The Intersection of Ethics and Artificial Intelligence in U.S. Federal Cybersecurity: An In-Depth Analysis

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

Kevin Michael Carlson

Spring 2024

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

Advisor

Gerard Fitzgerald, Department of Engineering and Society

Introduction

The United States government faces increasing cyber threats, occupying a threshold larger than human technical resources can handle (Norris et al., 2015). Recent advances in artificial intelligence (AI) have made it possible for defense organizations to execute offensive cyberattacks and mitigate threats autonomously. As with any new technological system, a systematic analysis must be acknowledged so that the subsequent utilization of such a system adheres to the ethical codes and customs held by the society in which it is created (de Almeida et al., 2021). Federal cybersecurity is a growing area of concern as superpowers worldwide become increasingly fluent in their cyber capabilities. While AI remains in its early years, professionals in both AI and ethics have begun outlining the necessary guardrails and guidelines for implementing AI in industry practices. As it pertains to cybersecurity, the applications for AI must be meticulously scrutinized due to the sensitivity and volatility of the field of national cybersecurity. The research completed in this project will provide insight into the current ethical frameworks for AI employed by the European Union and China. These policies will be compared against each other and current ethical literature for AI use in the U.S. national government. The paper will conclude with a suggestion for the U.S. federal government to instill an ethical framework for future advances in AI technology within their cybersecurity ventures.

Ethical Frameworks for AI Around the Globe

Current literature provides a wide range of insight on the ethical implications of artificial intelligence in the scope of cybersecurity with a common theme: As AI evolves, the ethical guidelines surrounding it must evolve in an equal proportion. AI and cybersecurity are complex fields in their own right, and the literature surrounding them individually proves the growing

amalgamation of the two. Among AI research in government, current literature points to the competitive global marketplace for advances in AI technology and the different utilization techniques each country employs (Kwon, Lim, 2021). As AI becomes increasingly relevant in various disciplines, researchers have observed the need for foundational guidelines to address the concerns of privacy, accountability, transparency, and bias (Shneiderman, 2021). The execution of this is already taking shape in a variety of different ways, and the examination of such guidelines is imperative for distinguishing potential policy reform for the U.S. to employ

Legislation is currently being processed in the European Union to pass the "EU AI Act", originally proposed by the European Commission in April 2021. The aforementioned act would help to ensure better conditions for the development and use of AI by establishing obligations for providers and users depending on the level of risk from artificial intelligence (*Regulatory Framework on AI | Shaping Europe's Digital Future*, 2022). The EU AI Act adopts a regulatory framework, defining 4 levels of risk in AI: Unacceptable risk, high risk, limited risk, and minimal risk. The levels of risk define the applications of AI that fall into each level as well as the strict obligations that must be met by the systems before being placed on the market. A forward-thinking aspect of this framework is its future-proof design; the proposal allows rules to adapt to technological change and requires ongoing quality and risk management analyses by providers. This framework addresses the use of AI principles in a variety of industries, allowing for a more comprehensive law. These principles are reflective of broader ethical standards society is held to as a whole. "Ethical discourse addressed both private virtue and public behavior, and an ethical citizen was expected to habitually behave in a manner that contributed to the public welfare" (Fried, 2011). While Fried was referencing human ethics in society, the statement can be overlaid with the implementation of AI in society. The behavior an AI conducts

itself should reflect that of the society it is built in, which is what the AI Act aims to do by regulating risk levels, just as a society would do for a human citizen.

The People's Republic of China (PRC or China) recently finalized its Interim Administrative Measures for Generative Artificial Intelligence in August 2023. These measures apply to the provision of generative AI services provided to the public in China and impose broad obligations on providers of generative AI services. These obligations include monitoring content, marking generated content, training data, data labeling, protecting personal information, and agreement between users and providers (Sheng et al., 2022). A unique feature of this regulation is the elastic approach to industry integration. Different industries will be required to strengthen their regulations of generative AI elastically to maintain proper regulatory compliance as AI is deployed. While this framework touches on certain positive aspects of responsible AI such as respecting IP rights, ensuring transparency, and minimizing discrimination, the key principle rooted in the framework's design is upholding core socialist values. While not surprising due to the strict history of China's media policy regarding information distribution, it remains a necessary anecdote for maintaining discretion when contrasting the framework as a whole. These AI measures in China were released as part of the State Council's (the chief administrative body within China) ongoing mission from their development plan released in 2017. The 'New Generation Artificial Intelligence Development Plan' acts as a unified document that outlines China's AI policy objectives. "By 2025, China aims to have achieved a 'major breakthrough' in basic AI theory and to be world-leading in some applications" (Roberts et. al., 2020, p. 60). Roberts and his colleagues describe the strategic focuses China outlines for AI as international competition, economic development, and social governance. These three characteristics of national growth may be vital to a country's global success and presence but

when put at the forefront of innovation for AI, the integrity of ethical guidelines in the future becomes tainted with the thought of secretism and ignorance. While China remains a comparable technological powerhouse to the U.S., the U.S. must differentiate itself from China in terms of integrity when developing AI as a whole.

Ethical Frameworks in the U.S.

The United States is at the forefront of technological innovation globally. The power it brings incites constant change to the ethical and political landscape of the country. Dr. Joan Fontrodona contends that "scientific and technical progress always raises new ethical questions. And it is the answers to these questions from the sphere of ethics that lead to scientific and technical tools" (Fontrodona, 2013, p. 27). It can be reasoned from this statement that AI will continue to progress, and it is the responsibility of the U.S. government to properly set ethical boundaries for maximizing innovation. Ethical change is already underway in national defense. In late 2023, Congress passed the Department of Defense Ethics and Anti-Corruption Act. The legislation has the goals of restricting contractor and foreign influences as well as ensuring defense contractor transparency.

The federal government has been an active player in artificial intelligence development and deployment for years. In 2019, President Donald J. Trump launched the American Artificial Intelligence Initiative, the United States' national strategy for maintaining American leadership in AI. The initiative emphasizes six key policies and practices: Invest in AI research and development, unleash AI resources, remove barriers to AI innovation, promote an international environment supportive of American AI innovation, embrace trustworthy AI for government services and missions, and train an AI-ready workforce (Saveliev & Zhurenkov, 2020). With the

government's support of AI innovation, we are likely to see vast gains in AI technologies at an exponential rate. Even though the US leads the global market with over 54% of AI start-ups being based in the US, regulatory engagement has taken a considerably longer time to catch up to innovation (Whyman, 2023). A unique aspect of American AI ethics has been the distribution of responsibility to government agencies for ensuring relevant ethical considerations in unique use cases. Air Force Lt. Gen. Jack Shanahan, director of the Joint Artificial Intelligence Center said in a press release, “We owe it to the American people and our men and women in uniform to adopt AI ethics principles that reflect our nation’s values of a free and open society.” Shanahan is referencing the Defense Department’s adoption of five principles for the ethical development of artificial intelligence capabilities in February 2020. The five AI ethical principles, based on recommendations from the Defense Innovation Board, are “Responsible, Equitable, Traceable, Reliable, and Governable” (Lopez, 2020). The Defense Department is a major contributor to U.S. federal cybersecurity and these five principles can be elaborated on in the future to benefit multi-agency cybersecurity ethics as the number of use cases for AI in federal cybersecurity expands.

Intersection of AI Ethics and Cybersecurity

The rapid growth of information technology (IT) around the world as a result of technological advancements in the internet and the presence of electronic data has been revolutionary. As a result, cybersecurity has become increasingly critical for business information management in all fields. All organizations search for appropriate policies and security measures to respond to and prevent cyberattacks, which means new technologies are bound to be examined and tested for applications in the field. Unfortunately, according to Gemalto’s Breach Level index, 60% of cybersecurity attacks lasted less than an hour and relied

on new forms of malware (Thales Group, 2018). The cybersecurity space is evolving rapidly which makes AI a fantastic aid in staying ahead of malicious entities. While the cross-section between AI and cybersecurity is relatively new, the combination of the two growing industries has a large number of potential applications together. As a result, it also becomes an issue of ethical policy. Similar to applying AI to other industries, certain ethical challenges arise when AI is introduced into cybersecurity. Timmers references such problems in his paper *Ethics of AI and Cybersecurity When Sovereignty is at Stake*, "Extensive monitoring and pervasive risk-prevention with the help of AI can be highly intrusive and coercive for people, whether employees or citizens. AI can also be so powerful that people feel that their sense of being in control is taken away. They may get a false sense of security too. Deep-learning AI is, as of today, not transparent in how it decides so many data points, yet an operator may blindly trust that decision. AI also can incite freeriding as it is tempting to offload responsibility onto 'the system'." (Timmers, 2021, p. 637). Should an AI system miss the presence of exploitations in critical infrastructure, such as telecommunications, the risk is substantial to the general public's privacy, welfare, and residual trust in these systems in the future.

Qualitative principles play a pivotal role in shaping the responsible development and deployment of artificial intelligence systems. These principles, as articulated by Shneiderman (2021), can be categorized into eight key domains: privacy, accountability, safety, transparency, non-discrimination, control of technology, responsibility, and promotion of human values. Each of these categories represents fundamental considerations that guide ethical decision-making in AI development and usage. Interestingly, these principles exhibit significant overlap with the ethical frameworks governing cybersecurity. This alignment underscores the connectedness of ethical principles across AI and cybersecurity domains, emphasizing the importance of

integrating ethical measures into both fields to ensure the responsible advancement of technology.

Enhancing AI ethics in U.S. federal cybersecurity

The United States government typically takes a decentralized approach when approaching regulation (Whyman, 2023). While different from the typical regulatory policy created in the EU, a bottom-up approach allows for agency-to-agency control in creating efficient and effective ethical guidelines. Due to the breadth of federal cybersecurity, allocating responsibility to individual federal agencies is the best route for policy creation. In this method, the role of synthesizing relevant ethical guidelines of the mission and use cases that each cybersecurity agency has falls on the agency itself. I propose a rights-based ethical framework for these agencies to employ. "Which option best respects the rights of all those who have a stake?" That is the question asked by a rights-based framework. Derived from the biggest potential issues with federal cybersecurity, a rights-based approach will prioritize the lives of the American public and ensure AI in cybersecurity endeavors remains ethically conscious of the rights and freedoms of those that such technology is affecting. Another framework typically used for technological innovation is the "Common Good" approach. This framework deals with optimizing what is good for the community at large. With public trust and transparency being key factors in AI development for cybersecurity, quantifying the common good would blur the ethical lines through conflicts of interest. When looking at AI ethics in its entirety, there are a plethora of ethical principles that can be associated. Introducing a framework for federal cybersecurity applications in AI must distinguish and elaborate on the most relevant principles for the most effective implementation. The enhancement of ethical guidelines on AI in U.S. federal cybersecurity should address three key concerns: privacy, bias, and transparency.

Privacy

Agency-related cybersecurity missions have been scrutinized for invasions of privacy to the American public. In 2013, a computer analyst named Edward Snowden released a substantial amount of information regarding NSA contracts dealing with the compilation of citizens' private communications. As a major catalyst for awareness about federal cybersecurity, the actions of such agencies have become increasingly criticized by the public. Following the Snowden leaks, a study by the Pew Research Center concluded that 63% of respondents felt their privacy was violated through the NSA's collection of personal data (Pew Research Center, 2013). With the introduction of AI in agency-related cybersecurity, maintaining public trust through privacy regulation in the federal government remains vital for continuing action. As an ethical principle in AI, privacy is the protection of personal data and the right to select who has access to it. Developing an AI system includes extensive monitoring of a system and comprehensive data collection to train the AI, which can greatly improve system resilience, but can expose users to unwanted risk should confidentiality be breached (Taddeo, 2020). It is imperative to establish robust privacy safeguards and transparent procedures to address this risk effectively. Clear policies on data storage, access, and usage are essential for ensuring the protection of individual privacy rights when deploying AI technologies. A broad ethical solution to implementing this policy is purpose-oriented data collection and analysis (Blanchard and Taddeo, 2023). Data used to extract intelligence-related information should only be collected and analyzed after being assessed for a given purpose to meet the principles of necessity and proportionality. The two colleagues argue that relevance assessment should be based on the likelihood of a specific type of data revealing relevant information for a given purpose and should be context-dependent. A key principle of this is that relevance assessment is conducted before, not after, the collection of

data. These standards would decrease the likelihood of marginally relevant data being collected solely for data acquisition and large-scale predictive analysis.

Bias

Minimizing social and statistical bias in AI systems is a problem developers face constantly. Social bias refers to human-created biases, such as stereotypes that may be reflected in AI systems while statistical bias refers to the systematic error in an AI system's predictions that arise from biased data or algorithms. Inherent biases exist everywhere, some with unethical foundations that constrain cognition and inhibit an individual's ability to make ethical decisions (Watts et al. 2019). As new AI technological systems are developed and deployed, the responsible parties involved in these steps must make a conscious effort to include discretionary decision-making within the systems to properly recognize, monitor, and mitigate biases. Overlooking bias in an AI system in a cybersecurity discipline has the potential to make discriminatory decisions and target innocent lives. Facial detection and recognition software (RDT), an AI technology used in the realm of cybersecurity, recently wrongfully identified Harvey Murphy Jr., an innocent African American male, in an armed robbery. The media from this event raised alarms about algorithmic bias and its ability to misidentify, racially discriminate, and produce other unintended consequences (Fung, 2024). To prevent situations like this from occurring in the future and to train AI systems more effectively, processes for evaluation and mitigation of bias must be actively present in such systems. Blanchard and Taddeo recommend "assessing your data" to best achieve this task. "Analysts relying on AI should be able to access the relevant data set and have adequate technical competencies to assess whether protected characteristics are present and how they are 'read' by the AI system. AI systems should also run on synthetic data to ensure that risks of training a system on biased data

are reduced to a minimum" (Blanchard & Taddeo, 2023, p. 18). The colleagues go on to mention the need for a diverse demographic in teams responsible for evaluating datasets and identifying bias-related risks.

Transparency

The implementation of transparency in federal cybersecurity AI practices is two-fold. The relevant parties must remain transparent with the public on necessary actions, and future AI systems must be developed and deployed with proper, well-documented ethical due process. While confidentiality is critical for many national security missions, public discourse of network information could lead to cyber threats, ultimately damaging public confidence and reputation. Therefore, "assuring a trusted and resilient information and communications infrastructure is needed. A reliable, resilient, trustworthy digital infrastructure for the future enhances online choice, efficiency, security, and privacy" (Yaghmaei et al., 2020, p. 34). With regards to transparent AI development, oftentimes the complexity of deep-learning AI technologies leads to black-box decision-making processes i.e. a process that can only be viewed in terms of input and output, and no interior elaborations are available. Guidelines should be put in place to document and analyze the training procedures, data, and results of AI technologies as they are developed. Vogel and her associates agree with this sentiment in their paper *The Impact of AI on Intelligence Analysis: Tackling Issues of Collaboration, Algorithmic Transparency, Accountability, and Management*. Relevant employees handling AI technology are challenged to behave ethically when dealing with black-box systems. "The intelligence analyst should possess the capacities to: (1) productively leverage these algorithmically produced assessments; (2) recognize the limitations of the technologies in terms of the data they handle and how they handle it, knowing just enough about the tool's inner workings; and (3) identify alternative (possibly traditional)

sources of data that should be leveraged to compensate for technology blind spots." (Vogel et. al., 2021, p. 840). These understandings will necessitate new ways to create testing, evaluative, and auditing processes so that employees and higher-level staff maintain sufficient knowledge of the AI systems they create.

The three foundational principles for instilling ethical AI in cybersecurity (privacy, bias, and transparency) are vital pillars of any agency-wide policy to come in the future. Lopez commented on very similar virtues in the five principles of AI ethics being adopted by the Department of Defense in 2020. Specifically highlighting privacy, bias, and transparency is a firm method for establishing public trust in cybersecurity agencies, and ensuring the rights of the employees of such agencies and those of American citizens are upheld in the decision-making process for using AI in federal cybersecurity. The modeling and deployment of a full-scale ethical framework in federal cybersecurity should include commentary and policy enhancement in areas other than the three principles listed to ensure a greater breadth of decision-making remains bounded by ethics. Additionally, the AI Act in the EU can be drawn on for its future-proof design. The framework allows legislation to adapt to technological change. It also ensures ongoing quality and risk management by providers to maintain the trustworthy design of AI.

Conclusion

Artificial intelligence has an application in limitless industries, with unforeseen ethical risks. Without proper guidelines and policies in place, AI is virtually certain to blur the lines of ethicality on issues such as privacy, transparency, control, and safety. National governments around the world are finalizing documentation consisting of rules and regulations to be set in the

development and deployment of AI systems to uphold national ethical codes. Leaders in the European Union adopted a top-down approach, prescribing rules in a risk-based regulatory framework. The U.S. has strayed behind in standardizing ethics in AI, where the majority of documentation surrounding AI in the U.S. focuses on standard principles such as innovation, investment of resources, and promoting trustworthy AI. Without a comprehensive law defining ethical principles for AI, it is necessary to synthesize an ethical framework for AI in federal cybersecurity. Generalizing the current cross-section of ethics in AI and cybersecurity consists of identifying the fundamental risks associated with both areas. While AI in cybersecurity alludes to guaranteed ethical risks, if the given bottlenecks for ethical longevity are identified and protected, the benefit of deploying new AI technology in the cybersecurity space is substantial. Privacy, responsibility, accountability, and transparency are crucial ethical principles when introducing AI to federal cybersecurity. I believe a bottom-up, rights-based framework is the best approach when handling AI in federal cybersecurity. This solution would offer varied guidelines on an agency-to-agency basis, allowing for a tailored fit depending on the mission and specific development of AI in each cybersecurity agency. As a rights-based framework, the guidelines will prioritize ethical solutions for protecting the rights of American citizens. In an industry that has been plagued with poor media involving privacy, such as the Snowden leaks, public trust must be held to a high standard concerning the implementation of AI in new federal cybersecurity endeavors.

References

de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: a Framework for Governance. *Ethics and Information Technology*, *23*(3), 505–525. https://doi.org/10.1007/s10676-021-09593-z

Blanchard, A., & Taddeo, M. (2023). The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations. *Digital Society, 2.*

Data breaches compromised 3.3 billion records in first half of 2018. Thales Group. (2018, October 23). https://www.thalesgroup.com/en/markets/digital-identity-and-security/press-release/data-breaches-compromised-3-3-billion-records-in-first-half-of-2018

Fontrodona, J. (2013). The Relation Between Ethics and Innovation. In T. Osburg & R. Schmidpeter (Eds.), *Social Innovation: Solutions for a Sustainable Future* (pp. 23–33). doi:10.1007/978-3-642-36540-9_3

Fried, J. (2011). Ethical Standards and Principles. In *Student Services: A Handbook for the Profession* (5th ed., pp. 96–119). essay, Jossey-Bass.

Fung, B. (2024, January 23). *Lawsuit: Facial recognition software leads to wrongful arrest of Texas man; he was in Sacramento at time of robbery*. CBS News. https://www.cbsnews.com/sacramento/news/texas-macys-sunglass-hut-facial-recognition-software-wrongful-arrest-sacramento-alibi/

Jones, P., Reid, G., Vogel, K., & Kampe, C. (2021). The impact of AI on intelligence analysis: tackling issues of collaboration, algorithmic transparency, accountability, and management. *Intelligence & National Security, 36.*

Lim, J., & Kwon, H. (2021). A Study on the Modeling of Major Factors for the Principles of AI

Ethics. In *DG.O2021: The 22nd Annual International Conference on Digital Government Research* (pp. 208–218). Association for Computing Machinery.

Lopez, C. T. (2020, February 25). *DOD adopts 5 Principles of Artificial Intelligence Ethics*. U.S. Department of Defense. https://www.defense.gov/News/News-Stories/article/article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/

Norris, D., Joshi, A., & Finin, T. (2015). Cybersecurity challenges to American state and local governments. *Proceedings of the European Conference on e-Government, ECEG, 2015, 196-202.*

Pew Research Center. (2013, June 17). Public split over impact of NSA leak, but most want Snowden prosecuted. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2013/06/17/public-split-over-impact-of-nsa-leak-but-most-want-snowden-prosecuted/

Regulatory framework on AI | Shaping Europe's digital future. (2022, September 29). Digital-Strategy.ec.europa.eu; European Commission. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & SOCIETY*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Saveliev, A., & Zhurenkov, D. (2020). Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China. *Kybernetes, ahead-of-print.*

Sheng, Jenny, Ko, J., Liu, J. Y., Farmer, S., Chunbin Xu, Wenjun Cai, & Fred Ji. (2023). China

Finalizes Its First Administrative Measures Governing Generative AI. *Intellectual Property & Technology Law Journal*, *35*(8), 17–19.

Shneiderman, B. (2021). Viewpoint Responsible AI: Bridging From Ethics to Practice: Recommendations for increasing the benefits of artificial intelligence technologies. *Communications of the ACM*, *64*(8), 32–35. https://doi.org/10.1145/3445973

Timmers, P. (2019). Ethics of AI and cybersecurity when sovereignty is at stake. *Minds and Machines*, *29*(4). https://doi.org/10.1007/s11023-019-09508-4

Whyman, B. (2023, October 10). *AI Regulation is Coming- What is the Likely Outcome?* www.csis.org; Center for Strategic & International Studies. https://www.csis.org/blogs/strategic-technologies-blog/ai-regulation-coming-what-likely-outcome

Yaghmaei, E., van de Poel, I., Christen, M., Gordijn, B., Kleine, N., Loi, M., Morgan, G., & Weber, K. (2017, December 28). *Canvas White Paper 1 – Cybersecurity and Ethics*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3091909#:~:text=%2D%20White%20Paper%201%20%E2%80%93%20Cybersecurity%20and,ethical%20discourse%20on%20cybersecurity%20are