A Proposal for the Establishment of an Artificial Intelligence Safety Agency

Thomas Dunn

**Abstract**

With the rapid and widespread adoption of Artificial Intelligence (AI) and machine learning (ML) algorithms in cyberspace and risks posed by their implementations, it is important to establish an agency to independently oversee developments in all industries affected by such systems, to protect the rights and safety of individuals and human civilization, guided by technical expertise.

**Background**

In the past decade, the field of Artificial Intelligence has seen rapid progress in a variety of disciplines, from game playing [1,2], computer vision [3], recommender systems [4,5], to autonomous vehicles [6]. Largely consequent of improvements in techniques such as Deep Learning [] or Generative Adversarial Networks [], as well as hardware improvements [],

What is Machine Learning?

Machine Learning, as characterized by

Public companies, private companies, universities, and government agencies have made substantial progress in the field, and the application of the technology has brought enormous amounts of wealth, power, and knowledge to their users.

**Concerns**

Many concerns have been raised over the future use of narrow and general artificial intelligent systems, as well as the negative consequences of systems already created.

Disinformation

COVID

During the coronavirus pandemic, many “bot” accounts, as they are called, flooded the internet with disinformation that can be attributed to the death of many Americans. According the National Institute of Health,

Climate Crisis

Deepfakes

Political Balkanization

Vehicle Control

Self-driving cars

Although safe now, it is important to establish liability structures, to determine whether an algorithm has responsibility or developer. If developer, then certification and licenses should be issued to programmers working on AVs.

Airplanes and UAVs

Civilian

Military

Surveillance

Government

Corporate

Information Hazards

Medicine & Biotech

With increased automation in the drug development process, human oversite is required to ensure

Moore’s law + Genetic code of Spanish Flu? Information Hazard

Cybersecurity

Besides malicious human actors, the release of a sufficiently intelligent self-preserving program onto the internet can pose a threat to internet infrastructure and communications networks.

Weapons

Trading bots

flash crash beyond human control, crypto

Job loss

Transportation

Factory

Lawyers

Insurance

Stock Broker

**Purpose**

Individual agents competing in a young and unstructured market should and cannot be trusted to independently verify the safety and long-term consequences of their systems in an unbiased manner. Whereas the field of medical ethics, for example, was able to establish norms and practices over centuries of implementation, as well as the direct interaction [**Paper from Brent Mittelstadt]**

There are agencies within the Federal Government dedicated to supporting the advancement of the capabilities of the technology, from JAIC to the NSA. There is yet to exist any organization dedicated to holding the AI developers to any significant level of scrutiny within our legal framework, to ensure the principles and ideals espoused are executed effectively.

Although some general values have been formulated, they are abstract to such an extent to leave open for internal interpretation of AI developers. Formalization is required to truly quantify such values.

There exists no serious consequences for violation of self-established ideals, or licensing structure for workers involved in their implementation, even if they were more explicitly defined, so violators can continue to operate unscathed.

Just as plumbers, electricians, pilots, drivers, lawyers, doctors, and other professions whose actions affect the public sphere require licensing to practice their respective fields, AI developers, who often create systems intended to replace these very areas of human expertise, **should** require equivalent forms of licensure.

Explicitly define privacy, safety, security, transparency, and responsibility for developers to be held accountable to.

Just as with the automobile industry before the NHTSA, we currently live in the age before seatbelts on our AI systems.

**Citing experts in the domain of computer science, from Elon Musk to Nick Bostrom, the existential risk of AI is too large to leave its development unregulated.**

Confidential code can remain confidential, tested in a black box sort of manner, but just with car safety testing, pushed to the worst-case scenario to prove robustness in the face of

This document hopes to build a federal agency, comprised of experts and consultants from the public and private sphere alike, from fields related to the Safety of Artificial Intelligence Systems, whose purpose is to monitor, audit, and, if systems are found that violate privacy, safety, or security of individuals, intervene in, public companies, private companies, universities, and government agencies whose work contributes to the development of Artificial Intelligence.

Between both previous presidential administrations, the NSTC has been tasked with monitoring the industry and helping boost development to ensure American dominance in the field of AI. They have specified, across party differences, that the support for the technologies should only continue so long as they provide economic and national security, as well as privacy, safety, and transparency for American citizens. As the concerns have outlined, unregulated artificial intelligence, has, on multiple occasions, violated these standards, despite developers often time holding these ideals as well.

As **this** paper suggests, when abstract principles without clearly outlined definitions are used to guide ethical behavior, discrepancies in the exact meanings of the values will lead to behavior that seems to contradict the vision at the high level. What, exactly does one mean by privacy? What metrics can be used to measure it? Who decides what this definition is? How much safer must a system be than a human at a task, and who has responsibility for mistakes that occur?

For jobs lost to automation, we have seen that scapegoating “offshoring” of jobs is an easy emotional play for political leverage, even though the United States has continued to be one of the largest manufacturing nations in the world. This is mostly due to robotic automation, in many factories where “humans need not apply”. Without wages or working conditions required for the same work, machines look great on the company ledger, and shareholder value, but what of the alienation of the worker? In the coming decade, there will be a rise of self driving cars, trucks, and busses. According to the BLS, the amount of jobs at risk for elimination based on automation in the next decade

**In 2016, NSTC laid**

**Agency**

**Structure**

The Agency shall work in conjunction with the National Science and Technology Council’s (NSTC) Special Committee on Artificial Intelligence to establish norms and practices.

**Funding**

**Consulting**

**On a regular, recurring basis, directors of the Agency will consult with experts in industry, government, and academia in order to keep up to date on the latest changes to the development of Artificial Intelligence systems.**

**Conclusion**

Overall

**References-**

1. Silver, D., Schrittwieser, J., Simonyan, K. et al. Mastering the game of Go without human knowledge. Nature **550,**354–359 (2017). <https://doi.org/10.1038/nature24270>
2. Niels Justesen and Philip Bontrager and Julian Togelius and Sebastian Risi. “Deep Learning for Video Game Playing”. In: arXiv preprint arXiv: 1708.07902 (2019)
3. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: Advances in neural information processing systems. 2012, pp. 1097-1105
4. Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. ACM Comput. Surv. 52, 1, Article 5 (February 2019), 38 pages. <https://doi.org/10.1145/3285029>
5. Heng-Tze Cheng, Levent Koc, et al. 2016. Wide & Deep Learning for Recommender Systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016). Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/2988450.2988454
6. Daniel J. Fagnant, Kara Kockelman, Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations, Transportation Research Part A: Policy and Practice, Volume 77, 2015, Pages 167-181, ISSN 0965-8564, <https://doi.org/10.1016/j.tra.2015.04.003>.

Signatories-