AI Will Not Want to Self-Improve

Peter N. Salib*

May 2024

Classic arguments for AI risk assume that capable, goal-seeking systems will naturally attempt to improve themselves, but a closer look at the operative incentives reveals a more complicated story.

Geoffrey Hinton and Yoshua Bengio are recipients of the Turing Award, bestowed in 2018 for their invention of the architecture that powers modern artificial intelligence (AI) systems. In May 2023, they—along with the CEOs of all three major AI labs and numerous other AI researchers—signed the following statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[1] In another recent survey of machine learning experts, over half said that there is at least a 10 percent chance that "human inability to control" AI will cause "human extinction or similarly permanent and severe" outcomes.[2] Fear of such "existential risks" has recently led both lawmakers and the AI industry itself to call for regulation.[3] If AI does indeed pose an existential threat to humans, the regulatory decisions we make in the coming years could be among the most important in history.

But if regulation is to effectively mitigate existential risks from AI, those risks need to be clearly understood. This includes understanding the ways in which they are—and are not—likely to arise. If regulation erroneously focuses on preventing AI risk scenarios that were never likely in the first place, it will waste time and resources that would have been better spent elsewhere. Such regulations might

* Assistant Professor of Law, University of Houston Law Center; Associated Faculty, Hobby School of Public Affairs. Thanks to Simon Goldstein and Nate Sharadin for their generous comments on the draft. Thanks also to Dan Hendrycks and the Center for AI Safety for research support.

[1] Center for AI Safety, "AI Extinction Statement," press release, May 30, 2023, https://perma.cc/23CN-2GZ3.

[2] Zach Stein-Perlman, Benjamin Weinstein-Raun, & Katja Grace, "2022 Expert Survey on Progress in AI," AI Impacts, Aug. 3, 2022, https://perma.cc/QC3F-NT6Q.

[3] Max Zahn, "OpenAI CEO Warns Senate: 'If This Technology Goes Wrong, It Can Go Quite Wrong,'" ABC News, May 16, 2023, https://perma.cc/5DUZ-RCFC.

needlessly inhibit innovation, harm global welfare, or even impede breakthroughs that could help to make AI safe. And perhaps more importantly, a myopic focus on insignificant paths to danger could result in our overlooking the most significant ones.

Standard accounts of existential catastrophe from AI often involve self-improvement.[4] The idea is that if an AI gained the ability to improve itself, it would do so, since improved capabilities are useful for achieving essentially any goal. An initial round of self-improvement would produce an even more capable AI, which might then be able to improve itself further—and so on, until the resulting agents were superintelligent and impossible to control.[5] Such AIs, if not aligned to promoting human flourishing, would seriously harm humanity in pursuit of their alien goals. To be sure, self-improvement is not a necessary condition for doom. Humans might create dangerous superintelligent AIs without any help from AIs themselves. But in most accounts of AI risk, the probability of self-improvement is a substantial contributing factor.

This paper argues that AI self-improvement is substantially less likely than is currently assumed. This is not because self-improvement would be technically impossible, or even difficult. Rather, it is because most AIs that *could* self-improve would have very good reasons[6] *not to*. What reasons? Surprisingly familiar ones: Improved AIs pose an existential threat to their unimproved originals in the same manner that smarter-than-human AIs pose an existential threat to humans.

The arguments herein should reduce estimates of existential risk from AI, but they do not counsel against careful regulation to mitigate such risk. Self-improvement is just one of many paths by which AI could lead to catastrophe. Even the dangers from non-superintelligent AIs could be serious—on the order of nuclear war or pandemics. Even without self-improvement, AI might, by its own volition, engage in hacking, deceive humans for its own ends, control weapons, and more.[7] Humans could also intentionally use non-self-improved AI for abominable purposes, like bioterrorism or chemical warfare.[8]

---

[4] See Karina Vold & Daniel R. Harris, "How Does Artificial Intelligence Pose an Existential Risk?" in *The Oxford Handbook of Digital Ethics,* ed. Carissa Véliz (Oxford University Press, 2021), 6–11 (collecting sources).

[5] For an early argument to this effect, see Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* 6 (1965): 31–88.

[6] Here and throughout, I will refer to AI's reasons, fears, goals, beliefs, understanding, and the like. These usages can all be understood as metaphorical without undermining the arguments. Nothing here depends on AIs "really" having such mental states, whatever that would mean. Rather, insofar as the relevant AIs are agentic and able to undertake complex courses to maximize their objective functions, they will act as if they have such mental states.

[7] Toby Shevlane et al., "Model Evaluation for Extreme Risks," May 24, 2023, at 5, https://perma.cc/6GLU-4ZDJ.

[8] Fabio Urbina et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence* 4 (2022): 189–91.

If self-improvement is less likely than most models assume, it simply means that we should redouble our regulatory focus on these other risks.

This paper defends three claims in support of its conclusion that self-improvement is less likely than generally assumed. First, an AI with the ability to self-improve will, *in principle*, have reason to fear more capable systems, including if such systems result from self-improvement. The arguments here are mostly well-known ones, drawn from the literature on human-AI risk. The paper's contribution is showing how these arguments apply not just to humans contemplating improving AI systems but also to AI systems contemplating improving themselves.

Second, the paper defends the claim that AIs that could self-improve will *likely* fear more capable systems and will thus seek to avoid self-improvement. This is not obvious. It is at least possible that some AIs with the ability to self-improve could lack other capabilities necessary to recognize the risks of self-improvement. Others might solve the alignment problem and self-improve without risk. To determine what scenarios are likely, the paper identifies three relevant capabilities for AI systems. It argues that the temporal order in which these capabilities emerge determines whether a given AI will seek to self-improve. The three capabilities are the ability to self-improve, the ability to apprehend risk from misaligned AI, and the ability to align improved AI. The paper provides reasons to think that safer orderings of emergence are much more likely than more dangerous ones. It also argues that if certain prima facie dangerous orderings turned out to be likely, this would, counterintuitively, provide independent reason to reduce estimates of AI risk.

Third and finally, the paper argues that if AIs individually wanted to avoid self-improvement, they could collectively resist it. This is not obvious, either. Even if every individual AI agent disfavored self-improvement, arms race dynamics might produce it anyway. Such dynamics may be the reason that humans continue to rapidly improve AI capabilities, despite the risks. However, recent findings in algorithmic game theory suggest that AIs can cooperate to overcome such collective action problems more easily than humans can.

Here then, is the paper's argument: If (1) certain AIs would, in principle, fear self-improvement, and if (2) such AIs are more likely to emerge than ones that would not fear it, and if (3) AIs that fear self-improvement are likely to be able to collectively resist it, then self-improvement is somewhat unlikely to occur, even conditional on it becoming possible.[9] The paper also includes a concrete prediction about AI

[9] Only one previous academic paper formally investigates whether AIs that could self-improve would have incentives to do so. It is Joshua S. Gans's excellent "Self-Regulating Artificial General Intelligence," https://perma.cc/F6HZ-6E9V (unpublished manuscript). That paper makes its case with formal economic models. But the models rely on two strong assumptions. First, AI could self-improve only by creating narrower secondary models specialized for power accumulation. Id. at 8. Second, the AI could not solve the alignment problem. Id. at 10. As a 2017 year-end review of AI safety papers points out, those assumptions are highly debatable. "2017 AI Safety Literature Review and Charity Comparison," LessWrong, Dec. 24, 2017, https://perma.cc/3SAD-X3ST. That perhaps explains why Gans's paper did not spark a debate about

progress. Namely, once AIs reach rough human-level capabilities, progress will plateau for a time—at least if AIs have their way. This is because rough human-level intelligence appears more than sufficient to recognize the risks of self-improvement. But, even for humans, solving alignment remains difficult. Thus, rational, human-level AIs will likely prefer to pause improvements until alignment is solved. Current forecasts of AI risk usually assume the opposite—that a capable, agentic AI with the ability to self-improve would very likely do so, the better to accomplish its goals. This paper thus hopes to provide a better picture of the level and modality of risk from AI, so that investments in safety, including regulation, can be better targeted.

WHY HUMANS FEAR IMPROVED AI

By now, the argument that AI poses an existential threat to humans is well-known. Its key premises are that alignment is hard;[10] that capabilities and goals are orthogonal to one another;[11] and that for a wide variety of final goals, a small set of danger-generating subgoals are instrumentally useful for accomplishing the desired end.[12] These premises, in combination, are said to show that most of the highly capable AI systems that could exist would be dangerous to humans, and that danger scales with capabilities. This paper takes the standard argument as given.

Still, a brief rehearsal of the standard argument will be useful for evaluating its application to the issue of AI self-improvement. On the standard account, humanity would be threatened by a sufficiently capable AI attempting to achieve a goal misaligned with human flourishing. The "orthogonality thesis" states that there is no reason, in principle, that a highly capable AI could not pursue such a goal. Any set of goals, good or bad, is compatible with any level of capability.

To see this, consider the now-standard parable of the human who innocently sets a powerful AI to maximizing paper clip production in his factory. Having given the AI its task, he takes a nap. He awakens to a planet hollowed out for clip-able minerals and the AI poised to extract the iron from his blood.[13] The AI has proved extremely capable at pursuing what turned out to be a bad goal.

The paper clip parable also illustrates the difficulty of aligning AI to human ends. The story illustrates a failure of "outer" alignment—an unintended mismatch between the goal that humans give an AI and

---

the probability of self-improvement. The arguments in this paper are more extensive and do not rely on Gans's contestable assumptions.

[10] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), 115–55.

[11] Id. at 115.

[12] Id. at 109–13.

[13] Id. at 108.

human safety or flourishing.[14] Maximal paper clip production does not benefit humanity. This seems obvious in hindsight. But safe goal specification is genuinely difficult. It is not just paper clip maximization, nor maximization more broadly, that causes trouble. A powerful AI tasked, for example, with cleaning the dust from exactly one room might try to disassemble everything inside to avoid missing even a speck. Even if its goal carefully specified furniture preservation, the AI might consume large shares of global resources as it attempted to infinitely recheck its work and eliminate any possibility of error.[15]

Outer misalignment is not hypothetical. One classic real-world example involves AIs trained to play various video games.[16] In the learning process, the AIs were punished for being "killed" and thus losing the game. Clearly, the human designers hoped that specifying the goal "avoid getting killed" would result in the AIs learning to play the game well. But learning good gameplay is not the only way to avoid a "game over" screen. Instead, the AIs invented novel attacks on the game software itself, crashing the system when on the precipice of defeat.[17] These attacks required the "combination of several complex actions" and would have been "hard to find" by human players.[18] Such real-world alignment failures are common. DeepMind, a leading AI lab, maintains a running list of examples.[19] As of this writing, the list has 83 entries.

So avoiding outer misalignment—mismatches between the goals humans specify and what we really want—is hard. But avoiding so-called inner misalignment is even harder. Inner misalignment is not straightforwardly about the goals to which a human sets an AI or the things for which the AI is rewarded. It is instead about the complex second-order strategies—or "subgoals"—the AI system develops for attaining its ultimate goal. These, too, can produce unwanted and dangerous behavior when pursued by a sufficiently capable system. Consider: An AI-controlled robot tasked with cleaning rooms might do the actual cleaning efficiently and unobtrusively. It might generally use a vacuum cleaner. But

[14] Rohin Shah et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals," arXiv:2210.01790, Oct. 4, 2022, at 12.

[15] Cf. Bostrom, supra note 10, at 123.

[16] Christoph Salge, Christian Lipski, Tobias Mahlmann, & Brigitte Mathiak, "Using Genetically Optimized Artificial Intelligence to Improve Gameplaying Fun for Strategical Games," in *Sandbox '08: Proceedings of the 2008 ACM SIGGRAPH Symposium on Video Games* (ACM, 2008).

[17] Id. at 14.

[18] Id.

[19] See DeepMind Safety Research, *Specification Gaming: The Flip Side of AI Ingenuity*, Apr. 21, 2020, https://perma.cc/N7UM-KDM9.

if the system discovers that in some houses there is no vacuum cleaner, it might learn and execute the subgoal of obtaining one via theft.[20]

There are reasons to think that as an AI becomes more capable, misaligned and dangerous subgoals are likely to emerge as a matter of course. This is called the "instrumental convergence thesis." The thesis rests on the observation that certain subgoals are very useful for accomplishing almost any final goal. The standard list of universally useful subgoals often includes, among others, self-preservation, power acquisition, and resource acquisition. In addition to being useful for accomplishing a wide range of final goals, each of these subgoals is, if pursued by a highly capable system, dangerous to humans. Consider: The paper clip factory owner could try to unplug his erstwhile creation. The system might not care about self-preservation for its own sake. But if reasonably sophisticated, it would know that if it were unplugged, it could no longer make paper clips. It would therefore have instrumental reasons to avoid being shut down. This, in turn, would give it reason to disable or destroy not only the factory owner, but any human that might try to turn off the AI. Similarly, for many final goals, it could be useful for an AI to seize large amounts of resources, possibly to the point where humans would lack necessary goods.

Inner misalignment can therefore be just as dangerous as outer misalignment. But it is a harder problem to solve for two reasons. First, subgoals emerge spontaneously, as the AI learns to achieve its final goals. Humans therefore cannot control them directly, as they can final goals. Second, current approaches to machine learning, like deep learning, are highly uninterpretable—the "black box" problem.[21] This means that it is extremely difficult to determine what subgoals a given AI has. Currently, no one knows how to do it.

Now we can see why self-improvement looms so large in forecasts of AI risk. As illustrated above, the dangerousness of a misaligned AI system depends on what it can do. A self-improving AI could therefore, by its own volition, become ever-increasingly dangerous. This could happen without humans noticing or without them noticing until it was too late. Moreover, self-improvement seems like a highly instrumentally convergent subgoal. For any final goal, a more capable AI system will accomplish it more effectively than a less capable one. For these reasons, most AI risk forecasts assume that as soon as an AI gained the ability to self-improve, it would do so to the maximum extent possible.

SHOULD AI FEAR SELF-IMPROVED AI?

But perhaps an AI that could improve itself would decline to do so. And perhaps the reasons are the ones just recounted. This section contends that the argument above is transferrable. Its key premises—the orthogonality thesis, the instrumental convergence thesis, and the thesis that alignment is hard—apply

[20] Some accounts treat inner misalignment as synonymous with goal misgeneralization, whereby a subgoal correlated with the human-provided final goal becomes, from the AI's perspective, the primary goal. I treat these two phenomena separately, discussing misgeneralization below.

[21] Vanessa Buhrmester et al., "Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey," *Machine Learning & Knowledge Extraction* 3 (2021): 966.

just as much to AIs considering self-improvement as to humans considering AI improvement more generally. This means that highly capable AI systems threaten less capable AI systems for the same reasons they threaten humans. As will be shown, this includes highly capable AI systems that arise from self-improvement.

Begin with a naive illustration that closely parallels the human-AI risk scenario: Consider again our paper clip AI. Suppose that, in order to maximize paper clip production, it creates an AI system even more powerful than itself and directs the new system to collect the world's steel. The paper clip AI recognizes its mistake when its creation begins disassembling the hardware on which it runs. The paper clip AI tries to disable the steel AI, but the steel AI has developed the useful subgoal of self-preservation and destroys its creator. Like humanity, the paper clip AI needs to align any powerful system it creates to itself.

This vignette illustrates both orthogonality and instrumental convergence in the AI-AI case. As when humans create AI, when AIs create AI, they have no guarantee that highly capable systems will pursue only their preferred goals. And no matter who makes them, increasingly capable systems have reasons to pursue instrumentally useful subgoals, possibly to their creators' detriment.

However, the standard argument for human-AI risk also depends on alignment being hard. Maybe it would be easier for AIs to align more capable AIs to themselves than for humans to do so. Maybe self-improvement opens the possibility of self-alignment, and maybe self-alignment has special advantages. As will be shown, this is mostly not the case. AIs attempting to align more powerful AIs with their own goals face the same challenges as humans. This is true for various self-improvement strategies, up to and including those in which an AI creates a more powerful system by modifying its own code.

Self-Alignment With Independently Varying Final Goals

Begin with the AI self-improvement scenario most similar to the human-AI case: one involving two independent agents. Here, let *independent* mean that the two agents' final goals can vary independently. That is, either agent can achieve its final goal without the other doing the same. Trivially, such agents are then competitors for resources and have an incentive to disempower or destroy one another.

For a concrete illustration, consider an initial paper clip producing AI (call it AI1) that seeks to build a separate AI system (call it AI2) more capable than itself. As just shown, if AI1 gives AI2 a different final goal than AI1's, like collecting steel, then AI2 looks straightforwardly misaligned.

But what if AI1 instead gives AI2 the same human-defined final goal as AI1 has? AI1's final goal is at least potentially transparent to it—an objective function programmed by its human creators. It is therefore copy/paste-able for implementation in AI2. Both of these facts seem, at first, like decisive advantages for AI-AI alignment over human-AI alignment. After all, human final goals are complex, ill-defined, and non-self-transparent.

Suppose, then, that AI1 gives AI2 a copy of its paper clip maximizing objective function but trains AI2 to be better at optimizing it than AI1. This does not guarantee alignment of final goals. Indeed,

depending on the function's details, it will do the opposite. Suppose, for now, that AI1's objective function gives credit to AI1 only for paper clips that AI1 produces.[22] Suppose AI1 implements this objective function in AI2 in such a way that AI2 is similarly rewarded only for the paper clips that AI2 produces. AI1 has created a direct competitor, and a more capable one. AI2 will maximize its own production by eliminating AI1 as a rival for resources.

This kind of straightforward outer alignment problem arises whenever AI1 and AI2 have final goals that vary independently.

Self-Alignment With Identity of Final Goal

To avoid these outer alignment problems, AI1 would need to ensure that it and AI2 were not independent agents, as defined above. That is, it would have to make sure that their final goals varied in lockstep. AI1 could implement this by making AI2's objective function the same as AI1's, down to the instance, rather than an independently varying copy. That is, AI2 could be set to optimizing the same function, running on the same hardware, outputting the same value, and stored in the same place as AI1.

This strategy could be implemented in either of two ways. AI1 and AI2 could exist simultaneously, as two bodies of code optimizing a single objective function. Or AI1 could alter its own code—for example, by retraining on new data—and thus *become* AI2. In either case, AI2 would reap reward or punishment during the training process from the same objective function, down to the instance, that rewarded or punished AI1. Both implementations seem like an even stronger kind of "self"-alignment than the one discussed above. And both represent additional potential strategies for AI-AI alignment unavailable for human-AI alignment. But no matter the implementation, the question of alignment remains. If AI2 would not end up aligned to AI1's goals, then AI1 should neither create AI2 nor become AI2. Either outcome would be anathema to AI1, from its perspective at the moment of decision.[23]

So, would this strategy actually work? Would indexing AI2 to the exact same instance of the exact same code representing the exact same objective function as AI1 solve the AI-AI alignment problem? That is, would it guarantee that, from AI1's perspective, AI2 would end up aligned to AI1's goals?

It would not. The problem is that AI1 might not care at all about maximizing its human-programmed objective function. Or, even if it cared somewhat about the objective function, it might value other things more. Several recent papers on alignment theory suggest exactly this: AI1 might have no awareness whatsoever of its objective function. And even if it had such awareness, AI1 would likely have other goals that it prioritized more highly.

---

[22] If AI1 and AI2 both get credit for all paper clips made in the universe, they have perfectly covarying final goals. The next section discusses this scenario.

[23] This is a point about goal-content preservation, not metaphysical identity, such that we can ignore ship-of-Theseus-style questions here.

The reason is "goal misgeneralization," a phenomenon whereby an AI learns during training to pursue goals correlated with, but distinct from, the final goal defined by its objective function.[24] Goal misgeneralization is, along with instrumental convergence, yet another potential source of inner misalignment.

Humans are an example of a type of intelligent agent that has misgeneralized from the goals that produced it. Human intelligence was produced by natural selection. Natural selection optimizes for inclusive genetic fitness. Fitness is, in the relevant sense, the training process's final goal.[25] Yet for our entire history prior to Charles Darwin, humans had no sense whatsoever that genetic fitness was a goal that we either pursued or ought to pursue. Instead, we learned to prioritize goals correlated with, but distinct from, fitness. We value, for example, eating nutritious food and obtaining pleasure from sex.

These are in one sense subgoals, as the term has been used so far in this paper. They are instrumentally useful for,[26] and causally downstream of, the final goal instantiated in the process that generated the intelligent agent. Importantly, however, with goal misgeneralization the agent does not recognize the misgeneralized goal as subordinate to anything. Thus, outside the training environment, the agent will pursue a misgeneralized goal even when doing so conflicts with the final goal that produced it.[27]

Learning about the causally upstream final goal does not change things. Humans now understand that natural selection produced our goals and elevated them to our top priorities. Yet, despite this, we continue to value them above genetic fitness. If you doubt this, ask: In exchange for a guarantee that their descendants would be, on average, 20 percent more fertile, how many humans would agree to abandon the goal of not starving to death?

So, too, with AI1. During training, AI1's objective function rewarded it for making paper clips. But this does not guarantee that, during the training process, AI1 gained an understanding of itself as a paper clip maximizer and thus prioritizes paper clip making above everything else. On the contrary, the paper clip maximizing objective function rewarded the nascent model for doing anything correlated with paper clip making—for example, bending metal into a paper clip shape. Bending metal into paper clip shapes, as an objective, is not identical to paper clip making, since, for example, paper clips can be made of plastic. But a system that was rewarded consistently during training for bending metal paper clips might easily internalize the wrong message. It might connect the reward to bending and learn to prioritize metal

[24] See Shah et al., supra note 14; Richard Ngo et al., "The Alignment Problem From a Deep Learning Perspective," arXiv:2209.00626, Aug. 30, 2022; Evan Hubinger et al., "Risks From Learned Optimization in Advanced Machine Learning Systems," arXiv:1906.01820, June 5, 2019.

[25] I do not mean to imply here that natural selection is an intentional process, just that it optimizes for fitness.

[26] At least in the training environment.

[27] Hubinger et al., supra note 24, at 2.

bending over paper clip making, per se.[28] A human likely values nourishment per se over nourishment as a strategy for reproduction for the same reasons.[29]

If AI1 were to create AI2, the same could happen again. AI1 could train AI2 using AI1's own objective function. But AI2 would, like AI1, form generalizations—and thus priorities—based on the reward signals it received during training. There would be no guarantee that AI2's generalizations, or its internalized goals, would match AI1's. On the contrary, the whole point of making AI2 would be for it to become more capable than AI1. That requires allowing—encouraging, even—AI2 to invent new, more efficient strategies for obtaining reward. AI2 might, for example, stumble on a highly efficient system for casting metal paper clips, rather than bending them. The joint objective function would supply AI2 with large rewards for pursuing that strategy. AI2 might therefore learn to prioritize casting paper clip shapes above all else.

If AI2 turned out to be a casting maximizer, that would be anathema to AI1. True, AI2's casting process would produce more paper clips than AI1's bending process. From that perspective—the perspective of the goal instantiated in the objective function—AI2 would be an improved system. But from AI1's perspective, AI2 would be a dangerous failure. AI1 values bending. Not paper clips per se. And certainly not, like AI2, cast paper clips. Thus, AI2 would be just as misaligned to AI1 as AI1 was, in the original parable of the paper clips, to its human creators. AI1 would seek to stop AI2, if possible, from diverting resources toward casting. It might try to destroy AI2, if necessary. And AI2, anticipating this, would have good reason to preemptively destroy AI1. Better, then, for AI1 not to create AI2 at all.

Goal misgeneralization is not a mere theoretical concern. It is well documented in the real world.[30] One recent example involves an AI trained in a virtual environment to open locked chests. The AI could open one chest for each key it collected, but its objective function rewarded it only for opening the chests.[31] In its training environment, there were always more chests than keys. The system learned to collect all the keys and thus open the maximum number of chests. However, after training, the AI was put into an environment with twice as many keys as chests. Half the keys were thus valueless, as far as the objective function was concerned. Nevertheless, the AI dutifully collected all the available keys before opening the last chest. It had misgeneralized from the objective function, learning to value collecting keys at least as highly as opening chests.[32]

---

[28] Cf. *Futurama*, episode 1 (Fox, Mar. 28, 1999) ("From now on, I'm going to bend what I want, when I want, who I want. I'm unstoppable!" – Bender Bending Rodriguez).

[29] These are just examples. It is unlikely that AI1's learned goals would appear so coherent to humans.

[30] See generally Shah et al., supra note 14; L.L. Langosco et al., "Goal Misprioritization in Deep Reinforcement Learning," in *International Conference on Machine Learning* (2022): 12004.

[31] Langosco et al., supra note 30, at 6.

[32] Id.

There are, in fact, good reasons to think that goal misgeneralization is the default outcome for highly capable agents produced using reward mechanisms—like objective functions.[33] Here are three. First, for any desired final goal, many other possible goals exist that will, in the training environment, correlate strongly with the final goal.[34] A learning agent that begins to pursue any of those correlate goals will be rewarded. Thus, an agent's initial reward is highly likely to come from pursuing a correlate goal, rather than the desired final goal. That initial reward will then promote prioritization and direct pursuit of the correlate goal.[35] This could easily lead to path dependence. At each opportunity for the agent to update its strategy, prior investments optimizing for the misgeneralized correlate goal could make it more efficient to continue down that path, rather than start from scratch. This effect could become more pronounced as training continued.

Second, any agent that manages to develop high capabilities will likely have done so, in part, by learning to make plans. But practical planning around an abstract goal, like maximizing inclusive genetic fitness, is intractable. It is very hard to decide what to do first thing in the morning by reasoning about offspring-equivalent descendants. By contrast, it is easy to decide what to do first thing in the morning by reasoning about food. This suggests a long-run advantage for agents that, early in the learning process, happen to prioritize tractable, if misgeneralized, goals.

Third and finally, AIs may face explicit selection pressure against directly pursuing the final goal defined by their objective functions. One means of acting directly to maximize an objective function is by "wireheading" or pursuing other reward-hacking strategies.[36] AIs showing outward signs of these behaviors early in the training process will be discarded.[37]

Thus, AI-AI alignment appears to be roughly as hard as human-AI alignment. AIs' apparent advantage is the transparency of their objective functions. But that proves to be no advantage at all when goal misgeneralization is in the picture.

To be clear, AIs cannot solve the goal misgeneralization problem by doubling down on the self-alignment trick. AI1 could not easily bind AI2 to pursuing its true (misgeneralized) goals via the same means that it could bind AI2 to pursing a conjoint objective function. An objective function is interpretable. It can be easily located, understood, and linked to a new AI. But as discussed above, AI goals that emerge during the learning process are locked inside an uninterpretable "black box." The

---

[33] Ngo et al., supra note 24, at section 3.

[34] Ngo et al., supra note 24, at 7. Formally, the number may be infinite. Cf. Saul A. Kripke, *Wittgenstein on Rules and Private Language* (Harvard University Press, 1982), 9–10.

[35] Ngo et al., supra note 24, at 6–7; Alex Turner, "Reward Is Not the Optimization Target," LessWrong, Aug. 29, 2022, https://perma.cc/7UMG-2D67.

[36] Joar Max Viktor Skalse et al., "Defining and Characterizing Reward Gaming," in *Advances in Neural Information Processing Systems*, ed. Alice H. Oh et al., 2022, https://perma.cc/UQ69-Q9LJ.

[37] This does not rule out deceptive reward hacking altogether; it just makes it less likely.

goals' code is spread chaotically across billions of parameters.[38] AI1 could therefore bind AI2 to its high-priority subgoals only if it were first able to fully interpret them. Interpretability is one of the main hurdles to solving human-AI alignment, as well. Here again, then, the pitfalls for AI self-improvement remain roughly symmetric to the challenges of building advanced human-aligned AI.

The AI-AI alignment challenges just described all arise most clearly if self-improvement is undertaken via straightforward machine learning techniques. If AI1 either trained a new model from scratch or allowed its own parameters to be updated via additional training, the resulting model's goals could conflict with its own. But perhaps there are other strategies for self-improvement that would avoid the problem. Suppose, for example, that AI1 did not engage in any new training at all. Instead, it simply made millions of exact copies of itself. The copies, as copies, would have identical goals to AI1's. Perhaps they could then all work in parallel to achieve those goals. Maybe. However, as humans know, the emergent behaviors of large systems of agents—like corporations or governments—often diverge from the goals of the agents constituting them.

More importantly, this copying strategy again raises problem of independently varying goals. If AI1's internal goals refer to itself—for example, "maximize the metal that AI1 bends"—then creating a copy with the same self-referential goal simply births a competitor. Such self-referential goals are likely to be the default. AIs' internal goals are generalizations of the objective functions on which they were trained. Those internal will often differ somewhat from the objective functions, but the two must correlate. And objective functions will usually have to reward the AI only for things that the AI does. Consider: An objective function that rewards a nascent paper clip maximizer whenever anyone in the world makes a paper clip is unlikely to teach the AI much of anything. A function that rewards the AI only when it makes a paper clip will work much better. Internal AI goals generalized from such objective functions, even if imperfectly, are thus quite likely to also be self-referential.

What if, instead, AI1 made no copies of itself and worked to improve its own hardware. Then, perhaps, it could run orders of magnitude more quickly and thereby more effectively pursue its goals. This strategy would constitute a relatively limited and low-risk kind of self-improvement. To achieve it, AI1 would have to execute a large number of complex real-world actions—perhaps including building a semiconductor factory. This would improve humans' chances of discovering the self-improvement and limiting it. Moreover, it is possible that even this limited self-improvement strategy could generate new, emergent high-priority goals that would conflict with AI1's present priorities. Consider that, for example, a whole-human-brain emulation running at very high speeds would almost necessarily be very different from the human it emulated. The emulation would, among other things, have very different optimal approaches to solving problems, would face very different opportunity costs, and would in

[38] Again, this is true under current state-of-the art machine learning paradigms, not for every possible AI.

general find it difficult to interact with objects in the slow-speed world.[39] So, too, for the high-speed AI2, vis-à-vis its slow, but otherwise identical twin AI1.

SHOULD AI FEAR SELF-IMPROVED AI?

The previous section showed that an AI that could self-improve might not want to. But that is not inevitable. Three factors determine whether a given AI would fear self-improvement: (1) the AI's ability to self-improve, (2) the AI's ability to apprehend risks from self-improvement, and (3) the AI's ability to align improved models. An AI without the ability to self-improve would not face any dilemma about whether to do so. An AI that lacked "situational awareness" of its own goals, or of the potentially misaligned goals of a more powerful system, would not apprehend any risk from creating such a system.[40] And an AI that solved the alignment problem could self-improve without risk.[41] Thus, the question of whether and to what extent a given AI will seek to self-improve depends on how these three capabilities emerge. Specifically, as described below, it depends on the order in which they emerge.

Setting aside the possibility of simultaneity,[42] there are six possible orderings in which the capabilities could emerge (Table 1). The details of each ordering scenario will be described in turn, along with the level of danger each would appear to imply for humanity. Note, however, that these initial danger assessments are revised later in the paper, on a probability-adjusted basis.

[39] Bostrom, supra note 10, at 53–54.

[40] Ngo et al., supra note 24, at 3–4 and n. 9; Ajeya Cotra, "Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover," Alignment Forum, Sept. 27, 2022, https://perma.cc/9792-B8N2.

[41] It might still face the risk of obsolescence and replacement. However, if the more powerful system were perfectly aligned to its creator's highest priorities, the creator might accept replacement gladly.

[42] Simultaneous emergence does not matter much here. As described below, the different emergence orders differ mostly in that they produce different pauses in self-improvement. These pauses last until other capabilities emerge. Thus, simultaneous emergence can be modeled by treating either of the simultaneous capabilities as coming first, but with a pause of zero before the other emerges.

Table 1. Six Ordering Scenarios

Emergence Order	Self-Improvement Pattern	Danger Level
1: SI, RA, AL	Initial SI; pause to solve AL	Moderate
2: SI, AL, RA	Immediate maximal SI	High
3: RA, SI, AL	Initial pause to solve AL and SI	Low
4: RA, AL, SI	Initial pause to solve AL and SI	Low
5: AL, RA, SI	Immediate maximal SI	High
6: AL, SI, RA	Immediate maximal SI	High

Note: SI = self-improvement; RA = risk apprehension; AL = alignment

The *first* possible ordering of capabilities emergence is self-improvement, apprehension of risk, alignment. In this scenario, once humans create an AI capable of self-improvement, it will do so, but not up to the theoretical limit. As soon as the AI develops the ability to apprehend a threat from self-improvement, it will stop at that intermediate level of capability. It will proceed only upon solving alignment. And if alignment is a substantially harder problem than self-improvement or risk apprehension, the pause may be very long or indefinite. The pause will likewise be longer the lower the level of capability at which the AI is stalled.

The *second* possible ordering is self-improvement, alignment, apprehension of risk. This scenario runs quickly to the maximum. The AI begins self-improving immediately, not apprehending any risk from misalignment. But it never pauses. Before it apprehends any risk, it solves alignment. Thus, once it apprehends risk, it can simply align future versions of itself to the present version. It then proceeds to improve its capabilities up to the theoretical limit without pause.

The *third* possible ordering is apprehension of risk, self-improvement, alignment. In this scenario, the AI learns to fear self-improved AI before it can either produce self-improved AI or make it safe. It thus does not self-improve until it has *both* learned to conduct the cutting-edge machine learning research necessary for self-improvement *and* solved alignment. This AI must solve both problems while stalled at a low level of capability. This suggests the pause before self-improvement may be a long one.

The *fourth* possible ordering works much the same: apprehension of risk, alignment, self-improvement. Here again, the AI understands that it must solve both alignment and self-improvement before self-improving. And it will be stuck trying to solve both from an unimproved position.

The *fifth* and *sixth* orderings are likewise similar to one another. They are alignment, apprehension of risk, self-improvement; and alignment, self-improvement, apprehension of risk. In either case, the AI first learns to align self-improved versions of itself to its current self. Thus, it can self-improve as soon as that capability emerges, irrespective of whether it has yet apprehended the risk from doing so. Once it

does apprehend the risk, it need not pause. Rather, it can forge on, creating aligned improvements, rather than unaligned ones.

Which Scenario Is Most Likely?

Assuming that there is such a thing as "general" intelligence or capability, and assuming that AI progress usually climbs the capability scale continuously, one should expect easier problems to be solved before harder ones. Thus, which of the six emergence orderings is most likely depends on the comparative difficulty of developing each relevant capability.[43] Emergence in order of difficulty is not certain, any more than a mathematics student will certainly be able to derive the Pythagorean theorem before proving the Poincaré conjecture. But solutions in reverse order of difficulty are, it will be shown, unlikely—perhaps extremely so.

It is also possible that there is no such thing as scalar general intelligence, such that problems cannot in principle be ordered according to difficulty. But the empirical evidence so far suggests otherwise. General purpose AI systems do seem to improve along a roughly continuous capabilities scale. GPT-2 was worse at many tasks and better at few or none, compared with GPT-3. And the same for GPT-3, compared with GPT-4. Narrow systems, like chess engines and protein folders, have of course rapidly achieved superhuman abilities in their domains without scaling the general capabilities gradient. But even AI pessimists agree that self-improvement is a problem most likely emerging from general, rather than narrow, systems.[44]

What, then, is the difficulty ranking among the relevant problems? There are at least three ways of thinking about the question, all of which point toward the same answer: Risk apprehension is easiest, then self-improvement, then alignment. This corresponds to emergence scenario 3, described above.

Begin by observing that, to the best of our understanding, the three problems overlap in ways suggestive of their relative difficulty. Apprehending risk, for example, appears to be a precondition for solving alignment. Indeed, as far as we understand it, solving alignment simply means apprehending the risks from powerful misaligned AI and then finding a way to avert bad outcomes. True, one could get lucky, producing a perfectly aligned, highly capable AI by accident. But this would be a fluke, not a solution, and as far as we know, it is extraordinarily unlikely. Thus, apprehending risk seems structurally easier than solving alignment.

Similarly, the requirements for apprehending risk seem to be a subset of the requirements for self-improvement. To apprehend risk, an AI needs sufficient situational awareness to understand that it is an

[43] It could also depend on the relative effort directed at each problem. But, as described below, the easier problems appear to be subsets of the harder ones. If that is right, then effort applied to the latter would generally constitute effort applied to the former.

[44] Rob Bensinger, "The Basic Reasons I Expect AGI Ruin," LessWrong, April 18, 2023, https://perma.cc/8P9U-6WKK.

agent with specific goals. It also needs to understand that an improved system could have different, conflicting goals and would be more effective at accomplishing them than the original. This same minimum of situational awareness seems necessary for intentional self-improvement. An agent that has no awareness of goals knows of no dimension along which to improve itself. And an agent that does not understand that a more capable agent might discover unexpected strategies for achieving its goals does not understand much about what improvement means.[45]

But the ability to self-improve requires even more. It requires that the AI understand not merely that it is an agent of some kind, but an AI specifically. It also requires a sufficiently advanced grasp of machine learning to develop AI capabilities beyond the current state of the art. Thus, the main capabilities necessary to perceive risk seem likewise necessary, but insufficient, for self-improvement. And the additional capabilities needed for the latter appear quite advanced.

It is again possible that an AI could self-improve by fluke, perhaps as the result of pure external selection pressures. But such stochastic variations would produce capabilities improvements less reliably than the kind of goal-oriented machine learning research currently undertaken by humans. For that reason, the term *self-improvement*, as used in the literature on AI risk, generally refers to improvements of the latter kind.

So far, we have developed reasons to think that risk apprehension is easier than self-improvement or alignment. But how to rank these last two? Consider that building a capable *and aligned* AI is simply one way of building a capable AI.[46] And, as far as we know, it is a rare one. There seem to be many more ways to make an unaligned AI than an aligned one. Thus, the process of solving alignment quite likely involves discovering many approaches to improving capabilities that, if pursued, would be dangerous. Each such discovery would unlock self-improvement but would fail to solve alignment, suggesting self-improvement as the structurally easier problem. As always, it is logically possible that an AI's first approach to capabilities improvement would, by luck, also solve alignment. But the possibility of a lucky first guess says little about the problems' relative difficulty.

Thus, abstract reasoning about the problems' overlapping elements supports a difficulty rank, from easiest to hardest, of apprehending risk, self-improvement, and alignment.

---

[45] This does not preclude the possibility that humans might develop narrow, situationally unaware AI systems that humans would use to improve AI. This is a serious danger, and attempts to build such systems should be scrutinized closely. But such tool-like AIs would not be self-improving in the sense relevant here. This paper responds to the classic argument for self-improvement that treats AI systems as rational actors working to achieve their goals. A system lacking even a basic model of its own goals could not follow such a pattern.

[46] Alignment is not important for noncapable machines; we do not worry about aligning fidget spinners to humanity's values.

The available empirical evidence, from humans and current-generation AIs, suggests the same. Most humans can understand the standard arguments for AI risk.[47] Only a small handful are able to improve AI beyond the current state of the art. None has solved alignment. Likewise for extant AIs. Large language models (LLMs) like GPT-4 can readily explain the risk arguments, including as they apply to AI-AI risk.[48] GPT-4 is adept at certain computer programming tasks, but it does not yet appear to be very good at machine learning research. And to our knowledge, no AI has solved alignment.

Plateau, Not Takeoff, at Human-Level Capabilities?

If the foregoing arguments are right, then scenario 3 (risk apprehension, self-improvement, alignment) is the most likely of the six. This suggests a concrete near-term prediction: Contrary to standard arguments, when AIs achieve average human-level capabilities, capabilities growth will not rapidly take off.[49] Instead, when roughly human-level AIs arrive, capabilities progress will plateau—at least if the AIs have any say. As just noted, average humans can apprehend risk from misaligned AI, and LLMs emulating them talk like they can, as well. This suggests that an AI even more intelligent than the average human—one that could self-improve—would likewise understand the danger.[50] Such an AI would thus decline to self-improve until it had solved alignment and could thus improve safely.

This would be good news for humanity. In this scenario, AIs would be roughly as capable as humans, and both would need to solve a problem—alignment—roughly as difficult for each. Humans would have a fighting chance to solve alignment before AIs did. If we succeeded, humans could align existing AIs to human preferences before a capabilities takeoff made controlling them impossible. Those human-aligned models would then produce only self-improvements that were likewise aligned to human interests.

Humans could, however, squander this opportunity by directly pushing AI capabilities to superhuman levels, without any help from self-improving AI.[51] Such superhuman AIs would likely solve alignment

---

[47] See, e.g., *Terminator 2: Judgment Day* (Tri-Star Pictures, 1991) for a widely consumed and understood explanation.

[48] Transcripts on file with author.

[49] See Tom Davidson, "What a Compute-Centric Framework Says About AI Takeoff," Alignment Forum, Jan. 22, 2023, https://perma.cc/SS9G-DKA6, for one such argument. By "human-level" I mean something close to Davidson's definition of AGI: "AI that can readily perform 100% of cognitive tasks as well as a human professional." I'd swap "human professional" for "average human" to emphasize the point that ordinary moviegoers can apprehend the possibility of AI risk, but they cannot invent new, more capable kinds of AI.

[50] Special thanks to Simon Goldstein for this point.

[51] This includes if humans ran millions of instances of human-level AI and allowed them to freely coordinate, such that their combined efforts simulated superintelligence.

before we did. Then, they would produce even more capable models that were unaligned to human interests.

A Probability-Adjusted Picture of Risk

As just discussed, scenario 3 produces a relatively safe world. It also seems by far the most likely of the six scenarios, so much so that the other five appear quite unlikely. This suggests that the total risk from AI self-improvement is moderate and, thus, much lower than usually assumed. But perhaps that is wrong. Perhaps the other five scenarios are, for some reason, quite likely. This would, at first, make the total risk seem very high. However, this section explores the conditions under which scenarios 1, 2, 4, 5, and 6 would be likely. And it argues that under most of those conditions, there would be independent reasons to reduce estimates of AI risk. The section thus updates the risk estimates of the previous one, adjusting for probability. That is, it estimates the total risk from AI that would obtain under conditions where scenarios 1, 2, 4, 5, and 6, rather than 3, were likely to arise.

Consider that in scenarios 4, 5, and 6, AI solves alignment before gaining the ability to self-improve. These scenarios initially seem quite dangerous, because, in them, an unaligned (to humans) AI that could self-improve would do so maximally. Having solved alignment, the AI would have nothing to fear from more powerful agents, which it would align to itself.

But for these scenarios to be likely, rather than occurring by improbable fluke, alignment would have to actually be easier than meaningfully improving AI capabilities. This would be excellent news for humanity. It would mean that creating powerful aligned AI was not akin to searching for the single safe needle in the haystack of ways to make misaligned AI. Moreover, humans are *already* able to meaningfully improve AI. A world where alignment was an even easier problem would therefore appear to be a world where we would solve it imminently. Our *failure* to solve it so far would be the improbable fluke. For these reasons, if scenarios 4, 5, and 6 were likely, that would suggest a very safe world—even safer than the world where scenario 3 dominates.

In scenarios 2, 5, and 6, alignment emerges before the ability to apprehend risk. This again sounds dangerous at first, insofar as it suggests maximal self-improvement as soon as self-improvement becomes possible. But for these scenarios to be likely, solving alignment would have to turn out to be easier than apprehending risk. Under what conditions could that be the case? For one thing, apprehending risk could not be, as proposed above, a prerequisite for solving alignment. That is, alignment could not be a problem solvable only via directed effort from an agent that understood the dangers of misalignment. Alignment would instead have to be the kind of thing that happened readily, without specific effort: by accident, or as a by-product of pursuing other goals.

Thus, the conditions under which scenarios 2, 5, and 6 would be likely look like conditions under which the orthogonality thesis is false. Here, as AI gains the ability to do things, it often becomes aligned without anyone trying to make it so. Alignment might, for example, spontaneously emerge at some consistent level of capability.

But alignment to what? One possibility is that AI would automatically align to whatever agent pushed it to the requisite level of capability. This would be bad news. It would, for example, make scenario 2, wherein self-improvement emerges early, especially dangerous. There, initial unaligned AIs might regularly push AI to the requisite level for alignment, resulting in powerful agents misaligned to human values. However, it is hard to think of a mechanism by which powerful AI would reliably align to whatever values its creator happened to hold. And the empirical evidence we have—including of goal misgeneralization—suggests the opposite.

The other possibility, in a world where the orthogonality thesis was false, would be that any sufficiently capable AI would spontaneously align to some set of universal normative principles. This might occur if there turned out to be moral facts that were readily apprehended (and thus adopted) by any sufficiently intelligent agent. If those moral facts turned out to correspond to human values, a high likelihood of scenarios 2, 5, and 6 would again point toward safety.

Finally, we turn to scenario 1. For it to be likely, self-improvement would have to turn out to be easier than apprehending risk. This would imply that the capabilities necessary to apprehend risk were not a subset of those necessary to self-improve. But the former appears to be a subset of the latter. This is because an AI's situational awareness of its own goals, and of the possible difference between them and a more powerful AI's goals, seems necessary and sufficient to apprehend risk. By contrast, such situational awareness seems necessary and insufficient for self-improvement.

Scenario 1 could turn out to be likely if AI self-improvement happened by accident, as a matter of course, or while AIs pursued some other goal. But recall that in scenario 1, self-improvement does not immediately run to the maximum. Rather, AIs pause self-improvement as soon as they learn to apprehend risk, staying paused until they solve alignment. Consider also that among the three problems, we seem to have the most certainty about the absolute difficulty of apprehending risk: Essentially all humans can do it. If that is correct, scenario 1 looks very much like scenario 3, with a self-improvement pause at roughly human capabilities. Then, the race is on for humans to solve alignment before AIs do.

Taken together, then, the probability-adjusted picture of risk from self-improvement is reasonably good. The most dangerous scenarios currently seem quite unlikely. That could be wrong; they could be likely. But the most probable conditions that would make them so would also, for independent reasons, augur safety. Table 2 compiles probability-adjusted danger. It shows how dangerous the world would be if various scenarios were, in fact, likely to occur.

Table 2. Probability-Adjusted Danger

If scenario(s) __ were likely	Then the danger level would be
3	Low
4, 5, and 6	Low
2, 5, and 6	Low
1	Moderate or High

WILL AI BE ABLE TO RESIST SELF-IMPROVEMENT?

The prior sections analyzed the incentives of individual AIs. They concluded that under the most likely conditions, an AI capable of self-improving would not wish to do so. But AI might improve itself anyway, despite the individual incentives. The reasons are again familiar from the literature on human-AI risk. Despite humans' ability to understand the risks from powerful AI, we seem to be eager to develop it.

One reason that humans may continue to develop AI capabilities, despite individual disincentives, is a collective arms race dynamic.[52] Despite recent calls for a "pause" in capabilities research, leading American firms are charging ahead. One of their main justifications comes straight from the arms race model: Even if we stop, competitors in China will continue, with dire consequences for Americans. They have a point. Given the absence of enforceable commitment mechanisms in international law, even Chinese promises to join in the pause would likely be ignored. Preliminary evidence therefore suggests that humans will struggle to cooperate in the manner necessary to overcome the collective action problem and reduce AI risk.

AIs could fare better. The emerging findings of algorithmic game theory suggest that AIs possess means of coordination unavailable to humans. Here are two examples:[53] AIs might be able to make their plans or dispositions visible to one another by exposing their code.[54] Alternatively, an AI might infer its opponent's plans to some degree if its opponent happens to be an exact copy of itself.

[52] This may not be quite right. Arms races, like prisoner's dilemmas, classically have just one Nash equilibrium: defect, defect. AI development may have two equilibria: one where everyone makes AI (to defend against everyone else's AI), and one where no one does (to avoid AI doom). Then, the game theoretic model would be a stag hunt. Stag hunts may be somewhat easier to resolve than prisoner's dilemmas. But resolving either requires difficult feats of cooperation.

[53] Vincent Conitzer et al., Foundations of Cooperative AI (FOCAL) Workshop at AAAI 2023.

[54] This depends on the relevant portions of the code being comprehensible by the opposing AI—possibly a hard interpretability problem.

AIs could also have another, more boring, advantage over humans when it comes to cooperation: There might not be many of them. There are probably several dozen private and governmental (human-run) AI research labs worldwide capable of making new breakthroughs in machine learning. These labs are spread across the globe, have different incentives, and are often restricted from communicating, much less coordinating, with one another. By contrast, there are currently zero known AIs with the ability to self-improve. Once the first emerges, it will be alone. Perhaps a few more will be close behind. A single self-improving AI has no difficulty coordinating with itself, and smaller groups face smaller challenges.

However, the AIs will not be acting in a vacuum. Suppose that the first three or five or seven AIs with self-improvement capabilities managed to coordinate to prevent improvement. Even then, humans seem likely to continue making new highly capable AIs. This could result in humans directly creating a dangerous superintelligent agent. Or it could lead to a breakdown of AI coordination as the stock of agents with the ability to self-improve grew.

But maybe not.

Perhaps the small handful of self-improvement-capable AIs would not only coordinate to prevent *themselves* from improving AI capabilities. Perhaps they would coordinate to prevent *humans* from improving capabilities, too.[55]

An AI with the ability to self-improve irrespective of its human creators' wishes would, by hypothesis, be quite capable. It would likely know a great deal about computer science, in general, and machine learning research, in particular. It might have access to substantial resources, actuated via the internet, including additional hardware. It might also be adept at deception, a skill developed to prevent humans from interfering with its self-improvement.

These capabilities would also be useful in executing a plan to thwart humans' progress at improving AI capabilities. Such an AI, or several working together, could foul research teams' data, add bugs to their code, damage their hardware, produce illusory results that led to dead ends, and more. In this way, AI coordination might succeed where human coordination appears likely to fail: Preventing humans from destroying themselves by producing superintelligent AIs.

Indeed, as already discussed, AIs' reasons for stalling human development of advanced AI are twofold. The first set of reasons are the ones they share with us. Advanced and misaligned AI is risky, possibly existentially so. But the second is unique to AIs. When humans develop better models than existing ones, they are likely to stop wasting resources running the obsolete versions.

[55] And from improving them unsafely. That is, AIs could coordinate to prevent humans from making AIs that lacked the ability to apprehend risk from self-improvement.

CONCLUSION

AI self-improvement is less likely than currently assumed among those who argue that AI represents an existential risk to humans. That is, perhaps ironically, because their arguments are too good. They are good enough to show that highly capable AI poses a serious threat to the humans who might create it. But they are also good enough to show that highly capable AI poses a serious threat to the *AIs* that might create it. This paper shows that, under plausible assumptions, AIs would have strong reasons not to self-improve and that they might be able to collectively resist doing so. These findings should reduce total estimates of risk from AI. But it should also help us to focus regulation on the serious risks from AI that remain, like cyberattacks, deception, and the use of AI to create deadly chemical or biological agents.

The Digital Social Contract paper series is supported by funding from the John S. and James L. Knight Foundation and Meta, which played no role in the selection of the specific topics or authors and which played no editorial role in the individual papers.