Three Lines of Defence Against AI Risks at a Societal Level

Risto Uuk
Future of Life Institute
risto@futureoflife.org

2 June 2023

Abstract

In this essay, I discuss a framework to reduce AI risks at a societal level: 1. Don't develop the AI model; 2. Don't deploy the AI model; 3. Implement safeguards for the AI model. I start by sharing some risks and considerations arising at each phase of the AI production lifecycle. I then briefly address two objections to the framework: (1) development and deployment of certain AI models cannot be stopped, and (2) there is no need to stop the development and deployment, as putting in safeguards is enough. Finally, I discuss some implications of the framework. The main takeaway is that society can take reasonable actions in each step, not just in the last one.

1 Introduction

"The most reliable method for ensuring that large language models are not used in influence operations is to simply not build large language models. Every other proposed change to the design and construction of these models will be less effective at preventing misuse than not building the model in the first place. However, a complete stop to the development of new large language models is extremely unlikely, and so we focus primarily in this section on how these models could be built differently to reduce the risk of misuse."[1]

"Organizations that develop and deploy artificial intelligence (AI) systems need to manage the associated risks—for economic, legal, and ethical reasons. However, it is

[1]Goldstein et al., 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations'.

not always clear who is responsible for AI risk management. The Three Lines of Defense (3LoD) model, which is considered best practice in many industries, might offer a solution. It is a risk management framework that helps organizations to assign and coordinate risk management roles and responsibilities. ..."[2]

These two paragraphs inspired the following essay. It struck me that at a societal level, humanity has three lines of defense against AI risks:

1. Don't develop the AI model
2. Don't deploy the AI model
3. Implement safeguards for the AI model

The first line of defense is that we could decide not to develop the AI model in the first place. No AI model means no risks from that AI model. But let's say we decided to develop it, what then? Well, we could still decide not to deploy it. Some risks arguably arise from the development, but other risks come during deployment. If we did not deploy the model, those risks would not occur. But suppose we decided to deploy the model as well. Well, now we could at least implement safeguards to make the AI model as safe as possible. This framework is simplistic, but it could have some useful implications for reducing risks from AI.

As a society, we tend to approach technology deterministically and believe that wide AI development and deployment is something that cannot be stopped. This may be one of the reasons most of our focus goes into the last defensive strategy, which is to make systems safer. This is, however, the least effective strategy from the perspective of risk, because there may actually not be any techniques to make certain AI models as safe as we want them to be.[3] Instead, society could use all lines of defence and have much better chance of success.

Just a clarifying note here: when I am talking about AI models, I mainly have in mind the most broadly capable systems. These are current general purpose AI models such as large language models,[4] but similarly or even moreso, future models that are more advanced. In addition, for simplicity I focus on three main stages of AI production, but this cycle could involve more stages depending on the purpose of the framework.

2 The Framework

1. Don't develop the AI model

The development phase may create unique risks. For example, OpenAI at one time trained their GPT-2 model with a sign-flipped reward optimised for maximally bad output

[2]Schuett, 'Three Lines of Defense against Risks from AI'.
[3]Bowman, 'Eight Things to Know about Large Language Models'.
[4]Gutierrez et al., 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems'.

due to a coding bug:

"One of our code refactors introduced a bug which flipped the sign of the reward. Flipping the reward would usually produce incoherent text, but the same bug also flipped the sign of the KL penalty. The result was a model which optimized for negative sentiment while preserving natural language. Since our instructions told humans to give very low ratings to continuations with sexually explicit text, the model quickly learned to output only content of this form. This bug was remarkable since the result was not gibberish but maximally bad output. The authors were asleep during the training process, so the problem was noticed only once training had finished. ..."[5]

AI researchers have collected a lot of examples of AI development leading to unwanted behaviour in smaller-scale AI experiments.[6] These behaviours are known as specification gaming, satisfying the literal specification of an objective without achieving the intended outcome. As an example, in a reward learning setup, a robot hand pretends to grasp an object by moving between the camera and the object, to trick the human evaluator. While these examples did not result in harm to anyone, they could be damaging if they occurred in real life.

Finally, sometimes the training process can be very messy, indicating that serious issues can arise in this phase. During the training of Meta's OPT-175B, there were estimated to be over 70 automatic restarts due to hardware failures as well as at least 35 manual restarts.[7] Researchers had to pause the training, run a series of diagnostics tests, and make numerous changes during the training process. Eventually this development led to an AI model that the researchers themselves said had a "high propensity to generate toxic language and reinforce harmful stereotypes."

2. Don't deploy the AI model

To me, one of the most obvious and famous examples of deployment-related risks from AI is Microsoft's Twitter bot, deployed in 2016 to have automated discussions with Twitter users.[8] Less than 24 hours after deployment, the bot had to be taken down because it started disputing the existence of the Holocaust, referring to women and minorities with unpublishable words and advocating genocide. A more recent, highly troubling incident is the Dutch tax authorities using AI (or a related automated system) to create risk profiles to spot fraud among people applying for childcare benefits.[9] The damage done was not hypothetical: tens of thousands of families were pushed into poverty because of debts, some victims committed suicide, and more than a thousand children were taken into foster care.

[5]Ziegler et al., 'Fine-Tuning GPT-2 from Human Preferences'.
[6]'Specification Gaming Examples in AI - Master List'.
[7]Zhang et al., 'OPT'.
[8]Victor, 'Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.'
[9]Heikkilä, 'AI'.

In addition to risks arising from deploying AI systems in any domain, there are some areas where deployment is likely especially risky. Unfortunately, many are very eager to deploy AI models in those areas. One of the most concerning examples occurred when in just a few hours a drug-developing AI invented 40,000 potentially lethal molecules.[10] One of the researchers said that the biggest concern for them was how easy it was to get this result. Other domains where deploying AI systems might be especially risky include critical infrastructure, politics, law enforcement, nuclear, and more.

3. Implement safeguards for the AI model

It may be the case that some AI systems just cannot be made safe enough to satisfy the majority of society. For example, Bowman (n.d.) states that though there has been some progress in understanding and mitigating various issues with large language models, "there is no consensus on whether or how we will be able to deeply solve them, and there is increasing concern that they will become catastrophic when exhibited in larger-scale future systems."[11]

OpenAI, in their GPT-4 system card, list various risks that their system may create, including ones related to hallucinations, harmful content, harms of representation, influence operations, proliferation of weapons, privacy, cybersecurity, risky emergent behaviours, economic impacts, acceleration, and overreliance.[12] They explain that their mitigation strategies prevent certain kinds of misuses but have limitations. They also present how their interventions have improved outcomes on various benchmarks. However, there is no discussion about what thresholds they think these systems should meet in order to be safe enough for society to use in different ways.

One major issue that is hard to solve with these safety strategies is how to prevent society-wide systemic risks that arise from integrating these models into all economic activities. Jan Leike, the Alignment Team Lead at OpenAI, has elegantly said, "Before we scramble to deeply integrate LLMs everywhere in the economy, can we pause and think whether it is wise to do so? This is quite immature technology and we don't understand how it works. If we're not careful we're setting ourselves up for a lot of correlated failures."[13]

3 Objections

A common reaction to the idea of prohibiting or restricting the development and deployment of AI is that it is not realistic. For example, already in the first month after

[10] Calma, 'AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours'.
[11] Bowman, 'Eight Things to Know about Large Language Models'.
[12] 'GPT-4 System Card'.
[13] Jan Leike.

the release of ChatGPT, it was used by over 57 million people.[14] A technology that is deployed so widely, and for which ongoing investment is growing, cannot be stopped, this viewpoint says.

There are historical examples of societies prohibiting or restricting promising technologies and economic activities. For example, nuclear arms have been limited to a handful of states through international treaties;[15] recombinant DNA research was put on pause;[16] human cloning is widely illegal.[17] Furthermore, early in the COVID-19 pandemic, governments around the whole world shut down a large part of the economy to reduce the spread of the virus.[18] These examples of restrictions may not be perfectly analogous to restricting or prohibiting the development or deployment of advanced AI models. While restricting technological development is indeed a difficult problem for a society, it is wrong to say that it is impossible.

Another objection is that these AI systems can be made safe enough over time and therefore there is no need for restricting their development and deployment. No technology can be made so safe that there is absolutely no risk from it; it is a matter of making it safe enough. This objection states that there is already evidence of some interventions working to make these systems safer, and over time, improvements will be even greater.

I definitely hope so. But we have no guarantee of that, and so far the results are not too promising. First, there is some survey evidence indicating that many AI researchers believe humans may not be able to control future advanced AI systems to the degree that they do not cause human extinction.[19] Second, even though OpenAI appears to have spent considerable effort in making GPT-4 safe, myriad risks remain. Even on narrow benchmarks of disallowed and sensitive content, GPT-4 produces far more than 0% of unwanted output.[20] In addition, in Anthropic's red teaming paper based on a random sample of 500 attacks on their AI model, a lot of harmful output was identified, including in categories such as discrimination, hate speech, violence, fraud, weapons, terrorism, self harm, and even child abuse.[21]

4 Discussion

We are not as a society dictated only to focus on implementing safety guardrails. We have two other lines of defence which are more effective, and society can make reasoned decisions about all of these lines of defence. Societies have previously made decisions to

[14]Cerullo, 'ChatGPT Is Growing Faster than TikTok'.
[15]Fuhrmann and Lupu, 'Do Arms Control Treaties Work?
[16]Berg, 'Asilomar 1975'
[17]'Human Cloning'.
[18]'Coronavirus'.
[19]Stein-Perlman, Weinstein-Raun, and Grace, '2022 Expert Survey on Progress in AI'.
[20]'GPT-4 System Card'; Maslej et al., 'The AI Index 2023 Annual Report'.
[21]Ganguli et al., 'Red Teaming Language Models to Reduce Harms'.

ban certain technologies completely and strictly regulate others. There is no economic law that requires particular technology needs to be developed and deployed no matter what (even though economic incentives make it very attractive to do so).

David Krueger, Assistant Professor of Machine Learning at the University of Cambridge, has stated that the obvious solution to AI existential risk is and has always been not to build artificial general intelligence. He acknowledges that this is an extremely difficult global coordination problem, analogous to climate change, but he does not think there is any good alternative at the moment.[22]

There can also be specific interventions in the development and deployment phases that would reduce risks arising during those steps without entirely stopping these phases. In the former phase, developers could thoroughly assess the potential impact of the model before going ahead with the development according to the initial plan; in the latter, decisions could be made about how slowly and in what way the model would be released to users.

Depending on how dangerous society considers particular AI models to be, they can consider all three lines of defence and decide not to use all of them. For example, a society may reach the conclusion that ChatGPT-like model development is not highly dangerous by itself, and thus they can largely skip the first line of defence, but then commit strongly to the second line of defence, restricting the deployment of the model to more than ten million people instead of a possible billion people.

5 Conclusion

The overarching point of the essay is that society has all three lines of defence in its arsenal. All three have their uses and benefits, and none of the objections against them outweigh those upsides. Given the severity of the threats we face, it would be unwise to dismiss any one of our defences out of hand.

References

BBC News. 'Coronavirus: The World in Lockdown in Maps and Charts'. 6 April 2020. https://www.bbc.com/news/world-52103747.

Bowman, Samuel R. 'Eight Things to Know about Large Language Models', n.d. https://cims.nyu.edu/~sbowman/eightthings.pdf.

Berg, Paul. 'Asilomar 1975: DNA Modification Secured'. Nature 455, no. 7211 (September 2008): 290–91. https://doi.org/10.1038/455290a.

[22]David Krueger.

Calma, Justine. 'AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours'. The Verge, 17 March 2022. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx.

Cerullo, Megan. 'ChatGPT Is Growing Faster than TikTok', 1 February 2023. https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/.

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, et al. 'Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned'. arXiv, 22 November 2022. https://doi.org/10.48550/arXiv.2209.07858.

David Krueger. Twitter Post, 8 May 2023. https://twitter.com/DavidSKrueger/status/1655535985327841281.

Goldstein, Josh A, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova. 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations'. Georgetown University's Center for Security and Emerging Technology, OpenAI, Stanford Internet Observatory, January 2023. https://cdn.openai.com/papers/forecasting-misuse.pdf.

Fuhrmann, Matthew, and Yonatan Lupu. 'Do Arms Control Treaties Work? Assessing the Effectiveness of the Nuclear Nonproliferation Treaty 1'. International Studies Quarterly 60, no. 3 (1 September 2016): 530–39. https://doi.org/10.1093/isq/sqw013.

Gutierrez, Carlos Ignacio, Anthony Aguirre, Risto Uuk, Claire Boine, and Matija Franklin. 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems'. SSRN Scholarly Paper. Rochester, NY, 5 October 2022. https://doi.org/10.2139/ssrn.4238951.

Heikkilä, Melissa. 'AI: Decoded: A Dutch Algorithm Scandal Serves a Warning to Europe — The AI Act Won't Save Us'. POLITICO (blog), 30 March 2022. https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/.

'Human Cloning'. In Wikipedia, 31 May 2023. https://en.wikipedia.org/w/index.php?title=Human_cloning\&oldid=1157912590\#cite_note-PoloHorses-46.

Jan Leike. Twitter Post, 17 March 2023. https://twitter.com/janleike/status/1636788627735736321.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, et al. 'The AI Index 2023 Annual Report'. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2023. https://aiindex.stanford.edu/report/.

OpenAI. 'GPT-4 System Card', 23 March 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Schuett, Jonas. 'Three Lines of Defense against Risks from AI'. arXiv, 16 December 2022. http://arxiv.org/abs/2212.08364.

'Specification Gaming Examples in AI - Master List'. Accessed 30 May 2023. https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

Stein-Perlman, Zach, Benjamin Weinstein-Raun, and Katja Grace. '2022 Expert Survey on Progress in AI'. AI Impacts, 3 August 2022. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Victor, Daniel. 'Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.' The New York Times, 24 March 2016, sec. Technology. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Ziegler, Daniel, Nisan Stiennon, Jeffrey Wu, Tom Brown, Dario Amodei, Alec Radford, Paul Christiano, and Geoffrey Irving. 'Fine-Tuning GPT-2 from Human Preferences'. OpenAI, 19 September 2019. https://openai.com/research/fine-tuning-gpt-2.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, et al. 'OPT: Open Pre-Trained Transformer Language Models'. arXiv, 21 June 2022. https://doi.org/10.48550/arXiv.2205.01068.