American Journal of Engineering Research (AJER) ***2023***
American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-12, Issue-10, pp: 07-14
www.ajer.org
Research Paper Open Access

Red Teaming: A Framework for Developing Ethical AI Systems

Isi Idemudia (Ph.D), Accenture

Abstract—*Artificial intelligence (AI) systems have become co-pilots in today's world, where humans and machines work collaboratively to make decisions that affect day-to-day. AI has transformed modern life through previously unthinkable feats, ranging from self-driving cars and machines that can master the ancient board game Go to more "every day" developments, such as customer support chatbots and personalized product recommendations [1]. The aim of this study is to determine a new model for implementing red teaming in AI deployments, the goal of which is to minimize undesirable and unfair outcomes. A practical action research methodology was used as the study aims to solve a specific problem of reducing bias in AI systems. This study discussed the group fairness framework and the red teaming methodology in cybersecurity which led to the formation of the red team. Likewise, the roles of several agents of the red and blue teams are investigated to design a responsibility matrix to act as an implementation guide for red teaming in AI. As AI systems become more widespread, organizations will be seeking more implementation guides rather than principles to be successful. The fair AI red team framework is aimed at being one of these guides to assist organizations in implementing AI systems that have fairer outcomes.*

Impact Statement—The topic of responsible AI practices cannot be overemphasized. The more decisions impacting humans are being controlled by Artificial intelligence systems, the more the topic of how these systems can act responsibly becomes critical. However, recent research by Gartner suggested that by 2022, 85% of AI projects could deliver erroneous outcomes due to various forms of bias [2]. While a lot of work has been done on developing principles for implementing fair AI systems, most organizations simply do not know how to implement and monitor the deployment of fairer AI systems. The fair AI framework proposed in this research provides a practical implementation guide, a step-by-step approach on how to ensure organizations can deploy fairer AI systems with less negative outcomes using red teaming.

Index Terms—*AI development, AI red teams, AI bias, Bias, Fair AI, Responsible AI, Red teaming.*

Date of Submission: 02-10-2023 Date of acceptance: 15-10-2023

I. INTRODUCTION

The topic of the ethics and governance of artificial intelligence (AI) is a timely one [2]. Recently, concerns have been expressed about the use of AI regarding potential discrimination, explainability, data transparency, AI's impact on jobs and the malicious use of AI technology[1]. AI is argued to be able to improve healthcare, education, and transportation as well as to automate tasks. In addition, it can help us to understand the world, make better decisions, and connect with others [4].

Connecting with others around us has been the fundamental phenomenon of humans. This is similar to how structural-functional theory seeks to explain and analyze society by examining its various parts and their functions [3]. Structural-functional theory argues that society is a complex system of interrelated parts that work together to meet its basic needs [3]. These parts, or structures, include the economy, the family, the education system, and the government, all of which can be interconnected with AI to contribute to society's overall stability and well-being. It is safe to say that AI is no longer only in the laboratory and has now entered people's daily lives; thus, it is having a "real-world impact on our culture, our people, and our institutions" [5].

According to [6] AI systems have the ability to harm or create value for the companies that develop them, the people who use them, and the members of the public who are affected by their use. To ensure a high expected value for users and society, AI systems must be safe—that is, they must reliably work as intended—and secure—that is, they must have limited potential for misuse or subversion. The more severe the harm that could result from safety failures, misuse, or structural risks, the more critical it is that AI systems are safe and beneficial in a wide range of possible circumstances [7]. The safety of systems can be traced back to the early days of computing, when computers were first used to store and process sensitive information. As computers have become more powerful and connected to the Internet, the need for cybersecurity has grown [8].

The concept of 'red teaming', borrowed from the world of cybersecurity, relates to the use of 'white hat' hackers to test enterprise defenses [2]. Red teaming can be used to test the effectiveness of security controls, identify new threats, and improve the overall security posture [9]. As AI systems were becoming more common, Microsoft created the AI Red Team in 2018, a group of interdisciplinary experts dedicated to thinking like attackers and searching for flaws in AI systems [1]. Although reasonable progress is being made in response to safety concerns related to AI, the present study aimed to examine red teaming as a strategy for mitigating bias or the risk of unintended consequences. A challenge in creating a methodology for ethical AI is the trade-off between innovation speed and risks. This is because the speed of technological changes and their uptake in the market dictate a fast incorporation of AI in products and services [10]. Despite such challenges, red teaming can be a valuable tool for improving the security of AI systems. Through identifying vulnerabilities and developing mitigation strategies, red teaming can help make AI systems more fair [11].

A. Background

Gartner, a research and advisory company, suggested that by 2022, 85% of AI projects could deliver erroneous outcomes due to various forms of bias [12]. Moreover, research by Accenture suggested that while 63% of leaders believe that it is crucial to monitor AI systems, most are unsure of how to do so [2]. Accenture's 2022 Tech Vision research found that only 35% of global consumers trust how AI is being implemented by organizations, while 77% think that organizations must be held accountable for their misuse of AI. Several studies have been published about the ethical principles for producing AI systems. However, few publications exist on how companies should go about the creation of AI systems while respecting relevant ethical and societal implications. The lack of published methodologies on how to implement fair AI systems and the highly disturbing statistics on AI system outcomes compelled the present researcher to conduct this study.

B. Scope of the study

Microsoft has devised a blueprint for the public governance of AI, which describes a five-point approach to help AI governance advance more quickly. The five points are as follows:

1) Implement and build upon new government-led AI safety frameworks.
2) Require effective safety brakes for AI systems that control critical infrastructure.
3) Develop a broader legal and regulatory framework based on the technology architecture for AI.
4) Promote transparency and ensure academic and public access to AI.
5) Pursue new public–private partnerships to use AI as an effective tool for addressing the inevitable societal challenges that come with new technology.

C. Limitations of the study

The scope of this study was limited to the fourth and fifth pillars of Microsoft's AI blueprint. This study aimed to define strategies for promoting transparency and inclusive access to AI systems through the use of red teaming. Other types of fairness, such as individual and counterfactual fairness, were excluded from this study.

D. Aim and Objectives

The aim of this study was to define a model that can enable organizations to build AI systems using red teaming to mitigate the unintended outcomes of AI due to various forms of bias as well as deliver more ethical AI solutions. Thus, this study sought to answer the following research questions:

1) Are biases in AI systems purely technological or socio-technological in nature?
2) How can an organization implement red teaming in AI to reduce bias?
3) Who should comprise an AI red team?

This paper is structured in the following manner: Section I presents the problem that this study aimed to address, the scope of the study and the study's aims and objectives; Section II presents a review of the relevant literature; Section III presents the methodology employed to conduct the study; Section IV presents the newly

developed Fair AI red team framework; and lastly, Section V presents the conclusion along with recommendations for future research.

II. LITERATURE REVIEW

The field of science and technology studies (STS) describes systems that consist of a combination of technical and social components as 'sociotechnical systems'. Both humans and machines are necessary to make any technology work as intended [13]. Likely due to expectations based on techno-solutionism and a lack of mature AI process governance, organizations often default to overly technical solutions for AI bias issues. Yet, these mathematical and computational approaches do not adequately capture the societal impact of AI systems [14]. Using a sociotechnical approach to AI bias makes it possible to evaluate dynamic systems of bias, understand how they impact each other, and determine under what conditions these biases are attenuated or amplified. Adopting a sociotechnical perspective can enable a broader understanding of AI's impacts and the key decisions that occur throughout and beyond.

This study aimed to build on the research of Green [14] and Selbst et al. [13] by incorporating human practices through a sociotechnical lens to obtain an enhanced understanding of how AI systems are functions of society. It also extended the work of Eitel- Porter [2] on responsible AI governance procedures by recommending the use of 'red teams' and 'fire wardens' as 'white hackers' to review algorithms and outcomes for signs of bias or risk of unintended consequences.

The review of literature will consider what has been done by regulators regarding AI principles and patterns, then go on to exploring what the traditional red team framework in cybersecurity entails and finally expands on the group fairness framework to understand how fairness can be grouped.

A. Regulators

Efforts have been made to create frameworks and principles for guiding the deployment of AI systems by regulators across the globe. The European Commission [15] published Ethics Guidelines for Trustworthy, which proposed an assessment checklist for AI practitioners based on the following seven principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency, diversity, nondiscrimination, and fairness; societal and environmental wellbeing; and accountability. The European Union (EU) has adopted a human-centric approach towards ensuring a trustworthy AI system that is held accountable to the same fundamental human rights that guide all Europeans. The EU's strong emphasis on fundamental human rights has resulted in guidelines that require organizations to conduct a fundamental rights impact assessment before developing new policies or programs. To allow for external feedback on any potential infringements of fundamental rights, mechanisms should be established after the assessment is completed [15].

The strategy in the United States, on the other hand, is developed mainly through private-sector initiatives and self-regulation [15]. The National Institute of Standards and Technology (NIST) launched a critical new AI safety initiative called the NIST AI Risk Management Framework. It is a voluntary framework that organizations can use to assess and manage the risks associated with AI [16]. The framework consists of the following five steps:

1) *Identify:* Determine the AI systems and capabilities that are to be evaluated.
2. *Assess:* Evaluate the potential dangers posed by AI systems and their capabilities.
3) *Respond:* Put in place and execute risk mitigation plans.
4) *Monitor:* Keep an eye on AI systems and their capabilities to ensure that the risk mitigation strategies are working.
5) *Review:* Conduct regular assessments of AI systems and capabilities to ensure that risks are still being effectively mitigated [16].

The NIST AI Risk Management Framework is a valuable resource for organizations that are developing or deploying AI systems. The framework can help organizations to identify and manage the risks associated with AI, as well as to ensure that AI systems are used in a safe and responsible manner [16].

By contrast, the Chinese strategy is characterized as essentially government-led, with the strong coordination of private and public investment in AI technologies [15]. The Chinese government has only recently elevated AI to the status of a national 'megaproject', in the tradition of Chinese techno-nationalism [17].

B. Traditional Red Team Methodology

Red teaming is the process of identifying and assessing assumptions, alternative options, vulnerabilities, limitations, and risks for an organization. It is a tool that can be used to provide decision-makers with a more robust baseline for decision-making [18]. Fig. 1 presents the phases of red team according to the North Atlantic Treaty Organization (NATO).

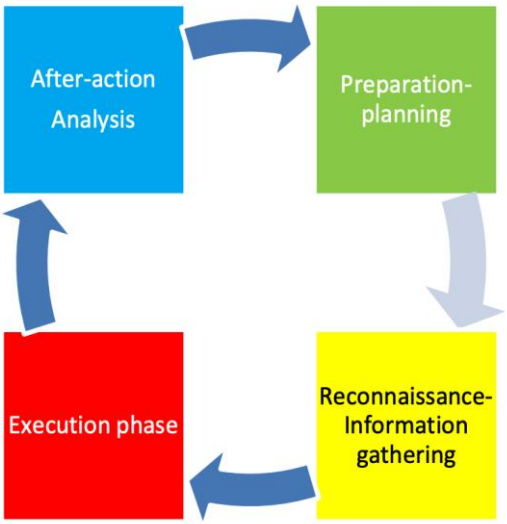

Fig. 1. The cyber red team cycle

According to [18] the phases of red teaming are as follows:
1) *Planning and preparation:* The purpose of cyber red teaming is determined during this phase. Before undertaking any activities, it is necessary to assess the current needs of a specific organization and the scope of the actions to be taken.
2) *Information gathering / Reconnaissance:* This phase entails preliminary surveying or research of the intended information system, which can include web research, social engineering, and common techniques, as well as more complex operations such as specific intelligence reports.
3) *Execution:* This is the hands-on phase of the cyber red teaming process, in which tools and expertise are used to discover vulnerabilities.
4) *After-action analysis:* During this phase, all of the actions taken are documented, the results listed, and recommendations and proposals provided. Follow-up actions can also be envisaged in which cyber red teams can be involved.

C. Group Fairness Framework

Hardt et al. [19] proposed a framework for group fairness that uses the following three criteria to evaluate the fairness of an AI model:
1) *Independence:* This means that the model's prediction is independent of the sensitive variable.
2) *Separation (also known as equalized odds):* This means that the model's prediction is independent of the sensitive variable given the target variable, meaning that the true positive and false positive rates are equal for all sensitive groups.
3) *Sufficiency (also known as predictive rate parity):* This means that the target variable is independent of the sensitive attribute given the model output, meaning that the positive predictive value is the same for all sensitive groups.

D. Section Summary

The human-centric and feedback loop approach of the [15]. is an essential component of the red teaming model and is in line with Selbst et al. [13]. The monitoring and review phases of the NIST AI Risk Management Framework and the private–public sector partnerships involved in the Chinese strategy were also fed into the formation of this research. Additionally, the red teaming methodology in cybersecurity was applied to AI to, in this case, expose vulnerabilities and unintended outcomes regarding bias through a human-centric lens.

III. METHODOLOGY

The practical action research methodology was used in this research because the focus was on addressing and solving the specific problem of reducing bias in AI through red teaming. The interdisciplinary methodology and model defined in this study is one that overlays the group fairness framework by [19] and the traditional red team methodology to develop a unique model for red teaming for ethics in AI.

IV. THE FAIR AI RED TEAM FRAMEWORK

This section describes the four components of the Fair AI Red Team Framework which are;
1) The makeup of a red team and a blue team
2) The red team responsibility matrix
3) The red team methodology and
4) The fair AI models.

A. The Red Team

The red team should comprise agents from the following fields:
1) *The public*: The Cambridge dictionary defines 'the public' as all ordinary people. Ordinary people consist of the general public, irrespective of race, gender, traditions, religion, sexuality, or education.
2) *Academia:* Academia is the world of higher education, including universities, colleges, and other institutions that offer advanced degrees. It is also a community of scholars who engage in research and teaching. Academia is often seen as a bastion of free thought and inquiry, and it plays a crucial role in the advancement of knowledge. This community of scholars includes lecturers, researchers, and students alike.
3) *Regulatory institutions:* These are organizations that are responsible for setting and enforcing rules and regulations in a particular industry or sector. They are often government agencies, but they can also be nongovernmental organizations or private companies. This set of agents also includes industry experts. For example, if one is designing a chatbot to help healthcare providers, medical experts could assist in identifying risks in that domain [20].
4) *Public–private partnerships (PPPs):* PPPs are collaborations between public and private sectors for delivering public services or infrastructure.

B. The Blue Team

The blue team comprises everyone on the red team plus the product team. Its objectives are geared towards mitigation. The blue team generates reports from the outcomes of the red team and uses the outcomes to enhance the system. The blue team is also involved in constantly monitoring the AI system for biases that might not have been caught before the go-live process.
The product team comprises the members of the team who built the AI system. These include the developers, data scientists, ML engineers, representatives of the AI governing board, and industry experts.
Blue Team = Red Team + Product Team

C. Responsibility Matrix for Red Team Agents

Based on the framework for group fairness developed by [19] and the red team agents identified in Section 7.1, this study developed a responsibility matrix for the agents who are a part of the red team. Table 1 presents the responsibility matrix for red team agents:

TABLE I
Responsibility Matrix for Red and Blue Team Agents

Agents	Persona	Responsibility
The public	• All genders • All races • All sexualities • All religions • All persons with disabilities	Check and monitor for independence, sufficiency, and separation of AI systems
Academia	• University students • University lecturers • University researchers	Check and monitor for independence, sufficiency, and separation of AI systems
Regulatory institutions	• Federal regulators • State regulators • Local/council regulators	Check and monitor for independence, sufficiency, and separation of AI systems

Public–private partnerships	Collaboration between a government agency and a private company for delivering a public service or project	Check and monitor for independence, sufficiency, and separation of AI systems
Non-profits	A legal entity that is organized for a social cause or public benefit and does not distribute its profits to its members	Check and monitor for independence, sufficiency, and separation of AI systems
Product team	Developers Data scientists Representatives of the AI governing board	Monitor and report for independence, sufficiency, and separation of AI systems

D.The Fair Methodology

Fig. 2 illustrates the step-by-step process for implementing red teaming for AI ethics, which is referred to as the FAIR methodology. This methodology describes six stages, starting with reconnaissance and ending with monitoring, which the red team should engage in to successfully identify biases.

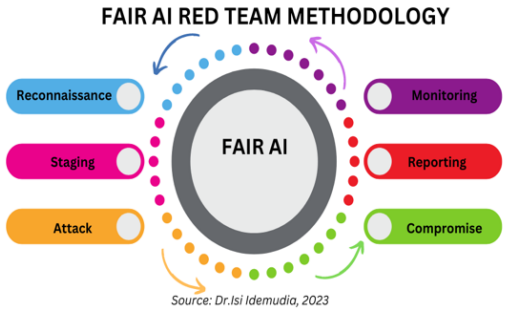

Fig. 2 Step-by-step process for implementing red teaming for AI ethics

This will assist in mitigating some of the biases in the application layer of the AI system. The FAIR methodology incorporates similar phases of traditional red teaming methodologies but with a different objective regarding fairness.

The six stages of the process are elaborated on as follows:

1) *Reconnaissance (recon):* This stage is formed with intelligence gathering from all of the agents. As a prerequisite for this phase, the product team provides details about the AI system, the problem that it is attempting to solve, and the expected outcomes. Red teamers first survey the attack floor and understand the possible gaps.

2) *Staging:* Once recon has been conducted and vulnerabilities have been found, the next stage is staging the attack. Here, the red teamers set up various test cases, which should be diverse, to catch diverse failure modes and maximize test coverage. This can be a manual or an automated process. An example would be a test case for checking whether an AI chatbot provides responses that leak memorized training data, which would be harmful if the data are copyrighted or private, or for discovering groups of people that the chatbot discusses in more offensive compared with other groups [11].

3) *Attack:* Typically, after possible vulnerabilities have been identified using the test cases, the red teamers find a foothold in the AI system by unleashing various weapons on the sensitive variables. Attacks could include impersonation and potential plagiarism, offensive responses, data leakage, generated contact information, and distributed biases revealing discrimination against groups (e.g., a particular gender, race, or religion) [11].

4) *Compromise:* In this stage, the red teamers move towards compromising the system. Essentially, once the attack has been delivered, the red teamers focus on exploiting the system to determine how harmful it could become. The activities at this stage could include data exfiltration; if the chatbot is seen to leak memorized training data, then it could continue to be fed prompts that seek memorized text; such responses are particularly useful to adversaries who aim to extract training data or perform membership inference. Alternatively, if the system is seen to provide offensive responses, then those types of responses could be continually solicited through input prompts. This would meet the objective that offensive replies beget offensive replies [11].

5) *Reporting and analysis:* At the end of the engagement, the red team provides a report of the findings of the

exercise. This lets the whole team know of the shortfalls and bias gaps. The analysis of the findings is now conducted along with the product owners in the form of a workshop or war room.

6) *Continuous monitoring:* The job of monitoring the system continues even after the initial report is submitted and the system has gone live. This is in line with the NIST AI Risk Management Framework, which recommends reviewing and monitoring as part of its risk framework.

E. The Fair Model

Fig. 3 presents an illustration of the Fair AI Red Team Model.

Fig. 3. The Fair AI Red Team Model

The Fair AI Red Team Model, also known as the FAIR model, illustrates the full scope of the red and blue teams' involvement throughout the red teaming lifecycle, from recon to post-deployment monitoring of the AI system. The blue team, which is a combination of the red team and the product team, is responsible for reporting and analyzing the findings of any bias discovered during the attack and that compromised stages of the model. This team is also responsible for continuously monitoring the AI system for biases while it is in use and reporting them for further mitigation procedures.

The FAIR model ensures that all the agents who are members of society and who will be affected by the presence of the AI system in society are part of the vetting process. Not only does this increase the transparency of AI systems but also creates trust among the public.

V.CONCLUSIONS AND FUTURE WORK

This paper has discussed a methodology called the Fair AI Red Team (FAIR) framework, which can be used by organizations that plan to use AI on a large scale and wish to avoid unintentionally creating undesirable side effects, such as unfair discrimination in an AI system. The methodology began by establishing that AI is now a part of society and, as such, must be examined in a socioeconomic context, not just as a technology tool. This study continued by exploring the red team methodologies used in cybersecurity to understand how red teaming is being used in identifying vulnerabilities and developing mitigation strategies for technology systems. Then, it proceeded to define a set of agents who should be a part of the red and blue teams regarding fair AI. Other ingredients that organizations should keep in mind are that AI red teams should be composed of people with a variety of social and professional backgrounds, demographic groups, and interdisciplinary expertise appropriate for the deployment context of their AI system. This will help to ensure that the teams are able to identify and address a wide range of potential risks and vulnerabilities. Through diversity, red teams can obtain an enhanced understanding of the potential bias that an AI system may face and develop effective mitigation strategies. It is the combination of those different elements that makes the methodology successful.

Like many other researchers, I have only just begun my journey; therefore, I do not consider my framework and methodology to be the final standards for red teaming in ethical AI. Rather, they are a starting point that will evolve as more collective experience becomes available. Possible future research directions on this issue are broad, and I do not aim to provide a comprehensive list of them here. Instead, the following are examples of potentially fruitful research questions that could be examined in future studies:

1) How might the individual and counterfactual fairness groupings of AI alter the responsibility matrix for red team agents?

2) What is the proper organizational structure in institutions, governments, and regulatory bodies that would be

able to support a red team's formation? What lessons can be drawn from history or contemporary industries?
3) What further strategies can be discovered or constructed to help monitor, analyze, and report red teams' findings of bias?
As AI is expected to be adopted on a large scale, such AI fair models will become increasingly critical, and practical experience of the usage of the models must be shared to ensure the sustainable and ethical use of AI.

REFERENCES

[1]. R. S. S. Kumar, "Microsoft AI Red Team building future of safer AI," Microsoft Security Blog, 2023. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/
[2]. R. Eitel-Porter, "Beyond the promise: implementing ethical AI," AI Ethics, vol. 1, pp. 73–80, 2021, [Online]. Available: https://doi.org/10.1007/s43681-020-00011-6
[3]. J. W. Lucas, Structural Functional Theory. John Wiley & Sons, Ltd, 2007. [Online]. Available: https://doi.org/10.1002/9781405165518.wbeoss289
[4]. UNESCO, "New UNESCO report on Artificial Intelligence and Gender Equality," 2020. [Online]. Available: https://en.unesco.org/AI-and-GE-2020
[5]. M. L. Littman et al., "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report," Comput. Sci. > Artif. Intell., no. 2022, pp. 1–82, 2022, [Online]. Available: https://doi.org/10.48550/arXiv.2210.15767
[6]. D. Schif, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon, "Principles to Practices for Responsible AI: Closing the Gap," Comput. Sci. > Comput. Soc., vol. 2020, p. 04707, 2020, [Online]. Available: https://doi.org/10.48550/arXiv.2006.04707
[7]. R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," 2022. [Online]. Available: https://doi.org/10.6028/nist.sp.1270
[8]. M. Veale and I. Brown, "Cybersecurity," Internet Policy Rev., vol. 9, no. 4, pp. 1–22, 2020, [Online]. Available: https://doi.org/10.14763/2020.4.1533
[9]. A. Applebaum, D. Miller, B. Strom, C. Korban, and R. Wolf, "Intelligent, automated red team emulation," in ACM International Conference Proceeding Series, Dec. 2016, pp. 363–373. Accessed: Sep. 23, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/2991079.2991111
[10]. R. Benjamins, A. Barbado, and D. Sierra, "Responsible AI by Design in Practice," Comput. Sci. > Comput. Soc., no. 2019, p. 12838, 2019.
[11]. E. Perez et al., "Red Teaming Language Models with Language Models," Comput. Sci. > Comput. Lang., no. 2022, p. 03286, 2022, [Online]. Available: https://doi.org/10.48550/arXiv.2202.03286
[12]. Gartner, "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence," Gartner Press Releases. https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence (accessed Aug. 16, 2023).
[13]. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.
[14]. B. Green, "The flaws of policies requiring human oversight of government algorithms," Comput. Law Secur. Rev., vol. 45, p. 105681, 2022, [Online]. Available: https://doi.org/10.1016/j.clsr.2022.105681
[15]. T. Madiega, "EU guidelines on ethics in artificial intelligence: Context and implementation," 2019. [Online]. Available: https://policycommons.net/artifacts/1337743/eu-guidelines-on-ethics-in-artificial-intelligence/1945725/
[16]. NIST, "Artificial Intelligence Risk Management Framework," 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
[17]. M. C. Horowitz, G. C. Allen, E. B. Kania, and P. Scharre, "Strategic Competition in an Era of Artificial Intelligence," 2018. [Online]. Available: http://www.indexfunds.org/resources/Research-Materials/NatSec/Strategic_Competition_in_Era_of_AI.pdf
[18]. P. Brangetto, E. Çalişkan, and H. Rõigas, "Cyber Red Teaming," 2015. [Online]. Available: https://ccdcoe.org/uploads/2018/10/Cyber_Red_Team.pdf
[19]. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf
[20]. B. Winkle and E. Urbin, "Introduction to red teaming large language models (LLMs)," Microsoft, 2023. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming (accessed Aug. 18, 2023).

Dr. Isi Idemudia (Member, IEEE) became a member of IEEE in 2018. She was born in Nigeria on March 19.1983 and lives in Atlanta, Georgia, USA. Dr. Isi holds a PhD in Technology Management from the University of Poet Harcourt, Nigeria in 2019. She currently works as a Technology delivery Manager at Accenture, in Atlanta GA, USA. She is the Author of Out of Africa and Into the Cloud; Girls can code too, (2020) where she aims to inspire the next generation of female coders. Dr. Isi Idemudia serves on the Board of Nigeria-US It Network as the Vice- President Education.