HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

How the Government Can Mitigate the Emerging Risks of Artificial Intelligence

Atharv Joshi[1] and Michael Chechelnitsky[#]

[1]St. Francis High School, USA
[#]Advisor

ABSTRACT

The astonishing pace of development in Artificial Intelligence (AI) has been welcomed for the benefits it brings to society and businesses. However, the malicious use of AI has raised alarms in academia and the government. The unbridled growth of AI technology has been portrayed as a risk to humankind. [7] The initial impact of the misuse of AI is shown in the form of deepfakes and sophisticated phishing attacks. The pace of AI development is unprecedented, and other issues that will come up are unknown. In this paper, I will assess how the "known unknown" risks of artificial intelligence should be handled and propose ways for the government to get involved in a constructive way, such that the risks of AI can be mitigated. I draw lessons from emergency preparedness and establish a need to have government involvement at the federal level to mitigate the growing risks from AI. This organization would bring other government agencies, businesses, and citizens together to take steps to deal with the risks from AI. Taking measures to counter deepfakes, facilitating the learning of AI models, and handling biases in AI models are three important goals for the government's agency.

Introduction

Prominent figures such as Bill Gates, Sam Altman, and Ray Kurzweil [1] [2] [7] have raised alarms by predicting that AI is our biggest existential threat that needs to be a priority to deal with. At the same time, AI is being used to solve numerous problems that benefit society, ranging from using it as a copilot in education, to screening and detecting cancer [11]. As humans, we have been here before. Since the Industrial Revolution, we have seen humans push the limits of science and technology, and at the brink of several big changes, the demise of society was predicted. We are at a similar moment with the spread of AI. If this would be considered an emergency of our generation, then we should understand how we handle emergencies generally.

Over long periods of time, we have learnt how to prepare for emergencies such as natural disasters. The government, citizens, and businesses all have a role to play. Although we do not know when these emergencies will occur, having a plan is important. The government takes the lead in delegating tasks and planning how to execute them. We can draw lessons from emergency preparedness and provide a concrete way the government can step in and play a role in addressing the risks from AI. I will also propose areas of technical changes that should be considered to mitigate and slow the risks of AI in the immediate future.

Background

Gartner [9] [10] defines Artificial Intelligence (AI) as applying advanced analysis and logic-based techniques to interpret events and suggest, support, and automate decisions or take actions. Today, humans can use AI to automate and accelerate many tasks and jobs. Going further, an artificial general intelligence (AGI) is defined by Gartner as a form of AI that possesses the ability to understand, learn, and apply knowledge across a wide

range of tasks and domains. It can be applied to a much broader set of use cases and incorporates cognitive flexibility, adaptability, and general problem-solving skills. Artificial general intelligence is an anticipated future of AI that can understand or learn any intellectual task that a person can do.

We humans have a desire to achieve victory over something that looks unachievable, from climbing Mount Everest to landing on Mars. The pursuit of artificial general intelligence is our next audacious goal. OpenAI's ChatGPT could not count until 10 in 2019. Now, the same technology can paint pictures and write poetry. What concerns people is how far and how fast will this go and what it means for humans. When AGI enables systems to be better and faster, it could increase unemployment which in turn leads to societal unrest. Experts have also identified issues with fairness, bias, and a lack of trust. Lastly, there is a what-if scenario where AGI makes better decisions and directs humans to take action.

While we are astonished by recent breakthroughs and the pace of innovation, we know how technology has failed to deliver many times. As the technology evolves, it is hard to predict what other problems and issues will arise that humans must debate and solve. This is why I refer to the possible future issues as known unknowns. We know there will be issues to address, but we don't know what the issues are and when we will encounter them. In such situations, focusing on preparing for the unknown is a must.

The Known Unknown: Preparing for a Disaster

If you stay in an area that is known for natural disasters such as earthquakes or hurricanes, a natural disaster is bound to strike at some point. Every U.S. state has a State Emergency Response Commission that instructs people to have a family emergency plan in place: a getaway bag, food and water to last for two weeks, and more. In an earthquake zone, the people do not know when an earthquake will strike, how strong it will be, what the impact will be, where the impact will be most severe, or what the impact will be. I call this known unknown.

We can observe a few things in emergency preparedness. First, the government takes the lead. There is an office responsible for preparing for a potential disaster which permeates all levels of government. At the federal level, there is a Federal Emergency Management Agency (FEMA). Every state has The Governor's Office of Emergency, and most counties have an office of emergency. Second, the agencies provide a concrete direction to people on what they need to do, establishing a clear role for them. Third, the agencies make sure the plan is flexible to adapt to any situation. Whether there is an earthquake, tornado, or landslide, the organizations know what they need to do.

Government invests billions of dollars through FEMA on emergency preparedness. It creates Flood maps and Earthquake maps for the country and manages the Flood Insurance program. FEMA monitors weather patterns and various signals to understand if an emergency is brewing in the near future. It maintains the National Incident Management System. To prepare, FEMA creates training and education programs and plans for risks from a variety of emergencies including biohazard, fires, avalanches, hurricanes, landslides etc. They have playbooks and are always prepared to engage with any type of emergencies.

This type of preparation to handle the known unknowns is what the government needs to have to manage the risk of AI. AI technology is rapidly evolving. In some fields, such strategy games, AI has equaled or surpassed humans. [6] We do not know which frontiers of AI will evolve and how quickly. Let's consider these as known unknowns.

The Government's Role in Handling AI Advancement

Many new and transformational technological advances, such as the Internet, were driven by government funded institutions such as DARPA and NASA. The government has also encouraged the development and

adoption of technologies (electric vehicles and clean energy initiatives with solar and wind) through favorable government regulations and tax laws. While tech companies are pouring billions of dollars in AI related R&D, the impact of the outcomes will not be limited just to businesses. It permeates through society in both positive and negative ways. On one hand, we will see advances in medicine, treatments, better education tools for students, and higher living standards for the population. But on the other hand, there is potential to threaten privacy, elections, and democracy using deepfakes [12]. Further, the use of modern AI technology could threaten a country's security, and AI is increasingly being viewed as a competitive differentiator among advanced countries. Therefore, governments should get involved with AI in some capacity. The challenge is to figure out the best way for governments to get involved in such a way that they do not hamper innovation, but also ensure that it has some direction, and that society is ready to deal with the outcomes of these innovations.

Like the Federal Emergency Management Agency, the U.S. needs an organization that is centrally responsible for working through the risks of AI. We can model the organization along the lines of the National Institute of Standards and Technology. This organization defines the standards of measurements for the U.S. Over time, every country has established a similar organization that coordinates with every other country's organization as required. I propose the federal organization for AI risk be called "The Office of AI Risk". This will be the federal organization that coordinates everything related to the risks of AI and coordinates with similar organizations of other countries to share knowledge and expertise. It should also engage with academia and businesses to gather the best knowledge and work on the implementation of standards and protocols to manage the risks of AI.

The federal agency is not to be misunderstood to be an innovation killer. The goal of the agency is to support the evolution of AI in a responsible way. It enables the industry and society to adopt AI in a safe manner, provides guardrails to use AI responsibly, and sets standards to make AI useful, not harmful.

Handling Fake Pictures, Videos, and Audio

As Chesney and Citron explain, the ability to distort reality has taken an exponential leap forward with deep fake technology [12]. With easy-to-use tools, it is possible to create a fake picture, fake audio, and even fake videos that make the viewer believe to be real. This has become possible with many pictures, audio samples, and videos that are now easily available. For example, in 2018, a deepfake video was created of President Barack Obama using a free tool called FakeApp. The longer the app was left to process it, more realistic the video became [17]. The number of newly reported AI incidents and controversies has grown 26 times between 2012 and 2021 [18].

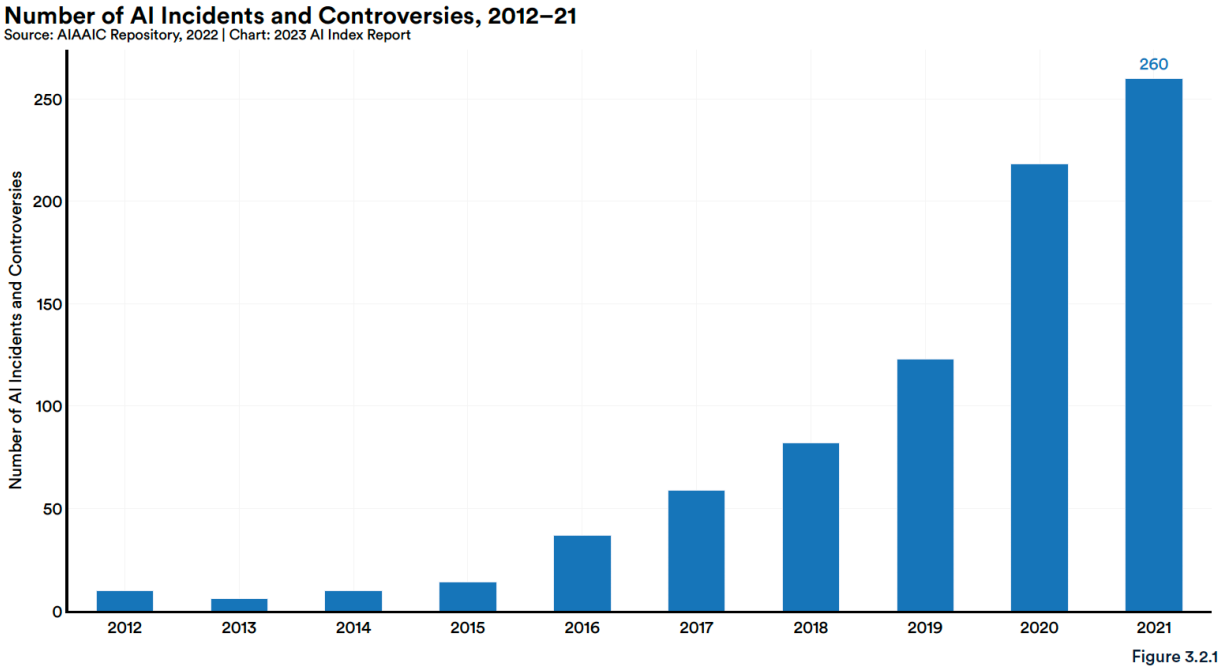

Figure 3.2.1

1 This figure does not consider AI incidents reported in 2022, as the incidents submitted to the AIAAIC database undergo a lengthy vetting process before they are fully added.

While I will not discuss the benefits and harms of deep fakes in this paper, I will explore the ways to thwart the risks of deep fakes.

Working with the industry, The Office of AI Risk could take steps to handle the risk. I propose that every picture, audio and video need to have a unique embedded stamp. The stamp encompasses who created the asset, where it was created, the IP address of the device used, and other attributes. HTML and JavaScript standards need to evolve to incorporate a check for the stamp every time before rendering it. The stamp needs to be authenticated by a server that is similar to a server that resolves IP addresses. If the stamp is missing, looks suspicious or tampered with, or cannot be authenticated, the page should not render the element. This is going to require exhaustive cooperation from the industry. Every vendor that plays a role in the value chain will need to act responsibly and incorporate the stamp. Vendors such as Google, Apple, Adobe, Meta will need to play a major role in this initiative. The Office of AI Risk will need to work with all the tech companies to ensure their software tools are updated to create and embed this stamp.

Explainability

As the reliance on the use of sophisticated and advanced models goes up, it becomes important to understand the reasons why a particular decision and prediction was made by the model. If humans fully understand how decisions are being made, the AI system can be controlled to not make a faulty decision, which will improve its decision-making capability. This in turn increases our trust in AI systems. Explainability becomes critical in the broad range of industries where life, death, finances, and personal wellness are in question. Any AI based system needs to have built-in explainability. If required, the AI system needs to be able to tell why the system took the action that it took.

We have seen early signs of this in the financial industry. On Sept 19, 2023, The Consumer Finance Protection Bureau (CFPB) published guidelines to lenders using AI in credit decisions. [19]. The crux of the guidelines is that lenders must go beyond merely using AI; they must also understand how it learns, processes

data, and ultimately makes decisions. Lenders need to be precise and pinpoint and convey the exact principal reason guiding the decision. For e.g., if the lender considers occupation while making a credit decision, instead of declining the consumer a loan by saying insufficient income, the lender needs to disclose to the consumer that their occupation was considered in their decision. As a result, lenders have started using AI responsibly by using machine learning algorithms that are inherently explainable. For example, algorithms such as decision trees and Bayesian classifiers provide traceability and transparency in their decision making.

Similar rules need to be adjudicated by every cabinet ministry in the government. The Office of AI Risk can play the coordinating role. It is important to acknowledge at this time that no one organization will be able to create explainability requirements by themselves for broad usage because we need the domain knowledge of the AI's use cases and require flexibility in the level of explainability. The CFPB can investigate the AI models used for credit decisioning because they have a deep understanding of the domain of credit decisioning. AI explainability built into defense systems and banking credit decisioning may look similar from a technology perspective, but their execution and parameters differ significantly, and a single organization cannot do justice [14].

Handling Biases

All model development relies on training data. The predictions and decisions by models are directly related to the training data. The data can represent various societal biases and worldviews that may not be representative of the user's intent or of widely shared values. For example, Joy Buolamwini and Timnit Gebru found that facial analysis technologies had higher error rates for minorities potentially due to unrepresentative training data [15]. This has practical implications. Security companies that use AI to determine a potential threat would introduce a high number of false positives if the training data were not robust and diverse. For e.g., if an AI based video surveillance solution is not trained on diverse data such for gender, race, ethnicity, age etc., it will be detrimental to minorities and women. It can also open security loopholes in surveillance systems at airports and critical infrastructure locations.

OpenAI CEO Samuel Altman, in his testimony to the Senate Judiciary Committee, acknowledged the existence of social biases and worldviews in the GPT models, and that the models have the potential to reinforce and reproduce specific biases and worldviews [14].

Ensuring we have a handle on biases needs to be one of the top agendas of The Office of AI Risk because businesses do not have great incentives to worry about minorities, underprivileged segments of the society. The Office of AI Risk can create test suites by industry to check for biases. This may look to be a daunting task at first. But we do have a precedent in the financial industry, more specifically banking. Because of regulatory requirements, banking industries go through testing of various kinds for risk management. Lenders go through their lending algorithms to prove there is no bias and discrimination against certain segments of the society. Regulatory steps will be required to check for biases in foundational models and a pre-deployment approval process will need to be created in partnership with industry. Part of the approval process will require disclosures on the data used for model development. The Office of AI Risk can play a central role in the creation and enforcement of regulations. It can also play intermediary between the industry and government.

Conclusion

The pace of development of artificial intelligence toward artificial general intelligence is bound to have twists and turns. We cannot predict the timelines, nor can we predict all the issues that will come about. We can prepare for the known unknowns. The government needs to play a key role in governing AI, similar to how it plays a key role in emergency preparedness. Deepfakes, explainability, and biases are key issues while handling AI. The government would have a role to bring all the parties together, set standards, and enforce the standards

and regulations. The Office of AI Risk would be the pivotal federal government agency that coordinates all the activities.

Further Research

The technology around artificial intelligence and artificial general intelligence is evolving rapidly. Beside the core algorithmic research, there are several areas for further research.

Given the proposed vast responsibilities for The Office of AI Risk, an entire proposal can be written on the governance structure of the organization, ways to effectively work with other government organizations, efficient ways to collaborate with the industry, industries the organization should focus on for AI risk, and the best ways to bring researchers together to solve problems.

Further research can also be conducted in bias identification. We need to find effective ways to identify biases in complex AI systems and standard ways to measure biases in the model. This is an area of ongoing academic research.

Building explainable models and AI systems is an area of ongoing research. Much progress has been made to build explainable toolkits, but they are specific to the types of models. Testing for explainability is a complex area, given that the interplay between AI and domain knowledge is a critical area of research.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] *Statement on AI Risk: Cais*. Statement on AI Risk | CAIS. (n.d.). https://www.safe.ai/statement-on-ai-risk
[2] Gates, B. (2023, July 11). The risks of AI are real but manageable. gatesnotes.com. https://www.gatesnotes.com/The-risks-of-AI-are-real-but-manageable
[3] WP Company. (2023, May 27). Opinion | congress wants to regulate AI. here's where to start. The Washington Post. https://www.washingtonpost.com/opinions/2023/05/26/ai-regulation-congress-risk/
[4] CFPB acts to protect the public from black-box credit models using complex algorithms. Consumer Financial Protection Bureau. (2022, May 26). https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/
[5] California, S. of. (n.d.). California governor's Office of Emergency Services: California's emergency services leader. California Governor's Office of Emergency Services | California's Emergency Services Leader. https://www.caloes.ca.gov/
[6] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," Science, vol. 365, no. 6456, pp. 885–890 Aug. 2019. DOI: 10.1126/science.aay2400.
[7] Cellan-Jones, R. (2014, December 2). Stephen Hawking warns artificial intelligence could end mankind. BBC News. https://www.bbc.com/news/technology-30290540
[8] Brooks, R. (2021, October 20). The seven deadly sins of AI predictions. MIT Technology Review. https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions/
[9] Definition of Artificial General Intelligence (AGI) - gartner information technology glossary. Gartner. (n.d.). https://www.gartner.com/en/information-technology/glossary/artificial-general-intelligence-agi
[10] Gartner report G00792868 Published May 31 2023 "The Future of AI: Reshaping Society"

[11] Ramesh R, Saluja S, "How Artificial Intelligence can aid screening and detecting lung cancer in lung cancer patients". Journal of Student Research, vol 12, issue 2 (2023) https://doi.org/10.47611/jsrhs.v12i2.4238

[12] Chesney, Robert and Citron, Danielle Keats, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (July 14, 2018). 107 California Law Review 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21, Available at SSRN: https://ssrn.com/abstract=3213954 or http://dx.doi.org/10.2139/ssrn.3213954

[13] Schmelzer, R. (2019, July 24). Understanding Explainable AI. Forbes. https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/?sh=4937bd697c9e

[14] Oversight of A.I.: Rules for Artificial Intelligence | United States Senate Committee on the Judiciary. (2023, May 16). www.judiciary.senate.gov. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence

[15] Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What Do We Do About the Biases in AI? Harvard Business Review; Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[16] DEEP Increasing Threat of AKE Identities. (n.d.). https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

[17] Silverman, C. (2018, April 17). How To Spot A DeepFake Like The Barack Obama-Jordan Peele Video. BuzzFeed. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

[18] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023

[19] https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/