**Analysis of AI Governance**

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

**Caroline Hickey**

Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Caroline Hickey

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

# Analysis of AI Governance

## Introduction

The term Artificial Intelligence has been around since the 1950's, when it was first coined by John McCarthy at the Dartmouth Conference. Seventy years later, artificial intelligence is just scratching the surface of its potential. As defined by McCarthy, Artificial Intelligence is, "the science and engineering of intelligent machines" (2007). Although this definition is broad, AI has evolved in a way that makes it difficult to pinpoint one universal definition. AI is no longer a single technology, but a subset of technologies and techniques that range from facial recognition to robotics. Although AI has been around for almost a lifetime, many argue that it is still an emerging technology. "AI today exhibits typical characteristics of an emerging technology such as radical novelty, relatively fast growth, prominent impact, and uncertainty and ambiguity" (Ulnicane et. al, 2020). It wasn't until 2009 that the discussion of AI took off. Due to an increase in computing speed, the potential of what AI can do had reached new feats, allowing the world to tackle problems that require "super" human computing skills (Fast & Horvitz, 2016). Alongside its potential for success is AI's potential for disruption and failure. Public fear of AI can be generalized into five main issues: justice and equality, use of force, safety and certification, privacy, and displacement of labor and taxation (Gasser & Almeida, 2017). The future of AI relies on how the world regulates, governs, and increases trust in this emerging technology. Unlike other technologies the world has seen, the full impact of AI is largely unknown, as it has the potential to be involved in nearly every field and every aspect of life. The rapid pace of innovation of AI signals a need for its governance. It is important to note that governance is different from government. Political policy and law-making will not suffice to control this potent technology. Governance is defined as, "the mechanisms whereby societal

actors and state actors interact and coordinate to regulate issues of societal concern" (Ulnicane et. al, 2020). The collaboration of state and non-state actors, developers and CEOs, nations and people, will lead us to the most comprehensive and effective governance strategy. Only recently, countries have become concerned with this very objective. While little law has been put into action, discussion and concern of these topics has increased. With international collaboration, standards and regulations of AI should be developed in order to keep pace with the fast rate of AI development in a safe and fair manner.

## Review of Some Proposed Policies

### Executive Order 13859

In 2019, the United States Government and President issued an Executive Order that promoted AI research and development in a way that is safe and trustworthy (Exec. Order No. 13859, 2019). This executive order intends to "ensure that technical standards…reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies…and develop international standards to promote and protect those priorities" (2020). This will be done by promoting AI R&D through funding, creating AI standards, producing and deploying AI systems, and developing supportive and regulatory policies. This Executive Order's main focus is on stimulating AI development so that the U.S. will become the AI hub of the world. There are many references to protecting American advancements in AI from other nations in order to increase national security. This aims to be done through federal funding, reducing barriers, and creating an American AI workforce. The Assistant to the President for National Security Affairs was assigned to create a plan, "to protect the United States advantage in AI and AI technology critical to United States economic and national security interests against strategic competitors and adversarial nations" (Exec. Order No. 13859,

2019). Much of the work of creating plans which can be implemented into action is delegated to other agencies. In addition to the NSA, NIST, the National Institute of Standards and Technology, was assigned to develop technical standards through consultation with the private sector, academia, and other non-government entities. Through delegation, people who are experts in the field of AI, and not politicians, will be able to create knowledgeable and thorough standards. Education in AI is also a high priority in the EO. Agencies are urged to create Federal fellowships, as well as increase grants and scholarships for AI research and development. Programs for education in AI technologies within school curriculums as well as independent programs are proposed for American citizens. Another important component of the Executive Order is the provision of data and computing resources for AI R&D. The US government plans on increasing the access and use of Federal data for non-federal AI researchers. Part of this includes ensuring that privacy and civil liberties are protected despite the increase in access.

While the Executive Order aims to create standards for AI, it is important to keep in that it was made in response to a Department of Defense report detailing the threat of other countries becoming leaders in the AI race (Department of Defense, 2018). The document's ultimate goal urges for more development and research of AI in order to win the AI race and advance US political and economic interests. Funding and the stimulation of AI R&D is helpful to speed up advancements in AI, but regulation and standards must be established first. The AI race is causing the United States, as well as other nations, to value quick development over safe development. "Under the assumption that the first AI will be very powerful and transformative, each [nation] is incentivized to finish first – by skimping on safety precautions if need be" (Brachtl, 2020). In fact, "if AI is developed with militaristic intentions in mind, there is a strong chance it will initially be purposed as a weapon" (Brachtl, 2020). The

U.S. stance on AI research and development seems to be guided more by geopolitics than actual interest in regulating and governing AI. Additionally, the lack of conversation about collaboration might cause further division of nations and create barriers for the need of international AI standards. While the United States' plan for AI research and development is a good start to begin the conversation about AI governance, it does not emphasize safe and fair development of AI or collaboration with international entities. The Executive Order does succeed in addressing one of the biggest bottlenecks in the advancement of artificial intelligence: data preparation. High quality, unbiased data is required to train and test these systems. A greater access to quality data will help researchers overcome this bottleneck and increase the time spent on the actual AI system itself.

**Wilson Briefs**

The Wilson Center, a non-partisan policy forum that champions independent research and open dialogue, published a set of recommendations for policymakers, advisory bodies, and funders in order to tackle the fears of AI while maximizing the benefits of AI for society. They focus on "incorporating human participation into complex socio-technical systems to ensure the safe and equitable development of automated intelligence" (Bowser et al, 2017). Because AI development does not take place in a void, the Wilson Center suggests approaching AI research and development in a way that involves all stakeholders and technologies, while promoting human participation. "The conscientious development of AI systems that carefully considers the coevolution of humans and technology in hybrid thinking systems will help ensure that humans remain ultimately in control, individually or collectively, as systems achieve superhuman capabilities" (Bowser et al, 2017). Not only should development continue, but deep investigation and research of AI by leading researchers from the private sector should be

conducted. This will help policymakers to better understand and anticipate what is to come in the field of AI. Once research is conducted, regulation is highly recommended. According to the Wilson Brief, safety and transparency requirements would "promote innovation yet force accountability" (Bowser et al, 2017). The Wilson Center champions innovation, so they recommend AI research and development funding through grants and prizes. This will promote collaboration by bringing together government and non-government actors as well as prompting healthy competition that will spur innovation.

The Wilson Briefs on AI Governance included recommendations that covered research, funding, standardization, and collaboration. Although their recommendations addressed many of the fears of AI, it did not provide the manner in which to execute their recommendations. Their framework lacked the depth and description needed for proper and realistic implementation. This is most likely because of their non-partisan stance, trying to please all parties. That being said, their ideas were sound and feasible. They highlighted the fact that policymakers know little about the technical aspects of AI making it difficult for them to properly guide discussions of policy. A major aspect of their recommendations is collaboration with leading researchers and other experts in the field. Another interesting recommendation made was to recognize other emerging technologies in the AI research and development process. This recommendation takes into account the fact that AI is not one singular technology, but an emerging set of technologies that are still growing. This helps to mitigate situations where new technologies that might fall under AI are not properly regulated.

**Singapore's AI Governance Framework**

Singapore, a top leader of AI research and development, created a framework for AI governance that, in theory, would be used by any company producing new artificial intelligence

systems. Although this framework is not legally binding, it is the start of government intervention in the safe and fair development of AI. The framework, called the Model AI Governance Framework, provides a guide and set of expectations for companies to address ethical and governance issues when creating AI. The framework is based off of two guiding principles: “the decisions of AI should be explainable, transparent, and fair”, and “AI systems should be made human-centric” (Singapore, 2020). In order to bring their guiding principles to fruition, they focus on four key areas that they believe are comprehensive in covering the main principles: internal governance structures and measures, determining the level of human involvement in AI-augmented decision making, operations management, and stakeholder interaction.

Part of having a transparent and fair AI system is creating internal systems to ensure the oversight of AI development. The Model for AI Governance suggests setting up internal governance systems for establishing clear roles and responsibilities as well as risk management systems (Singapore, 2020). Singapore suggests creating clear and district roles made for appropriate personnel who are involved in AI development. This includes proper training and guidance so that the AI development is seamless, ethical, and transparent. Another part of creating internal governance is risk management and internal controls. This would include understanding the bias in the datasets and addressing the risk that comes with the bias. It is also recommended to establish monitoring and reporting systems to ensure that the development is transparent and that issues can be detected and resolved quickly. The Model AI Governance framework also focuses on operations management in order to create a transparent and explainable process. This includes data preparation and algorithm development and selection. It is recommended that good data accountability practices are put in place, which includes ensuring

the quality of data, understanding where the data came from, minimizing bias in data, and periodic reviews of datasets. Algorithms must also be assessed based on measures such as explainability, repeatability, robustness, regular tuning, reproducibility, traceability, and auditability in order to promote transparency and fairness of the algorithms.

One of the guiding principles emphasizes human involvement in AI systems. Not only does this help increase trust in AI systems, it also provides checks and balances so that AI decisions are monitored. The framework suggests three different forms of human involvement that should be able to fit any AI system: human-in-the-loop, human-out-of-the-loop, and human-over-the-loop. Human-in-the-loop would require humans to make final decisions based on AI recommendation and input. Human-out-of-the-loop signifies no human involvement in executive decision making and no option for human override. Human-over-the-loop allows AI be the decision maker, while human actors provide monitoring and supervising of the system with the ability to take control when needed. Companies should choose their human involvement based on the probability of harm versus the severity of harm caused by the system. For example, a system with high probability of harm and high severity of harm should consider human-in-the-loop intervention. Human-out-of-the-loop should only be used in low severity, low probability systems. Companies are also advised to build trust with their stakeholders, the other human component of AI research and development. This includes disclosing to consumers that AI is being used, creating transparency in a meaningful and easy-to-understand way, and creating communication channels for customers.

Included with the framework is an Implementation and Self-Assessment Guide for Organizations, or an "ISAGO". This provides a guide for companies to identify potential gaps in their development process and to fix the gaps (Singapore, 2020). ISAGO is a series of questions

based on the Model AI Governance Framework for companies to consider as well as examples on how companies could implement these new AI standard practices. Singapore's framework supports its call for explainable, transparent, and fair AI decision-making as well as creating solutions that are human-centric.

Singapore's Model AI Governance Framework is a comprehensive and easy-to-use guide for what AI development should look like (Singapore, 2020). Unfortunately, this framework is entirely voluntary. No company that is creating AI systems is required to follow or even consider the guidelines set forth in the framework. Considering only companies who care about the ethical and safe development of AI will follow the guide, the effectiveness of the Model AI Governance Framework is questionable. Any barrier, financial or other, that companies face during the implementation of the framework might cause them to abandon their efforts and resort to easier and cheaper options. Nevertheless, Singapore has provided an incredibly comprehensive and accessible framework that has potential to change the way that AI is developed. This framework was created through collaboration with various nations as well as input from consumers and companies. The risks and benefits of AI was well researched and it is evident that experts in AI and business contributed to the development of the framework. In conversations surrounding AI, the fear of displacement of labor and the lack of trust of autonomous systems is a considerable concern. This framework addresses these fears by emphasizing the role of humans in AI development and decision-making. This will not only increase public trust; it also creates a system of checks and balances to ensure the artificial intelligence is being used properly. In turn, this addresses other fears such as injustice and safety. Possibly the only flawed aspect would be the lack of conversation about funding. Singapore's approach to AI Governance is angled towards private sector responsibility. While funding only helps speed up how quickly AI is

developed, grants, scholarships, or funding smaller companies would help create some much-needed diversification as, “the vast majority of the development of AI and all its associated elements (development platforms, data, knowledge and expertise) is in the hands of the “big five” technology companies” (Ulnicane et. al, 2020).

**Best Option**

There is not one “best” option proposed as of right now that would sufficiently undertake the responsibility of governing artificial intelligence. Despite that fact, there are organizations and governments that are having the right conversations and leading the world in the right direction. Creating standards and regulations all while stimulating AI research and development is not a simple task and requires a complex and multilayered solution. Taking note from policy analysis, there are a few key aspects that must be met in order to produce safe and fair AI. For one, humans must be involved in the development and decision-making of AI. Singapore’s risk analysis of human involvement model is a good example of promoting human participation in situations that are beneficial. Another key aspect is funding, grants, and scholarship. It must be noted that throwing money at large companies that already invest millions into AI research will not necessarily help create better AI or policy. Increasing the flow of money into AI R&D with the intent of diversifying the market, however, will incentivize smaller companies and researchers to focus on AI systems and could create more healthy competition. The best option to execute these tasks would be to do so over time. Immediate implementation of these policies is not practical. A multi-year approach would best fit the current world of AI. “It is important to note that any such emerging model must be situated in and interact with existing institutional frameworks of applicable laws and policies, particularly human rights, as the development and deployment of AI does not take place in a vacuum” (Gasser & Almeida, 2017). Because there is

no existing framework in place for AI as it is different from any prior technology, it is important to start implementing a framework that can meld with other existing institutional frameworks. Currently, countries such as the United States govern technologies without precaution; that is, allowing for the development of systems regardless of risk and only providing regulation after a risk has come to fruition. Europe, on the other hand, governs using the precautionary principle. This principle states that, "when an activity raises threats of harm to human health or the environment… the proponent of an activity, rather than the public, should bear the burden of proof" (European Commission, 2017). The idea of the principle is that any course of action that has a chance of causing damage, especially when there is scientific uncertainty on the matter, should be regulated first regardless of the cost of regulation or lack of evidence. In other words, it is better to be safe than sorry. This approach would prove useful in the discussions of AI. As of now, many are leaping into the benefits of AI without fully uncovering the risks. It is important to regulate these emerging technologies before implementing them into our daily lives. It is safer to regulate than to do damage control after the fact.

AI governance is a difficult topic to tackle, in part because of how complex and intertwined a solution would be. Nonetheless, it is incredibly necessary in order for the world to continue its development of AI systems. Without governance, the fears associated with AI may materialize. By developing standards and creating a framework for AI research and development, the world can lessen the consequences of AI failures.

## References

Bowser, A., Sloan, M., Michelucci, P., & Pauwels, E. (2017). *Artificial Intelligence: A Policy-Oriented Introduction*. 18.

Brachtl, C. (2020). *A Battle of Mutual Undoing: The AI Arms Race*. Journal of Sino-American Affairs, 2-10.

United States Department of Defense. (2018). *Summary of the Department of Defense Artificial Intelligence Strategy*.

European Commission. Directorate General for the Environment. & University of the West of England (UWE). Science Communication Unit. (2017). *The precautionary principle: Decision making under uncertainty.* Publications Office.

Exec. Order No. 13859, 3 C.F.R. 2-10 (2019).

Fast, E., & Horvitz, E. (2016). *Long-Term Trends in the Public Perception of Artificial Intelligence*.

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, *21*(6), 58–62.

McCarthy, J. (2007). *What is Artificial Intelligence?* 15.

Singapore, Personal Data Protection Commission. (2020). *Companion to the Model AI Governance Framework: Implementation and Self-Assessment Guide for Organizations*.

Singapore, Personal Data Protection Commission. (2020). *Model AI Governance Framework* (2nd ed.).

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2020). Framing governance for a contested emerging technology: insights from AI policy. *Policy and Society*.