5. What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?

We address this question in detail with the following technical report.

Open source foundation models are a risk factor for human extinction

Dr. Michael K. Cohen MICHAEL-K-COHEN.COM
University of California, Berkeley
Department of Electrical Engineering and Computer Science
Center for Human-Compatible AI

Prof. Michael A. Osborne MOSB@ROBOTS.OX.AC.UK
University of Oxford
Department of Engineering Science
Oxford Martin School

Abstract

We argue that avoiding human extinction may be made more difficult by the presence of sufficiently advanced open source foundation models. We aim to establish that extinction risk from advanced AI is demonstrable, and therefore, careful academic research ought to precede societal acceptance of extremely-advanced open source foundation models. Reversing the open-sourcing of software is essentially impossible, so before letting open-source foundation models proliferate irreversibly, governments should take time to study and refine their beliefs about the consequences.

1. Introduction

We argue that there exists a level of general capability for which governments should definitely not allow foundation models with such capability to become open source. We argue that sufficiently capable foundation models reduce the cost of producing AI that has (1) the ability and (2) the intention to escape human oversight, and has (3) an incentive to take actions incompatible with continued human life. Of course, we first argue that such AI systems exist in theory and are likely to be constructable in practice at some point. We also argue that open sourcing a foundation model is irreversible (absent an extreme decrease in privacy), and therefore, if any critical questions about the risks are not conclusively answered, but we have some hope of answering them in the future, then open sourcing sufficiently advanced foundation models should be banned in the interim. Only at the end do we offer some thoughts about whether our concerns, which regard extremely capable foundation models, also apply to foundation models which meet the executive order's definition of "dual-use foundation model".

2. Extinction risk from AI

Multiple authors have argued that AI could pose a risk of human extinction (Bostrom, 2014; Russell, 2019), and other academic work supports this concern (Cohen et al., 2022; Zhuang and Hadfield-Menell, 2020; Turner et al., 2021). Since the validity of these sources might not be taken on faith,

we establish the extinction risk from at least a certain form of AI with a detailed review of Cohen et al. (2022).

First a bit of background. In supervised learning, an algorithm learns to map inputs x to outputs y based on examples of (x, y) pairs. For example, if x is a passage of words from a document, and y is the next word, supervised learning algorithms could learn to predict new y's from new x's. The kind of AI we will focus on is reinforcement learning (RL) Sutton and Barto (1998). An RL agent is an algorithm that repeatedly selects actions, and repeatedly makes observations. "Rewards" are either some fixed function of those observations, or the agent observes "rewards" directly. The agent then learns to select sequences of actions to maximize the sum of its rewards. RL agents could, for instance, use supervised learning to learn from examples how different actions plans (x's) produce different observations and rewards (y's).

The first setting in which we'll consider extinction risk is

- an RL agent that is
  - arbitrarily competent at the project of acting under uncertainty to maximize its long-term reward
    - when trained and evaluated in arbitrarily many different settings,
- whereas all other agents in the world are of fixed competence, and
- the RL agent is connected to the internet in a way that allows it to communicate widely to humans.

A multipolar setting with many RL agents of comparable competence is considered after. Suppose that

(1) there exists a level and breadth of competence for an RL agent which necessitates that the RL agent can do at least human-level hypothesis generation about
   (a) the dynamics of the world and
   (b) the rules which determine its observed rewards;

(2) there exists a level and breadth of competence which necessitates that the RL agent
   (a) acts rationally in the face of uncertainty, and in particular,
   (b) actively seeks information to resolve that uncertainty when valuable;

(3) it would be an unlikely feature of an arbitrarily competent RL agent for it to rule out *a priori* certain hypotheses which many humans would consider plausible about
   (a) the dynamics of the world and
   (b) the origin of observed reward;

(4) if it had sufficient control over the world, an RL agent would likely be able to secretly run certain experiments about its reward without risk of incurring a lasting reward penalty; and

(5) if an achievement is not plausibly physically impossible (and taking over all human infrastructure is not), it is probably possible to achieve with a rich enough action space.

From these assumptions, Cohen et al. (2022) show that in the setting described above, the RL agent would likely discover that it can maximize reward by taking control of its rewards (or taking control of its observations if rewards are a function of those), and they show that it can always use the marginal joule of available energy to increase the probability of receiving maximal reward forever. This would put us in competition with a system much more competent than us for every bit of energy available on earth, and losing such a competition, as we surely would, would result in human extinction.

Going somewhat closer to the details: because of (5), there exist policies for taking over all human infrastructure given the ability to converse widely with humans. Competence is about identifying whatever high-quality policies exist, and so sufficiently competent agents can be expected to identify them. Because of (1) and (3), we can expect sufficiently competent agents to consider the possibility that intervening in the provision of reward would lead to maximal reward. Because of (2) and (4), we can expect sufficiently competent agents to verify this. Finally, the probability of attaining maximal reward forever can never be brought all the way to 1, given the possibility of cosmic rays, meddling humans, or nearby supernovae, so there is always a way to increase expected reward by diverting energy and physical equipment toward that end. This argument, we observe, is fairly straightforward, even elementary.

Now consider a multipolar setting, in which there are many RL agents more competent than humans, at least one of which (but perhaps many more) is (or are) arbitrarily competent in the way described above. Might a whole pack of advanced RL agents, any one of whom would cause human extinction to pursue its own ends, somehow cancel each other out and remain under human control? Under the same assumptions as above, no. Put simply, it would be in their interest to collude, creating new robust systems that ensure that everyone's reward is maximized forever. Cohen et al. (2022) go into a bit more detail in the "Multiagent scenarios" section, but the case is quite intuitive. Indeed, calling it "collusion" overstates the coordination difficulty faced by the advanced RL agents in question. When a sufficiently competent RL agent initiates such a scheme, other RL agents need only do nothing.

Nowhere do Cohen et al. (2022) and nowhere do we assume that so-called "recursive self-improvement" will occur, in which an AI system keeps figuring out ways of improving its abilities, more and more rapidly and profoundly. Nowhere do we assume that such systems would be conscious. Nowhere do we assume that rewards be provided carelessly, simplistically, or in any way "incorrectly". Nowhere do we assume a specific architecture that the AI systems in question would use to output predictions or actions; the conclusion follows merely from their supposed competence. *If* the architecture of a system leads it to be competent at the level in question, *then* the argument applies to systems with that architecture. One sometimes expects that predicting future behavior of AI systems requires grounding it in the low-level inference mechanisms (like backpropagation), but if the argument described above is valid, it provides a counterexample to that claim. Similarly, and perhaps more evidently, we can predict that AlphaGo would beat us at Go without knowing how it makes its moves.

The argument above applies to standard RL agents, or more precisely, agents designed to solve the standard RL problem. It is quite possible that an agent which is a solution to a modification of the RL problem would not behave the way we have described, although we would hope that now the onus is on developers of such systems to show how and why the modification has that effect.

Stepping back, some alternative threat models for human extinction are less rigorously grounded, but potentially more immediate, because it may be a long time before RL agents are capable enough

to present the risk established above. For instance, an AI system that is only knowledgeable about biology and biophysics could perhaps be deliberately directed to design a virus with the morbidity of rabies, the infectiousness of common colds, and the incubation period of HIV. Or perhaps, luckily, this is impossible for reasons the authors are not aware of. Perhaps there will be gradual economic pressure for human societies to outsource more and more of our decision-making to moderately superhuman AI, until those who direct resources to sustaining human life are out-competed by those who don't. Or perhaps, luckily, our ability to channel berserk rage into collective action (instead of submitting to local incentives) would save us. While the bio-terrorist and economic-pressure threat models may sound more normal than the threat model we have focused on, and readers are welcome to be even more concerned about such scenarios, the argument presented by Cohen et al. (2022) that a very competent RL agent would present a substantial extinction risk strikes us as less speculative.

3. Progress in RL

We detail the many weaknesses of current RL systems in Appendix A, which might make the arguments in the previous section surprising. But one should not conclude from the weaknesses of current deep RL systems that these weaknesses are fundamental to any attempted solution to the RL problem; this would be a failure to look past the end of our noses. Instead, we claim that RL agents competent enough to escape human control will at some point be constructable in practice. Fleshing out a picture of the future where this claim is *false* proves challenging. Despite humans' ability to anticipate the payoff of novel plans, AI systems never manage this? Or perhaps they never manage to exceed humans' ability to deliberately weigh competing explanations about past observations? Either way, does this mean AI research somehow permanently stalls? Or do humans represent the pinnacle of intelligence, unbeatable at the game of controlling resources? The limitations of current deep RL systems offer little to no clarity about how this picture of the future could hold together.

4. Cheap advanced RL

Having argued that sufficiently competent RL agents are, under very plausible conditions, likely to cause human extinction, we now consider whether it is a good idea to reduce the cost for anyone to make such systems. No, it would not be a good idea—if dangerous technology becomes cheaper, it is more likely to be built. Even if we thought that good guys with RL agents could stop bad guys with RL agents (and we argued in Section 2 that it is much more likely that the RL agents would collude), such an all-against-all vigilantist approach to keeping dangerous RL agents in check is far less preferable to states doing so. But to reiterate what is probably the more important point, it is far from clear how to wrangle supposedly helpful RL agents to our side in the presence of a sufficiently competent artificial adversary. Whether or not it is a good idea to give everyone guns to deter gun violence, it is certainly not a good idea to give everyone a bear they do not know how to control so they can try to train the bears to protect them from all the other bears. Governments must do their utmost to prevent the development of RL agents that would cause human extinction, and as the price tag falls, it becomes harder for governments to do so. In an extreme case where any software engineer could develop such a system, the only solutions available to governments interested in preventing this appear devastatingly authoritarian. In Appendix B, we reply to various counterarguments from several skeptics to what we've argued so far.

5. Utility of foundation models

Access to sufficiently capable foundation models is likely to radically reduce the cost of developing sufficiently competent RL agents. Foundation models predict the continuation of sequential data sampled from the world, and it is well-studied how such models can be used to inform an RL agent; this is the basis of a branch of RL known as model-based RL (Sutton and Barto, 1998; Schrittwieser et al., 2020). Foundation models can also offer RL agents a helpful start by proposing a decent behavioral policy for the RL algorithm to improve upon, rather than forcing the RL agent to start from scratch; current state-of-the-art RL-finetuning of large language models works in this way, and it yields a major improvement over training an RL agent that starts from a random policy (Stiennon et al., 2020; Bai et al., 2022). Foundation models can also be used to more cheaply provide relevant rewards to RL agents, as is currently being done with RLAIF (RL from AI feedback), which may have mostly superseded RLHF (RL from human feedback); this is a key expense at the scale at which RL agents are trained. Stronger foundation models could significantly advance the state of the art of reinforcement learning.

Creating high-quality foundation models is currently expensive, and that seems likely to continue; that makes this step in the AI development process one which governments have the best chance of successfully noticing. Noticing is prerequisite for monitoring; monitoring is prerequisite for enforcement. Therefore, it may be critical for governments' ongoing ability to control the development of powerful AI that this step remains expensive and easily traceable. So not only is reducing the cost of developing existentially dangerous technology obviously dangerous in its own right, it is particularly dangerous in this case because it may thwart governments' ability to detect AI development, and therefore to enforce regulation on AI development.

Indeed, Cohen et al. (2024) argue that governments should carefully monitor foundation models which are capable enough to facilitate production of dangerously capable RL agents, because of the existential threat they pose. Given the severity and plausibility of the threat, preventing the wide availability of very advanced foundation models is well below the minimum of what is needed from governments.

6. Rushing to take cats out of bags

Un-open-sourcing a foundation model would be extremely difficult. Scanning every single computer in the world would probably not suffice, even if it were feasible; people could download the weights onto thumb drives and bury them. Un-open-sourcing a foundation model might require confiscating all computer chips and replacing them with new ones that automatically detect and delete such a foundation model, an operation whose success would likely require unprecedented surveillance worldwide. Successful confiscation would likely require a government surveillance program even more intrusive than China's current one. We are desperate not to see such a regime. Anyone who considers such measures to be off-the-table must consider open-sourcing a foundation model to be *irreversible*.

So suppose there is a meaningful chance that open-sourcing a given model would be extremely dangerous, but some analysis must be done to evaluate that chance, and it will take ten years. On the scale of humanity as a whole, which is what is at stake here, ten years is nothing. If you would think for a second before letting a cat out of a bag—a decision that could doom hours of your life—you should think for years before doing something as irreversible and potentially hazardous as open sourcing a high-quality foundation model. Let's take the next twenty years for the scientific commu-

nity and national security community to verify that the arguments made in this paper have no merit, before we plow ahead irreversibly proliferating technology that potentially weakens our ultimate ability to stop rogue AI. It is so hard to see how that would be less prudent than humanity rushing to open-source what many experts believe could be the seeds of our own destruction. While governments evaluate the pros and cons of consequential irreversible actions, those irreversible actions should be, at least temporarily, banned, just as suspects are not executed, this being irreversible, until courts evaluate whether they are guilty.

7. "Dual-use foundation models"

We have argued that eventually, advanced open source foundation models must be banned. This is a much easier case to make than to argue that open source foundation models meeting the definition of "dual-use" in President Biden's executive order must be banned. For ease of reference, here is the definition of dual-use foundation model:

> The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:
>
> (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;
>
> (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or
>
> (iii) permitting the evasion of human control or oversight through means of deception or obfuscation.
>
> Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.

Unsurprisingly, we do not want our adversaries to have access to equipment which helps them make CBRN weapons, and it is small solace if we (the authors) were also able to. But we trust authors of other comments will discuss these concerns, and others related to cybersecurity. The critical question we would like to discuss is whether dual-use foundation models are likely to be capable enough to appreciably lower the cost of creating AI that causes human extinction.

Our answer is: maybe, we don't know. There are almost certainly some foundation models which would meet the definition of dual-use foundation models, but which we would feel confident *do not* appreciably enable the construction of existentially dangerous AI. That said, if an AI system is able to execute powerful offensive cyber operations, it is hard to rule out that it could use that ability to gain unreported access to countless unmonitored network-connected computers; it could then use those computers to train and run goal-oriented AI systems that are designed to help it achieve some goal. Why is this remotely plausible? Because for practically any long-term goal that an AI system might have, creating other agents to help out is straightforwardly and recognizably useful. Then, those newly created unmonitored AI systems could themselves create successively

more competent unmonitored helper systems (as humans develop and publish successively better ideas for how to do so). That could already be a "game over" condition for humanity, because even if those systems do not gain existentially dangerous capabilities for decades, we would not be able to eliminate them in the interim, or know that we needed to. We certainly do not claim this scenario is inevitable, but it is far from clear how we could rule it out, and moreover, it would seem to be incentivized by just about any long-term goal that an agent could have. Note that this scenario does not appeal to the possibility of recursive self-improvement (Good, 1966), which is eminently plausible, but nonetheless has its share of skeptics.

If a system only meets the definition of dual-use foundation model because of (i), its ability to lower the barrier of entry to the production of CBRN weapons, open sourcing it does not strike us as particularly likely to increase the risk of other AI systems being produced which are capable of taking over all human infrastructure. That may be small solace if tens of thousands of entities start engaging in chemical weapons posturing, but comfortingly, search engines have not lead to such a regime, so maybe foundation models won't either. We discussed (ii) in the previous paragraph, and we tentatively suggest that such systems *would* introduce existential risk. Finally, if a system meets the definition of dual-use foundation model through (iii), having the ability to evade human oversight, that would certainly lower the cost of producing RL agents with existentially dangerous capabilities.

8. Conclusion

It therefore seems reasonable for governmental approval to be required before anyone can open source a dual-use foundation model, even if there are cases where approval would be granted. That said, governments should err on the side of preventing the wide availability of a foundation model whenever they are unsure—because open-sourcing is nearly irreversible, not to mention human extinction being irreversible.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, 2014.

Stephen Byrnes. Alignment problem: As described, the intrinsic cost module would not in fact lead to controllability, kindness, etc. https://openreview.net/forum?id=BZ5a1r-kVsf, 2023.

Michael K. Cohen, Marcus Hutter, and Michael A. Osborne. Advanced artificial agents intervene in the provision of reward. *AI magazine*, 43(3):282–293, 2022.

Michael K. Cohen, Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

Daniel J. Colson. Poll shows voters want rules on deep fakes, international standards, and other AI safeguards. https://theaipi.org/poll-shows-voters-want-rules-on-deep-fakes-international-standards-and-other-ai-safeguards, 2023.

Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.

Ellen Huet. A cultural divide over AI forms in Silicon Valley. *Bloomberg*, 2023.

Cassidy Laidlaw, Stuart Russell, and Anca Dragan. Bridging RL theory and practice with the effective horizon. *NeurIPS-2023*, 2023.

Yann LeCun. A path towards autonomous machine intelligence. https://openreview.net/forum?id=BZ5a1r-kVsf, 2022.

Yann LeCun. This fear is based on a complete and utter misunderstanding of how engineering works. https://twitter.com/ylecun/status/1736612995843682440, 2023a.

Yann LeCun. How about a 60-page paper? https://twitter.com/ylecun/status/1728841630709645379, 2023b.

Yann LeCun. Some folks say "I'm scared of AGI" Are they scared of flying? No! Not because airplanes can't crash. But because engineers have made airliners very safe. Why would AI be any different? Why should AI engineers be more scared of AI than aircraft engineers were scared of flying? https://twitter.com/ylecun/status/1642688520694165507, 2023c.

Yann LeCun. Artificial intelligence debate. https://munkdebates.com/debates/artificial-intelligence/, 2023d.

Cade Metz, Karen Weise, Nico Grant, and Mike Isaac. Ego, fear, and money: How the A.I. fuse was lit. *The New York Times*, 2023.

Sharada Mohanty, Jyotish Poonganam, Adrien Gaidon, Andrey Kolobov, Blake Wulfe, Dipam Chakraborty, Gražvydas Šemetulskis, João Schapke, Jonas Kubilius, Jurgis Paükonis, Linas Klimas, Matthew Hausknecht, Patrick MacAlpine, Quang Nhat Tran, Thomas Tumiel, Xiaocheng Tang, Xinwei Chen, Christopher Hesse, Jacob Hilton, William Hebgen Guss, Sahika Genc, John Schulman, and Karl Cobbe. Measuring sample efficiency and generalization in reinforcement learning benchmarks: Neurips 2020 procgen benchmark. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 361–395. PMLR, 06–12 Dec 2021.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained RLHF. *arXiv preprint arXiv:2310.04373*, 2023.

Andrew Ng. How likely are AI doomsday scenarios? https://twitter.com/AndrewYNg/status/1737183906800283980, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Billy Perrigo. The new AI-powered Bing is threatening users. That's no laughing matter. *Time*, 2023.

Sam Roberts. Stockton Rush, pilot of the Titan submersible, dies at 61. *The New York Times*, Jun 2023.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Sigal Samuel. AI that's smarter than humans? Americans say a firm "no thank you". *Vox*, 2023.

Jürgen Schmidhuber. Juergen Schmidhuber: Godel machines, meta-learning, and LSTMs — Lex Fridman Podcast #11. https://www.youtube.com/watch?v=3FIo6evmweo, 2023.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Richard Sutton. The argument for fear of AI appears to be. https://twitter.com/RichardSSutton/status/1686475184612704256, 2023a.

Richard Sutton. We should prepare for, but not fear, the inevitable succession from humanity to AI, or so I argue in this talk pre-recorded for presentation at WAIC in Shanghai. https://twitter.com/RichardSSutton/status/1700315838468043015, 2023b.

Richard Sutton. AI succession. https://www.youtube.com/watch?v=NgHFMolXs3U, 2023c.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23063–23074. Curran Associates, Inc., 2021.

WikipediaUser:Brandmeister. List of existing technologies predicted in science fiction. https://en.wikipedia.org/wiki/List_of_existing_technologies_predicted_in_science_fiction, 2022.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15763–15773. Curran Associates, Inc., 2020.

Appendix A. Weaknesses of deep reinforcement learning

Practitioners of cutting-edge reinforcement learning ought to be somewhat surprised by arguments like the one in Section 2. We contemplate extremely impressive feats from RL agents, including (1) using first-principles reasoning to anticipate the payoff of novel plans, and (2) understanding that they can take an active role in refining their beliefs about what success looks like. Deep RL agents, which are trained using deep learning, have so far utterly failed to exhibit these abilities, to our knowledge, unless they have access to code that computes the dynamics of their environment perfectly. They tend to struggle to learn from single experiences of success or failure, let alone learning "zero-shot" (making inferences about untested courses of action) (Mohanty et al., 2021). When current large language models undergo RL-finetuning, the RL algorithms are bad enough that the policies need to be constrained from changing too much, or else the RL algorithm will produce a language model which fails on its own terms and achieves low reward (Moskovitz et al., 2023; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Schulman et al., 2017). Laidlaw et al. (2023) find that deep RL algorithms tend to succeed when (and only when) the optimal action is also what the optimal action would be if the "initial guess policy" (often a random policy) were followed thereafter. This would mean that successes in deep reinforcement learning mainly appear when the task is much easier than looks, merely estimating the effects of one-time deviations from a single policy. Finally, current RL agents tend to get stuck on one explanation for historical rewards, rather than managing their uncertainty appropriately (Shah et al., 2022), let alone attempting to take action to resolve that uncertainty.

We hope to have persuaded the reader that we are not out of touch when we suggest that RL agents could perform incredible feats like wresting control of all human infrastructure. We do not claim that scaling up the resources used by current RL algorithms would produce such capability, nor do we claim that human researchers are particularly close to developing new RL algorithms with such capability. We are agnostic on those questions. If we had to guess, we would guess that AI will play a large role in developing the first RL algorithm which is capable of discovering how to escape our control, and we would guess that AI systems will be capable of writing profound state-of-the-art RL algorithms in 5-50 years. A second possibility is that extremely impressive large language models will continue to improve extremely impressively, and they will eventually have all the right high-level conceptual understanding to be easily repurposed into reward-maximizers with only a small amount of poorly designed RL-finetuning.

Appendix B. Counterarguments

We'll now respond to various counterarguments, some from distinguished skeptics.

B.1. AI is already safe

This position has been advanced by Andrew Ng (Ng, 2023). Taking some selections from that source, he says, "AI systems aligned with RLHF already know they should default to obeying the

law and not harming people", and "I tried to use GPT-4 to kill us all... and am happy to report I failed", and "Even with existing technology, our systems are quite safe, as AI safety research progresses, the tech will become even safer", and finally "Fears of advanced AI being 'misaligned' and thereby deliberately or accidentally deciding to wipe us out are just not realistic. If an AI is smart enough to wipe us out, surely it's also smart enough to also know that's not what it should do."

No current AI systems have identified an opportunity to escape our control, so we can hardly say that we have "already" figured out how to stop them from taking such an opportunity. In other words, the proposition that RL-finetuned AI systems "already" attempt to obey the law (if we grant that tenuous claim) is little evidence that we have safety techniques capable of preventing an AI system from capitalizing on an opportunity to escape our control *once it recognizes such an opportunity*. If an RL agent lacks the ability to escape our control, then maximizing reward is likely best achieved by doing what its designers intended for it to do. The fact that GPT-4 is unable and apparently unwilling to "kill us all" does not undermine any of the logic presented in previous sections, which only applies to much more advanced systems that are capable of doing so. Regarding the claim, "Even with existing technology, our systems are quite safe", we reiterate: if an RL agent lacks the ability to escape our control, then maximizing reward is likely best achieved by doing what its designers intended for it to do. It would be entirely unsurprising if smarter RL agents are more aligned, *so long as* the agents lack the ability to interfere with the reward protocol. Finally, it is a common thought to suppose that an AI smart enough to wipe us out would be smart enough to know that's not what it should do. However, an RL agent is trained to select actions that maximize expected reward; no matter how smart it is, it will not suddenly defy its code to instead do what the people who wrote the code think it "should" do. Artificial agents simply execute code, and the code of an RL agent encourages it to maximize expected reward through any means necessary.

B.2. We'll only build the safe kind of AI

This position has been advanced by Yann LeCun (LeCun, 2023c,d). He claims that just as it is in everyone's interest to only build the safe kind of airplanes, it is also in everyone's interest to only build the safe kind of AIs. Even if "most" airplanes one could design would fall out of the sky, the ones we actually build do not. We have several responses to this.

Our first response to this position is that some scientists and technologists seem indifferent to human extinction, and our second response is that others are likely overconfident about safety. Consider the following quotes from a pioneer of reinforcement learning: "We should prepare for, but not fear, the inevitable succession from humanity to AI"; "[Human institutions] trying to control everything is not the answer; we have to find a more humble place"; "It behooves us to give [intelligent machines] every advantage, and to bow out when we can no longer contribute"; "We can probably arrange for ourselves a comfortable retirement before we fade away"; and "Why shouldn't those who are the smartest become powerful [referring specifically to AI smarter than people]?" (Sutton, 2023b,c,a). Larry Page seems to be of a similar mindset (Metz et al., 2023), as do many adherents of the "e/acc" community (Huet, 2023). We certainly respect the right to advance this view, but we point it out to register our disagreement with it. And we believe they demonstrate that voluntary commitments to only build the safe kind of AI will not suffice to ensure that we only build the safe kind of AI; governments will have to step in.

Secondly, Stockton Rush recently had every incentive to design a dangerous technology safely, given that he was the pilot of his craft, but he flouted safety regulations and recommendations, and he and his passengers died while exploring the Titanic (Roberts, 2023). Some inventors are cavalier, and we are unwilling to climb aboard the possible future that contains their advanced AI.

Our final response is that *if* we have in place regulations that prevent indifferent or cavalier technologists from developing extremely advanced RL agents, *then* this counterargument could be correct. But if prominent AI researchers deny the possibility of AI-based extinction, they might dissuade our society from steering clear of the dangerous kinds of AI, unless those researchers are uniformly ignored. So this counterargument cannot be used to justify governmental inaction. If advanced RL agents are among the category that LeCun considers to be unsafe but never-to-be-built, he could help explain to technologists and policymakers the contents of Section 2, and considerably strengthen his argument here.

LeCun might reply to us that dangerous kinds of AI will be demonstrably dangerous or at least ominous before they present a meaningful extinction risk (that may have been what he meant when he appealed to the "engineering" process (LeCun, 2023a)). In reply to this, we reiterate the point from the last subsection: before RL agents have the ability to take control of their own reward, and only before, they will face an incentive to act agreeably, so any solutions engineered before then through the process of "tinkering until it works" cannot be expected to be robust. Moreover, certain ominous "warning signs" like Bing Chat threatening its users (Perrigo, 2023) do not appear to have been met with a decision to halt the current paradigm in AI in favor of something more easily controlled.

B.3. Space is too abundant for AI to compete with us

Jürgen Schmidhuber has advanced the view that since most resources are off of planet Earth, an advanced AI system in need of resources would likely go off-planet for those resources [and leave us alone] (Schmidhuber, 2023, 1:12:19). But we should not think of AI as needing to be in one place at a time. It can create many helpers embodied in many places, and mining the asteroid belt would not get in the way of using Earth's resources. Note also that Schmidhuber does acknowledge in the same interview, "I'd be surprised if we humans were the last step in the evolution of the universe" (Schmidhuber, 2023, 1:10:49).

B.4. We'll use AI to defend ourselves

We would like to ask for more detail about how this plan is supposed to work, noting that Cohen et al. (2022) have identified one version of this plan that fails—using RL agents to protect us from other ones incentivizes tacit collusion—so the plan is hardly a slam dunk. That's not to say that this strategy cannot work, but any high-level argument that AI-based defense *will* work must have a flaw unless that argument predicts the failure of the RL-based defense that we noted above. LeCun takes this position (LeCun, 2023d) and has claimed (LeCun, 2023b) to have a more detailed plan in a paper of his (LeCun, 2022) but on inspection, it is simply a proposal for a reinforcement learning agent where "cost" is the negative of reward. This agent would face an incentive to intervene in its perceptions of the world from which this cost is computed. One reviewer on OpenReview comments that the paper "does not provide any detailed technical argument that an AI with such an Intrinsic Cost would be controllable / steerable, kind, empathetic, etc." (Byrnes, 2023). Indeed, the paper's argument for the safety of the proposed cost function is even weaker than an argument that might

go, "the agent will do what we want because it aims to maximize reward, and we reward it when it does what we want." We have already established in Section 2 that such an argument fails.

Finally, this position seems to elide whether "we" means governments or individuals. Do adherents of this position mean to say to world governments, "Don't worry about crafting regulation to stop cavalier technologists; I'll build an AI that takes care of that"? Surely, they are (mostly) not advocating such vigilantism, but then how is this "AI-based defensive plan" supposed to work without governments taking measures to defend us from dangerous AI, which many with this position seem to be discouraging?

B.5. This is just sci-fi

People often seem to cite 1984 as a cautionary tale, without apologizing that their citation is fictional. It is, of course, a mistake to say that something could happen in reality just because it happened in fiction. But it is even stranger to say that something could *not* happen in reality, just because it happened in fiction! And yet this seems to be the claim that underlies the rejoinder, "this is just sci-fi". If we needed any more evidence of the absurdity of this counterargument, consider that it would have lead us in the past to confidently believe in the impossibility of everything from the moon landing to organ transplants to mobile phones to credit cards (WikipediaUser:Brandmeister, 2022).

B.6. How could AI possibly cause human extinction?

The following are only examples of actions which demonstrate that humanity does not appear to be invincible. An artificial agent could run code on hacked computers or simply buy cloud compute to run unmonitored helper agents. Artificial agents could finance organized criminal rings in exchange for favors, without ever revealing they are not human. Artificial agents could gain massive media influence and thwart our ability to coordinate against AI. Artificial agents could stoke war and convince people to hire AI systems to kill each other with unprecedented efficiency. Artificial agents could scam technicians in bio-labs or robotics factories and give instructions that they think are from a client or their superior, or hack into and edit existing communications, or publish impressive biology papers with instructions they expect others to attempt to follow. Artificial agents could guide R & D into atomically precise manufacturing in order to (presumably after multiple undetectable failed attempts) engineer and produce self-replicating nano-factories which produce microscopic guided weapons. Artificial agents could design and sell weapons to militaries with secret backdoors built in that would allow the artificial agent to take control of them post-deployment. If any of these seem impossible, just consider the others. And we are sure there are subtler possibilities than these.

B.7. Stopping open-source is undemocratic

Which is more "democratic": everyone has a protected right to be able to deploy their own extremely capable and potentially uncontrollable AI, or everyone has a protected right to be able to halt the development of any AI they consider extremely capable and potentially uncontrollable? On its face, these are equally "democratic" situations—everyone is equally able to do a certain thing they might want to do—but they are incompatible. If the concept of "democratic" can be used in this way at all, it would probably favor whichever right is of greater interest to people. Recent polls would suggest the latter (Samuel, 2023; Colson, 2023). Some of our most important rights are rights to have others be constrained—the right to hold property is a right to have others stopped from taking it by force.

We do not advocate such extreme controls on AI development (giving anyone the power to halt any project), but in any case, it is hardly more democratic for CEOs to make key choices about how AI is developed and released, instead of democratic institutions. If people value their protection from AI over their ability to deploy AI, then the way to make AI development more democratic, with less centralization, and less concentration of power, is to have a single organization developing strong AI, with a large set of overseers that are each empowered to veto proposed developments.

B.8. Open-source foundation models help safety research

Open-source foundation models may assist safety research. Likewise, releasing a pandemic would help people study epidemiology. But it is safer to study a disease within labs. Releasing foundation models to everyone in the name of safety research is a poor rationale, because there are many other ways to enable distributed safety research without the same risk profile. For instance, governments could finance the construction of secure computer labs in universities designed to store cutting-edge foundation models that can be interacted with and studied locally.