SIPRI Background Paper
December 2024

BIAS IN MILITARY ARTIFICIAL INTELLIGENCE

ALEXANDER BLANCHARD AND LAURA BRUUN*

SUMMARY

● **To support states involved in the policy debate on military artificial intelligence (AI), this background paper provides a deeper examination of the issue of bias in military AI. Three insights arise.**

First, policymakers could usefully develop an account of bias in military AI that captures shared concern around unfairness. If so, 'bias in military AI' might be taken to refer to the systemically skewed performance of a military AI system that leads to unjustifiably different behaviours—which may perpetuate or exacerbate harmful or discriminatory outcomes—depending on such social characteristics as race, gender and class.

Second, among the many sources of bias in military AI, three broad categories are prominent: bias in society; bias in data processing and algorithm development; and bias in use.

Third, bias in military AI can have various humanitarian consequences depending on context and use. These range from misidentifying people and objects in targeting decisions to generating flawed assessments of humanitarian needs.

Artificial intelligence (AI) systems are biased. In various ways and degrees, they reflect and reproduce existing human biases around, for example, gender, race, age or ethnicity.[1] States have increasingly expressed concerns about the presence of such bias in their intergovernmental discussions on the governance of military AI, such as in the policy debate on autonomous weapon systems (AWS).[2] Yet bias in military AI is rarely discussed in depth nor is it reflected in the outcome documents of these meetings.[3] This contrasts with the civilian domain, where multinational efforts are well under way to address bias in AI.[4]

The issue of bias in the military domain is not unique to AI. The presence of bias, especially in targeting decisions, has long been studied and discussed.[5] However, given the potentially transformative effects of AI, and given that it has the potential to exacerbate bias, a deeper understanding of the challenges and risks posed by bias in military AI is needed.[6]

This background paper is intended as a common reference document for policymakers in intergovernmental discussions on military AI. It explores

[1] E.g. Crawford, K., *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press: New Haven, CT, 2021); Leavy, S., O'Sullivan, B. and Siapera, E., 'Data, power and bias in artificial intelligence', arXiv 2008.07341, 28 July 2020; and Buolamwini, J. and Gebru, T., 'Gender shades: Intersectional accuracy disparities in commercial gender classification', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (MLResearch Press: 2018).

[2] E.g. Certain Conventional Weapons (CCW) Convention, Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS), Report of the 2021 session, CCW/GGE.1/2021/3, 22 Feb. 2022; CCW Convention, GGE on LAWS, 'Addressing bias in autonomous weapons', Working paper submitted by Austria, Belgium, Canada, Costa Rica, Germany, Ireland, Mexico, Panama and Uruguay, CCW/GGE.1/2024/WP.5, 8 Mar. 2024; United Nations, General Assembly, 'Lethal autonomous weapons systems', Report of the Secretary-General, A/79/88, 1 July 2024; and United Nations, General Assembly, First Committee, 'Artificial intelligence in the military domain and its implications for international peace and security', Draft resolution A/C.1/79/L.43, 16 Oct. 2024.

[3] Mohan, S. and Cho, D., 'Gender and lethal autonomous weapons', UN Institute for Disarmament Research (UNIDIR), [Aug. 2024].

[4] E.g. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), *Official Journal of the European Union L*, 12 July 2024, articles 10, 15.

[5] On the example of assuming men of certain age and in a specific location as lawful targets see Shoker, S., *Military-age Males in Counterinsurgency and Drone Warfare* (Palgrave Macmillan: London, 2021); Heller, K. J., '"One hell of a killing machine": Signature strikes and international law', *Journal of International Criminal Justice*, vol. 11, no. 1 (Mar. 2013); and International Committee of the Red Cross (ICRC) et al., *International Humanitarian Law and a Gender Perspective in the Planning and Conduct of Military Operations* (ICRC: Geneva, 2024).

[6] Certain Conventional Weapons Convention, CCW/GGE.1/2024/WP.5 (note 2); Chandler, K., *Does Military AI Have Gender?* (UN Institute for Disarmament Research: Geneva, 2021); and Bhila, I., 'Putting algorithmic bias on top of the agenda in the discussions on autonomous weapons systems', *Digital War*, vol. 5 (2024).

* SIPRI and the authors are grateful for the generous financial support for this project received from the German Federal Foreign Office.

Box 1. Military artificial intelligence

Artificial intelligence (AI) has various military applications, from logistical support to cyberwarfare.[a] This background paper focuses on uses that could have humanitarian implications.

These uses include the integration of AI into the targeting process of, for example, autonomous weapon systems (AWS), which are commonly defined as weapon systems that, once activated, can select and engage targets without human intervention.[b] These uses also include AI to assist consequential decision-making at the tactical, operational and strategic levels, such as in an AI-enabled decision-support system (DSS) used to collect and analyse battlefield information for operational intelligence assessments.[c] The paper also considers the use of AI for humanitarian services, including forecasting instability and conflict, and for aid allocation during disaster relief.[d]

A variety of different computing techniques fall under the category of 'AI'. Unless otherwise specified, this paper focuses on contemporary statistical learning methods in AI (e.g. machine learning and deep learning). This is an approach to AI development that entails pattern detection via supervised learning, reinforcement learning or unsupervised learning.[e]

[a] Geiß, R. and Lahmann, H. (eds), *Research Handbook on Warfare and Artificial Intelligence* (Edward Elgar: London, 2024).

[b] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), pp. 24–27; and International Committee of the Red Cross (ICRC), 'ICRC position on autonomous weapon systems', 12 May 2021.

[c] Klonowska, K., 'Article 36: Review of AI decision-support systems and other emerging technologies of warfare', eds T. D. Gill et al., *Yearbook of International Humanitarian Law 2020* (Asser Press: The Hague, 2021); and Nadibaidze, A., Bode, I. and Zhang, Q., *AI in Military Decision Support Systems: A Review of Developments and Debates* (University of Southern Denmark, Center for War Studies: Odense, Nov. 2024).

[d] Beduschi, A., 'Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks', *International Review of the Red Cross*, no. 919 (June 2022).

[e] Boulanin and Verbruggen (note b), p. 16.

what 'bias in military AI' refers to (section I), outlines sources of bias (section II) and details some of the potential humanitarian consequences (section III). The paper is based on a review of relevant literature and on consultations and interviews with experts from academia, civil society, governments and industry. While AI has a wide range of military applications, this paper focuses on the use of military AI to inform decisions that have humanitarian implications (see box 1).

I. What does 'bias in military AI' refer to?

To fruitfully discuss, identify and respond to the challenges presented by bias in military AI, states would benefit from a shared understanding of the issue. However, there is currently no consensus on a definition of bias, either in the expert literature on bias in AI generally or in the intergovernmental debate on military AI.

Disagreement in the expert literature

In the expert literature, bias is understood in various neutral or value-laden ways.[7]

[7] E.g. Coeckelbergh, M., *AI Ethics* (MIT Press: Cambridge, MA, 2020); Silberg, J. and Manyika, J., 'Notes from the AI frontier: Tackling bias in AI (and in humans)', McKinsey Global Institute, June 2019; Danks, D. and London, A. J., 'Algorithmic bias in autonomous systems', *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)* (IJCAI: Melbourne, 2017); Tsamados, A. et al., 'The ethics of algorithms: Key problems and solutions', *AI & Society*, vol. 37 (2022); and Crawford (note 1), pp. 123–49.

According to a neutral account, bias refers to the ways in which an AI system may be skewed towards some characteristic, environment or behaviour in its operation.[8] For example, an AI self-driving car that is trained on data from London will perform better in London. According to this account, bias is not intrinsically negative; sometimes, biasing a system towards certain features is desired in order to achieve optimal performance.

In the value-laden account, bias is taken to be inherently negative, referring to a system's inclination to treat certain groups of people in a way that is considered unfair.[9] Here, bias is often described as the systemic skewing of a system against individuals or groups of people—often according to characteristics such as gender, age or ethnicity—meaning that the performance of the system is worse for particular demographics. Referring to bias as systemic unfairness is meant to indicate that inequitable treatment of or harm to people is not the result of incidental technical errors, but rather reflects the inequitable practices inherent to society.[10]

To a large extent, the lack of consensus in the expert literature results from diverging motivations for addressing bias. The neutral account is motivated primarily by improving system reliability and optimal performance. The value-laden account is motivated primarily by moral, political and social concerns about the discriminatory effects of AI, particularly demonstrated in domains such as healthcare and law enforcement.[11] Indeed, for some, it is the 'fairness' element that distinguishes the issue of bias from issues of reliability generally.[12]

A common reference point in the policy debate: Bias as systemic unfairness

In the (still nascent) policy debate on bias in military AI, states have yet to agree on a consensus definition. This lack of a common understanding of what 'bias' refers to has been reflected in, for example, debates about whether bias is wanted or unwanted or whether it is intended or unintended.[13] This debate boils down to a lack of clarity among states about what bias is and whether or not bias is inherently negative.

However, unlike in the expert debate, states that have addressed bias in military AI in the policy debate appear to be broadly aligned in terms of their expressed motivations for raising the issue. Their shared concern is that bias in military AI can reflect and exacerbate the inequitable treatment of people

[8] Crawford (note 1); and Danks and London (note 7).

[9] Ziosi, M., Watson, D. and Floridi, L., 'A genealogical approach to algorithmic bias', *Minds and Machines*, vol. 34 (2024), p. 2; Allen, R. B., Friedman, B. and Nissenbaum, H., 'Bias in computer systems', *ACM Transactions on Information Systems*, vol. 14, no. 3 (1996); Ferrera, E., 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies', *Sci*, vol. 6, no. 1 (Mar. 2024); Coeckelbergh (note 7); Silberg and Manyika (note 7); and Buolamwini and Gebru (note 1).

[10] Noble, S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press: New York, 2018).

[11] Ziosi et al. (note 9), p. 2; Allen et al. (note 9); Ferrera (note 9); Coeckelbergh (note 7); Danks and London (note 7); and Silberg and Manyika (note 7).

[12] US Department of Commerce, National Institute of Standards and Technology (NIST), 'AI risks and trustworthiness', Trustworthy & Responsible AI Resource Center, [n.d.].

[13] Certain Conventional Weapons Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 1st session, 6 Mar. 2024, Interventions by the USA, 01:59, South Korea, 02:15, and Austria, 02:23; and Varella, L., 'Other measures to ensure compliance with IHL', *CCW Report*, 5 Sep. 2024, pp. 45–46.

based on traits such as race, gender or class, and it thus relates to an issue of fairness. For example, a number of states (e.g. Argentina, France, Palestine and Sierra Leone) have highlighted the need to address risks posed by the reliance of AWS on data sets 'that can perpetuate or amplify unintentional social biases, including gender and racial bias'.[14] Likewise, other states (e.g. Austria, Belgium, Canada, Costa Rica, Germany, Ireland, Mexico, Panama and Uruguay) have highlighted the growing documentation of 'examples of gender and racial biases in AI' and that 'data-based systems reproduce existing inequalities'.[15] Similar concerns about bias have also been expressed in national policy statements. For example, the United States Department of Defense emphasizes the 'equitable' use of AI and has made a commitment to 'take deliberate steps to minimize unintended bias in AI capabilities'.[16] Similarly, the British Ministry of Defence has highlighted that 'the risk of discriminatory outcomes resulting from algorithmic bias or skewed data sets' is a particular concern with AI-enabled military systems.[17] These accounts of bias thus largely reflect that part of the expert literature that treats bias as an issue of inequitable treatment.

To advance from this shared concern around unfairness, policymakers could usefully develop an account of bias in military AI that captures such expressed concerns. If so, 'bias in military AI' might be taken to refer to the systemically skewed performance of a military AI system that leads to unjustifiably different behaviours—which may perpetuate or exacerbate harmful or discriminatory outcomes—depending on such social characteristics as race, gender and class.

II. The sources of bias in military AI

There are many different ways in which bias can arise in the development and use of military AI. This section outlines three principal sources of bias: society, data processing and algorithm development, and use (see figure 1).[18] The role of data is important here: data, which is essential for AI systems to function, can reflect historical biases in society as well as the biased preferences of individuals and organizations that collect, process and use it.[19]

[14] Certain Conventional Weapons Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 'Roadmap towards a new protocol on autonomous weapons systems', Working paper submitted by Argentina, Costa Rica, Guatemala, Kazakhstan, Nigeria, Panama, Philippines, Sierra Leone, Palestine and Uruguay, CCW/GGE.1/2022/WP.3, 8 Aug. 2022, para. 17. See also Certain Conventional Weapons Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Working paper submitted by Bulgaria, Denmark, France, Germany, Italy, Luxembourg and Norway, CCW/GGE.1/2024/WP.3, 4 Mar. 2024.

[15] Certain Conventional Weapons Convention, CCW/GGE.1/2024/WP.5 (note 2), para. 4.

[16] US Department of Defense (DOD), 'Autonomy in weapon systems', DOD Directive no. 3000.09, 25 Jan. 2023, para. 1.2(f).

[17] British Ministry of Defence (MOD), *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence*, Policy paper (MOD: London, June 2022), p. 11.

[18] Allen et al. (note 9).

[19] Researcher on algorithmic bias, Online author interview, 14 Aug. 2024; and Software engineer, Online author interview, 21 Aug. 2024. See also Ziosi et al. (note 9); and Kostopoulos, L., *The Role of Data in Algorithmic Decision-Making: A Primer* (UN Institute for Disarmament Research: Geneva, 2019).

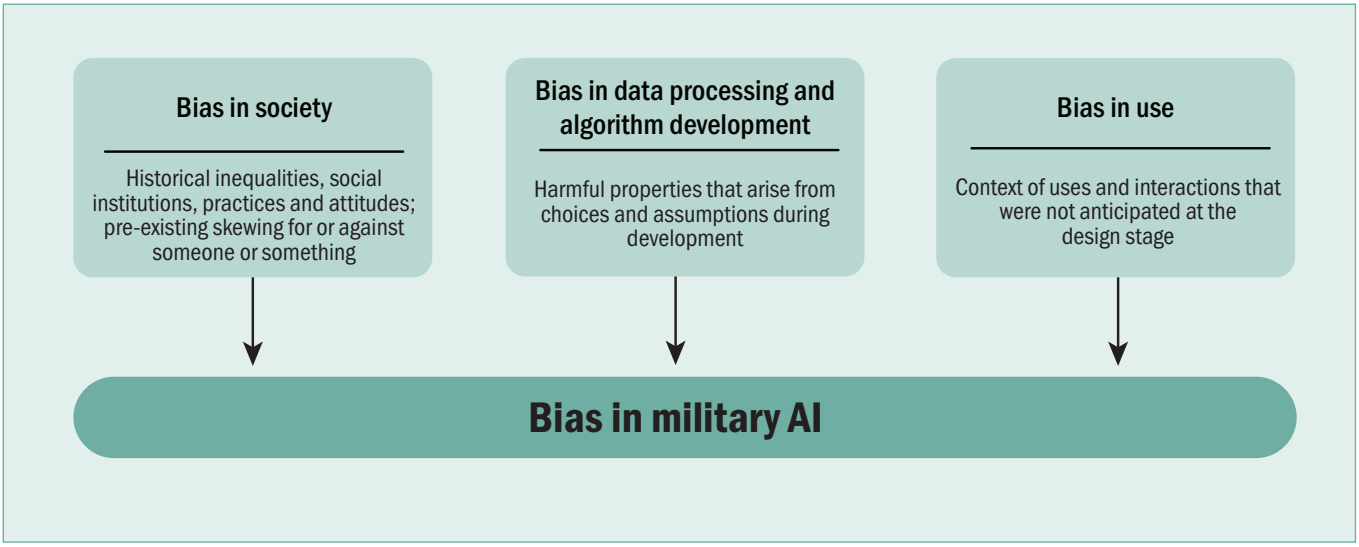

Figure 1. Sources of bias in military artificial intelligence

Bias in society

Bias in society, also described as pre-existing bias or historical bias, refers to the way in which societies have historically been skewed towards or against certain individuals or groups of people, typically along lines such as ethnicity, gender, class or ability.[20] Societal bias is reflected in all stages of the life cycle of an AI system, and so it can be seen as the ultimate source of bias in AI.

Bias in society can be introduced in military AI systems in multiple, often subtle ways. At the earliest stages, societal bias can influence choices about which military AI systems to develop and about where and against whom they should be used.[21] However, it primarily appears in the underlying training data sets, notably as a failure to capture a real-world distribution of factors (sometimes referred to as 'selection bias').[22] For example, societal bias can result in data sets that either under-represent certain populations, environments or scenarios or over-represent them.[23] This includes when data sets over-represent particular characteristics that are specific to a certain context but are then taken as universal representations. For example, Western architectural styles (e.g. of churches) could be taken as representative of all civilian objects in an AI decision-support system (DSS) used to identify objects that are protected under International Humanitarian Law (IHL);

[20] E.g. Allen et al. (note 9); Bhila (note 6); Holland, A., *Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-making* (International Committee of the Red Cross: Geneva, 2024); Suresh, H. and Guttag, J., 'A framework for understanding sources of harm throughout the machine learning life cycle', *EAAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery: New York, 2021); Joseph, S., Schultz, J. and Castan, M., *The International Covenant on Civil and Political Rights: Cases, Materials, and Commentary*, 2nd edn (Oxford University Press: Oxford, 2004), p. 735; and Dvaladze, G., *Equality and Non-Discrimination in Armed Conflict: Humanitarian and Human Rights Law in Practice* (Edward Elgar: Cheltenham, 2023), p. 118.

[21] Allen et al. (note 9).

[22] Milaninia, N., 'Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications', *International Review of the Red Cross*, no. 913 (Apr. 2020).

[23] Ferrera (note 9); Chandler (note 6); Bartoletti, I. and Xenidis, R., *Study on the Impact of Artificial Intelligence Systems, Their Potential for Promoting Equality, Including Gender Equality, and the Risks They May Cause in Relation to Non-discrimination* (Council of Europe: Strasbourg, 2023); Milaninia (note 22); and Bhila (note 6).

or the male body could be taken as representative of all body types in AI systems used to assess the physical effects of weapons on people.

Moreover, even when data sets are representative of real-world conditions, they can still contain bias. This is because fully representational data presents the world as it is, and the world itself contains bias. For example, a data set may correctly represent the historical distribution of targeted terrorist cells in a region but, if certain locations or groupings have historically been disproportionally subject to surveillance (and thus targeted at higher frequency), such historical skewing will be reflected in the data set.[24]

Indeed, AI systems are adept at identifying patterns in data distribution, including those patterns that developers are unlikely to perceive. As such, AI can reinforce existing social inequalities or stereotypes even when training data is scrubbed of sensitive characteristics. The reproduction of societal biases need not be conscious or explicit, and it may occur despite the best intentions of the developers. Indeed, the reproduction of societal bias in data sets can be taken as evidence of existing bias in wider society, its subcultures, institutions and institutional practices.[25]

Bias in data processing and algorithm development

Bias in data processing and algorithm development refers to the potentially harmful properties of military AI that arise from the choices and assumptions of the various actors who are involved in developing the AI systems.[26] It covers a broad set of activities where bias can arise, including the labelling, modelling and preprocessing of data; the design of algorithms; the learning processes; and the development of training objectives and performance metrics.

In the context of military AI, bias in data processing could, for instance, arise if programmers emphasize certain outcomes, behaviours or information over others (also referred to as reporting bias). This could lead to a skewed representation in data sets of, for example, enemy characteristics or civilian movements.[27] Meanwhile, bias in algorithm development could be introduced via proxy indicators, which are indirect ways of measuring something when a direct measurement may not be available (e.g. postcodes as proxies for income). Proxy indicators are important components of many machine learning systems. While many are harmless (e.g. using a picture of a cat as a proxy for an animal), others are more problematic. Notably, bias may arise if factors such as age, gender or race are used as proxy indicators for,

[24] On the civilian example of criminal proceedings see e.g. National Immigration Project, 'Bias in the criminal legal system', Stanford Law School, Immigrants' Rights Clinic, June 2024.

[25] Researcher on algorithmic bias (note 19); Government AI ethics and bias specialist, Online author interview, 15 Aug. 2024; and Software engineer (note 19). See also Ziosi et al. (note 9), p. 2; Garcia, M., 'Racist in the machine', *World Policy Journal*, vol. 33, no. 4 (winter 2016/17); and Crawford (note 1), p. 135.

[26] Allen et al. (note 9); Suresh and Guttag (note 20); Schelter, S. and Stoyanovich, J., 'Taming technical bias in machine learning pipelines', *Bulletin of the Technical Committee on Data Engineering*, vol. 43, no. 4 (Dec. 2020); Bartoletti and Xenidis (note 23); UN Institute for Disarmament Research (UNIDIR), *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer* (UNIDIR: Geneva, 2018); and Holland (note 20).

[27] Milaninia (note 22).

for example, combatant status. This can lead to what has been described as 'proxy discrimination'.[28]

The reproduction of bias in data processing and algorithm development can be explained by many (often unintentional) factors, including deficient workplace practices or institutional blind spots. This can result from a lack of diversity among teams of developers in terms of, for example, gender, ethnicity, economic status and motivations to develop AI systems.[29]

Bias in use

Bias in use refers to bias that arises during the deployment of an AI system; that is, changes in the behaviour of the system that occur over time due to new contexts, uses or interactions that were not anticipated during its initial design and training.[30]

This includes transfer-context bias, which refers to bias resulting from a mismatch between a model's training and its environment of use.[31] Thus, transfer-context bias may arise when the system is deployed outside the intended context or because developers did not account for all intended uses of the system.[32] It can also arise due to a biased understanding of the intended environment of use that may fail to account for, for example, local practices or physical disabilities of those in the environment.[33] Transfer-context bias can degrade the performance of an AI system. For instance, an AI-supported threat-perception tool trained on data from rural environments may be inaccurate when used in urban settings.

Bias in use may also result from human–machine interaction during deployment. For example, a military AI system that uses positive feedback loops (e.g. reinforcement learning) may adopt the preferences of individual users.[34] This may mean that algorithmic outputs come to reflect a specific user's biased preferences or choices or that latent biases in the algorithm are revealed by interactions with the user.[35] During use, the presence of bias in the system may also be compounded or reinforced by the cognitive biases of the users.[36] For instance, automation bias may lead humans to believe that AI systems are necessarily objective, thereby overly trusting (and so acting upon) their outputs.[37]

[28] Bartoletti and Xenidis (note 23); Milaninia (note 22); and Holland (note 20).

[29] Chandler (note 6); Ramsay-Jones, H., 'Intersectionality and racism', Campaign to Stop Killer Robots, *Campaigner's Kit* (Campaign to Stop Killer Robots: Geneva, 2020); and Certain Conventional Weapons Convention, CCW/GGE.1/2024/WP.5 (note 2).

[30] Allen et al. (note 9); International Committee of the Red Cross (ICRC), 'Autonomy, artificial intelligence and robotics: Technical aspects of human control', 6 June 2019; and Holland (note 20).

[31] E.g. UN Institute for Disarmament Research (note 26); and International Committee of the Red Cross (note 30).

[32] Tsamados et al. (note 7); and Allen et al. (note 9).

[33] UN Institute for Disarmament Research (note 26); and Díaz Figueroa, M. et al., 'The risks of autonomous weapons: An analysis centred on the rights of persons with disabilities', *International Review of the Red Cross*, no. 922 (Nov. 2022).

[34] International Committee of the Red Cross (note 30); Bartoletti and Xenidis (note 23); and Holland (note 20).

[35] Ferrera (note 9); and Lai, K. et al., 'Assessing risks of biases in cognitive decision support systems', 28th European Signal Processing Conference (EUSIPCO), 18–21 Jan. 2020.

[36] Holland (note 20); Milaninia (note 22); Ferrera (note 9); and UN Institute for Disarmament Research (note 26).

[37] Software engineer (note 19).

III. The humanitarian consequences of bias in military AI

The harmful outcomes of bias in AI are well-documented in the civilian domain. For example, a lack of attention to skewed data has resulted in AI recruitment tools disqualifying otherwise qualified female candidates; and facial recognition systems have failed to correctly recognize people with darker skin at the same rate as those with lighter skin.[38] In the military domain, the humanitarian consequences of bias in AI depend on the nature of the military AI application and the context. While this is a relatively underexplored topic, the following provides a non-exhaustive overview of some of the ways in which bias in military AI—notably for targeting, as well as broader applications related to, for example, humanitarian services, intelligence and surveillance—could expose certain groups of people to greater risk of harm or unfair treatment.

Misidentification of targets

Bias in AI used for targeting (e.g. AWS and AI-enabled DSS) poses risks of target misidentification. This includes instances of false positives, whereby non-threats are misidentified as threats, and instances of false negatives, whereby threats are misidentified as non-threats.

Machine learning models often rely on pattern recognition and proxy indicators. If informed by biased training data sets, they may draw incorrect conclusions, with characteristics such as race, gender or ability improperly influencing the identification and selection of targets.[39] For example, machine learning systems could infer threats based on racial and gender stereotypes. This exacerbates the risks already associated with militaries conducting operations in locations where they have a poor understanding of sociocultural context and traditions.[40]

Such misidentification, if not corrected through adequate human oversight, may lead to harmful outcomes for civilians and civilian objects. This has the potential to contravene IHL principles, notably the principle of distinction, but also provisions on adverse distinction.[41]

Disproportionate incidental civilian harm

Bias in military AI systems that are used to assess collateral damage in relation to an attack can lead to instances where systems fail to adequately account for certain contexts, people or objects. This can potentially lead to disproportionate harm being inflicted on the civilian population.

[38] Chen, Z., 'Ethics and discrimination in artificial intelligence-enabled recruitment practices', *Humanities and Social Science Communications*, vol. 10 (2023); and Buolamwini and Gebru (note 1).

[39] Díaz Figueroa et al. (note 33); Milaninia (note 22); and Bhila (note 6), p. 204.

[40] Chilcot, J., *The Report of the Iraq Inquiry: Executive Summary* (HM Stationery Office: London, 6 July 2016).

[41] The principle of distinction appears in articles 27, 41, 48, 52 of Protocol I Additional to the 1949 Geneva Conventions, and Relating to the Protection of Victims of International Armed Conflicts, opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978; and rules 1, 6, 7, 13, 14, 47 of the International Committee of the Red Cross (ICRC), Customary IHL Database,. The prohibition against adverse distinction is expressed, among other places, in Common Article 3 of the Geneva Conventions I–IV of 12 Aug. 1949; Article 9 of the Additional Protocol I; and Rule 88 of the ICRC Customary IHL Database. See Dvaladze (note 20), pp. 223–32.

This might arise if an AI-enabled DSS, used to calculate proportionality in an attack, were to rely on data sets that do not capture the complexities of civilian presence in the conflict zone.[42] For example, data sets that reflect the male body as the archetypical body could skew assessments of harm for people with other body types. This phenomenon is described in the literature as the 'one-size-fits-men' approach.[43] Equally, if people with physical disabilities are not reflected in the data sets, the system may fail to identify—and account for—people in wheelchairs, for example.[44]

Overall, this could result in people, objects and environments being inadequately protected, potentially leading to disproportionate civilian casualties and damage to civilian objects.

Disproportionate surveillance and profiling

AI-enabled surveillance and intelligence-gathering tools that exhibit bias may lead to discriminatory practices such as over-surveillance or profiling of certain groups.[45]

Reliance on biased data could reinforce or exacerbate stereotypes. For example, based on data from previous conflicts, certain demographics could be associated with insurgency or threats. This could perpetuate cycles of suspicion and intrusive practices, including inequitable monitoring and scrutiny of particular ethnic or religious groups, leading to invasions of their privacy and risks to their human rights.[46] It could also lead to pre-emptive military action based on probabilistic assessments, rather than verified intelligence.

Stigmatization of and discrimination against vulnerable populations in relief actions

If AI systems are used to support humanitarian services during armed conflict, bias in those systems could inadvertently reinforce stigmatization and discrimination and contribute to differential treatment of vulnerable populations.[47] In particular, an AI model biased towards particular population indicators may overlook communities that do not conform to those indicators, particularly in conflict areas where conventional indicators of

[42] Chandler (note 6); Ferrera (note 9); British Parliament, House of Lords, AI in Weapon Systems Committee, *Proceed with Caution: Artificial Intelligence in Weapon Systems* (House of Lords: London, 1 Dec. 2023), p. 11; Bode, I., 'Falling under the radar: The problem of algorithmic bias and military applications of AI', Humanitarian Law and Policy Blog, International Committee of the Red Cross, 14 Mar. 2024; Holland (note 20); Certain Conventional Weapons Convention, CCW/GGE.1/2024/WP.5 (note 2); and Schmitt, M. N., 'Targeting and international humanitarian law in Afghanistan', *International Law Studies*, vol. 85, no. 1 (2009), p. 311.

[43] Criado Perez, C., *Invisible Women: Exposing Data Bias in a World Designed for Men* (Vintage Books: London, 2019); and Tengroth, C. and Lindvall, K. (eds), *IHL and Gender—Swedish Experiences* (Swedish Red Cross: Stockholm, 2015), p 112.

[44] Guo, A. et al., 'Toward fairness in AI for people with disabilities: A research roadmap', *ACM SIGACCESS Accessibility and Computing*, no. 125 (Oct. 2019); and United Nations, General Assembly, Human Rights Council, 'Rights of persons with disabilities', Report of the special rapporteur on the rights of persons with disabilities, A/HRC/49/52, 28 Dec. 2021.

[45] Rowe, M. and Muit, R., 'Big data policing: Governing the machines?', eds J. McDaniel and K. Pease, *Predictive Policing and Artificial Intelligence* (Routledge: London, 2021).

[46] Blanchard, A. and Taddeo, M., 'The ethics of artificial intelligence for intelligence analysis: A review of the key challenges with recommendations', *Digital Society*, vol. 2, no. 1 (Apr. 2023).

[47] Pizzi, M., Romanoff, M. and Engelardt, T., 'AI for humanitarian action: Human rights and ethics', *International Review of the Red Cross*, no. 913 (Apr. 2020).

need may be unavailable or unreliable.[48] For instance, an AI system trained on data derived from social media or mobile phone usage might prioritize areas where such usage is high and may then misclassify populations that lack digital infrastructure as being in less need of assistance.[49]

Failures to identify the needs of vulnerable populations could lead to delayed response or resource misallocation, leaving certain populations under-supported and potentially exacerbating existing inequitable treatment.[50] It could even leave vulnerable populations excluded from relief action altogether. Such discrimination could risk violating humanitarian principles of impartiality and neutrality and could also deepen societal divides, perpetuating cycles of exclusion and hardship.

Exacerbation of the difficulty of spotting, correcting and attributing bias

While the possibility that bias motivates or influences military decision-making is not an issue novel to AI, AI adds complexity and scale to existing concerns around bias. An appeal of AI systems for militaries is that they quickly identify, process, filter and analyse large volumes of data to increase the speed of decision-making. However, decision-making at pace, compounded with the risks of automation bias, could diminish the opportunity to spot and correct bias.[51]

AI also makes it more difficult to scrutinize bias. Among the factors complicating scrutiny of AI systems are the involvement of multiple actors in generating data sets; the malleability of algorithms; a lack of transparency around data practices; and the use of proprietary systems.[52] Inscrutability of AI outputs can make it harder to assign responsibility and accountability when bias in military AI contributes to harmful outcomes.

IV. Conclusions

In order to identify and respond to the humanitarian implications of bias in military AI, policymakers need a deeper, shared understanding of the issue. This background paper, intended as a common reference document, aims to support such efforts. It makes three contributions.

First, based on states' expressed concerns about discriminatory risks from bias in military AI, 'bias in military AI' might be taken to refer to the systemically skewed performance of a military AI system that leads to unjustifiably

[48] United Nations Development Programme (UNDP), *Conflict Sensitivity and Monitoring & Evaluation Toolbox* (UNDP: New York, May 2024) ; and Organisation for Economic Co-operation and Development (OECD), *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility: Improving Learning for Results*, DAC Guidelines and Reference Series (OECD Publishing: Paris, 2012), chapter 4.

[49] Kim, K. and Boulanin, V., *Artificial Intelligence for Climate Security: Possibilities and Challenges* (SIPRI: Stockholm, 2023), pp. 6–11.

[50] Wilton Park, 'Risks when humanitarians use AI', May 2024; and Lewis, D. A., 'AI and machine learning symposium: Why detention, humanitarian services, maritime systems, and legal advice merit greater attention', Opinio Juris, 28 Mar. 2020.

[51] Bartoletti and Xenidis (note 23); Holland (note 20); Researcher on algorithmic bias (note 19); and Software engineer (note 19).

[52] Researcher on algorithmic bias (note 19); Software engineer (note 19); Bartoletti and Xenidis (note 23); Tsamados et al. (note 7); and Milaninia (note 22).

different behaviours depending on social characteristics such as ethnicity, gender, ability, age, class and religion.

Second, three broad sources of bias in military AI are prominent: bias in society; bias in data processing and algorithm development; and bias in use. Bias can thus be introduced at multiple junctures during the development and use of an AI system, often through the ways in which the underlying data are collected, processed and used.

Third, the humanitarian consequences of bias in military AI relate principally to its use in targeting, but broader applications such as surveillance and humanitarian services should not be overlooked. The presence of implicit assumptions around gender, ethnicity, ability and other sensitive characteristics in military AI systems can result in misidentification of threats and non-threats, flawed assessments of humanitarian needs, and invasive surveillance and monitoring practices.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

SIPRI BACKGROUND PAPER

BIAS IN MILITARY ARTIFICIAL INTELLIGENCE

ALEXANDER BLANCHARD AND LAURA BRUUN

CONTENTS

ABOUT THE AUTHORS

Dr Alexander Blanchard is a Senior Researcher in the SIPRI Governance of AI Programme. His work focuses on issues related to the development, use and control of military applications of AI. Before joining SIPRI, Blanchard was the Defence Science and Technology Laboratory (DSTL) Digital Ethics Fellow at the Alan Turing Institute, United Kingdom. He was also previously assistant lecturer in the Department of Political Science of the University of Copenhagen, Denmark.

Laura Bruun is a Researcher in the SIPRI Governance of AI Programme. Her focus is on how autonomous weapon systems (AWS) and broader applications of military AI affect compliance with—and interpretation of—international humanitarian law. Before joining SIPRI, Bruun worked at Airwars in London, UK, where she monitored and assessed civilian casualty reports from airstrikes by Russia and the United States in Iraq and Syria.

DOI: https://doi.org/10.55163/CJFT9557