HUMANITARIAN
LAW & POLICY

The problem of algorithmic bias in AI-based military decision support systems

September 3, 2024, Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / New Technologies / Special Themes / Weapons

13 mins read

Ingvild Bode
Professor of International Politics, University of Southern Denmark

Ishmael Bhila
Doctoral Researcher, Paderborn University

Algorithmic bias has long been recognized as a key problem affecting decision-making processes that integrate artificial intelligence (AI) technologies. The increased use of AI in making military decisions relevant to the use of force has sustained such questions about biases in these technologies and in how human users programme with and rely on data based on hierarchized socio-cultural norms, knowledges, and modes of attention.

In this post, Dr Ingvild Bode, Professor at the Center for War Studies, University of Southern Denmark, and Ishmael Bhila, PhD researcher at the "Meaningful Human Control: Between Regulation and Reflexion" project, Paderborn University, unpack the problem of algorithmic bias with reference to AI-based decision support systems (AI DSS). They examine three categories of algorithmic bias – preexisting bias, technical bias, and emergent bias – across four lifecycle stages of an AI DSS, concluding that stakeholders in the ongoing discussion about AI in the military domain should consider the impact of algorithmic bias on AI DSS more seriously.

As the international debate around AI technologies in the military domain diversifies, AI-based decision support systems (AI DSS) have increasingly become an important focus point. AI DSS, as *defined by Klonowska*, are "tools that use AI techniques to analyse data, provide actionable recommendations, and assist decision-makers situated at different levels in the chain of command to solve semi-structured and unstructured decision tasks". This wider integration of AI technologies into military decision-making raises concerns, thereby increasing the likelihood of such processes being subject to algorithmic bias – *"the application of an algorithm that compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability, or sexual orientation"*.

The issue has been addressed in the international debate on AI in the military domain. The *2023 REAIM (Responsible AI in the Military Domain) Call to Action*, for example, draws attention to *"potential biases in data"* as an issue for military personnel to be aware of. This is a good start, but the phenomenon of bias extends far beyond its presence in data.

Further, albeit focused on autonomous weapons systems (AWS), *the 2023 report* of the Group of Governmental Experts (GGE) at the Convention on Certain Conventional Weapons (CCW) *used "unintended bias" instead of "algorithmic bias"*. There may be a technical reason for this language, in that all AI systems work based on bias, in the sense of putting different levels of emphasis on particular modulations/putting more emphasis on one modality than another. However, such a purely technical understanding underappreciates the significance of bias as a political and empirical phenomenon affecting particular groups of people more adversely than others and therefore severely impacting decision-making processes integrating AI technologies.

Rule 88 of customary international humanitarian law (on non-discrimination, Part V, Chapter 32) and *Additional Protocol I, Article 85 (4)(c)* clearly prohibit adverse distinction in the application of international humanitarian law (IHL), i.e. (military) practices that are based on race, colour, sex, language, religion or belief, political or other opinion, national or social origin, wealth, birth or other status, or on any other similar criteria (such as apartheid and other inhuman or degrading practices). Such practices tend to define algorithmic bias. As such, preexisting biases are inherent to AI systems and can be systemic. This is because AI systems remain rooted in social institutions and society through how they process the data they are fed. Biases therefore do not need to be intended, explicit, or conscious to be regarded as unlawful and morally deplorable.

To begin to grasp the comprehensiveness of this issue, it is useful to consider how algorithmic bias occurs *across the entire lifecycle* of an AI DSS from pre-development to re-use/retirement. We focus our illustration of bias on four lifecycle stages of an AI DSS*:* (1) in curating the data set, (2) in design and development, (3) in use, and (4) in post-use review. We proceed by *describing the basic occurrence of bias* at these four stages and then move towards analysing problems arising from bias, particular to AI DSS concerning military use-of-force decisions. Examples are drawn from use of force decision-making at the operational and tactical levels, as this is where *we find most (current) use cases.*

Bias is a broad category, with some general types occurring across various AI systems and others being *specific to, for example, image recognition*. To refine understanding of bias at these different stages, we distinguish between preexisting, emergent, and technical bias and focus our examples chiefly on biases exhibited by AI DSS featuring image recognition.

(1) Bias in data sets

Bias in data is perhaps the best documented, with multi-faceted research identifying both explicit and implicit forms. Such preexisting bias is part of *"social institutions, practices, and attitudes"* and thus also of data sets. The training data provided to an algorithm is based on developers setting a certain statistical standard (for example, an assumption that a certain category/identity group of a population is more likely to be a threat) that may be morally, ethically, and legally unacceptable. Yet only through this process of curating datasets is raw data turned into actionable information.

There is no transparency around such data sets and the assumptions these contain, especially in the military domain. Explicit forms of bias include direct stereotypical language or imagery in data sets. Implicit forms, that are typically harder to address and mitigate, introduce bias in the form of over- or under-representing particular data points. Such forms of *sampling bias* lead to well-reported *higher misidentification rates* of darker-skinned versus lighter-skinned people.

Further, the selection, collection, and preparation of data by developers *can introduce bias* into a system. This includes pre-processing where a data set is prepared for training, a step that involves removing what are considered unimportant data points. In this way, pre-processing is prone to introducing bias into the data. Generally, an algorithm is only as *good as the data it is fed*, and data collection, storage, and use practices can *lead to discriminatory outcomes*.

In the case of AI DSS, the development of "*kill lists*" is highly problematic, as this process has been shown to depend on data inputs that conform to preexisting societal biases. This data will contain labels, for example particular characteristics identifying terrorist suspects. Such characteristics are likely to contain preexisting bias, such as racial and identity stereotypes, both explicitly and implicitly. It is possible, for example, for an AI DSS to be developed with the biased assumption that any Muslim who is devout is 'extreme', as the whole idea of counter-terrorism cannot be separated from its *racial and ethnic* origins.

(2) Bias in design and development

Choices and practices at the stage of design and development can exacerbate bias in data. At this lifecycle stage, preexisting forms of bias coalesce with technical bias that *"arises from technical constraints or technical considerations"*. This category of bias extends to both the data work that humans perform and internal, often opaque, technical processes within, for example, *neural networks*.

A useful example for human-steered processes is the process of iterative data annotation, labeling, classifying, as well as evaluation of outputs produced that occurs throughout the training process. In performing these tasks, *human cognitive biases*, often unconscious, come in. At a more basic level, the very process of constructing human and social categories so that they are amenable to computer processing may also be productive of bias. In this way, AI algorithms may also reinforce bias, for example, through *overfitting into categories* that are too "rough" which, when combined with high variance in the data set, may make the AI model unable to recognise relevant trends.

Furthermore, the opaque functioning of how neural networks process data may well insert their own sets of biases – potentially amplifying the biases that already existed in the data sets. An example is *class-imbalance* bias whereby an AI algorithm exhibits lower recognition rates for classes of data points that appear less frequently in the data set. The occurrence of this bias is well-known but requires active mitigation techniques, e.g. supplementing the data set with synthetic data.

Both specific examples of bias mentioned above, overfitting and class-imbalance, are relevant for AI DSS used in military decision-making. Military decision-making contexts are characterized by high uncertainty and elements of chaos. In such situations, AI DSS risk having too few points of comparison and therefore lacking relevant categories or using the wrong categories to recognize the situation appropriately. A *lack of appropriate qualitative and quantitative training data for many military decision-making contexts* has been recognized as a particular problem.

The fact that AI DSS are purposefully intended to identify certain groups of people means that developers must therefore evaluate the cultural, religious, ethnic, and other identity biases that influence the decisions not only of the system, but also of themselves. For example, the US Project Maven was developed with the *intention initially to aid data-labelling efforts for the Defeat-ISIS campaign*, which meant that it was developed with certain groups or identities of people in mind. The efficacy of this system in identifying the correct targets has been *questioned*, and it is, indeed, essential to question how these different sources of bias can inform how these systems are designed and developed, particularly when human targets are concerned.

(3) Bias in use

At the point of use, preexisting and technical forms of bias that are already embedded in AI DSS combine with *emergent bias*. This arises out of how particular users interact with AI DSS in particular contexts of use. In a use-of-force context, this recognizes that using AI DSS involves value-based sensemaking between military strategic, operational, and tactical decision-makers, all of whom may inject their value judgements into interacting with the system outputs.

A well-known category of bias introduced at this use phase is automation bias, describing an unquestioning trust on the part of human users in the outputs produced by an AI DSS. This unquestioning trust can facilitate algorithmic bias in practice, allowing decisions that could otherwise have been questionable if made solely by humans simply because a machine is perceived as more trustworthy and reliable.

Additionally, bias in an AI DSS can be *negatively self-reinforcing*, leading to a loop of increased bias generated by a system as it goes uncorrected. For example, if an AI DSS consistently identifies people of a particular gender with a particular physical appearance from a certain neighbourhood as suspected threats, it can reinforce its bias by assuming that all people matching these characteristics from that neighbourhood are 'threats actors'. Instead of correcting the bias, the system reinforces itself, especially if the decision-makers do not identify the bias in time.

One of the most unique features of many data-driven AI techniques relates to the ability to continuously 'learn' from contexts and interaction with technical and social environments. Consider for example, marketing data collected by tech companies. Computer systems gather data based on recognizing patterns, searches, and what captures the user's attention, increasing the frequency and intensity of recommendations the more one engages with platforms.

While such algorithmic functions may be particular to a marketing context, AI DSS in the context of military use-of-force decision making may be used to increase the number of potential targets. Such systems may therefore start by identifying modest numbers of suspected threat actors, but their purpose is to increase the numbers by associating and connecting more and more people. Even in and during use, AI DSS may therefore also continue to learn and be trained by human users. This process can lead both to old biases being reinforced and new biases being learned. Bias can *re-enter* a system as humans interact with the final output, interpret the data, and give feedback to the system.

Who is involved in this process, how this process is monitored and by whom are key questions to ask. The basic adaptability of continuous learning algorithms makes these attractive choices for use in military decision-making but also fundamentally hard to predict.

(4) Bias in post-use review

Reviewing AI DSS after use covers examining whether particular systems performed as intended and expected by system developers at the design stage, as well as including forward-looking suggestions for improvements. This can be both considered as a distinct lifecycle stage but also an ongoing activity that should follow after each use case and precede each use case, especially when continuous learning AI DSS are used.

In principle, this stage could be crucial to identify and mitigate faulty decision-making resulting from bias. Yet, if that does not occur at this stage, the biased outputs produced by AI DSS throughout the lifecycle will become data used to sustain further decision-processes. Importantly, recent research has found *evidence that humans may inherit the bias exhibited by AI DSS*. Even in situations where humans no longer interact with AI DSS, they may therefore replicate the bias they learned from an AI DSS.

Conclusion

In sum, AI DSS risk spreading the impact of algorithmic bias through military decision-making processes related to the use-of-force. Preexisting and technical forms of bias enter AI DSS from early stages of the lifecycle onwards and unfold their impact throughout its entirety, while emergent bias comes in at the point of use. Much needs to be done to increase awareness of these biases in AI DSS, their potentially disastrous effects, and ways towards their mitigation. These could, for example, include standardizing methods of developing AI DSS after use and *therein implementing bias mitigation strategies*.

Author's note: We gratefully acknowledge the feedback and input provided about the technical dimensions of algorithmic bias by Thea Riebe. This research has been supported by two sources of funding: Ingvild Bode has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 852123 (AutoNorms). Ishmael Bhila has received funding from the Collaborative Research Centre (CRC) 1187 "Media of Cooperation" at the University of Siegen.

The *Artificial intelligence in military decision-making* series outlines and investigates the manifold challenges, risks, and the potential that pertain to the use of artificial intelligence-based decision support systems (AI DSS) in military decision-making.

See also:

- Wen Zhou, Anna Rosalie Greipl, *Artificial intelligence in military decision-making: supporting humans, not replacing them*, August 29, 2024
- Ingvild Bode, *Falling under the radar: the problem of algorithmic bias and military applications of AI*, March 14, 2024
- Ruben Stewart & Georgia Hinds, *Algorithms of war: The use of artificial intelligence in decision making in armed conflict*, October 24, 2023

Tags: AI, algorithmic bias, armed conflict, artificial intelligence, conduct of hostilities, ethics, IHL, international humanitarian law, modern warfare

You may also be interested in:

Artificial intelligence in military decision-making: supporting humans, not replacing them

13 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / New Technologies / Special Themes / Weapons

Wen Zhou & Anna Rosalie Greipl

The desire to develop technological solutions to help militaries in their decision-making processes is not ...

Beyond the rubble: eight overlooked ways that urban warfare is affecting children

14 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / New Technologies / Special Themes / Weapons

Timothy P. Williams, Alexandra Jackson & Vanessa Murphy

In cities from Gaza to those in Sudan and Ukraine, childhoods are irrevocably changed by ...