Falling under the radar: the problem of algorithmic bias and military applications of AI

Ingvild Bode

Associate Professor at the Center for War Studies, University of Southern Denmark

Last week, states parties met for the first session of the Group of Governmental Experts (GGE) on lethal autonomous weapon systems (LAWS) in 2024. This debate featured the GGE's most substantive discussion to date about bias under the topic "risk mitigation and confidence building", including around a working paper dedicated to bias by Canada, Costa Rica, Germany, Ireland, Mexico, and Panama. In this post, Dr. Ingvild Bode, Associate Professor at the Center for War Studies (University of Southern Denmark) argues that bias is as much a social as a technical problem and that addressing it therefore requires going beyond technical solutions. She holds that the risks of algorithmic bias need to receive more dedicated attention as the GGE's work turns towards thinking around operationalization. These arguments are based on the author's presentation at the GGE side event "Fixing Gender Glitches in Military AI: Mitigating Unintended Biases and Tackling Risks" organized by UNIDIR on 6 March2024.

Algorithmic bias, which can be defined as "the application of an algorithm that compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability, or sexual orientation" has long occupied significant space in academic research and policy debate about the social implications of artificial intelligence (AI). Perhaps surprisingly, this does not extend to research and debate about AWS and AI in the military context. Beyond some noteworthy exceptions, chiefly UNIDIR's 2021 report "Does Military AI Have Gender?" as well as policy briefs published by the Observer Research Foundation and the Campaign to Stop Killer Robots, issues of bias have not been covered at length.

We can nonetheless draw on insights produced by established literature in the civilian space for thinking about bias in AWS and other military applications of AI for two reasons. First, much of the innovative potential on AI technologies comes from civilian tech companies who are increasingly collaborating with military actors. Second, and more fundamentally, the types of techniques used in civilian and military applications of AI, such as machine learning, are the same and will therefore be subject to similar concerns regarding bias.

Bias in AI technologies and its consequences

We can think about algorithmic bias in three main ways: 1) bias in data, 2) bias in design and development, and 3) bias in use. This means that bias can occur throughout the entire lifecycle of algorithmic models from data collection, training, evaluation, use, and archiving/disposal.

1. **Bias in data used for machine learning models.** Any potential training data is a limited snapshot of the social world. This snapshot likely contains both direct biases, for example stereotypical language and images, but also indirect biases in the shape of the frequency of occurrence. An image set may, for example, contain more pictures of physicists that are men than are women. Bias in data therefore results from unrepresentative data leading to unrepresentative output. In other words, "bias occurs

when certain types of data are missing or more represented than others, often deriving from how the data was obtained and sampled”. Both over- and underrepresentation are key problems. Considering bias in data is a good starting point but the issue of algorithmic bias extends beyond this stage. This is expressed, for example, in the popular notion of ‘garbage in, garbage out’ where the quality of input determines the quality of output. Whatever biases are part of, either implicitly or explicitly, the training data will have “a ripple effect throughout the rest of model development, as the training data itself is the only information that a supervised model can learn from”.

2. **Bias in design & development.** Bias in the data can be amplified in various stages of processing the data as part of, for example, machine learning techniques. The training process of AI technologies is a value-laden process. Human tag workers, programmers, and engineers make several choices here, such as annotating/labeling/classifying data samples, feature selection, modelling, model evaluation and post-processing after training. Algorithmic bias can therefore be the outcome of, often unconscious, value judgements involved in the machine learning lifecycle through the different tasks they perform. But biases may also come in through ‘black boxed’ processes associated with how algorithms function. At this point, **AI technologies mirror bias inherent in training data and the biases of its developers.**
3. **Bias in use.** Finally, AI technologies gain new meanings and functions – as well as potentially biases – through being used repeatedly and in an increasingly widespread way. This can happen in two ways: first, simply through employing systems featuring AI technologies, any biases that these contain will be amplified. Second, people will act on the outputs that AI systems produce. They may create “more data based on the decisions of an already biased system”. Users of AI technologies may therefore find themselves in **“negative feedback loops”** that become the basis for future decisions. In this way, biased outputs produced by AI technologies may also be used as further justification to continue existing (biased) practices. At the point of use, we also must account for bias in how humans interact with AI technologies, more prominently automation bias. This describes a tendency for humans to depend excessively on automated systems and to defer to outputs produced by such technologies. There is considerable evidence for automation bias from studies outside of the military domain. We can therefore, unfortunately, easily think about situations where human users place too much trust in AI systems in a military context.

Knowing that AWS and other military applications of AI will likely contain algorithmic biases brings serious consequences. Biases can lead to legal and moral harms as people of a certain age group, gender, or skin tone may be wrongfully assessed to be combatants. These harms are well-summarised in, for example, UNIDIR’s 2021 report, mentioning a full range of problematic consequences from such misrecognition. Biases also impact system functionality and predictability. This has to do with a lack of transparency and explanation. It is often unclear which features of the data a machine learning algorithm has assigned to its output: **“this means that no explanation can be given for why a particular decision has been made”**. In addition, bias in datasets used for military applications of AI may be exacerbated. This is because the available data that is suitable to train military applications may be more restricted in scope than data used to train civilian applications. Data available may, for example, represent a specific conflict and type of operation that is not applicable to broader applications. In other words, both the quantity and quality of applicable datasets for military applications of AI could be trained on may be limited.

Bias as a socio-technical problem

Research on algorithmic bias, and in particular gender bias, can be grouped into studies focused either on bias perpetuation or on bias mitigation. A prominent example of studies documenting gender bias perpetuation is the gender shades project conducted by Joy Buolamwini and Timnit Gebru. The authors examined three kinds of facial recognition software and found that all three recognise male faces far more accurately than female faces and were generally better at recognising people with lighter skin tones. The system's worst-performing model did not recognise faces of dark-skinned women 1/3 of the time. Other studies examine how and by what methods bias can be mitigated. This research is mostly technical in nature and focuses on specific techniques that can be used in machine learning models and facial recognition systems such as rebalancing data or regularization or the design of 'fair' algorithms through adversarial training. Technical mitigation strategies are not straightforward. The systems studied have later been identified as problematic in studies focusing on getting bias perpetuation, for example, had performed in functional ways during testing. The problem of algorithmic bias is not easily solved.

Thinking around bias underlines, once again, that technology is not neutral. Technologies are "products of their time", they mirror our societies. Rather than seeing technology as an object and technological development as a process that is happening on trajectories distinct from our societies, we need to recognise technology and its development process as social in nature. In other words, "bias is inherent in society and thus it is inherent in AI as well". For this reason, technical solutions will not be sufficient to resolve bias.

Addressing the problem of algorithmic bias requires fundamentally changing discriminatory attitudes. Even at the design stage, for example, mitigation strategies would also have to be mainstreamed into how AI programmers think about (initial) modelling parameters. Here, it matters to have a closer look at the tech companies who are dominating investment in and development of AI technologies and their particular interests because these interests are likely to have a direct impact on choices made at the design stage. To change this, we would need to address and change the "bias embedded in work cultures" or professions that are particularly crucial in the design of AI technologies – in other words, STEM professions. At present, STEM professions are dominated by a limited set of people that are not representative of our wider societies: "tech firms employ few women, minorities, or those aged over 40". Addressing this issue of representation and diversity requires capacity-building among under-represented groups. But it also requires fundamental changes to the professional cultures of, for example, engineering and IT, that rest on, for example, long-standing, still often implicit associations of technical knowledge and expertise with masculinity as well as particular ethnic backgrounds.

To conclude, algorithmic biases have become firmly recognised as a major risk factor associated with AI technologies in the military context. Bias and harm mitigation are, for example, listed in many of the emerging lists of responsible AI principles in the military domain. At the same time, a lot of the logic that appears to be driving states to integrate AI technologies into weapon systems and in the broader military context rests on the argument that using such technologies can make the conduct of war more rational and predictable. But this idea that AI technologies may be 'better than human judgement' ignores the ways in which AI technologies that are potentially used in weapon systems are both shaped by and shape (human) decision-making. The problem of algorithmic bias demonstrates we should think about AI technologies not as something separate from human judgement, but as deeply enmeshed in forms of human judgement throughout the entire lifecycle of AI technologies – for good and for bad.