"Don't Blame the Robots" – Artificial Intelligence Bias & Implications for Nuclear Security

Anna Pluff and Sneha Nair
The Stimson Center, Washington, D.C.

Abstract

Innovation and evolution are critical elements for surety in nuclear security. New threats require innovative approaches to mitigation efforts and emerging technologies are a crucial part of nuclear security infrastructure. But embracing new tools comes with new considerations and risks. Responsible nuclear security practitioners must ask: what happens when the tools for protection begin working against those they are meant to protect? Such is the concern with artificial intelligence (AI) and machine learning (ML) for nuclear security. Bias remains a pervasive issue that those in the nuclear security field grapple with. Practitioners in the AI/ML fields face similar ethical quandaries and challenges – with bias being reflected in data, learning trends, and system outputs. These potential implications of bias in AI/ML and nuclear security can pose some benefits and serious concerns for new technology integration into existing systems without due caution. Biased people produce biased products: racial, gender, accessibility, and heteronormativity biases are just a few examples of ways that system designs can unintentionally prejudice a technological system against a particular group of people. Without sufficient safeguards in place, AI/ML systems can reproduce and exacerbate biases in the nuclear security field at every level – from hiring to firing. This paper will examine the issue of bias in nuclear security and new technological systems as separate issues, then explore the potential overlaps to understand how bias can affect implementation of AI/ML technologies in the nuclear security field. In examining the potential risks of bias in implementation of AI/ML being used in nuclear security, potential areas in need of safeguarding will be identified as well as best practices for responsible implementation. Change is inevitable, and embracing new solutions is essential for addressing evolving threats – but new solutions cannot reproduce past mistakes. Mitigating bias in nuclear security is an ongoing process and considering the implications for emerging technologies is only one step towards achieving comprehensive and sustainable solutions.

Introduction

Artificial intelligence (AI) and machine learning (ML) methods have ushered in a technological renaissance when it comes to enhancing security protection measures. AI/ML provides useful tools for generating models from datasets or logic-based algorithms that imitate or even improve human decision-making and performance. A myriad of areas and fields have already seen the benefits of AI technology – from clinical research, finance, transportation, epidemiology, nutrition, medical imaging, and war fighting.[1] Applications of AI methods to nuclear

[1] IAEA "CHAPTER 11. NUCLEAR SECURITY" in *Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology* (Vienna: IAEA, 2022) https://www-pub.iaea.org/MTCD/Publications/PDF/ART-INTweb.pdf ; Jill Hruby and M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapons Systems," *Nuclear Threat Initiative* (August 2021) https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/

technologies have also seen the optimization of agricultural production, food product development, supply chain management, and safety and authenticity control.[2] AI in recruitment has also become increasingly commonplace, with bots and algorithms being used to review application materials, conduct asynchronous virtual interviews, background and reference checks for prospective applicants. In the field of nuclear security, potential applications of AI include the analysis of spectroscopic and geospatial data to improve detection of nuclear material outside of regulatory control, improvements to nuclear material accounting and control systems, and the possibility of identifying threats – both internal and external – at nuclear facilities.[3]

Many in the nuclear field are embracing AI and seek to harness its cutting-edge techniques to accelerate seemingly visionary technological development and improved security detection methods. However, despite the efficient and enhanced security measures AI might offer, these technologies can also exacerbate existing ethical concerns in the nuclear field. AI/ML technologies introduce a host of risks and uncertainty as human operators might not fully recognize potential vulnerabilities or become too reliant upon AI results. When technologies have the potential to be deployed widely, there is a critical necessity to understand their limitations. When it comes to AI in nuclear security, practitioners must move forward cautiously to mitigate risks and avoid compounding present inequalities. Given the historic marginalization of minority communities in the nuclear security field, exacerbating these exclusionary trends must be avoided at all costs.[4] One major area where experts must converge is the analysis of bias in AI, in order to avoid inserting social and cognitive biases into AI machines. The paper will proceed to define AI, ML, and bias, provide examples of bias in AI performance, address its implications in nuclear security, and offer best practices moving forward as operators work to implement new technologies.

Definitions

For the purpose of this paper, AI can broadly refer to a collection of software-based technologies that produce systems capable of tracking complex problems that mimic human intelligence – including the ability to think logically, engage, and learn through applications such as voice recognition, knowledge capture, robotics and motion, and natural language processing. To perform these tasks, AI receives signals from its environment and takes subsequent actions that then affect the environment by generating outputs such as content, predictions, recommendations, classifications, or decisions.[5]

---

[2] Ibid.

[3] Ibid. AI can also improve nuclear power by combining digital data simulations of real nuclear facilities with AI systems, optimize complex procedures and improve reactor design, performance, and safety. AI applications in safeguards can also help nuclear inspectors examine satellite imagery, environmental sampling, gamma ray spectroscopy, and video surveillance. See Artem Vlasov and Matteo Barbarino, "Seven Ways AI Will Change Nuclear Science and Technology," *International Atomic Energy Agency* (September 22, 2022) https://www.iaea.org/newscenter/news/seven-ways-ai-will-change-nuclear-science-and-technology for more information.

[4] Sneha Nair, "Converging Goals: Examining the Intersection Between Diversity, Equity, and Inclusion and Nuclear Security Implementation," in *Nuclear Threat Initiative's 16th Global Dialogue on Nuclear Security Priorities* (April 2023).

[5] Reva Schwartz, Apostal Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication 1270* (March 2022) https://doi.org/10.6028/NIST.SP.1270

Systems achieve AI capabilities by using machine learning (ML) algorithms. Everyday machine learning (ML) applications include facial and voice recognition as well as predictive suggestions for books, movies, and shows, people might like based on other purchases and reviews.[6] ML refers more specifically to the "field of study that gives computers the ability to learn without being explicitly programmed."[7] In contrast to AI, ML mathematical models can only perform based on the inputs given or what it was trained to do – it cannot adapt to the learning process itself nor can it apply the results to understand a problem.[8] Training data must be collected before an ML model is built, and it is thus necessary that all the data included will be the phenomena the model will need to interpret. Thus, the quality of data directly impacts model performance.

The caliber of data given to AI/ML models directly relates to the problem of bias in AI. Bias exists in numerous forms and is omnipresent in society, but it can also become ingrained in our automated systems. Human decisions and AI decisions can generate bias, which for the purpose of this paper, is defined as AI outcomes which are systematically less favorable to individuals within a particular group.[9] There may be no relevant difference between the groups that justifies such harms.[10] Typically, these biased outputs follow traditional societal biases like race, gender, sex, etc. Bias in algorithms is caused by under-representative or incomplete training data that is fed into ML models. Most attempts to address the harmful effects of AI bias remain focused solely on computational factors such as representative datasets or fair ML algorithms.[11]

It is critical to note that AI systems do not operate in isolation; AI is usually deployed to help humans make decisions that hold direct consequences on other humans' lives. Many practitioners acknowledge that AI systems can exhibit biases that stem from their programming and data sources, i.e., a machine learning software may be fed information that underrepresents a

---

[6] Jill Hruby and M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapons Systems," *Nuclear Threat Initiative* (August 2021) https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/

[7] Reva Schwartz, Apostal Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication 1270* (March 2022) https://doi.org/10.6028/NIST.SP.1270

[8] Shannon Eggers and Char Sample, "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data," *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

[9] Another type of bias often associated with AI is the "automation bias," which occurs when human decision makers place too little or too much trust in AI results. See Jill Hruby and M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapons Systems," *Nuclear Threat Initiative* (August 2021) https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/ for more information.

[10] Nicol Turner Lee, Paul Resnick, Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (May 22, 2019) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ ; "Understanding Bias in Algorithmic Design," *Impact Engineered* (September 5, 2017) https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

[11] Reva Schwartz, Apostal Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" *NIST Special Publication 1270* (March 2022), https://doi.org/10.6028/NIST.SP.1270

particular gender or ethnic group.[12] But to better understand and mitigate these biases, it is necessary look beyond just data-specific perspectives and take into account *human* and *systemic* biases. *Systemic biases* – also referred to as historical or institutional bias – results from institutions operating in ways that harm or disadvantage certain social groups, such as race.[13] Operators tend to rely on flawed information that reflects deeply ingrained historical inequalities. *Human biases* relate to how people use data to fill in or replace incomplete information. For example, where someone lives, or their neighborhood might influence how likely authorities would consider them to be a crime suspect.[14] When human, systemic and computational biases combine, AI models can perpetuate negative effects on individuals.

It is important to consider how bias can multiply as more individuals engage with AI technology. Bias enters AI through the teams involved in AI system and design. These individuals can bring their own cognitive biases into the design process. Systemic biases enter at the institutional level and affect how organizations and teams are structured and who controls the decision-making processes, and individual and group heuristics and cognitive/perceptual biases throughout the AI life cycle.[15] Decisions made by end users, downstream decision makers, and policy makers are also impacted by these biases, which can lead to biased outcomes or limited points of view.[16] These biases in AI application have to be taken into consideration when applied to the nuclear security field.

Harmful Effects of Bias in AI

So, what does harm in AI bias look like? Applications that utilize AI are often utilized across sectors and contexts for decision-making and decision-support. Thus, AI systems that replace human processes for high-impact decisions can directly affect the lives and fates of other humans. AI development teams tend to have unrealistic expectations of how the technology will be applied and what it can accomplish, especially when deployed to the general public. More concerning is that ML models tend to exhibit "unexpectedly poor behavior when deployed in real world domains" without domain-specific constraints supplied by human operators.[17] Examining these "poor behaviors" will require addressing the social contexts in which AI/ML models are developed and deployed within.

For example, after the 2020 murder of George Floyd, an unarmed Black man, by Minneapolis police offers, attention was brought to bias in AI law enforcement techniques. Police use past information about a crime as material for ML algorithms to make predictions about future crimes.[18] However, the data used to "teach" software systems is embedded with historical and institutional biases. Predictive algorithms are skewed by arrest rates, and black

[12] National Institute for Science and Technology, "There's More to AI Bias Than Biased Data, NIST Report Highlights" (March 16, 2022) https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights
[13] Ibid.
[14] Ibid.
[15] Ibid.
[16] Ibid.
[17] Alexander D'Amour, *et al.*, "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *Journal of Machine Learning Research* 23 (2022): 1-61 https://arxiv.org/abs/2011.03395
[18] Hope Reese, "What Happens When Police Use AI to Predict and Prevent Crime?" *JSTOR* (February 23, 2022) https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

people are more likely than white people to be reported for a crime—whether the reporter is black or white. Unfortunately, this leads to Black neighborhoods being marked by AI/ML as "high risk" at a disproportionate rate.[19]

In general, automated policing approaches have exhibited high rates of inaccuracy. Facial recognition tools have been shown to be especially faulty. Issues begin during the process of labelling headshot photographs in the training dataset. As Edward Santow writes, "when these labels include information that can more easily be mistaken (like an individual's age or gender), or where these labels involve subjective judgements (like relative attractiveness or happiness), the computer essentially will learn to take on the subjective beliefs of the people who are assigning the labels."[20] One notable example of the problems associated with facial recognition software can be seen in the 2018 London Metropolitan Police trial, which used facial recognition technology to identify previously unknown people who were suspected of committing crimes. Only 2 out of the 104 results were accurate.[21] Studies have shown that human prejudices are often baked into these tools and ML is simply trained on biased data that perpetuates inequalities in society.[22] If facial recognition becomes a more commonplace AI tool, especially in policing, it may mean that Black people or other marginalized groups may be more frequently identified and tracked since most of these individuals are already enrolled in law enforcement databases based on larger social trends that incarcerate and arrest Black people at a higher rate.[23]

Researchers have noted that facial recognition systems are mostly likely to demonstrate bias against people of color and have exhibited different accuracy rates for various demographic groups. For example, Facebook's facial recognition algorithm labeled Black people as "primates" which it told BBC "was clearly an unacceptable error."[24] In 2018, researchers from MIT and Microsoft generated news with a report showing that gender classification algorithms – which are related, though distinct from face identification algorithms – had error rates of just 1% for white men, but almost 35% for dark-skinned women.[25] In response to MIT's findings both

---

[19] Ibid.

[20] Edward Santow, "Can Artificial Intelligence Be Trusted with Our Human Rights?" *AQ: Australian Quarterly* 91, no. 4 (2020): 10–17. https://www.jstor.org/stable/26931483.

[21] Ibid. ; Robert Booth, "Police Face Calls to End Use of Facial Recognition Software," *The Guardian* (July 3, 2019) https://www.theguardian.com/technology/2019/jul/03/police-face-calls-to-end-use-of-facial-recognition-software ; Hope Reese, "What Happens When Police Use AI to Predict and Prevent Crime?" *JSTOR* (February 23, 2022) https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

[22] Will Douglas Heaven, "Predictive Policing Algorithms Are Racist. They Need to Be Dismantled," *MIT Technology Review* (July 17, 2020) https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

[23] William Crumpler, "The Problem of Bias in Facial Recognition," *Center for Strategic and International Studies* (May 1, 2020) https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition

[24] "Facebook Apology as AI Labels Black Men 'Primates,'" *BBC* (September 6, 2021) https://www.bbc.com/news/technology-58462511

[25] William Crumpler, "The Problem of Bias in Facial Recognition," *Center for Strategic and International Studies* (May 1, 2020) https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition ; Larry Hardesty, "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems," *MIT News* (February 11, 2018) http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

IBM and Microsoft issued commitments to improving the accuracy of their recognition software for darker-skinned faces.[26]

Bias in AI and Nuclear Security

Existing and near-term AI security solutions includes technologies such as behavior monitoring for insider threat identification and enhanced security tools for information and communications technology (ICT) and operational technology (OT) environments.[27] Some longer-term developments include data-fusion applications for physical protection and nuclear materials accounting and control (NMAC) and future use of AL/ML tools in nuclear power plants for condition monitoring.[28] When it comes to analyzing insider threats, AI also represents a sought-after pathway for mitigating threats. Insider threats have attracted greater attention in recent years, as they have access to and authority within the facility. This means they have more opportunities to choose vulnerable targets and time to plan out malicious behavior.[29] Most of the known incidents of nuclear material theft and sabotage at nuclear facilities were carried out by insiders.[30]

To reduce insider threats, insider mitigation programs may use AI/ML-based behavioral recognition programs to pinpoint suspicious employee behavior. These programs track and monitor employee computer-based actions. Examples include: file browsing, usage, and downloads, USB usage, and application/system logins. Physical actions are also monitored – such as facility entries and exits – to identify normal vs. abnormal behavior.[31]

Two examples of AI/ML deployments within physical protection systems include facial recognition software and abnormal behavior identification. ML techniques have been studied in the literature as a promising and innovative solution for insider threats by using these tools.[32] However, they can be biased and/or inaccurate when the associated dataset is imbalanced. It is critical that nuclear security practitioners and computer programmers take into account the biases that have been shown to emerge in other facial recognition tools to avoid reproducing societal biases at nuclear facilities. The nuclear field continues to struggle to reach gender and racial parity among marginalized groups and it is important to address the potential effects of AI

[26] Nicol Turner Lee, Paul Resnick, Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (May 22, 2019) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

[27] Shannon Eggers and Char Sample, "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data," *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

[28] Shannon Eggers and Char Sample, "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data," *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

[29] IAEA, "Preventive and Protective Measures against Insider Threats," *International Atomic Energy Agency Nuclear Security Series* no. 8 https://www-ns.iaea.org/downloads/security/security-series-drafts/implem-guides/nst041.pdf

[30] Matthew Bunn and Scott Sagan, *A Worst Practices Guide to Insider Threats: Lessons From Past Mistakes* (Cambridge: American Academy of Arts and Sciences, 2014) ; Jung Hwan Kim, Chul Min Kim, and Man-Sung Yim, "An Investigation of Insider Threat Mitigation Based on EEG Signal Classification," *Sensors* 20, no. 21 (November 2020): doi: 10.3390/s20216365

[31] Shannon Eggers and Char Sample, "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data," *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

[32] Shi Chen, Kazuyuki Demachi, "Proposal of an Insider Sabotage Detection Method for Nuclear Security Using Deep Learning," *Journal of Nuclear Science and Technology* 56, no. 7 (2019): 599-607. https://doi.org/10.1080/00223131.2019.1611501

mechanisms in exacerbating existing harms. These effects might be caused by AI technologies mislabeling or misidentifying an individual of a marginalized group as a threat, similar to failures in AI-powered policing.

This leads into the larger discussion of existing bias in the nuclear field. It is clear that it is important to fully understand the social context in which an acquired technology will be used, and how the social context itself might be biased.[33] When it comes to the nuclear security community, only 33 percent of the National Nuclear Security Administration's total workforce self-identify as members of historically under-represented groups, and members of "underrepresented minorities" and "other people of color" together made up about 32 percent of the combined workforce of the 17 National Laboratories in 2022. The importance of large and representative datasets has already been described, but the situational contexts in which data is applied, given past racialized harm and biases that data has magnified, is critical to address. Even beyond direct nuclear security applications, applications of AI in hiring and recruitment are being used to review application materials, conduct interviews and practical assessments, background, and reference checks for prospective applicants. The aforementioned biases in AI can have trickle down effects on how candidates' qualifications are assessed, and the criteria used by the algorithm, thus influencing and exacerbating the existing homogeneity of practitioners the field.

Tackling bias in AI and developing best practices can help the nuclear security community better embrace an effective diversity, equity, and inclusion (DEI) culture. Indeed, reforming bias in AI also offers the chance to examine DEI practices at nuclear facilities and wider organizations. Researcher Nicol Turner Lee has argued that the lack of diversity among programmers or teams designing the training sample for AI/ML can lead to the underrepresentation of demographic groups.[34] Research has also shown that that bias may in fact be *reduced* when the team making the decisions is representative of the affected populations, in that the team members would presumably have a better understanding of those populations because of the characteristics team members share with them.[35] The makeup of the team creating algorithms is thus imperative for shaping what biases may get introduced or mitigated. Algorithm developers must consider the role of DEI within their work teams, training data, and the level of cultural sensitivity within their decision-making processes.[36] Diverse AI development teams can potentially avoid harmful discriminatory effects on certain protected groups and avoid what could be detrimental consequences if underrepresented groups were

[33] Douglas Yeung, Inez Khan, Nidhi Kalra, Osonde A. Osoba, "Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement," *RAND Corporation* (January 2021): 1-24. https://www.jstor.org/stable/resrep29576

[34] Nicol Turner Lee, "Detecting Racial Bias in Algorithms and Machine Learning," *Journal of Information, Communication and Ethics in Society* 16, no. 3 (August 2018): 252-260 DOI:10.1108/JICES-06-2018-0056

[35] Andrew R. Todd, Galen V. Bodenhausen, Jennifer A. Richeson, and Adam D. Galinsky, "Perspective Taking Combats Automatic Expressions of Racial Bias," *Journal of Personality and Social Psychology* 100, no. 6, (2011): 1027–1042.

[36] Nicol Turner Lee, Paul Resnick, Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (May 22, 2019) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

identified as a threat by AI technology as a result of poor design, rather than misbehaviors of concern.

Strategies for Mitigation

The Intelligence and National Security Alliance (INSA) offers up strategies for addressing bias in insider threat programs by recognizing that biases can enter both human and machine processes. Its recommendations include effectively addressing cognitive bias by raising awareness through training and transparency, as well as structured decision-making that integrates strategies such as purposeful questioning to explore alternatives.[37] When hiring individuals to work at nuclear facilities, the inclusion of certain types of data must be carefully considered, as it can introduce a selection bias into insider threat programs. For example, the collection of arrest records during background checks may introduce bias into employee profiles given substantial evidence of racial disparities in arrests.[38] Similarly, incorporating analysis of travel to countries of concern can flag individuals of certain ethnicities who have innocuous family ties to such nations and may travel there frequently.[39]

Finally, when it comes to analyzing data, we must consider cultural and social interpretations of levels of risk. Human cognitive and social biases, as well as cultural influences, determine the likelihood of a certain indicator being associated with an insider threat and thus impacts how we judge and assess people. For example, in the U.S., from the period of 1940-1995, the government considered homosexuality to present a security risk because of societal biases and fear of communist influence, even though no linkage between sexual orientation and espionage was ever found to exist.[40] This example shows that nuclear facilities must take cognitive and social biases seriously when implementing AI/ML technologies. To reduce bias moving forward, organizations must engage in robust DEI practices to widen the breadth of ideas, considerations, understanding, and sensitivity when adopting AI programs to mitigate insider threats.

Role of Civil Society in Addressing AI/ML Risks

Civil society organizations can be the fulcrum of information exchange and promote collaborative opportunities among practitioners, stakeholders, lawmakers, and policymakers to address the question of bias in AI. Civil society organizations are often best equipped to engage in investigative research to better address the positive and negative aspects of AI in nuclear security. Their work often facilitates information exchange to support a common understanding of how AI works and how to best utilize AI tools. The first step is often understanding the scope

[37] "Strategies for Addressing Bias in Insider Threat Programs," *INSA's Insider Threats Subcommittee, Presentation* (January 2022) https://www.insaonline.org/docs/default-source/default-document-library/2022-white-papers/bias-and-insider-threat-programs-paper.pdf

[38] Benjamin Mueller, Robert Gebeloff, and Sahil Chinoy, "Surest Way to Face Marijuana Charges in New York: Be Black of Hispanic," *New York Times* (May 13, 2018) https://www.nytimes.com/2018/05/13/nyregion/marijuana-arrests-nyc-race.html

[39] "Strategies for Addressing Bias in Insider Threat Programs," *INSA's Insider Threats Subcommittee, Presentation* (January 2022) https://www.insaonline.org/docs/default-source/default-document-library/2022-white-papers/bias-and-insider-threat-programs-paper.pdf

[40] Government Accountability Office, Security Clearances: Consideration of Sexual Orientation in the Clearance Process, GAO/ NSIAD-95-21, March 24, 1995, p. 15. At https://www.gao.gov/assets/ nsiad-95-21.pdf.

of AI bias and its associated risks.[41] One useful tool for organizations to adopt is a bias impact statement. These statements help operators systematically address assumptions about how the algorithm will work prior to its deployment. Civil society organizations can help operators design bias statements to address and avert potential biases and respond and adapt when biases emerge.[42] Moreover, civil society works best when it can communicate with and engage stakeholders. Stakeholder engagement can also help those programming the AI by applying a diversity of perspectives and concerns. One proposed solution by researchers at the Brookings Institute is the establishment of an advisory council of civil society organizations that can work alongside companies and help define the scope of procedures and predict biases based on their experiences and research.[43] Civil society can also host events and workshops to identify key questions or concerns and help enhance guidelines for operators and regulators. These events can serve as meeting grounds and in-roads for experts, regulators, and industry partners to widen the discussion surrounding AI and promote pathways to cooperation. Finally, these opportunities will help the field better iterate positive policy development, avoid repeating similar mistakes, and develop coordinated responses.

The following summarizes the steps civil society organizations can take to promote responsible AI:

1. Publish investigative research to better understand risk;
2. Assist in developing templates for bias impact statements;
3. Engage stakeholders and support interdisciplinary cooperation in the nuclear community;
4. Widen the discussion by holding workshops/events on benefits and challenges of AI; and
5. Exchange ideas to establish common goals, coordinate responses, and avoid redundant mistakes

Conclusion

AI advancements have certainly offered faster and improved data insights, more efficient and automated processes, and a reduction in common-place errors. Nuclear security applications have seen its benefits from behavior analysis for insider threat mitigation, source tracking of stolen nuclear material, and facial recognition software for physical protection.[44] Nonetheless, with any new technological advancement, there are serious vulnerabilities to be considered. As human decision makers continue to navigate their increasingly complex relationships with machines, it is important that nuclear security practitioners continue to follow best practices in AI applications to reduce amplifying biases or inadvertently introducing new threats and vulnerabilities. Understanding the limitations of AI can help mitigate risks early in the research

[41] Anja Kaspersen and Chris King, "Mitigating the Challenges of Nuclear Risk While Ensuring the Benefits of Technology," edited by Vincent Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives* (Stockholm International Peace Research Institute, 2019) http://www.jstor.org/stable/resrep24525.20.
[42] Nicol Turner Lee, Paul Resnick, Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings* (May 22, 2019) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
[43] Ibid.
[44] Shannon Eggers and Char Sample, "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data," *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

and development process. It is important to hold conversations with civil society organizations about how these risks can be addressed and improved when AI technologies are put into practice.[45]

Human and systemic institutional and societal factors will continue to remain one of the most significant sources of AI bias and will require operators and practitioners to expand their perspective beyond the ML pipeline to recognize how this technology is created within and impacts our society.[46] AI is neither built nor deployed within a perfect silo, sealed off from societal realities of discrimination or unfair practices. It is thus imperative that AI models are viewed as more than simply mathematical and computational inputs.[47] Examining bias in current AI models will help nuclear practitioners effectively combat serious security threats without aggravating existing discriminatory practices. With every new technology comes a serious reckoning of its societal implications and how it will make individuals, organizations, and society safer.

[45] Vincent Boulanin, "Promises and Perils of Artificial Intelligence for Strategic Stability and Nuclear Risk Management: Euro-Atlantic Perspectives," *Stockholm International Peace Research Institute* (2019) http://www.jstor.com/stable/resrep24525.21

[46] Reva Schwartz, Apostal Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication 1270* (March 2022), https://doi.org/10.6028/NIST.SP.1270

[47] National Institute for Science and Technology "There's More to AI Bias Than Biased Data, NIST Report Highlights" (March 16, 2022) https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

Bibliography

"Artificial Intelligence: Using Standards to Mitigate Risks." Analytics Exchange Program (AEP), 2018.

Booth, Robert. "Police Face Calls to End Use of Facial Recognition Software." *The Guardian* (July 3, 2019) https://www.theguardian.com/technology/2019/jul/03/police-face-calls-to-end-use-of-facial-recognition-software

Boulanin, Vincent. "Promises and Perils of Artificial Intelligence for Strategic Stability and Nuclear Risk Management: Euro-Atlantic Perspectives." *Stockholm International Peace Research Institute* (2019) http://www.jstor.com/stable/resrep24525.21

Bunn, Matthew and Scott Sagan. *A Worst Practices Guide to Insider Threats: Lessons From Past Mistakes.* Cambridge: American Academy of Arts and Sciences, 2014.

Chen, Shi, Kazuyuki Demachi. "Proposal of an Insider Sabotage Detection Method for Nuclear Security Using Deep Learning." *Journal of Nuclear Science and Technology* 56, no. 7 (2019): 599-607. https://doi.org/10.1080/00223131.2019.1611501

Crumpler, William. "The Problem of Bias in Facial Recognition." *Center for Strategic and International Studies* (May 1, 2020) https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition

D'Amour, Alexander, *et al.* "Underspecification Presents Challenges for Credibility in Modern Machine Learning." *Journal of Machine Learning Research* 23 (2022): 1-61. https://arxiv.org/abs/2011.03395

Douglas Heaven, Will. "Predictive Policing Algorithms Are Racist. They Need to Be Dismantled." *MIT Technology Review* (July 17, 2020) https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

Eggers, Shannon and Char Sample. "Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data." *Report Prepared for the NNSA Office of International Nuclear Security Emerging Threats and Technologies Working Group* (Idaho National Laboratory: 2020).

"Facebook Apology as AI Labels Black Men 'Primates." *BBC* (September 6, 2021), https://www.bbc.com/news/technology-58462511

Government Accountability Office. "Security Clearances: Consideration of Sexual Orientation in the Clearance Process," GAO/ NSIAD-95-21 (March 24, 1995), https://www.gao.gov/assets/ nsiad-95-21.pdf.

Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." *MIT News* (February 11, 2018) http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Hruby, Jill and M. Nina Miller. "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapons Systems." *Nuclear Threat Initiative* (August 2021) https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/

Hwan Kim, Jung, Chul Min Kim, and Man-Sung Yim. "An Investigation of Insider Threat Mitigation Based on EEG Signal Classification." *Sensors* 20, no. 21 (November 2020): doi: 10.3390/s20216365

IAEA. "Chapter11. Nuclear Security." In *Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology.* Vienna: IAEA, 2022. https://www-pub.iaea.org/MTCD/Publications/PDF/ART-INTweb.pdf

IAEA. "Preventive and Protective Measures against Insider Threats/" *International Atomic Energy Agency Nuclear Security Series* no. 8 https://www-ns.iaea.org/downloads/security/security-series-drafts/implem-guides/nst041.pdf

Johnson, Melanie. "7 Effective Uses of AI in Recruitment," *Unleash* (March 30, 2022) https://www.unleash.ai/artificial-intelligence/7-effective-uses-of-ai-in-recruitment/

Kaspersen, Anja and Chris King. "Mitigating the Challenges of Nuclear Risk While Ensuring the Benefits of Technology." Edited by Vincent Boulanin *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives* (Stockholm International Peace Research Institute, 2019) http://www.jstor.org/stable/resrep24525.20.

Mueller, Benjamin, Robert Gebeloff, and Sahil Chinoy. "Surest Way to Face Marijuana Charges in New York: Be Black of Hispanic,. *New York Times* (May 13, 2018) https://www.nytimes.com/2018/05/13/nyregion/marijuana-arrests-nyc-race.html

Nair, Sneha. "Converging Goals: Examining the Intersection Between Diversity, Equity, and Inclusion and Nuclear Security Implementation." *Nuclear Threat Initiative's 16th Global Dialogue on Nuclear Security Priorities.* April 2023.

National Institute for Science and Technology. "There's More to AI Bias Than Biased Data, NIST Report Highlights" (March 16, 2022) https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

Reese, Hope. "What Happens When Police Use AI to Predict and Prevent Crime?" *JSTOR* (February 23, 2022) https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

Santow, Edward. "Can Artificial Intelligence Be Trusted with Our Human Rights?" *AQ: Australian Quarterly* 91, no. 4 (2020): 10–17. https://www.jstor.org/stable/26931483.

Schwartz, Reva, Apostal Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." *NIST Special Publication 1270* (March 2022), https://doi.org/10.6028/NIST.SP.1270

"Strategies for Addressing Bias in Insider Threat Programs." *INSA's Insider Threats Subcommittee, Presentation* (January 2022) https://www.insaonline.org/docs/default-source/default-document-library/2022-white-papers/bias-and-insider-threat-programs-paper.pdf

Todd, Andrew R., Galen V. Bodenhausen, Jennifer A. Richeson, and Adam D. Galinsky. "Perspective Taking Combats Automatic Expressions of Racial Bias." *Journal of Personality and Social Psychology* 100, no. 6, (2011): 1027–1042.

Turner Lee, Nicol. "Detecting Racial Bias in Algorithms and Machine Learning." *Journal of Information, Communication and Ethics in Society* 16, no. 3 (August 2018): 252-260. DOI:10.1108/JICES-06-2018-0056

Turner Lee, Nicol, Paul Resnick, Genie Barton. "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms." *Brookings* (May 22, 2019) https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

"Understanding Bias in Algorithmic Design." *Impact Engineered* (September 5, 2017) https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Vlasov, Artem and Matteo Barbarino. "Seven Ways AI Will Change Nuclear Science and Technology." *International Atomic Energy Agency* (September 22, 2022) https://www.iaea.org/newscenter/news/seven-ways-ai-will-change-nuclear-science-and-technology

Yeung, Douglas, Inez Khan, Nidhi Kalra, Osonde A. Osoba. "Identifying Systemic Bias in the Acquistion of Machine Learning Decision Aids for Law Enforcement." *RAND Corporation* (January 2021): 1-24. https://www.jstor.org/stable/resrep29576