technological safety problems we've faced in the past.

6 Deployment

Let's suppose, then, that ensuring practical PS-alignment is difficult. Should we expect to see practically PS-misaligned APS systems actually deployed?

One might think: no. After all, if we couldn't figure out how to build planes that don't crash, we wouldn't expect to see people dying in plane crashes all the time. Rather, we'd expect to see people not flying. What's more, plenty of mundane commercial incentives favor safety (safety fail-

[67] Bostrom (2014) calls this a "treacherous turn." See also Cotra (2021) on the "training game."

[68] And whatever their present safety, most current technologies involved many errors (plane crashes, rocket explosions, etc) along the way.

[69] See Ord (2020: 130), for some discussion of biological accidents in particular.

ures can result in significant social/regulatory backlash and economic cost), and the incentives to prevent harmful, large-scale forms of misaligned power-seeking seem especially clear (since sufficiently severe failures can result in the direct disempowerment of the relevant decision-makers, their loved ones, and so on).

Faced with such incentives, why would anyone deploy a strategically aware AI agent that will seek power in unintended ways? A simple reason is: the decision-maker might just act stupidly, and contrary to the evidence and incentives a rational decision would reflect. But I think there are more specific reasons for concern. Here I'll focus on four.[70]

The first is the familiar phenomenon of *externalities.* It might be individually rational for a less-than-fully-altruistic actor to deploy a possibly PS-misaligned system, even if that's very bad in expectation for society, if society's interests (let alone: the interests of all future generations) are inadequately reflected in the actor's incentives.[71] Indeed, even if the actor recognizes the risks to society's interests, and gives them some weight, he/she might not give them *enough* weight to outweigh the prospect of personal profit, power, or prestige.[72]

The second and related reason is *race dynamics.* The time and effort that an AI developer devotes to ensuring practical PS-alignment trades off against the speed with which she can scale up its capabilities. And there are significant rewards to deploying such a system before others do.[73] In order to beat her competitors, therefore, a developer might choose speed over safety. And this might then incentivize other developers to take on increased risk as well, creating further incentives for the initial developer to move speedily and riskily. The result might be an ongoing feedback loop of increasing pressure on all parties—altruistic and not-so-altruistic—to ratchet up their risk tolerance or drop out of the race.

The third reason is that there will be *many relevant actors.* That is, once *some* actors can create APS systems, then over time (and absent active efforts to the contrary) a larger and larger number of actors around the world will likely become able to do so as well. And even if some actors are sufficiently conscientious, intelligent and responsible to avoid problematic de-

[70] An additional risk here is the possibility of "unintentional deployment." Suppose, for example, that a power-seeking system meant to be contained in some training environment, or limited in its means of influencing the outside world, manages to break out of that environment, or to obtain other forms of influence. It may then succeed in gaining various forms of real-world power, although it was never intentionally deployed.

[71] Here we might think of analogies with climate change.

[72] Of course, an actor who deploys a possibly misaligned AI risks harm to *themselves,* as well. But especially if the risk of misalignment-and-resulting-catastrophe is sufficiently small, then it may be in the actor's narrow self-interest to deploy the AI, although the expected disvalue to wider society is very large.

[73] See Askell et. al. (2019) for discussion of first-mover examples. Bostrom's notion of a "decisive strategic advance" (2014) is an extreme example.

ployment (this alone is far from guaranteed), others plausibly won't be. For example, they might be overconfident in the degree of PS-alignment of the systems they've created; or overly dismissive of the amount of harm that misaligned systems would cause; or insufficiently incentivized to register risks to larger society; or insufficiently constrained by legal and regulatory mechanisms, liability regimes, and so on that apply elsewhere. Of course, we can look for ways of setting up coordination and incentive mechanisms that would apply to *all* the relevant actors, around the world (including e.g. China, Russia, North Korea, etc)—but efforts in this vein seem challenging.[74]

The fourth reason is that practically PS-misaligned APS systems can still be *extremely useful*, at least initially. Obviously, if an APS system is blatantly failing to behave in the way that its designers intend, including during testing and evaluation, then it's much less likely to get deployed (compare with e.g. a house-cleaning robot that routinely kills a user's pets). But practically PS-misaligned APS systems need not behave like this. For one thing, as discussed above, they might actively deceive designers about their degree of alignment.[75] But even absent deception/manipulation of this kind, it's hard to predict/test the AI's behavior on the full range of post-deployment inputs—especially in a rapidly changing world, in the absence of deep understanding of how the system works, and if the AI system might gain new knowledge and capabilities post-deployment. Indeed,

[74]Indeed, to the extent that resources invested in ensuring practical PS-alignment trade off against resources invested in increasing the capabilities of the systems one builds, over time we might expect to see actors who invest less in alignment, and who take more risks, to scale up the capabilities of their systems faster. This could result in the competitive dynamics discussed above (e.g., other actors cut back on safety efforts to keep up, and/or deploy systems that wouldn't meet their own safety standards, but which are safer than the ones they expect competitors to deploy); but if other actors *don't* cut back on safety as a result, the most powerful systems might end up increasingly in the hands of the least cautious and socially responsible actors (though there are also important correlations between social responsibility and factors like resource-access, talent, etc).

[75]In particular, we should expect suitably sophisticated and strategically aware systems to *understand* what sorts of behavior humans are looking for during training/testing, even if their objectives don't intrinsically motivate such behavior. So if they are optimizing for getting deployed (for example, because deployment grants greater power), they will have strong instrumental incentives to behave well, to demonstrate the type of usefulness (described above) that will pull us towards deploying them, and to convince us that their objectives are fully (or at least sufficiently) aligned with ours. Indeed, they'll even have incentives to appeal to ethical concerns about how it is morally appropriate to treat them—incentives that will apply *regardless* of the legitimacy of those concerns (though I also expect such concerns to *be* legitimate in at least some cases). This isn't to say that humans will be actually fooled; and some AI systems might themselves be able to help with our efforts to detect deception in others. But unless we can develop deep understanding of and control over the objectives our AI systems are pursuing, evidence like "it performs well on all the tests we ran, including tests designed to detect deceptive/manipulative behavior" and "it clearly knows how to behave as we want" may tell us much less about its ultimate objectives, or about how it will behave once deployed, than we wish. And in the context of such uncertainty, some humans will be more willing to gamble than others.

I think that one of the central reasons we should expect to see practically PS-misaligned AI systems getting used/deployed is precisely that they will *demonstrate* a high degree of usefulness during training/testing—and consequently, it will be hard to resist deploying them.

Here's an analogy. Suppose that scientists create a new, genetically-engineered species of chimpanzee, whose cognitive capabilities significantly exceed those of humans. Initially, scientists confine these chimps in a laboratory environment, and incentivize them to perform various low-stakes intellectual tasks using rewards like food and entertainment. And suppose, further, that these chimpanzees are clearly capable of generating things like vaccine designs, prototypes for new clean energy technology, cures for cancer, highly effective military/political/business strategies, and so forth—and that they will in fact do this, if you set up their incentives right (even though they don't intrinsically value being helpful to humans, and so are disposed, in some circumstances, to seize power for themselves—for example, if they can get more food and entertainment by doing so).

In such a context, it would become increasingly difficult for various actors around the world to resist drawing on the intellectual capabilities of the chimps in a manner that gives the chimps real-world forms of influence. If a new Covid-19 style pandemic started raging, for example, and we knew that the chimps could rapidly design a vaccine, there would be strong pressure to use them for doing so. If the chimps can help "users" win a senate race, or end climate change, or make a billion dollars, or achieve military dominance, then some people, at least, will be strongly inclined to use them, even if there are risks involved—and those who *don't* use them will end up losing their senate races, falling behind their business and military competitors, and so forth.

And even if the chimps, at the beginning, are appropriately contained and incentivized to be genuinely cooperative, it seems unsurprising if, as people draw on their capacities in more and more ways around the world, they get exposed to opportunities and circumstances that incentivize them to seek power for themselves, instead.

Something similar, I think, might apply to APS AI systems. Indeed, even if people *know*, or strongly suspect, that such systems would seek power in misaligned ways in some not-out-of-the-question circumstances, the pull towards using them for goals that matter a lot to us may simply be too great. When pandemics are raging, oceans are rising, parents and grandparents are dying of cancer, rival nations are gaining in power, and billions (or even trillions) of dollars are sitting on the table, concerns about science-fictiony risks from power-seeking AI systems may, especially for *some* relevant actors, take a backseat.

Overall, then: I don't think we should expect obviously non-useful, practi-

cally PS-misaligned APS systems to get intentionally deployed. But I think practically PS-misaligned APS systems might well get deployed regardless. Let's turn to what happens then.

7 Correction

In many contexts, if an AI system starts seeking to gain/maintain power in unintended ways, the behavior may well be noticed, and the system prevented from gaining/maintaining the power it seeks. Let's call this "correction."

Some types of correction might be easy (e.g., a lab notices that an AI system tried to open a Bitcoin wallet, and shuts it down). Others might be much more difficult and costly (for example, an AI system that has successfully hacked into and copied itself onto an unknown number of computers around the world might be quite difficult to bring under control).

Confronted with post-deployment PS-alignment failures, will humanity's corrective efforts be enough to avert full-scale human disempowerment? I think they might well; but it seems far from guaranteed.

A few initial points bear emphasis. First, the possibility of human disempowerment doesn't rest on any particular view about how *quickly* or *dramatically* the transition to advanced AI capabilities will occur. There needn't be a "fast take-off" (i.e. a rapid escalation to advanced capabilities); or a "discontinuous take-off" (i.e. an escalation that proceeds much more rapidly than some historical extrapolation would have predicted); or an "intelligence explosion" (i.e. an AI-driven feedback loop that propels explosive growth in capabilities).[76] The risk of human disempowerment seems to me *greater* in such scenarios (indeed, substantially greater); but it persists regardless.

Second, and relatedly, the emergence of a single artificial intelligence that dominates the whole world—an AI "singleton," in the language of Bostrom (2014)—is one possible route to human disempowerment, but not the only one.[77] In particular, human disempowerment might instead result from the deployment of *many* PS-misaligned systems, engaged in complex patterns of cooperation and competition. Here we might think of the relationship between humans and chimpanzees: no single human or human institution rules the world, but the chimps are still disempowered relative to humans.

Third, the success or failure of a given instance of misaligned power-seeking depends both on the absolute capability of the power-seeking sys-

[76]On different take-off scenarios, see e.g. Bostrom (2014: 75-95), and Davidson (2021) for a more quantitative analysis. The canonical citation for the idea of an "intelligence explosion" is Good (1966); see also Yudkowsky (2013) for more detailed discussion.

[77]See Bostrom (2014: 95-110).

tem, *and* on the strength of the constraints and opposition that it faces.[78] And in this latter respect, the world that future power-seeking AI systems would be operating in would likely be importantly different from the world of 2023. In particular, such a world would likely feature substantially more sophisticated capacities for detecting, constraining, responding to, and defending against problematic forms of AI behavior—capacities that may themselves be augmented by various types of AI technology, including non-agentic AI systems, specialized/myopic agents, and other AI systems that humans have succeeded in eliciting aligned behavior from, at least in some contexts.

In general, a large number of factors and dynamics are relevant to the success of a given instance of power-seeking on the part of an APS system. I won't discuss these in detail here, but they include: whether and to what extent the system is in a position to enhance its capabilities, what degree of secrecy it's able to maintain, how well it can hack other computer systems, how well it can get access to additional computing power, what options it has for making money, what sort of automated infrastructure it has access to, how easily it can make use of human labor, how easily it can wield social influence, how well the system can develop novel technologies and scientific breakthroughs (factoring in equipment requirements, serial time bottlenecks, and so on), how easily it can coordinate with other APS systems, and how much direct destructive and coercive capacity (weapons, drones, surveillance, options for attacking background conditions of human survival) it has available.

Obviously, there are huge uncertainties about these and other relevant dynamics, and about how they will interact. And with respect to a given sort of power-relevant task, like hacking or social persuasion, we should be careful to distinguish between "better than human" and "arbitrarily capable." Still, if we remain unable to ensure the PS-alignment of deployed, frontier AI systems, then as frontier capabilities increase, it seems plausible that humans will be at an increasing disadvantage. And if we reach a point where power-seeking, misaligned AI systems represent a large majority of the world's quality-weighted cognitive labor, the situation seems dire.

People sometimes argue that "warning shots"—evident instances of misaligned power-seeking, observed in early strategically aware AI systems—will prevent us from reaching such a point. Perhaps so, but there are reasons for pessimism.

For one thing, it's possible that frontier capabilities will escalate rapidly. This would leave little time for warning shots to inform humanity's AI research and decision-making, before it must confront highly capable, sophisticated, strategically-aware AI agents.

[78] See e.g. Drexler's (2019, Chapter 31) distinction between "supercapabilities" and "superpowers."

But even if frontier capabilities escalate more slowly, warning shots may not be enough. For example, even with the attention prompted by a large warning shot, the problems may prove too difficult to solve before it's too late. And certain sorts of "solutions" may function as band-aids: they correct a system's observed behavior, but not the underlying issue with its objectives. For example, if you train a system by penalizing it for lying, you may incentivize "don't tell lies that would get detected," as opposed to "don't lie" (and the training process itself might provide more information about which lies are detectable).

Moreover, there are reasons to expect fewer warning shots as the strategic and cognitive capabilities of frontier systems increase, *regardless* of whether techniques for ensuring practical PS-alignment have adequately improved. This is because more capable systems, regardless of their PS-alignment, will be better able to model what sorts of behavior humans are looking for, and to forecast what attempts at power-seeking will be detected and corrected—a dynamic that could lead to a misleading impression that earlier problems have been adequately addressed, or even an impression that those problems stemmed from lack of intelligence rather than lack of alignment.[79]

Finally, even if there is widespread awareness that existing techniques for ensuring practical PS- alignment are inadequate, various actors might still push forward with scaling up and deploying highly-capable AI agents, for the reasons discussed in the previous section (e.g., because they have lower risk estimates; because they are willing to take more risks for the sake of profit, power, short-term social benefit, competitive advantage; and so on).

Overall, future scenarios in which global civilization grapples with practical PS-alignment failures in advanced AI agents, especially on a widespread scale or with escalating severity, are difficult to analyze in any detail, because so many actors, factors, and feedback loops are in play. Such scenarios need not, in themselves, spell the full-scale disempowerment of humanity, if we can get our act together enough to correct the problem, and to prevent it from re-arising. But an adequate response will likely require addressing one or more of the basic factors that gave rise to the issue in the first place: e.g., the difficulty of ensuring the practical PS-alignment of APS systems (especially in scalably competitive ways), the strong incentives to

[79]See e.g. the roll-out scenario described in Bostrom (2014). It's worth noting that in principle, if the ability to accurately assess whether a given instance of misaligned power-seeking will succeed arises sufficiently early in the AI system's we're building/training, the period of time in which we see widespread and overt misaligned power-seeking from such agents could be quite short, or even non-existent. That is, it could be that the type of strategic awareness that makes an AI agent aware of the benefits of seeking real-world forms of power, resources, etc is closely akin to the type that makes an AI agent aware of and capable of avoiding the downsides of getting caught. If the former is in place, the latter may follow fast—even if the trajectory of AI capability development in general is more gradual.

use/deploy such systems even if doing so risks practical PS-alignment failure, and the multiplicity of actors in a position to take such risks. It seems unsurprising if this proves difficult.

8 Catastrophe

A final premise of this article's overarching argument is that the permanent and involuntary disempowerment of humanity would be an *existential catastrophe*. Is that true?

Precise definitions can matter here, but speaking loosely, and inspired by the definition in Ord (2020), I'll think of an existential catastrophe as an event that drastically reduces the value of the trajectories along which human civilization could realistically develop.[80] Readers should feel free, though, to substitute in their own preferred definition—the broad idea is to hone in on a category of event that people concerned about what happens in the long-term future should be extremely concerned to prevent.

It's possible to question whether humanity's permanent and involuntary

[80] Admittedly, there are complexities here that I won't broach. Ord's definition of existential catastrophe—that is, "the destruction of humanity's longterm potential"—invokes "humanity," but he notes that "If we somehow give rise to new kinds of moral agents in the future, the term 'humanity' in my definition should be taken to include them" (see Ord 2020: 39); and he notes, too, that "I'm making a deliberate choice not to define the precise way in which the set of possible futures determines our potential. A simple approach would be to say that the value of our potential is the value of the best future open to us, so that an existential catastrophe occurs when the best remaining future is worth just a small fraction of the best future we could previously reach. Another approach would be to take account of the difficulty of achieving each possible future, for example defining the value of potential as the expected value of our future assuming we followed the best possible policy. But I leave a resolution of this to future work" (Ord 2020: 37, footnote 4). That is, Ord imagines a set of possible "open" futures, where the quality of humanity's "potential" is some (deliberately unspecified) function of that set. One issue here is that if we separate a future's "open-ness" from its probability of occurring, then very good futures are "open" to e.g. future totalitarian regimes, or future AI systems, to choose "if they wanted," even if their doing so is exceedingly unlikely—in the same sense that it is "open" to me to jump out a window, even though I won't. But if we try to incorporate probability more directly (for example, by thinking of an existential catastrophe simply as some suitably drastic reduction in the expected value of the future), then we have to more explicitly incorporate further premises about the current expected value of the future; and suitably subjective notions of expected value raise their own issues. For example, if we use such notions in our definition, then getting bad news—for example, that the universe is much smaller than you thought—can constitute an existential catastrophe; and I expect we'd also want to fix a specific sort of epistemic standard for assessing the expected value in question, so such that assigning subjective probabilities to some event being an existential catastrophe sounds less like "I'm at 50% credence that my credence is above 90% that X," and more like "I'm at 50% credence that if I thought about this for 6 months, I'd be at above 90% that X". Like Ord, I'm not going to try to resolve these issues here. Hilary Greaves has an unpublished draft, "Concepts of existential catastrophe," examining various of these issues in more detail.

disempowerment at the hands of AI systems would qualify. In particular, if you are optimistic about the quality of the future that practically PS-misaligned AI systems would, by default, try to create, then the disempowerment of all humans, relative to those systems, will come at a much lower cost to the future in expectation.[81]

One route to such optimism is via the belief that all or most cognitive systems (at least, of the type one expects humans to create) will converge on similar objectives in the limits of intelligence and understanding—perhaps because such objectives are "intrinsically right" (and motivating), or perhaps for some other reason.[82] My own view, shared by many, is that "intrinsic rightness" is a bad reason for expecting convergence,[83] but other possible reasons—related, for example, to various forms of cooperative game-theoretic behavior and self-modification that intelligent agents might converge on[84]—are more complicated to evaluate.[85] And we can imagine other routes to optimism as well—related, for example, to hypotheses about the default consciousness, pleasure, preference satisfaction, or partial alignment of the AI systems that disempowered humans.

I'm not going to dig in on this much. I do, though, want to reiterate that my concern in this article is with the *involuntary* disempowerment of humanity. That is, sharing power with AI agents—especially conscious and cooperative ones—may ultimately be the right path for humanity to take. But if so, it should be a path we *chose*, on purpose, with full knowledge of what we were doing and why: we don't want to build AI agents who force such a path upon us, whether we like it or not.

I think the moral situation here is actually quite complex. Suitably sophisticated AI systems may be moral patients; morally insensitive efforts to use,

[81] And perhaps it would come at lower cost to the present, as well, if advanced AI systems are more generally benevolent.

[82] Note that this could be compatible with Bostrom's (2014) formulation of the "orthogonality thesis"—e.g., "Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with any final goal." That is, Bostrom's formulation only applies to the "in principle" possibility of combining high intelligence and any final goal. But there could still be strong correlations, attractors, etc in practice (this is a point I first heard from David Chalmers).

[83] If, for example, you program a sophisticated AI system to try to lose at chess—see, e.g., suicide chess—it won't, as you increase its intelligence, start to see and respond to the "objective rightness" of trying to win instead, or of trying to reduce poverty, or of spreading joy throughout the land—even after learning what humans mean when they say "good," "right," and so forth. See discussion in Russell (2019: 166).

[84] For an especially exotic version of this, see Oesterheld (2017).

[85] That said, the history of atrocities committed by strategic and intelligent humans does not seem comforting in this respect. And note that the incentives at stake here depend crucially on an agent's empirical situation, and on its power relative to the other agents whose behavior is correlated with its own. In a context where misaligned AI systems are much more powerful than humans, it seems unwise to depend on their having and responding to instrumental, game-theoretic incentives to be particularly nice.

contain, train, and incentivize them risk serious harm; and such systems may, ultimately, have just claims to things like political rights, autonomy, and so forth. In fact, I think that part of what makes alignment important, even aside from its role in making AI safe, is its role in making our interactions with AI moral patients ethically acceptable.[86] It's one thing if such systems are intrinsically motivated to behave as we want; it's another if they aren't, but we're trying to get them to do so anyway. More generally, once you build a moral patient, you come under strong moral reasons to treat it well. What "treating artificial moral patients well" involves seems to me a crucial question for humanity as we transition into an era of building systems that might qualify. At present, as far as I can tell, we have very little idea how to even identify what artificial systems warrant moral concern. In a deep sense, I think, we know not what we do.

But some moral patients—and some agents who might, for all we know, be moral patients, but aren't—will also try to seize power for themselves, and will be willing to do things like harm humans in the process. So building new, very powerful agents who might be moral patients is, not surprisingly, both a morally and prudentially dangerous game: one that humanity, plausibly, is not ready for. My assumption, in this report, has been that unfortunately, we—or at least, some of us—are going to barrel ahead anyway, and I fear we will make many mistakes, both moral and prudential, along the way.

The point, then, is not that humans have some deep right to power over the AI systems we build. Rather, the point is to avoid losing control of our AI systems before we've acquired the maturity to truly understand our different paths into the future—including paths that involve sharing power with AI systems—and to choose wisely amongst them.

9 Conclusion

This has been a comparatively brief discussion of an extremely important topic. I'm conscious, in particular, of how little I've said about timelines, take-off speeds, existing ideas for aligning advanced systems, the routes to AI takeover, and the broader ethics of sharing the world with (or perhaps, ceding the world to) digital minds. And my own views on these topics continue to evolve in important ways.

Still, the basic case for concern seems to me strong. At a high-level, we—or at least, some of us—are currently pouring resources into learning how to build something akin to a second advanced species;[87] a species potentially

[86]Thanks to Katja Grace for discussion of this point.

[87]I'm borrowing the term "second advanced species" from Holden Karnofsky, though see also Bostrom (2015), Russell (2019: Chapter 5), Ord (2020), and Ngo (2020) for similar framings.

much more powerful than we are; that we do not yet understand, and that it's not clear we will be able to control. In this sense, we are playing with a hotter fire than we have ever tried to handle. We are doing something unprecedented and extremely dangerous; with very little room for error; and the entire future on the line.

More specifically: within my lifetime, I think it more likely than not that it will become possible and financially feasible to create and deploy powerful AI agents. And I expect strong incentives to do so, among many actors, of widely varying levels of social responsibility. What's more, I find it quite plausible that it will be difficult to ensure that such systems don't seek power over humans in unintended ways; plausible that they will end up deployed anyway, to catastrophic effect; and plausible that whatever efforts we make to contain and correct the problem will fail.

That is, as far as I can tell, there is a disturbingly high risk (I think: greater than 10%) that I live to see the human species permanently and involuntarily disempowered by AI systems we've lost control over. What we can and should do about this now is a further question. But the issue seems extremely serious.[88]

References

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. *arXiv:1907.04534*. http://arxiv.org/abs/1907.04534

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. *Eighth International Conference on Learning Representations*. https://iclr.cc/virtual_2020/poster_SkxpxJBKwS.html

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

[88]Thanks to Asya Bergal, Alexander Berger, Paul Christiano, Ajeya Cotra, Tom Davidson, Daniel Dewey, Owain Evans, Ben Garfinkel, Katja Grace, Jacob Hilton, Evan Hubinger, Jared Kaplan, Holden Karnofsky, Sam McCandlish, Luke Muehlhauser, Richard Ngo, David Roodman, Rohin Shah, Carl Shulman, Nate Soares, Jacob Steinhardt, and Eliezer Yudkowsky for input on the longer report on which this article is based; thanks to Leopold Aschenbrenner, Ben Garfinkel, Daniel Kokotajlo, Eli Lifland, Neel Nanda, Nate Soares, Christian Tarsney, David Thorstad, David Wallace, Ben Levinstein, and two other anonymous reviewers, for writing public reviews of that report; thanks to Nick Beckstead for guidance and support throughout that investigation; thanks to Sara Fish for formatting and bibliography help; thanks to Ketan Ramakrishnan for helping with the process of editing the longer report into this shorter form, and for input and discussion more broadly; and thanks to Hilary Greaves for comments on this article. Research on this project was originally conducted for Open Philanthropy, but the views expressed here are my own.

Bostrom, N. (2015). *What happens when our computers get smarter than we are?* https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are/transcript

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. *ICLR 2019*. https://openreview.net/forum?id=H1lJJnR5Ym

Carlsmith, J. (2020, September 11). *How much computational power does it take to match the human brain?* https://www.openphilanthropy.org/brain-computation-report

Carlsmith, J. (2022, June 16). *Is power-seeking AI an existential risk?* arXiv. http://arxiv.org/abs/2206.13353

Cegłowski, M. (2016, October 29). *Superintelligence: The idea that eats smart people*. https://idlewords.com/talks/superintelligence.htm

Cellan-Jones, R. (2014). Stephen hawking warns artificial intelligence could end mankind. *BBC News*. https://www.bbc.com/news/technology-30290540

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W.W. Norton.

Christiano, P. (2018, April 7). *Clarifying "AI alignment"* [Medium]. https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6

Christiano, P. (2019). *Paul Christiano: Current work in AI alignment*. https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv:1810.08575*. http://arxiv.org/abs/1810.08575

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Clark, J., & Amodei, D. (2016, December 22). *Faulty reward functions in the wild* [OpenAI]. https://openai.com/blog/faulty-reward-functions/

Cotra, A. (2021, March 5). *The case for aligning narrowly superhuman models - LessWrong*. https://www.lesswrong.com/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-models

Davidson, T. (2021, March 25). *Report on semi-informative priors*. Open Philanthropy. https://www.openphilanthropy.org/blog/report-semi-informative-priors

Drexler, K. E. (2019, January). *Reframing superintelligence*. Future of Humanity Institute. University of Oxford.

Gabriel Goh, Nick Cammarata, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, & Chris Olah. (2021). *Multimodal neurons in artificial neural networks* [Distill]. https://distill.pub/2021/multimodal-neurons/

Garfinkel, B. (2018). *Ben Garfinkel: How sure are we about this AI stuff?* https://ea.greaterwrong.com/posts/9sBAW3qKppnoG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff

Garfinkel, B., Brundage, M., Filan, D., Flynn, C., Luketina, J., Page, M., Sandberg, A., Snyder-Beattie, A., & Tegmark, M. (2017). On the impossibility of supersized machines. *arXiv:1703.10987*. https://doi.org/10.48550/arXiv.1703.10987

Garfinkel, B., & Dafoe, A. (2019). How does the offense-defense balance scale? *Journal of Strategic Studies*, *42*, 736–763. https://doi.org/10.1080/01402390.2019.1631810

Garfinkel, B., Lempel, H., Wiblin, R., & Harris, K. (2020, July 9). *Ben Garfinkel on scrutinising classic AI risk arguments* [80,000 hours]. https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/

Good, I. J. (1966, January 1). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Rubinoff (Eds.), *Advances in computers* (pp. 31–88). Elsevier. https://doi.org/10.1016/S0065-2458(08)60418-0

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Greaves, H. (n.d.). *Concepts of existential catastrophe* [unpublished].

Hubinger, E. (2020, November 9). *Clarifying inner alignment terminology - AI alignment forum*. https://www.alignmentforum.org/posts/SzecSPYxqRa5GCaSF/clarifying-inner-alignment-terminology

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv:1906.01820*. http://arxiv.org/abs/1906.01820

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv:1805.00899*. http://arxiv.org/abs/1805.00899

Karnofsky, H. (2016, May 6). *Some background on our views regarding advanced artificial intelligence* [Open Philanthropy]. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence

Karnofsky, H. (2021, September 24). *The "most important century" blog post series* [Cold Takes]. https://www.cold-takes.com/most-important-century/

Karnofsky, H. (2022). AI strategy nearcasting. https://www.alignmentforum.org/posts/Qo2EkG3dEMv8GnX8d/ai-strategy-nearcasting

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020, April 21). *Specification gaming: The flip side of AI ingenuity* [DeepMind]. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Krueger, D., Maharaj, T., & Leike, J. (2020). Hidden incentives for auto-induced distributional shift. *arXiv:2009.09153*. http://arxiv.org/abs/2009.09153

Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2023, January 9). *Goal misgeneralization in deep reinforcement learning*. arXiv. http://arxiv.org/abs/2105.14111

LeCun, A. Z., Yann. (2019, September 26). *Don't fear the terminator* [Scientific American Blog Network]. https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv:1811.07871*. http://arxiv.org/abs/1811.07871

Manheim, D., & Garrabrant, S. (2019). Categorizing Variants of Goodhart's Law. *arXiv:1803.04585*. http://arxiv.org/abs/1803.04585

Ngo, R. (2020, September 28). *AGI safety from first principles: Introduction - AI alignment forum*. https://www.alignmentforum.org/posts/8xRSjC76HasLnMGSf/agi-safety-from-first-principles-introduction

Ngo, R., Chan, L., & Mindermann, S. (2023, February 22). *The alignment problem from a deep learning perspective*. arXiv. http://arxiv.org/abs/2209.00626

Oesterheld, C. (2017, August 10). *Multiverse-wide cooperation via correlated decision making*. Center on Long-Term Risk. https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/

Olah, C., Cammarata, N., Ludwig Schubert, Gabriel Goh, Michael Petrov, & Shan Carter. (2020). *Zoom in: An introduction to circuits* [Distill]. https://distill.pub/2020/circuits/zoom-in/

Omohundro, S. M. (2008). The basic AI drives. *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf

Ord, T. (2020, March 3). *Precipice*. Hachette Books.

Pace, B. (2019). Debate on instrumental convergence between LeCun, Russell, Bengio, Zador, and More. https://www.alignmentforum.org/posts/WxW6Gc6f2z3mzmqKs/debate-on-instrumental-convergence-between-lecun-russell

Pinker, S. (2018, February 13). *Enlightenment now: The case for reason, science, humanism, and progress* (Illustrated edition). Viking.

Russell, S. (2019, October 8). *Human compatible: Artificial intelligence and the problem of control*. Penguin Books.

Selsam, D. (2018, July 21). *The general intelligence hypothesis*. https://dselsam.github.io/posts/2018-07-08-the-general-intelligence-hypothesis.html

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022, November 2). *Goal misgeneralization: Why correct*

specifications aren't enough for correct goals. arXiv. http://arxiv.org/abs/2210.01790

Stein-Perlman, Z., Grace, K., & Weinstein-Raun, B. (2022, August 4). *2022 expert survey on progress in AI*. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

Tegmark, M. (2017, August 29). *Life 3.0: Being human in the age of artificial intelligence*. Penguin Books Limited.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*, 350–354. https://doi.org/10.1038/s41586-019-1724-z

Yudkowsky, E. (2008, July 3). Artificial intelligence as a positive and negative factor in global risk. In Nick Bostrom & Milan M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Machine Intelligence Research Institute. https://doi.org/10.1093/oso/9780198570509.003.0021

Yudkowsky, E. (n.d.-a). *AI safety mindset*. https://arbital.com/p/AI_safety_mindset/

Yudkowsky, E. (n.d.-b). *Omnipotence test for AI safety*. https://arbital.com/p/omni_test/

Yudkowsky, E. (n.d.-c). *Querying the AGI user*. https://arbital.com/p/user_querying/

Yudkowsky, E. (n.d.-d). *Strong cognitive uncontainability*. https://arbital.com/p/strong_uncontainability/