AI Safety Agency Proposal

**Abstract**

With the rapid and widespread adoption of Artificial Intelligence (AI) and machine learning (ML) algorithms in cyberspace, 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 the individuals in firm legal frameworks, guided by technical expertise.

**Background**

What is Artificial Intelligence?

What is Machine Learning?

What is Ethics?

In the past decade, 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.

Climate Crisis

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

Addictive social media algos

Mass government or corporate Surveillance

Deepfakes

Medicine

Cybersecurity

AI-weaponry

**Purpose**

As with any industry, individual agents competing in a market should and cannot be trusted to independently verify the safety and long-term consequences of their systems in an unbiased manner.

There are agencies within the Federal Government dedicated to supporting the advancement of the capabilities of the technology, but 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 so often espoused are executed effectively.

Although general values have been formulated, they are abstract to such an extent to leave open for internal interpretation of AI developers.

There exists no serious consequences or licensing structure to enforce such value, 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 time create systems intended to replace these very areas of human expertise, **should** require equivalent forms of licensure.

Explicitly define privacy, safety, security, transparency, responsibility in developers.

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 still remain that way, tested in a black box sort of manner, but just with car safety testing, pushed to the worst-case scenario to prove robustness. **Twitter bot becoming a Nazi**

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.

**Agency**

**Structure**

The Agency shall work in conjunction with the National Science and Technology Council (NSTC)

**Funding**

**Consulting**

**On regular, recurring basis, the Agency will consult with experts in industry, government, and academia in order to keep up to date on the latest changes to the practice.**

**Conclusion**

Overall

**References-**

1)

2)

3)

Signatories-