

Digital Tools: Safeguarding National Security, Cybersecurity, and AI Bias

Gaudys L. Sanclemente

Abstract: This article explores the critical challenge of biases in artificial intelligence (AI) and its potential implications for national security. It discusses types of biases in AI systems, their consequences on national security and outlines potential mitigation strategies. The paper examines case studies, regulatory measures, and the evolving landscape of AI's role in shaping national security, emphasizing the need for ethical and responsible use.

Keywords: national security; artificial intelligence; bias; digital tools; emerging technologies.

Ferramentas digitais: salvaguardando a segurança nacional, a cibersegurança e o tendenciosismo na IA

Resumo: Este artigo explora o desafio crítico representado pelo tendenciosismo na inteligência artificial (IA) e as suas potenciais implicações para a segurança nacional. Discute tipos de vieses nos sistemas de IA, suas consequências para a segurança nacional e potenciais estratégias de mitigação. O artigo examina estudos de caso, medidas regulatórias e o cenário em evolução do papel da IA na formação da segurança nacional, enfatizando a necessidade de um uso ético e responsável.

Palavras-chave: segurança nacional; inteligência artificial; viés; ferramentas digitais; tecnologias emergentes.


In the contemporary world, safeguarding national security is a paramount concern for governments worldwide, with emerging technologies assuming an ever more significant function in the defense strategies of nations. One of the most exciting and potentially transformative of these technologies is artificial intelligence (AI). Literature defined it as: the exploration of entities that gather information from their surroundings and execute actions (Russell & Norvig 2016); the automation of activities associated with human thinking (Bellman 1978); machines that execute tasks that necessitate human intelligence when carried out by individuals (Kurzweil 1990); the study of mental faculties through computational model use (Charniak & McDermott 1985); computations enabling the capability to observe, rationalize, and respond (Winston 1992); and intelligence behavior such as perception and reasoning in artifacts (Nilsson 1998).

In 1984, a scholar predicted that computer scientists and experts in AI would eventually create hardware and programs comparable to human brains and minds (Searle 1984). AI's significance stems from its ability to simulate human intelligence processes through computer systems (Russell & Norvig 2016), handling and extracting information from large datasets and big data to produce new data handling (Kitchin 2014). As AI technology advances rapidly, scholars' forecasts are increasingly coming to fruition, yet the utilization of AI can introduce biases that affect both effectiveness and fairness.

UNDERSTANDING BIAS IN AI

AI bias pertains to consistent mistakes or imprecisions in the choices made by AI algorithms, which unjustly promote or prejudice specific individuals or groups. These biases might arise intentionally or inadvertently, stemming from a range of causes. Biases arise from flawed algorithm design, training data skew, or system architecture, leading to unintended discriminatory decisions (Barocas, Hardt & Narayanan 2023). Thus, it is the unfair treatment of certain groups or individuals resulting from an AI algorithm's design or training data.

Bias signifies slanted information concerning computer systems that systematically and unfairly discriminate in favor of certain individuals or groups,

Gaudys L. Sancllemente  *Ph.D. in International Studies, is a mixed methods social scientist and research professional focusing on health, cybersecurity, and intelligence, bridging science, technology, and security. She is also an award-winning journalist, writer, poet, and artist, holding degrees in Master of Arts, Juris Doctor, and Master of Laws.*

while disadvantaging others (Friedman & Nissenbaum 1996). Biases encompass a wide spectrum ranging from inherent cognitive tendencies to societal influences. They can shape individuals' perceptions, interactions, and decisions. Numerous biases are present in various contexts and domains (Sanclemente 2021; Fleischmann et al. 2014), and their influence can extend across different phases of development. They can be introduced into every stage of the deployment of systems, from the intention that governs the algorithm's development, the code creation, executable code, and in the context of maintenance and execution (Défenseur des droits and Commission Nationale Informatique & Libertés 2020; Barocas & Selbst 2016). Similarly, in machine learning, bias can manifest during the construction of an application, encompassing data collection, processing, and inputting information into a machine-learning model.

The following paragraph offers a simplified and high-level depiction of an AI workflow's data collection and design process. Figure 1 illustrates the sequential stages in an AI workflow. While typically commencing with selecting a model, wherein the most suitable algorithm is chosen, it is worth noting that, in some instances, the process might initiate with data collection, which can subsequently influence the model's development. This step is succeeded by collecting relevant data, followed by data preparation involving cleaning and formatting the data for analysis. Subsequently, attention turns to model training and improvement enhancing the algorithm's performance and may be repeated. As the workflow progresses, the deployment of the model entails its integration into practical applications. The cycle continues with improvements or ongoing enhancements which can also repeat and highlight the iterative nature of ongoing AI enhancement.

Biases arise from flawed algorithm design, training data skew, or system architecture, leading to unintended discriminatory decisions. Thus, it is the unfair treatment of certain groups or individuals resulting from an AI algorithm's design or training data. Bias signifies slanted information concerning computer systems that systematically and unfairly discriminate in favor of certain individuals or groups, while disadvantaging others.

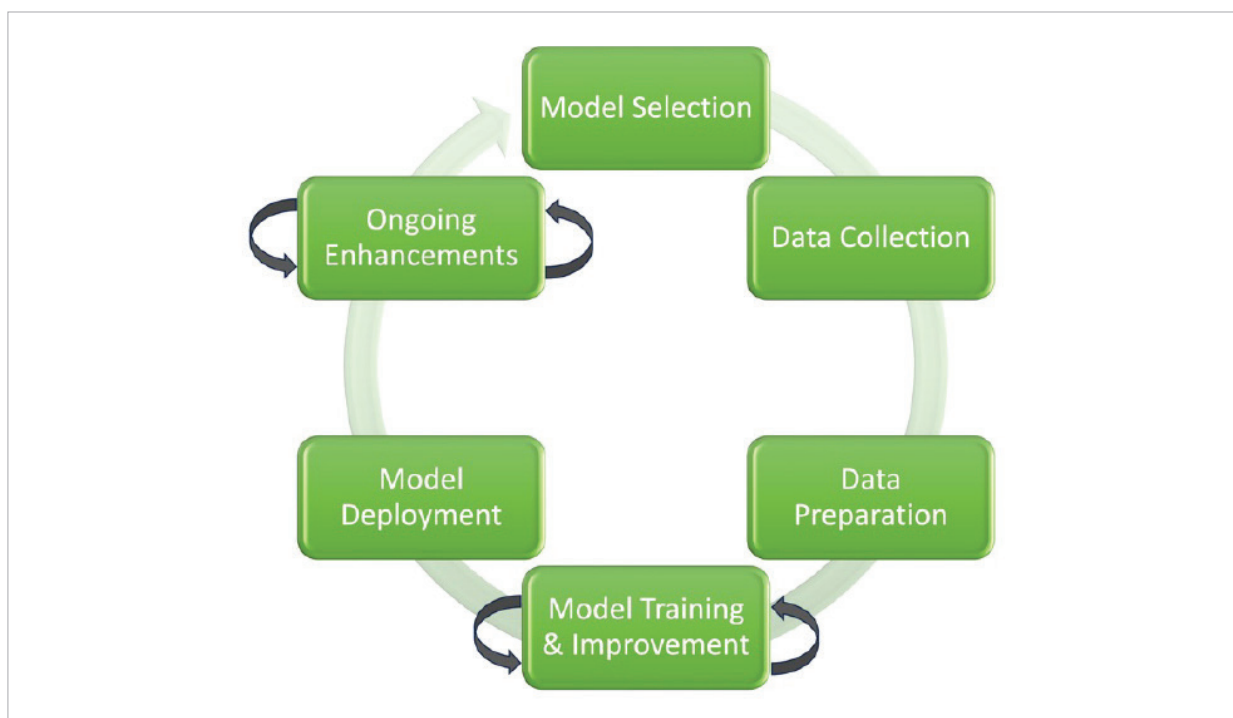


Figure 1. Sequential Phases of AI Workflow: From Model Selection to Continuous Improvement. Created by Gaudys L. Sanclemente.

It is important to highlight that bias can potentially manifest at multiple stages throughout this process. Factors such as data collection methods, algorithm design, and the context of application can all contribute to bias in AI systems. Vigilance and comprehensive evaluations are critical to addressing and mitigating these biases effectively.

Likewise, the definition of bias differs across various academic fields of study, from computer science and engineering to law, psychology, philosophy, and biology, typically involving aspects of uneven treatment, disparate impact, and unfair representation. From a philosophical perspective, social scientists examine this issue through a theoretical framework that is either already in existence or can be anticipated. On the contrary, data scientists and programmers label AI biases as glitches, classifying the problem as a technical issue akin to security, which requires rectification (Belenguer 2022). Data bias can manifest in various ways, potentially resulting in discrimination (Belenguer 2022). Thus, there can exist several forms of biases. For instance, sampling bias occurs when the data set used to train an AI algorithm does not represent the population, leading to inaccurate or unfair decisions (Sun, Nasraoui & Shafto 2020). Another bias includes confirmation bias, which occurs when an AI algorithm is programmed to confirm preexisting beliefs or assumptions rather than providing an objective and accurate analysis (Fleischmann et al. 2014; Evans 2007).

Similarly, implicit bias occurs when an AI algorithm incorporates societal biases, such as racial or gender biases, into its decision-making processes (Levendovski 2018; Barocas & Selbst 2016). Datasets containing implicit bias during the training phase can lead to unbalanced data that drive incorrect identifications of false positives or false negatives. Consequently, other forms of bias that can emerge in the design of computer systems are preexisting biases rooted in social institutions or individuals with significant input into the system design. Likewise, technical biases may emerge due to constraints within computer technology's software and hardware and challenges within the technical design (Friedman & Nissenbaum 1996). These technical biases underscore the need for comprehensive and ethical technological development practices.

Moreover, algorithmic bias might surface while selecting the appropriate algorithm for crafting the training model stemming from a problem within the algorithm that performs the calculations powering the machine learning computations. AI models might exhibit algorithmic bias due to the biases in the data they were trained on (Hoadley & Sayler 2020). One category of machine learning algorithms is the process of information filtering, which results in algorithmic bias and inclines individuals to predominantly encounter information that aligns with their existing beliefs (MIT Technology Review Insights 2022; Peralta et al. 2021). Other classifications include the neural network algorithm, which consists of interconnected units, as well as linear regression, support vector machines, decision trees, or random forests (Russell & Norvig 2016). While not an exhaustive compilation, these are a few examples of the most widely used machine learning algorithms.

AI biases within the realm of national security can engender discriminatory practices, violate human rights, adversely impact communities, and undermine the effectiveness of national security efforts. AI's application in national security spans a wide range of tasks, including threat detection, border control, addressing national security threats, and conducting intelligence analysis (Dorton, Harper & Neville 2022; Schmidt 2022; Gibert, Mateu, and Planes 2020). Nevertheless, as reliance on emerging technologies like quantum computing and AI is poised to intensify within the cyberspace domain, the implementation of AI may inadvertently introduce biases that compromise fairness and accuracy (Cavelty & Wenger 2020; Caliskan, Bryson & Narayanan 2017). An erroneous algorithm choice can culminate in biased predictions. Embracing a "one size fits all" methodology is less than ideal, given the distinct applications inherent in each algorithm; a tailored selection is imperative to suit specific contexts. Nevertheless, effectively navigating these diverse algorithms and biases culminates in the realization of a more justifiable application. Therefore, these biases can influence any developmental stage of the machine learning application.

NAVIGATING THE IMPACT OF BIASED AI ON NATIONAL SECURITY

National security strategy involves assessing the strategic landscape, skillfully using expertise and tools for better decision-making, and continually refining the strategic blueprint through iterative re-evaluation. AI technology presents both opportunities and challenges for policymakers, fundamentally impacting the scope of military force development and deployment (U.S. White House Office 2022). Ultimately, the influence of AI on national security strategy underscores the need for insightful and adaptive decision-making in an increasingly complex landscape.

Nevertheless, the impact of biased AI on national security can have serious consequences, including limiting the effectiveness of security measures, impinging on individual rights, and perpetuating discrimination. On the one hand, AI equips decision makers with the means to thwart artificially generated, nonsensical interpretations (Kahneman 2011). Conversely, AI algorithms imbued with bias can engender erroneous or unjust decisions, thereby introducing flaws into the fabric of national security endeavors. To illustrate, prejudiced algorithms deployed in border control settings might erroneously apprehend or expel innocent individuals, or worse, facilitate the unchecked entry of potentially hazardous persons into a country's borders (Laupman, Schippers & Papaléo Gagliardi 2022). These digital borders rely on machine learning, automated algorithmic decision-making systems, and predictive data analytics (UN General Assembly 2020). Similarly, while AI can be judiciously employed to adopt a preemptive stance against terrorism, the presence of bias in counterterrorism algorithms could also yield unfounded allegations and wrongful convictions, posing risks to diplomatic relationships and eroding public confidence (McKendrick 2019; Osoba & Welser IV 2017). Consequently, the influence of biased AI significantly contributes to lopsided outcomes that stand in stark contrast to the bedrock principles of justice and equitability.

The impact of biased AI on national security can have serious consequences, including limiting the effectiveness of security measures, impinging on individual rights, and perpetuating discrimination. (...) AI algorithms imbued with bias can engender erroneous or unjust decisions, thereby introducing flaws into the fabric of national security endeavors.

Furthermore, biased AI infringes individual rights, leading to privacy violations and discriminatory practices (Chouldechova 2017). For instance, facial recognition technology used in surveillance could result in the encroachment upon individuals' privacy, while biased predictive policing algorithms may unfairly target communities, eroding their due process rights (Ludwig & Mullainathan 2021; Ensign et al. 2018; Lum & Isaac 2016). Research has demonstrated biases in facial recognition technology against specific groups, such as people of color or women (Palmer 2023; Gentzel 2021; Levendovski 2018). Furthermore, biased AI perpetuates societal discrimination, potentially exacerbating pre-existing social and political tensions (Barocas, Hardt & Narayanan 2023; Hoadley & Sayler 2020). Hence, when facial recognition technology misidentifies individuals based on their race or ethnicity, it may undermine trust not just in law enforcement but also in the technology itself.

Additionally, an AI algorithm biased against certain activities or behaviors might overlook potential threats or generate false positives, carrying significant risks. Biases in AI can erode trust in national security institutions and diminish public support. Errors arising from biased algorithms can compromise security effectiveness by overlooking threats or mistakenly targeting innocent individuals (Raji & Buolamwini 2019; McKendrick 2019). In a rapidly evolving landscape, proactive measures to address these biases sustain the integrity and effectiveness of national security strategies. As we navigate this complex landscape, the imperative arises to guide AI solutions that enhance security and uphold the values upon which societies stand. Therefore, ensuring that AI employed for national security lacks bias and undergoes ethical development and implementation becomes imperative.

CASE STUDIES OF RESPONSIBLE AI: STRATEGIES AND PRINCIPLES IN MITIGATING AI RISKS

Emerging technologies, especially AI, have garnered attention due to their transformative potential in reshaping defense strategies (Hoadley & Sayler 2020). As demonstrated by the United States' ongoing efforts to enhance capabilities in cyber, artificial intelligence, and quantum systems (U.S. White House Office 2022), the integration of AI continues to be a significant factor in shaping this strategy. In the United States, in 2020, the Central Intelligence Agency had undertaken nearly 140 projects to utilize AI for tasks like image recognition and predictive analytics (Hoadley & Sayler 2020; Tucker 2017). By 2023, the U.S. government had invested in research and development to mitigate AI-associated risks. The government prioritized investment in the next generation of responsible AI by reaffirming eight strategies, focusing on perception, representation, learning, and reasoning (U.S. White

House Office 2023). Other strategies include developing human-AI collaboration, approaches to mitigating ethical AI risks, guaranteeing the safety and security of AI systems, cultivating communal public datasets for AI training and testing, and prioritizing international collaboration in AI research and development to tackle global challenges such as in healthcare and manufacturing (U.S. White House Office 2023; U.S. Department of Defense 2020). In particular, the Department of Defense (DoD) adopted five principles for the ethical development of AI—responsible, equitable, traceable, reliable, and governable (U.S. Department of Defense 2020).

By 2023, the DoD also introduced the foundations of defense AI systems, including the five strategic initiatives and the establishment of a generative AI task force towards responsible, strategic, and trusted AI development (U.S. Department of Defense 2023a; U.S. Department of Defense 2023b). Furthermore, the nation undertook initiatives to establish a structure for ensuring accountability, fairness, privacy, and the mitigation of bias concerning the ethical utilization of AI (U.S. White House Office 2023). Therefore, emphasizing the design phase becomes crucial for implementing safety precautions.

In 2019, the United States introduced the Algorithmic Accountability Act to enhance transparency and accountability in AI utilization, empowering the Federal Trade Commission to prompt companies to address potential biases in computer algorithms (Congress.gov 2022). While the bill encountered difficulties in passing during the 117th Congress, there's an optimistic outlook as it holds the potential for reintroduction and reconsideration in both the House and Senate chambers during the 118th Congress.

These legislative developments are part of a larger tapestry of progress the government is weaving to advance responsible AI. Notably, the Government Accountability Office crafted an accountability framework for AI within federal agencies and other entities (U.S. Government Accountability Office 2021). Likewise, several federal agencies have undertaken various other initiatives to ensure the responsible development and deployment of AI across sectors (U.S. Equal Employment Opportunity Commission 2021; U.S. Food and Drug Administration 2021; U.S. Department of Defense 2020). These actions encompass collaborations with industries, international partners, academia, and other agency departments, collectively working towards the advancement of responsible AI research and development. Therefore, the collective efforts highlight the unwavering commitment of the government to nurturing responsible AI practices.

Concurrently, other countries have proactively tackled automated systems and adopted measures to guarantee the ethical utilization of AI in matters of

national security. The European Union pioneered the General Data Protection Regulation (GDPR) to safeguard citizens' privacy rights and to establish a framework ensuring the conscientious and open handling of their personal data (European Union 2016). Correspondingly, the report from the European Union Agency for Fundamental Rights emphasizes the critical significance of upholding high-quality data and refined algorithms in the realm of AI and machine learning systems (European Union Agency for Fundamental Rights 2019). As these initiatives underscore, responsible AI employment remains paramount in the ever-evolving landscape of technology and security.

Likewise, in 2018, the Government of Canada introduced its guiding principles for the conscientious utilization of AI (Government of Canada 2019). Canada's strategy for embracing responsible AI closely resonates with its dedication to upholding human rights, inclusivity, safety, transparency, and accountability (Government of Canada 2018). These principles delineate a comprehensive framework meticulously crafted to lay down an ethical and responsible bedrock for the integration of AI across governmental domains, encompassing even national security functions.

More specifically, the principles encompass the government's commitment to a multifaceted approach that ensures the ethical use of AI. This involves evaluating the impact of AI utilization, developing and sharing approaches, promoting transparency in AI applications, providing meaningful explanations regarding AI decision-making processes, embracing openness through the sharing of source code and training data, and offering AI design training for government employees (Government of Canada 2018). In this manner, Canada's comprehensive approach to responsible AI underscores its dedication to a future where technology is intertwined with accountability and ethical considerations.

The government of Canada mitigates issues such as lack of explainability, bias, and automated decisions in conventional decision-making laws such as the Canadian Charter of Rights and Freedoms (Canadian Charter of Rights and Freedoms 1982). Likewise, the country's Treasury Board Directive on Automated Decision-Making policy requires federal institutions to ensure the responsible use of automated decision systems and AI (Treasury Board of Canada Secretariat 2019). Thus, safety and security are paramount, highlighting the necessity to safeguard both individuals and collective interests.

Likewise, accountability forms the cornerstone, holding government agencies and institutions responsible for AI applications' ethical and just outcomes. The guiding principles outlined in Canada's framework emphasize the values

that underpin AI implementation within governmental operations and services. In Canada, governmental bodies and private enterprises have acknowledged the necessity for standardized frameworks that guide the creation and implementation of AI (Martin-Bariteau & Scassa 2021). Therefore, they underscore the imperative of inclusiveness, ensuring that AI technologies serve a diverse array of citizens without bias.

The principles resonate as a blueprint for fostering ethical AI integration across various governmental functions, including national security. This comprehensive approach aligns AI deployment with respect for human rights, inclusiveness, safety, transparency, and accountability, setting a precedent for AI's responsible and ethical utilization within government contexts (Government of Canada 2018). Thus, by adhering to these principles, the government of Canada acknowledges the importance of aligning AI technologies with ethical considerations, which fosters responsible AI use and ensures just national security efforts.

CHALLENGES OF MITIGATING BIAS IN AI

The regulation of AI and mitigating bias presents intricate challenges for nation-states as governance traverses a wide array of legal domains and jurisdictions that span diverse sectors, including human rights and health (Martin-Bariteau & Scassa 2021). One significant hurdle is the detection of biases embedded within AI algorithms. Biases, often concealed within data or algorithms, can be elusive due to their unintentional nature. Unearthing these biases requires a comprehensive grasp of the data and algorithms in use to identify potential disparities. Thus, a thorough evaluation is crucial to ascertain whether responsible use of AI guidelines comprehensively tackles the intricate and potentially perilous ramifications of AI integration within this domain.

Subsequently, the process of addressing these biases once identified presents another challenge. Rectifying biases may necessitate substantial adjustments to algorithms, posing implementation challenges without jeopardizing the accuracy of the AI system. Moreover, the effort to remedy biases may demand access to more inclusive and diverse data, which is often arduous. Developing training data that genuinely represents all demographic groups can be intricate. Ethical considerations are paramount in mitigating biases in AI. Addressing biases might entail a trade-off, potentially sacrificing accuracy and impacting national security. Striking a balance between bias mitigation and accuracy maintenance is crucial.

Challenges inherent in addressing AI bias encompass the subtleties of bias identification and the intricate process of bias rectification (Hardt, Price & Srebo

2016; Rastogi, Agrawal & Ajai 2015). Detecting bias demands a deep understanding of data and algorithmic interplay (Dressel & Farid 2018). Thus, rectifying such biases requires access to comprehensive and diverse data and substantial algorithmic adjustments, a challenging amalgamation to execute.

The utilization of biased data to educate AI algorithms holds the potential to raise concerns pertaining to privacy, human rights implications, and aspects related to the protection of consumer interests—each of these concerns falls under the jurisdiction of distinct legislative agencies (Martin-Bariteau & Scassa 2021). Thus, while principles establish a bedrock for fostering the responsible implementation of AI in national security, in this multifaceted landscape, a holistic approach to AI governance is imperative for upholding ethics and accountability.

Biases, often concealed within data or algorithms, can be elusive due to their unintentional nature. Unearthing these biases requires a comprehensive grasp of the data and algorithms in use to identify potential disparities. Thus, a thorough evaluation is crucial to ascertain whether responsible use of AI guidelines comprehensively tackles the intricate and potentially perilous ramifications of AI integration within this domain.

SOLUTIONS TO MITIGATING BIAS IN AI FOR NATIONAL SECURITY

Developing more robust regulatory frameworks to accommodate the evolution of the technology system includes the emergence of new AI data models, increased transparency, and fostering greater collaboration while preserving national security. For instance, one potential solution to mitigate bias in AI for national security purposes is to use diverse data sets (Caliskan, Bryson & Narayanan 2017; Raji & Buolamwini 2019). Mitigating sampling bias and ensuring that the algorithm makes decisions based on accurate and unbiased information require using diverse and representative data sets to train AI algorithms. Administrating a varied training data set ensures that information equitably represents all groups of individuals.

Likewise, increasing algorithmic transparency and accountability measures is a crucial solution. Requiring government agencies to publicly disclose their use of AI algorithms and the data sources used to train them contributes to establishing

reliance and assurance in AI implementation, guaranteeing that these technologies are employed in manners aligned with the public welfare. Algorithmic transparency facilitates bias detection and rectification (Kossow, Windwehr & Jenkins 2021; Dressel & Farid 2018; Lepri et al. 2017). Increasing transparency in decision-making algorithms helps identify and address biases in the system design or the data sets used to train the algorithm. Thus, this transparency can foster public trust in the decision-making process and lead to more effective bias correction.

Nevertheless, to balance transparency and security in a potentially zero-sum scenario, it is important to reveal enough to address jurisdictional security issues while withholding certain AI algorithm actions for national security reasons. This precaution is necessary to protect against potential counterintelligence adversaries. By adopting a preemptive strategy toward the ethical utilization of AI in national security, we can secure the application of these technologies in a manner that serves the greater good of society, all while mitigating the risk of adverse outcomes or unintended repercussions.

Moreover, fostering increased collaboration between government agencies and the private sector could become imperative to ensure that AI technologies align with paramount practices and ethical directives. By incorporating these supplementary measures, we can strengthen the assurance that AI technologies adhere to national principles and human rights standards, concurrently optimizing the advantages these innovations bring to the increase of national security.

In addition, creating independent regulatory oversight bodies and redress mechanisms for those adversely affected by its use can provide clear guidelines for using AI in national security and penalties for non-compliance. Comprehensive regulation is essential to manage AI's evolving landscape (Boden et al. 2017). Thus, establishing an independent regulatory agency for AI ensures that these technologies adhere to nation-state values and human rights in their usage. Oversight mechanisms and redress avenues can ensure compliance and accountability, bolstering ethical AI use in national security (Caplan et al. 2018). Therefore, human oversight can identify and mitigate biases the algorithm might miss and consider ethical considerations comprehensively.

Developmental diversity presents another solution to AI bias. Promoting diversity within the development and testing teams can yield a favorable outcome by decreasing or preempting bias. Diverse teams, composed of ethicists, data scientists, and regulatory experts, can collaboratively address bias (Holstein et al. 2019). Likewise, a diverse group can include machine learning engineers, subject matter experts, human factor specialists, diversity and inclusion professionals,

lawyers, social scientists, linguists, and privacy and security experts. Thus, collaboration between diverse disciplines can address potential biases in AI and safeguard countries' national security.

Another potential solution is to reprogram existing AI tools by conducting an algorithmic adjustment that corrects for bias by re-weighting certain data points, retraining data to remove biases, or adjusting the thresholds for certain decision-making criteria (Sun, Nasraoui, and Shafto 2020; Dressel & Farid 2018; Hardt, Price & Srebo 2016; Kamishima et al. 2012) or incorporating counterfactual examples into the training data (Guidotti 2022; Thiagarajan et al. 2022; Wachter, Mittelstadt & Floridi 2017). Diversely, transfer learning involves repurposing a pre-trained AI model for a new task or domain, effectively mitigating biases in a different context (Hosna et al. 2022; Pan & Yang 2010). While re-programming existing AI tools can demand significant time and adjustments to the algorithm and the data sets used for training (Larkin et al. 2016), a proactive approach can ensure that national security efforts remain effective, ethical, and inclusive.

It is essential to acknowledge that no one-size-fits-all solution exists to mitigate bias in AI for national security purposes. Incorporating varied data sources enhances algorithmic fairness and leads to a more representative model. (...) However, a proactive approach to identifying and addressing AI bias contributes to practical, ethical, and inclusive national security efforts.

It is essential to acknowledge that no one-size-fits-all solution exists to mitigate bias in AI for national security purposes. Incorporating varied data sources enhances algorithmic fairness and leads to a more representative model (Barocas, Hardt & Narayanan 2023). The specific approach adopted depends on the algorithm's nature and context. However, a proactive approach to identifying and addressing AI bias contributes to practical, ethical, and inclusive national security efforts.

CONCLUSION

The future of AI in national security is vast and holds many opportunities, as it enhances decision-making efficiency and threat anticipation while raising concerns about cyber vulnerabilities (Laupman, Schippers & Papaléo Gagliardi

2022; U.S. Government Accountability Office 2022). Navigating AI's evolving role in national security will be crucial for harnessing its benefits while ensuring adherence to ethical standards.

AI holds the potential to revolutionize national security operations, facilitating rapid decision-making and mitigating the risk of human error. Recognizing the potential influence of biases on decisions and outcomes and effectively managing these biases to achieve impartial and equitable results remain crucial. Moreover, AI's capability to preemptively identify potential threats before they materialize further underscores its significance as an invaluable tool in countering terrorism and cybercrime. Therefore, it is essential to approach its use cautiously and ensure proper measures are in place to reap AI's benefits while providing nations with safety and security.

However, there are also potential risks associated with increased reliance on AI in national security. One significant concern arises from the potential for malicious actors to hack or manipulate AI systems, thereby leading to the dissemination of sensitive national security information. Moreover, the growing dependence on AI could result in job losses in the national security sector, as nation-states could rely on AI systems to perform many tasks previously handled by humans. Therefore, it is crucial to ensure responsible and ethical AI use and to strike a balance between using AI as assistance rather than replacement.

Another risk involves the emergence of "killer robots" or Lethal Autonomous Weapons Systems (LAWS), AI-powered weapons capable of identifying and attacking targets without human intervention (Khan, Imam & Azam 2021; Elliott 2019). The creation of these weapons gives rise to ethical issues, sparking continuous discussions regarding their acceptability. This moral dilemma necessitates international discourse and collaboration (U.S. White House Office 2022, 2023). The advancement of automated systems, exemplified by Lethal Automated Weapons Systems, can potentially eradicate human errors from warfare, including issues like battle fatigue or post-traumatic stress disorder (PTSD). This assumption stems from the belief

AI holds the potential to revolutionize national security operations, facilitating rapid decision-making and mitigating the risk of human error. Recognizing the potential influence of biases on decisions and outcomes and effectively managing these biases to achieve impartial and equitable results remain crucial.

that machines are less prone to errors. Nonetheless, heavy reliance on automated systems can lead to automation bias, a tendency to believe that these systems are flawless. Hence, it is essential to appreciate the influence of human psychology during weapons testing and certification processes.

As we progress toward an increasingly AI-driven world, it becomes imperative to contemplate AI's role in the future of national security. While AI holds the promise of enhancing national security endeavors, it also introduces potential risks, including errors, discrimination, and privacy concerns. Consequently, striking a balance between AI's advantages and risks, while ensuring its responsible application in national security, emerges as a pivotal concern. Vigilance remains essential, and the ongoing exploration of strategies to address AI biases and promote its responsible use in national security is crucial. Both individuals and organizations bear the responsibility to advocate for the responsible utilization of AI and to champion the development of ethical and impartial AI systems for the benefit of society. ■

References

Barocas, Solon, Moritz Hardt & Arvind Narayanan. 2023. *Fairness and Machine Learning*. Massachusetts: The MIT Press.

Barocas, Solon & Andrew D. Selbst. 2016. "Big Data's Disparate Impact". *California Law Review* 104 (3): 671-732. <https://www.jstor.org/stable/24758720>.

Belenguer, Lorenzo. 2022. "AI Bias: Exploring

Discriminatory Algorithmic Decision-making Models and the Application of Possible Machine-centric Solutions Adapted from the Pharmaceutical Industry". *AI Ethics* 2 (4): 771-787. <https://doi.org/10.1007/s43681-022-00138-8>.

Bellman, Richard. 1978. *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser Publishing Company.

- Boden, Margaret, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby & Alan Winfield. 2017. "Principles of Robotics: Regulating Robots in the Real World". *Connection Science* 29 (2): 124-129. <https://doi.org/10.1080/09540091.2016.1271400>.
- Caliskan, Aylin, Joanna J. Bryson & Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases". *Science* 356 (6334): 183-186. <https://doi.org/10.1126/science.aal4230>.
- Canadian Charter of Rights and Freedoms. 1982. "Part 1 of the Constitution Act, 1982, being Schedule B to the Canada Act 1982 (UK), c 11". <https://www.justice.gc.ca/eng/csj-sjc/rfc-dlc/crf-ccdl/pdf/charter-poster.pdf>.
- Caplan, Robyn, Joan Donovan, Lauren Hanson & Jeanna Matthews. 2018. "Algorithmic Accountability: A Primer". *Congressional Progressive Caucus: How Algorithms Perpetuate Racial Bias and Inequality*. Washington, DC. https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf.
- Cavelty, Myriam Dunn & Andreas Wenger. 2020. "Cyber Security Meets Security Politics: Complex Technology, Fragmented Politics, and Networked Science". *Contemporary Security Policy* 41 (1): 5-32. <https://doi.org/10.1080/13523260.2019.1678855>.
- Charniak, Eugene & Drew McDermott. 1985. *Introduction to Artificial Intelligence*. Massachusetts: Addison Wesley.
- Chouldechova, Alexandra 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". *Big Data* 5 (2): 153-163. <http://doi.org/10.1089/big.2016.0047>.
- Congress.gov. 2022. *S.3572 - 117th Congress (2021-2022): Algorithmic Accountability Act of 2022*. <https://www.congress.gov/bill/117th-congress/senate-bill/3572>.
- Défenseur des droits and Commission Nationale Informatique & Libertés. 2020. *Algorithms: Preventing Automated Discrimination*. Défenseur des droits (Paris, France). https://www.defenseurdesdroits.fr/sites/default/files/atoms/files/836200280_ddd_synthalagos_access.pdf.
- Dorton, Stephen L., Samantha B. Harper & Kelly J. Neville. 2022. "Adaptations to Trust Incidents with Artificial Intelligence". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66 (1): 95-99. <https://doi.org/10.1177/1071181322661146>.
- Dressel, Julia & Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism". *Science Advances* 4 (1): 1-5. <https://doi.org/10.1126/sciadv.aao555>. <https://doi.org/10.1126/sciadv.aao5580>.
- Elliott, Anthony. 2019. "Automated Mobilities: From Weaponized Drones to Killer Bots". *Journal of Sociology* 55 (1): 20-36. <https://doi.org/10.1177/1440783318811777>.
- Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger & Suresh Venkatasubramanian. 2018. "Runaway Feedback Loops in Predictive Policing". Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.48550/arXiv.1706.09847>.
- European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). O.J., L. 119/1.
- European Union Agency for Fundamental Rights. 2019. *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights*. Publication Office of the European Union (Vienna, Austria). https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf.
- Evans, Jonathan. 2007. *Hypothetical Thinking: Dual Processes in Reasoning and Judgement Essays in Cognitive Psychology*. New York: Taylor & Francis Routledge.
- Fleischmann, Marvin, Miglena Amirpur, Alexander Benlian & Thomas Hess. 2014. "Cognitive Biases in Information Systems Research: a Scientometric Analysis". In *ECIS 2014 Proceedings*.
- Friedman, Batya & Helen Nissenbaum. 1996. "Bias in Computer Systems". *ACM Transactions on Information Systems* 14 (3): 330-347.
- Gentzel, Michael. 2021. "Biased Face Recognition

- Technology Used by Government: A Problem for Liberal Democracy". *Philos Technol* 34 (4): 1639-1663. <https://doi.org/10.1007/s13347-021-00478-z>.
- Gibert, Daniel, Carles Mateu & Jordi Planes. 2020. "The Rise of Machine Learning for Detection and Classification of Malware: Research Developments, Trends and Challenges". *Journal of Network and Computer Applications* 153 (102526). <https://doi.org/https://doi.org/10.1016/j.jnca.2019.102526>.
- Government of Canada. 2018. "Responsible Use of Artificial Intelligence (AI): Guiding Principles". <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1>.
- Government of Canada. 2019. "Ensuring Responsible Use of Artificial Intelligence to Improve Government Services for Canadians". March 4, 2019. <https://www.canada.ca/en/treasury-board-secretariat/news/2019/03/ensuring-responsible-use-of-artificial-intelligence-to-improve-government-services-for-canadians.html>.
- Guidotti, Riccardo. 2022. "Counterfactual Explanations and how to Find them: Literature Review and Benchmarking". *Data Mining and Knowledge Discovery*. <https://doi.org/10.1007/s10618-022-00831-6>.
- Hardt, Moritz, Eric Price & Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning". Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.
- Hoadley, Daniel S. & Kelley M. Saylor. 2020. *Artificial Intelligence and National Security*. Report No. R45178. Congressional Research Service. <https://apps.dtic.mil/sti/pdfs/AD1170086.pdf>.
- Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík & Hanna Wallach. 2019. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Glasgow, Scotland, UK. ACM, New York, NY. <https://doi.org/10.1145/3290605.3300830>.
- Hosna, Asmaul, Ethel Merry, Jigme Gyalmo, Zulfikar Alom, Zeyar Aung & Mohammad Abdul Azim. 2022. "Transfer Learning: A Friendly Introduction". *Journal of Big Data* 9 (102): 1-19. <https://doi.org/10.1186/s40537-022-00652-w>.
- Kahneman, Daniel. 2011. *Thinking, Fast, and Slow*. New York: Farrar, Straus and Giroux.
- Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh & Jun Sakuma. 2012. "Fairness-Aware Classifier with Prejudice Remover Regularizer". Joint European Conference on Machine Learning and Knowledge Discovery in Databases. https://doi.org/10.1007/978-3-642-33486-3_3.
- Khan, Ahmad, Irteza Imam & Adeela Azam. 2021. "Role of Artificial Intelligence in Defence Strategy". *Strategic Studies* 41 (1): 19-40. <https://www.jstor.org/stable/10.2307/48732266>.
- Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. London: Sage.
- Kossow, Niklas, Svea Windwehr & Matthew Jenkins. 2021. "Algorithmic Transparency and Accountability". *Transparency International*. <http://www.jstor.org/stable/resrep30838>.
- Kurzweil, Ray. 1990. *The Age of Intelligent Machines*. Cambridge, Mass.: MIT Press.
- Larkin, Nathan, Andrew Short, Zengxi Pan & Stephen van Duin. 2016. "Automatic Program Generation for Welding Robots from CAD". IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Canada. <https://ieeexplore.ieee.org/document/7576827>.
- Laupman, Clarisse, Laurianne-Marie Schippers & Marilia Papaléo Gagliardi. 2022. "Biased Algorithms and the Discrimination upon Immigration Policy". In *Law and Artificial Intelligence. Information Technology and Law Series*, 187-204. The Hague: T.M.C. Asser Press.
- Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland & Patrick Vinck. 2017. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes". *Philosophy & Technology* 31: 611-627. <https://doi.org/10.1007/s13347-017-0279-x>.
- Levendovski, Amanda. 2018. "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem". *Washington Law Review* 93 (2): 579-630. <https://digitalcommons.law.uw.edu/cgi/>

viewcontent.cgi?article=5042&context=wlr.

Ludwig, Jens & Sendhil Mullainathan. 2021. "Fragile Algorithms and Fallible Decision-makers: Lessons from the Justice System". *The Journal of Economic Perspectives* 35 (4): 71-96.

Lum, Kristian & William Isaac. 2016. "To Predict and Serve?" *Significance* 13 (5): 14-19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>.

Martin-Bariteau, Florian & Teresa Scassa. 2021. *Artificial Intelligence and the Law in Canada*. Toronto: LexisNexis Canada.

McKendrick, Kathleen. 2019. "Artificial Intelligence Prediction and Counterterrorism". London: The Royal Institute of International Affairs Chatham House.

MIT Technology Review Insights. 2022. "Building a Better Society with Better AI". *MIT Technology Review Insights*, 7 June 2022". <https://www.technologyreview.com/2022/06/07/1053031/building-a-better-society-with-better-ai/>.

Nilsson, Nils J. 1998. *Artificial Intelligence: A New Synthesis*. San Francisco: Morgan Kaufman.

Osoba, Osonde & William Welser IV. 2017. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. RAND Corporation (Santa Monica, CA). https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

Palmer, Emily. 2023. "Artificial Intelligence Led to a False Arrest When She Was 8 Months Pregnant". *People*, September 18, 2023, 50-51.

Pan, Sinno Jialin & Qiang Yang. 2010. "A Survey on Transfer Learning". *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345-1359. <https://doi.org/10.1109/TKDE.2009.191>.

Peralta, Antonio F., Matteo Neri, János Kertész & Gerardo Iñiguez. 2021. "The effect of Algorithmic Bias and Network Structure on Coexistence, Consensus, and Polarization of Opinions". Preprint, submitted in 2021. <https://arxiv.org/abs/2105.07703>.

Raji, Inioluwa Deborah & Joy Buolamwini. 2019. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products". Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.

<https://doi.org/10.1145/3306618.3314244>.

Rastogi, Gunjan, Ritesh Agrawal & Ajai. 2015. "Bias Corrections of CartoDEM Using ICESat-GLAS Data in Hilly Regions". *GIScience & Remote Sensing* 52 (5): 571-585. <https://doi.org/10.1080/15481603.2015.1060923>.

Russell, Stuart J. & Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, N.J.: Prentice Hall.

Sanclemente, Gaudys L. 2021. "Reliability: Understanding Cognitive Human Bias in Artificial Intelligence for National Security and Intelligence Analysis". *Security Journal*. <https://doi.org/10.1057/s41284-021-00321-2>.

Schmidt, Eric. 2022. "AI, Great Power Competition & National Security". *Daedalus* 151 (2): 288-298. <https://www.jstor.org/stable/48662042>.

Searle, John. 1984. *Minds, Brains and Science*. Cambridge: Harvard University Press.

Sun, Wenlong, Olfa Nasraoui & Patrick Shafto. 2020. "Evolution and Impact of Bias in Human and Machine Learning Algorithm Interaction". *PLOS ONE* 15 (8): e0235502. <https://doi.org/10.1371/journal.pone.0235502>.

Thiagarajan, Jayaraman J., Kowshik Thopalli, Deepta Rajan & Pavan Turaga. 2022. "Training Calibration-based Counterfactual Explainers for Deep Learning Models in Medical Image Analysis". *Scientific Reports* 12 (597). <https://doi.org/10.1038/s41598-021-04529-5>.

Treasury Board of Canada Secretariat. 2019. "Directive on Automated Decision-making (5 February 2019)". Online: Government of Canada. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592§ion=html>.

Tucker, Patrick. 2017. "What the CIA's Tech Director Wants from AI". *Defense One*. <http://www.defenseone.com/technology/2017/09/cia-technology-director-artificial-intelligence/140801/>.

U.S. Department of Defense. 2020. "DOD Adopts Ethical Principles for Artificial Intelligence". February 24, 2020. <https://www.defense.gov/News/Releases/release/article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>.

- U.S. Department of Defense. 2023a. "DOD Announces Establishment of Generative AI Task Force". August 10, 2023. <https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/>.
- U.S. Department of Defense. 2023b. "Chief Digital & Artificial Intelligence Office Celebrates First Year". July 19, 2023. <https://www.defense.gov/News/Releases/Release/Article/3464012/chief-digital-artificial-intelligence-office-celebrates-first-year/>.
- U.S. Equal Employment Opportunity Commission. 2021. "EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness". October 28, 2021. <https://www.eeoc.gov/newsroom/eec-launches-initiative-artificial-intelligence-and-algorithmic-fairness>.
- U.S. Food and Drug Administration. 2021. "Good Machine Learning Practice for Medical Device Development: Guiding Principles". October 2021. <https://www.fda.gov/media/153486/download>.
- U.S. Government Accountability Office. 2021. *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*. Report No. GAO-21-519SP. U.S. Government Printing Office. <https://www.gao.gov/assets/gao-21-519sp.pdf>.
- U.S. Government Accountability Office. 2022. "How Artificial Intelligence is Transforming National Security". GAO. April 19, 2022. <https://www.gao.gov/blog/how-artificial-intelligence-transforming-national-security>.
- U.S. White House Office. 2022. *National Security Strategy*. The White House (Washington, DC). <https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf>.
- U.S. White House Office. 2023. "National Artificial Intelligence Research and Development Strategic Plan 2023 Update". Washington, D.C.: The White House. <https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>.
- UN General Assembly. 2020. *Seventy-fifth Session: Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance*. United Nations. <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/304/54/PDF/N2030454.pdf?OpenElement>.
- Wachter, Sandra, Brent Mittelstadt & Luciano Floridi. 2017. "Transparent, Explainable, and Accountable AI for Robotics". *Science Robotics* 2 (6). <https://ssrn.com/abstract=3011890>.
- Winston, Patrick Henry. 1992. *Artificial Intelligence*. Massachusetts: Addison-Wesley.
- Como citar:** Sanclemente, Gaudys L. 2023. "Ferramentas digitais: salvaguardando a segurança nacional, a cibersegurança e o tendenciosismo na IA". *CEBRI-Revista* Ano 2, Número 7: 137-155.
- To cite this work:** Sanclemente, Gaudys L. 2023. "Digital Tools: Safeguarding National Security, Cybersecurity, and AI Bias." *CEBRI-Journal* Year 2, No. 7: 137-155.
- DOI:** <https://doi.org/10.54827/issn2764-7897.cebri2023.07.03.07.137-155.en>

Recebido: 25 de agosto de 2023

Aceito para publicação: 18 de setembro de 2023

Copyright © 2023 CEBRI-Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original article is properly cited.