# A Multi-Faceted Approach for Trustworthy AI in Cybersecurity

**Xiang (Michelle) Liu**
**Marymount University**

**Diane Murphy**
**Marymount University**

*Artificial intelligence (AI) is an increasingly used technology in today's society due to its seemingly limitless use cases, from automation to augmentation and beyond. However, some questions remain unanswered, including what the consequences of these implementations are, how such implementations impact humans, and how trustworthy and secure they are. This article focuses on the implications of AI for the field of cybersecurity: new applications to secure, new attack vectors for malicious actors, and new ways to protect our systems with AI. The authors argue that it is imperative to incorporate ethics and related responsibility principles as central elements in the design and operation of AI systems for effective cyber defense. This paper proposes a multi-faceted approach to respond to the emerging challenges associated with AI and emphasizes what is needed now to expose students to AI in their curriculum. Due to the exploratory nature of this study and the newness of the field, our goal is to invite further discussion and investigation of this important subject, and to begin developing a curriculum to introduce trustworthy AI throughout the curriculum.*

*Keywords: Artificial Intelligence (AI), cybersecurity, trustworthy-AI, ethics, education*

## INTRODUCTION

Technology is rapidly changing and shaping how organizations work and people live. As computing has become more powerful and algorithms more sophisticated, various business processes are being automated to achieve specific goals, such as making predictions and performing tasks without being specifically programmed (Grove & Meehl, 1996; Koren, Bell, & Volinsky, 2009; Yeomans, 2015, July 07). More software is incorporating the ability to learn and predict, which embodies the essence of artificial intelligence (AI). Although such technological advances improve efficiency and enhance business processes, the consequences of their implementations, how such implementations impact humans, and how trustworthy and secure they are, remain open and riveting questions in the field (Jago, 2019). For instance, there has been a substantial body of recent work on Trustworthy AI design, as well as the security implications/issues of AI technologies, including the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2019), a collaborative initiative to form a global network of professionals and academics to develop AI responsibly (The Institute for Ethical AI and Machine Learning, 2020), and a Trustworthy AI Framework proposed by industry leaders (Saif & Ammanath, 2020).

As a recent technological advance that is growing at an unprecedented rate, AI and its applications are at the forefront of these questions. In particular, AI raises major concerns due to its "black box" characteristics (Accenture Federal Services, 2018) and "dual-use" nature (Brundage et al., 2018). In this article, the authors focus on the implications of AI and its impact on the cybersecurity field: new applications to secure, new attack vectors for malicious actors, and new ways to protect our systems with AI. We also consider what curriculum is needed to expose all students to AI as it will impact everyone. Due to the exploratory nature of this study and the newness of the field, our major goal is to draw attention to the areas discussed and propose trustworthy AI, which integrates multiple extant frameworks, as a potential solution to AI's cybersecurity threat. The aim of this article is not to provide a hasty resolution, but to invite further discussion and investigation of this important subject and to begin the development necessary to introduce AI into education to improve cybersecurity.

## BACKGROUND

While the field of AI originally surfaced in the 1950s, it has recently rapidly regained momentum and matured beyond its initial academic focus. The recent revival of AI results from several factors, including the exponential growth of computing power, the availability of improved algorithms, faster iteration and replication of experiments, the ready availability of large data sets (so-called big data), pervasive and ubiquitous networking connectivity, and extensive commercial investment (Jordan & Mitchell, 2015). These factors enable practitioners to implement AI that can "learn" and solve problems in increasingly complex settings (Hester et al., 2017; Jaderberg et al., 2016).

AI is an important technology in today's society due to the seemingly limitless use cases, from automation to augmentation and beyond. For example, AI has been implemented in a wide range of domains to create products, manage business processes, conduct fraud detection, vet resumés, approve loan applications, assist doctors in diagnosing some health conditions, and identify diseased crops (Kamps, 2016; Lotman & Viigimaa, 2020; Mann & O'Neil, 2016; McFarland, 2017; Yu, Beam, & Kohane, 2018). Additional applications of AI with significant recent progress include speech and image recognition, machine translation, spam filtering, language comprehension, driverless cars, and AI-enabled drones for expediting disaster relief operations (Brundage et al., 2018; Gil & Selman, 2019).

However, there seems to be no unanimous agreement on the definition of AI. Different entities have defined it in different ways for a variety of purposes (e.g., Department of Defense, 2018; Vinuesa et al., 2020). Before any further discussion, it is important to have a good working definition of AI. In this article, we adopt the definition of AI from the U.S. National Science and Technology Council (NSTC), which considers AI to be technology that *"enables computers and other automated systems to perform tasks that have historically required human cognition and what are typically considered human decision-making abilities"* (National Science and Technology Council (NSTC), 2019, June). This definition is consistent with that of the 2019 U.S. Executive Order on Maintaining American Leadership in Artificial Intelligence (White House, 2019, February 11) and the National Institute of Standards and Technology (NIST) document for prioritizing federal agency engagement in the development of standards for AI (National Institute of Standards and Technology (NIST), 2019, November 18).

## A CLOSER LOOK AT AI AND CYBERSECURITY

AI techniques are changing the landscape of current battles between defender and adversary in cybersecurity. Of particular concern is the risk that AI could be used as an attack tool, or even an attack surface, to enable larger-scale and more autonomous attacks by an adversary.

### AI as a New Defense Tool

Fighting cybercrime and securing cyberspace is a global mission. Cybersecurity researchers and practitioners have turned to AI to create innovative defense approaches or techniques to improve cybersecurity and fight cyberattacks. As very large amounts of traffic data are constantly being generated, traditional cybersecurity defense measures (e.g., signature-based antivirus software or firewalls) are becoming less effective at monitoring current levels of data volume, velocity, and variety, and thus, are failing to analyze and detect patterns, anomalies, or intrusions in traffic data (Zeadally, Adi, Baig, & Khan, 2020). AI, on the other hand, stands out for application in this field due to its ability to analyze data from millions of incidents and predict potential threats based on this analysis, such as a "zero-day" malware variant. Many vendors are now promoting AI-based products in the cybersecurity space.

### AI as a New Attack Tool

AI can also be manipulated for nefarious purposes. For example, hackers have created intelligent agents to automatically click advertisements, play online games, and buy and resell tickets for concerts, negatively affecting business models (Neal, Kouwenhoven, & Sa, 2015). AI has also been used to manipulate public opinion in Venezuela by retweeting political content (Forelle, Howard, Monroy-Hernández, & Savage, 2015) and has affected the US presidential election by spreading tailored news (Shao et al., 2017).

Deepfakes are another attack vector enabled by AI. Attackers train an algorithm on images, videos, and/or audio of someone's face and voice, and then use software to map that face and voice onto recordings of someone else to create an impersonation to achieve personal power or financial advantage (Zeadally et al., 2020). One recent case involved criminals using AI-generated audio to impersonate a CEO's voice and trick employees into transferring over €220,000 ($243,000) to them (Tung, 2019, September 4). These examples demonstrate the ability of attackers to change the outcomes of business processes to their advantage, and this issue will become more significant as AI becomes more prevalent in mission-critical applications such as in defense, medicine, and transportaion.

### AI as a New Attack Surface

A traditional cybersecurity posture generally overlooks an unprecedented area of attack surfaces: adversarial AI. Adversarial AI targets AI models themselves, allowing attackers to create "adversarial examples" that resemble normal inputs, but which repeatedly make minor changes to the model input to eventually break the model and produce incorrect results (Accenture Labs, 2019).

Adversarial AI became very public in 2016, when the Microsoft Tay chatbot was released via Twitter. It took less than a day for the bot to begin to post inflammatory and offensive tweets, causing Microsoft to shut down the service. Trolls on Twitter had made replies which influenced the chatbot's learning and showed how algorithms can be modified based on adversarial input, even when such input is provided by regular users (Liu, 2017). This approach is also referred to as "data poisoning" today.

Much of the examination of the impact of adversarial AI is currently being performed in research labs. For example, threat modeling and adversarial learning libraries have been developed and used in generating simulations and experiments, such as SecML (Melis et al., 2019) and AdversariaLib (Corona, Biggio, & Maiorca, 2016). In another example, AI researchers were able to modify an AI-based system that generated automated responses from email messages to output personally identifiable information (PII) such as credit card numbers. The data was included in the emails used in the training set. These findings were later used by Google to prevent exploitation of Gmail's smart compose feature, which autogenerates responses to emails (Carlini et al., 2018). Another experiment involved misleading the computer-vision AI-based tools used in autonomous vehicles by slightly modifying road signs by attaching stickers (Eykholt et al., 2018). Adversarial AI is also increasingly important to countries' defenses. A growing number of military applications are using AI for both defensive and offensive purposes (Biggio & Roli, 2018).

## IMPORTANCE OF TRUSTWORTHY AI

As AI has become ubiquitous and pervasive in today's technology-focused world, so has our need to trust the systems we are using, whether it being related to healthcare, energy, financial services, education, transportation, or other sectors. Trust is a major component of credibility and it is important to evaluate trust in the relevant context—for example, the safety requirement imposed on self-driving cars to minimize loss of life. We must be confident that an AI application is making the right decision at the right time for the right reason. This might be trusting the accuracy and effectiveness of the original AI application as well as detecting whether it is still trustworthy after continued use.

Trust is complex and includes components such as confidence, beliefs, willingness to be vulnerable, and expectancy (Halpern & Moses, 1992). Trust can also be considered an ethical concept, reflecting each individual's moral reasoning (Huang & Nicol, 2010). Today, in the world of global networks, there is more recognition of the risk of trusting other users or software applications—so much so that zero-trust architecture is becoming accepted as the necessary way forward in cybersecurity (Rose, Borchert, Mitchell, & Connelly, 2020). Zero-trust architecture provides a model to solve cybersecurity challenges inside and outside the network perimeter by automatically trusting nothing (Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020).). Consequently, we need to prove that any AI application is trustworthy throughout its lifecycle, even though there can be many stakeholders, such as organizations and individuals, involved in the supply chain. This has not been fully accepted; for instance, many developers still use pre-programmed models in readily available tools without a full understanding of their derivation.

The first important component in proving that an AI system is trustworthy is its reliability and verifiability. Many machine learning (ML) algorithms are based on training datasets and stationary environments that lack resilience to uncertain and adversarial events (Nelson, Biggio, & Laskov, 2011, October). A second important concept is explainable AI (referred to as XAI), the ability to explain the knowledge and decision-making process used to develop the AI application, including the algorithms used and the rationale for their use (PWC, 2019). Another major component is fair and unbiased AI. One challenge is that the mathematical concepts of bias used in AI algorithms do not always match to human understandings of bias (Lee, Resnick, & Barton, 2019). Humans review bias through the lens of fairness which is very subjective. For this reason, it is very difficult for a developer to write a predictive AI algorithm that will be fair to all parties, particularly when the developer may not be aware of their own biases or may not be aware of the extent of the population that will be affected by the AI algorithm. Furthermore, AI applications are often determined to be biased because the algorithms are trained with historical data. For example, in the development of Google Translate, it was determined that when translating from languages that had no gender context, nurses were always female, and doctors always male on translation to English. While developers typically try to exclude known sensitivities from the training set, the resulting algorithms may still reflect biases.

Much work still needs to be performed to develop models and techniques to validate AI applications as trustworthy, especially given the wide variety of use contexts. Recent government initiatives highlight that enhancements are needed in verification and validation for AI (National Science and Technology Council (NSTC), 2019, June). The plan calls for methods to measure and evaluate AI technologies via standards and benchmarks. It also notes the need for testbeds for AI. In terms of validation, they note that AI systems "may need to possess capabilities for self-assessment, self-diagnosis, and self-repair to be robust and reliable." In another example, the Defense Innovation Board was tasked to develop AI Principles for the Ethical Use of AI by the Department of Defense (Defense Innovation Board, 2020). This document underscores a critical link between trust in AI and ethics.

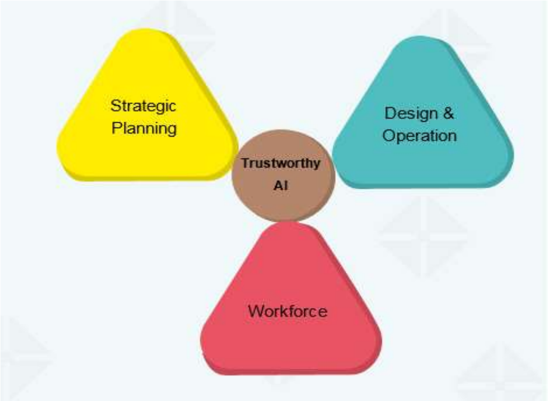
## RESPONSE TO AI AS AN ATTACK SURFACE

A large body of literature has focused on the social implications of AI misuse and policy responses (e.g., Brynjolfsson & McAfee, 2014; Chessen, 2017; Crawford & Calo, 2016). This manuscript fits into the stream of literature; however, it differentiates itself by paying more attention to ethical aspect of AI design and implementation rather than technical aspects. The motivation of this focus and inquiry is to ensure that AI applications achieve desired goals while conforming to relevant ethical and legal standards.

One recent survey of 250 senior business executives found that the level of understanding and application of responsible and ethical AI practices among respondents varied significantly across organizations, and was immature in most cases (PWC, 2019). Another global survey of more than 2,400 business leaders and managers revealed a persistent gap between respondents' practices to incorporate more AI-based technologies and the small amount of "right" data available to make correct decisions (SAS, 2019, January 08). Therefore, it is imperative to incorporate ethics and related principles of responsibility as central elements in the design and operation of AI systems for effective cyber defense.

This paper proposes a multi-aspect approach to respond to the emerging challenges associated with AI. We will discuss the approach from the strategic planning aspect, the system design and operation aspect, and the AI workforce readiness aspect. Figure 1 illustrates the multi-faceted approach.

**FIGURE 1**
**MULTI-FACETED APPROACH TO TRUSTWORTHY AI**



### Trustworthy AI by Strategic Planning

A number of government entities and private sector companies have recognized the importance of governance, policy, and ethics in the development of safe and societally beneficial AI. We examine those principles and frameworks, focusing on the ethical components related to trustworthy AI that are applicable to the field of cybersecurity. The metrics used to define the "trustworthiness" of AI are identified and summarized in Table 1.

**TABLE 1**
**SUMMARY OF MAJOR STRATEGIC INITIATIVES ON TRUSTWORTHY-AI**

| Initiative | Components of Trustworthy-AI | Implications in Strategic Planning |
|---|---|---|
| OECD Principles on AI (Organisation for Economic Co-operation and Development (OECD), 2019)<br>G20 Human Centered AI Principles (G20, 2019) | Transparency<br>Explainability/ Traceability | • AI systems should be robust, secure and safe throughout entire lifecycle.<br>• AI systems should ensure traceability in relation to datasets, processes and decisions.<br>• For adverse conditions, AI systems should function appropriately and not pose an unreasonable safety risk. |

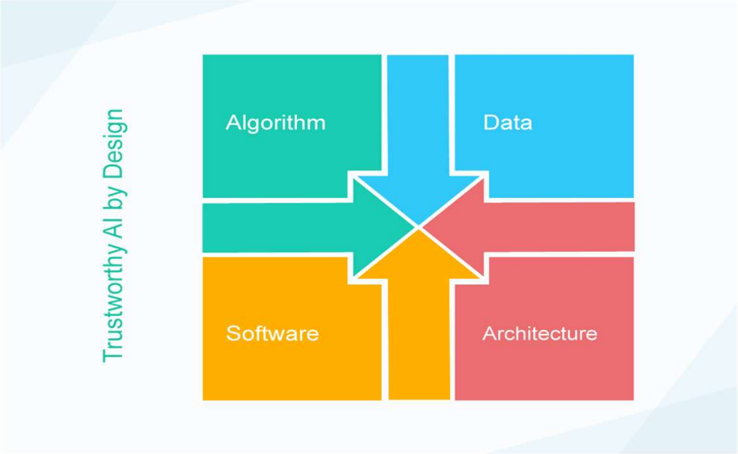
| | | |
|---|---|---|
| European Commission Ethics Guidelines for Trustworthy AI (European Commission, 2019, April 8) | Respect for human autonomy<br>Prevention of harm<br>Fairness<br>Explicability | • The allocation of cybersecurity functions should follow human-centric design principles.<br>• Decision-making processes should be explicable.<br>• Ensure traceability, auditability and transparent communication on system capabilities. |
| AI Principles for (Department of Defense, 2018) | Responsible<br>Equitable<br>Traceable<br>Reliable<br>Governable | • Avoid unintended bias during the development and deployment of AI<br>• The safety, security, and robustness of AI systems should be tested and assured.<br>• AI systems should detect and avoid unintended harm or disruption. |
| AI R&D Strategic Plan by NSTC (National Science and Technology Council (NSTC), 2019, June)<br>and<br>Cybersecurity R&D Strategic Plan by NITRD (The Networking and Information Technology Research and Development (NITRD), 2019) | Fairness<br>Transparency<br>Accountability<br>Explainability<br>Auditability<br>Creditable<br>Reliable<br>Recoverable | • AI system designs and decision-making is transparent and examined for any bias.<br>• Design architectures for AI systems should incorporate ethical reasoning.<br>• Recognize the vulnerability surface of AI/ML and take it into account in AI/ML implementations. |
| Responsible AI by PWC (PWC, 2019) | Governance<br>Interpretability<br>Explainability<br>Fairness | • AI governance starts with strategy and planning.<br>• Considers existing capabilities and compliance.<br>• Stakeholder must be enabled to look at underlying models and the data used to train them. |

When comparing the strategic initiatives proposed by different entities, it is noteworthy that despite the different sources and contexts, they all share similar components in terms of "trustworthiness". Organizations can safely follow any of the above guidelines when creating their strategic plans to incorporate trustworthy AI as part of their cybersecurity solutions.

## Trustworthy AI by Design

Similar to the concept of "security by design", we argue that a robust, "trustworthy" AI system starts with its design and development; trustworthiness should not be a posterior consideration. Four major factors in design may compromise "trust" in an AI system (Figure 2): data (McKinsey Global Institute, 2019; Stanely, 2017, June 2), algorithm (Danks & London, 2017; Kirkpatrick, 2016; Mann & O'Neil, 2016), architecture, and software (National Science and Technology Council (NSTC), 2019, June).

FIGURE 2
TRUSTWORTHY AI BY DESIGN



### Preparing the Cybersecurity Workforce

As AI becomes more ubiquitous, everyone must become more informed about it, especially the concept of its trustworthiness. Education across many fields will be affected by its use, including those involved in the development of AI-based apps and those cybersecurity professionals protecting our systems from misuse.

At the broadest level, individuals who are impacted, directly or indirectly, by use-inspired AI applications (e.g., when applying for a loan or getting a medical diagnosis), must be informed about the potential for bias in the decision. They must be prepared to ask for a description of the "rules" being used (XAI) and to be able to understand the explanations given.

Of more importance, however, is the enhanced education needed for developers and data scientists, not just in how to use the many AI/ML tools and techniques, but the ethical consequences of their actions. They must be educated about AI ethics and trustworthy AI, and the processes needed to ensure the development of quality AI apps—understanding the limitations of the data used to train their models and the basis of the model algorithms being used. Most significant will be the extent of testing, not just whether AI works, but whether its results are trustworthy, including factors such as fairness, transparency, explainability, and reliability. Testing must occur not only upon initial implementation, but must continue throughout the lifecycle of the application, as use may change the AI's trustworthiness. Developers may need to take advantage of AI-based techniques, such as automated static analysis, to improve their efficiency, given the shorter timelines dictated by the wide adoption of Agile methodologies and DevOps (Wang & Liu, 2018).

Finally, education is necessary for the cybersecurity workforce, who must extend the "zero-trust" model to AI applications developed inside and outside the network. This includes knowledge of new attack vectors, such as data poisoning or adversarial AI. Cybersecurity professionals must also be aware of the ways attackers are using AI to improve their attack profile, and how this might violate the standard cybersecurity methods of network monitoring and malware detection. There are also positives for cybersecurity, in that new AI-based tools are becoming available that will improve the efficiency of anomaly detection and enable a larger percentage of potential attacks to be detected automatically.

## CONCLUSION

There has been extensive progress in AI applications, which are making valuable contributions to many fields. However, more attention should be paid to the ways in which AI is increasingly becoming an ethical issue, whether intentionally or not. In this work, we examined several emerging AI ethics frameworks and proposed an integrative framework for trustworthy AI. The goal of this framework is to prevent and mitigate potential harms associated with the many beneficial applications of AI. We recognize the urgent need for education to ensure AI continues to be of benefit to society, and the need to educate users, developers, and the cybersecurity workforce protecting the nation of its features.

It is hard to dispute the importance of trustworthy AI. However, it is challenging for organizations to understand the implications of poor applications of AI and their impact on business and society. The executives and top-level management are urged to contextualize these principles into specific guidelines for the front-line workforce and to begin the process of educating their workforce and helping their users understand how decisions affecting them are made.

## REFERENCES

Accenture Federal Services. (2018*). Responsible AI: A Framework for Building Trust in Your AI Solutions*. Retrieved from https://www.accenture.com/acnmedia/PDF-92/Accenture-AFS-Responsible-AI.pdf

Accenture Labs. (2019). *AI is the New Attack Surface*. Retrieved from https://www.accenture.com/_acnmedia/accenture/redesign-assets/dotcom/documents/global/1/accenture-trustworthy-ai-pov-updated.pdf

Biggio, B., & Roli, F. (2018). Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning. *Pattern Recognition*, *84*, 317-331.

Brundage, M., Avin, S., Clarck, J., Toner, H., Eckersley, P., Garfinkel, B., . . . Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Machines*. New York: W.W. Norton & Company, Inc.

Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., & Song, D. (2018). *The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets*. ArXiv e-prints, 1802.08232. Retrieved from https://research.google/pubs/pub46702/

Chessen, M. (2017). *The AI Policy Landscape*. Medium. Retrieved from https://medium.com/artificial-intelligence-policy-laws-and-ethics/the-ai-landscape-ea8a8b3c3d5d

Corona, I., Biggio, B., & Maiorca, D. (2016). *AdversariaLib: An Open-source Library for the Security Evaluation of Machine Learning Algorithms Under Attack*. arXiv:1611.04786. Retrieved from https://arxiv.org/abs/1611.04786

Crawford, K., & Calo, R. (2016). *There is a Blind Spot in AI Research*. Nature. Retrieved from https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805

Danks, D., & London, A.J. (2017). *Algorithmic Bias in Autonomous Systems*. Paper presented at the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia.

Defense Innovation Board. (2020). *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*. Retrieved from https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Department of Defense. (2018). *The 2018 Department of Defense Artifical Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*. Retrieved from https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

European Commission. (2019, April 8). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., . . . Song, D. (2018). *Robust Physical-World Attacks on Deep Learning Models*. arXiv:1707.08945. Retrieved from https://arxiv.org/abs/1707.08945

Forelle, M., Howard, P., Monroy-Hernández, A., & Savage, S. (2015). *Political Bots and the Manipulation of Public Opinion in Venezuela*. arXiv:1507.07109. Retrieved from https://arxiv.org/abs/1507.07109

G20. (2019). *G20 AI Principles*. Retrieved from https://www.mofa.go.jp/files/000486596.pdf

Gil, Y., & Selman, B. (2019). *A 20-year Community Roadmap for Artificial Intelligence Research in the US.* Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI). Retrieved from https://cra.org/ccc/wp-content/uploads/sites/2/2019/08/Community-Roadmap-for-AI-Research.pdf

Grove, W.M., & Meehl, P.E. (1996). Comparative Efficiency of Informal (subjective, impressionistic) and Formal (mechanical, algorithmic) Prediction Procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323.

Halpern, J.Y., & Moses, Y. (1992). A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief. *Artificial Intelligence*, *54*(3), 319-379.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., . . . Leibo, J.Z. (2017). *Deep Q-learning from Demonstrations*. arXiv preprint server. Retrieved from https://arxiv.org/abs/1704.03732

Huang, J., & Nicol, D. (2010). A Formal-Semantics-Based Calculus of Trust. *IEEE Internet Computing*, *14*(5), 38-46.

IEEE. (2019). *Ethically Aligned Design, First Edition: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Jaderberg, M., Mnih, V., Czarnecki, W.M., Schaul, T., Leibo, J.Z., Silver, D., & Kavukcuoglu, K. (2016). *Reinforcement Learning with Unsupervised Auxiliary Tasks.* rXiv preprint server. Retrieved from https://arxiv.org/abs/1611.05397

Jago, A.S. (2019). Algorithms and Authenticity. *Academy of Management Discoveries*, *5*(1), 38-56.

Jordan, M.I., & Mitchell, T.M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, *349*(6245), 255-260.

Kamps, H.J. (2016). *Logojoy Turns Design into An AI-powered, Iterative Process*. Techcrunch. Retrieved from https://techcrunch.com/2016/12/01/logojoy-makes-designersunemployed/

Kirkpatrick, K. (2016). Battling Algorithmic Bias. *Communications of the ACM, 59*(10), 16-17. Retrieved from https://cacm.acm.org/magazines/2016/10/207759-battlingalgorithmic-bias/abstract

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30-37.

Lee, N.T., Resnick, P., & Barton, G. (2019). *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Eonsumer Harms*. Retrieved from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Liu, Y.X. (2017). *The Accountability of AI — Case Study: Microsoft's Tay Experiment*. Retrieved from https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f

Lotman, E., & Viigimaa, M. (2020). Digital Health in Cardiology: The Estonian Perspective. *Cardiology*, *145*, 21-26.

Mann, G., & O'Neil, G. (2016). Hiring Algorithms are Not Neutral. *Harvard Business Review*. Retrieved from https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

McFarland, M. (2017). Farmers Spot Diseased Crops Faster With Artificial Intelligence. *CNNBusiness*. Retrieved from https://money.cnn.com/2017/12/14/technology/corn-soybeanai-farming/index.html

McKinsey Global Institute. (2019). *Tackling Bias in Artificial Intelligence (and in humans)*. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

Melis, M., Demontis, A., Pintor, A., Sotgiu, A., & Biggio, B. (2019). *secml: A Python Library for Secure and Explainable Machine Learning*. arXiv:1912.10013. Retrieved from https://arxiv.org/abs/1912.10013

National Institute of Standards and Technology (NIST). (2019, November 18). *Plan Outlines Priorities for Federal Agency Engagement in AI Standards Development*. Retrieved from https://www.nist.gov/news-events/news/2019/08/plan-outlines-priorities-federal-agency-engagement-ai-standards-development

National Science and Technology Council (NSTC). (2019, June). *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*. Retrieved from https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

Neal, A., Kouwenhoven, A., & Sa, O. (2015). *Quantifying Online Advertising Fraud: Ad-Click Bots vs Humans*. Oxford Bio Chronometrics. Retrieved from http://oxford-biochron.com/downloads/OxfordBioChron_Quantifying-Online-Advertising-Fraud_Report.pdf

Nelson, B., Biggio, B., & Laskov, P. (2011, October). *Understanding the Risk Factors of Learning in Adversarial Environment.* Paper presented at the AISec '11: Proceedings of the 4th ACM workshop on Security and artificial intelligence, Chicago, Illinois.

Organisation for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. Retrieved from https://www.oecd.org/going-digital/ai/principles/

PWC. (2019). *A Practical Guide to Responsible Artificial Intelligence (AI)*. Retrieved from https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture (2nd Draft), SP 800-207(Draft)*. NIST. Retrieved from https://csrc.nist.gov/publications/detail/sp/800-207/draft

Saif, I., & Ammanath, B. (2020). 'Trustworthy AI' is a Framework to Help Manage Unique Risk. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/

SAS. (2019, January 8). Data, Analytics, & AI: How Trust Delivers Value. *MIT Sloan Management Review*. Retrieved from https://sloanreview.mit.edu/sponsors-content/data-analytics-and-ai-how-trust-delivers-value/

Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., & Menczer, F. (2017). *The Spread of Fake news by Social Bots*. arXiv:1707.07592, 96-104. Retrieved from https://arxiv.org/abs/1707.07592

Stanely, J. (2017, June 2). *Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case*. Retrieved from https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case

The Institute for Ethical AI and Machine Learning. (2020). *The Responsible Machine Learning Principles*. Retrieved from https://ethical.institute/principles.html

The Networking and Information Technology Research and Development (NITRD). (2019). *Federal Cybersecurity Research and Development Strategic Plan*. Retrieved from https://www.nitrd.gov/pubs/Federal-Cybersecurity-RD-Strategic-Plan-2019.pdf

Tung, L. (2019, September 4). *Forget email: Scammers use CEO voice 'deepfakes' to con workers into wiring cash*. ZDNet. Retrieved from https://www.zdnet.com/article/forget-email-scammers-use-ceo-voice-deepfakes-to-con-workers-into-wiring-cash/

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., . . . Fuso Nerini, F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 233. Retrieved from https://doi.org/10.1038/s41467-019-14108-y

Wang, C., & Liu, C. (2018). *Adopting DevOps in Agile: Challenges and Solutions.* (Master of Science in Software Engineering). Blekinge Institute of Technology, Karlskrona, Sweden. Retrieved from https://www.diva-portal.org/smash/get/diva2:1228684/FULLTEXT02.pdf

White House. (2019, February 11). *Executive Order on Maintaining American Leadership in Artificial Intelligence*. Retrieved from https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

Yeomans, M. (2015, July 07). What Every Manager Should Know about Machine Learning. *Havard Business Review*. Retrieved from https://hbr.org/2015/07/what-every-manager-should-know-about-machine-learning

Yu, K.H., Beam, A.L., & Kohane, I.S. (2018). Artificial Intelligence in Healthcare. *Nature Biomedical Engineering*, *2*(10), 719-731. doi:10.1038/s41551-018-0305-z

Zeadally, S., Adi, E., Baig, Z., & Khan, I. (2020). Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity. *IEEE Access*, *8*, 23817-23837.