The issue of bias

Whitepaper on algorithmic bias in (Lethal) Autonomous Weapons Systems by Alycia Colijn* and Heramb Podar*

Why do we need to address bias when we speak about (Lethal) Automated Weapons Systems, (L)AWS in short? In this report, we set out **which types** of bias should be taken into account, **when** they occur in the lifecycle of AWS, and **what this means** for policymakers.

Types of bias

Although the rolling text on the moment of writing[1] (September 2024) refers to unwanted bias in data sets, and unwanted automation bias, these two types of biases **do not cover the full spectrum of bias**. To clarify, bias is considered to be any type of flaw in an algorithmic system that leads to a statistic estimate that does not equal the true value[2]. For example, an unfair over- or underrepresentation of specific groups of people based on gender[3], ethnicity, sexual orientation, religion or location, or the misclassification of objects (e.g. hospitals, places of worship, etc.).

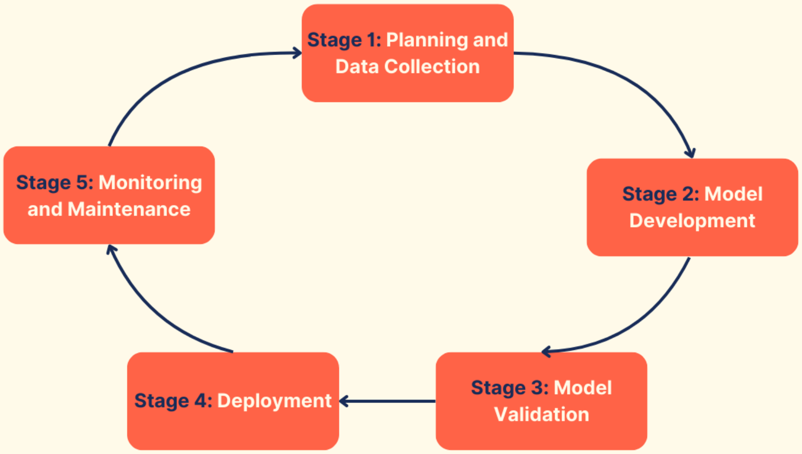
Pre-existing bias	Technical bias	Emergent bias
The first type of bias we recognize is the pre-existing bias[4]. This refers to any type of bias that **already exists in society** and is thus replicated, and often enlarged, by algorithms. This would include biased data sets as included in the rolling text.	Technical bias refers to any kind of bias that occurs from the **limitations of a system**. This could include a system drawing from an alphabetic list, unintentionally favoring options first up in the alphabetic order.	The last type of **bias occurs over time**, as a result from changing societal knowledge, population or cultural values. This could also include algorithmic decay, model drift or degradation.

Encode Justice

*Equal Contribution

Alycia Colijn (The Netherlands) ✉| alycia@encodejustice.nl
Heramb Podar (India) ✉| podar_hd@cy.iitr.ac.in

Bias over the system’s life-cycle


Stage 1: Planning and Data Collection
Stage 2: Model Development
Stage 3: Model Validation
Stage 4: Deployment
Stage 5: Monitoring and Maintenance

Bias in data collection

The first stage of development and deployment of systems is the **selection of data** that will be used for the training and testing of the system. These data sets are incredibly vulnerable for bias, mostly **pre-existing bias**. This can originate from the data sets, lack of available data (especially in military use[5]), but also from the data selection by developers. Where a developer could be held accountable for biased data selection, **who is accountable for data sets that reflect a certain societal status quo** - including the bias that already exists in society?

Bias in design and development

When the required data is collected, the model development (also referred to as model training stage) takes off. During this stage, the parameters of the models are fine-tuned. This includes decisions like: will the system make a final decision when it’s 90% sure, 95% sure of 99% sure? **Each parameter that is set, comes with the risk of further enlarging bias that already exists in the data that the training stage started with** (e.g. class imbalance[6] - a preference for bigger ‘classes’, for example caused by network forming in collaborative filtering algorithms[7] and overfitting[8] - the model recognizing random noise as a trend) and include (unconscious) bias of the developers.

It is common practice for data scientists to randomly split the initial dataset into two parts: one for training the model (model development) and the other for testing it (model validation), a process referred to as cross validation[9]. However, when the original data set contains certain bias **the model is then validated with a biased data set as well** (and thus not really validated for the 'real world').

Additionally, this is the stage where **technical bias** comes in. To spot bias that is caused by technical limitations, it is vital that **developers with different backgrounds** analyze the process of model development and evaluation.

Bias in deployment and monitoring

Once in use, emerging bias is the greatest risk. This happens when the world changes, and the model is not re-trained. The loss of accuracy can be referred to as **degradation**[10], **model drift**[11], **data drift**[12] or **decay**[12]. Data drift, degradation or decay occurs when the data that was used to train (develop) and test (validate) the algorithm, no longer reflect the situation in which the model takes decisions which is sometimes referred to as a distributional shift in environments. In military context, this for example happens when a system is trained in a specific environment, which changes the longer an armed conflict continues. Model drift includes data drift, but includes other types of drift that lead to a change between the input and output variables, e.g. changing (legal) definitions or changes in military uniforms that challenge the recognition and classification of combatants.

Next to degradation, **the self-learning capacities of AI can cause a negative self-reinforcing feedback loop**[13], which could be considered a form of overfitting over time. The model then identifies noise as a pattern, labeling for example individuals with a certain physical appearance or geographical location as targets.

Automation bias[14], as referred to in the rolling text, then occurs when **these fallacies are not corrected by human decision-makers because they place a higher degree of trust in the system than in human decision-making**.

What does this mean for policy-makers?

An unbiased system does not exist. However, there are ways to mitigate these risks as much as possible. For policy-makers, the different risks of bias mean that there are several aspects to take into account when moving to a next step in the discussion around (L)AWS.

1. Bias requires a lens that sees **beyond just social and automation bias** that is currently included in the rolling text.

2. Although there has been no attention to the **actors** and **teams** that develop the algorithms used in (L)AWS, they greatly affect the decision-making by (L)AWS. According to a 2022 global survey with over 70.000 respondents, over 90% of developers are male[15], such a widespread survey has not yet been conducted for ethnicity - underlining the problem of lacking attention for the broad range of dimensions diversity is required in - but smaller surveys show 60[16]-75[17]% being white. The lack of diversity in these areas, signal a similar homogeneity for other dimensions of diversity, e.g. sexual orientation, religious background, etc.

   Additionally, the private actors that currently focus on the development of algorithms for military use, all have **specific interests** in the process. This can lead to over-stating the accuracy, only training (developing) and testing (validating) in specific circumstances and environments and a lack of transparency on the development process of the final algorithm. It is thus vital to consider the public-private relationships that occur in the context of military use

3. Although the current rolling text refers to rigorous testing and evaluation of how the weapons system will perform, **periodic reassessment of these evaluation measures** is the absolute minimum to mitigate the risks of algorithmic decay, drift or degradation, and increase the chances of anticipated effects and predictability of the algorithmic decision-making.

4. The current rolling text refers to traceable and explainable effects of the use of LAWS, but does not yet operationalize these requirements. To operationalize, policy-makers could consider requiring (L)AWS – or parts of these systems – to be developed open source, open core or with source available[18] **to avoid so-called black box algorithms**.

Further readings

ICRC's Blog Series on AI in the military

The Risks and Inefficacies of AI-systems in Military Targeting Support by Jimena Sofía Viveros Álvares, see: https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/

Falling Under the Radar: The Problem of Algorithmic Bias and Military Applications of AI by Ingvild Bode, see: https://blogs.icrc.org/law-and-policy/2024/03/14/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai/

The Problem of Algorithmic Bias in AI-based Military Decision-Support Systems by Ingvild Bode and Ishmael Bhila, see https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/

UNIDIR Report on AI and Gender

Does Military AI Have Gender? By Katherine Chandler, see: https://unidir.org/files/2021-12/UNIDIR_Does_Military_AI_Have_Gender.pdf

Our shared future is more precarious now than ever before.

That is why Encode Justice calls on world leaders to join our intergenerational coalition for global AI action.

Find our Calls to Action and signatories, including the **first woman President of Ireland** and **former UN High Commissioner for Human Rights** Mary Robinson, former **employee of OpenAI** Daniel Kokotajlo, **Executive Director of the Somali Human Rights Organization** Abdullahi Akbar and a range of the **most renowned AI scholars around the world** from Stanford University, NYU, Yale, Princeton University, UC Berkely, Chinese Academy of Sciences, Université de Montréal, Indian Institute of Technology, University of Oxford, Australian National University, NOVA University of Lisbon, Universidad de Los Andes and the Technical University of Valencia on **ai2030.encodejustice.org**

References

1. GGE on LAWS. Rolling text. *Convention on Certain Conventional Weapons - Group of Governmental Experts on Lethal Autonomous Weapons System*.

2. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health*. 2004;58(8):635–641. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1732856/. Accessed Oct 1, 2024. doi: 10.1136/jech.2003.008466.

3. Acheson R. Gender and bias. *Women's International League for Peace and Freedom*. 2021. https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Gender-and-Bias.pdf. Accessed Oct 1, 2024.

4. Friedman B, Nissenbaum H. Bias in computer systems. *ACM Transactions on information systems (TOIS)*. 1996;14(3):330–347.

5. Corrected oral evidence: Artificial intelligence in weapons systems. . 2023(2). https://committees.parliament.uk/oralevidence/12983/html/. Accessed Oct 1, 2024.

6. Bauder RA, Khoshgoftaar TM, Hasanin T. An empirical study on class rarity in big data. . 2018-12:785–790. https://ieeexplore.ieee.org/document/8614150. Accessed Oct 1, 2024. doi: 10.1109/ICMLA.2018.00125.

7. Google. Collaborative filtering . https://developers.google.com/machine-learning/recommendation/collaborative/basics. Accessed October 1, 2024.

8. Ying X. An overview of overfitting and its solutions. . 2019;1168:022022.

9. Scikit Learn. 3.1. cross-validation: Evaluating estimator performance. https://scikit-learn/stable/modules/cross_validation.html. Accessed Oct 1, 2024.

10. Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*. 2022;245:108632. https://www.sciencedirect.com/science/article/pii/S0950705122002854. Accessed Oct 1, 2024. doi: 10.1016/j.knosys.2022.108632.

11. Holdsworth J, Belcic I, Stryker C. What is model drift? | IBM. https://www.ibm.com/topics/model-drift. Updated 2024. Accessed Oct 1, 2024.

12. Stihec J. Understanding data decay, data entropy, and data drift: Key differences you need to know. https://shelf.io/blog/understanding-data-decay-entropy-and-drift-key-differences-you-need-to-know/. Updated 2024. Accessed Oct 1, 2024.

13. Hagen A. Negative feedback loops: Using an economic model to inspect bias in AI. . 2020. https://www.microsoft.com/en-us/research/blog/when-bias-begets-bias-a-source-of-negative-feedback-loops-in-ai-systems/. Accessed Oct 1, 2024.

14. Automation bias. Databricks Web site. https://www.databricks.com/glossary/automation-bias. Updated 2019. Accessed Oct 1, 2024.

15. Vailshery LS. Software developers: Distribution by gender 2022. Statista Web site. https://www.statista.com/statistics/1126823/worldwide-developer-gender/. Accessed Oct 1, 2024.

16. McEnvoy D. How ethnically diverse is the tech workforce? . https://cord.co/techhub/working-culture/articles/ethnic-minority-representation. Updated 2022. Accessed October 1, 2024.

17. Weststar J. Developer satisfaction survey
2021. *Western University, Ontario, Canada*. 202. https://igda-website.s3.us-east-2.amazonaws.com/wp-content/uploads/2021/07/31184838/IGDA-DSS-2021-COVID-Report_July-18-2021-1.pdf. Accessed Oct 1, 2024.

18. Langhammer J. Black box security software can't keep up with open source | authentik. Authentik Web site. https://goauthentik.io/blog/2023-09-14-black-box-security-software-cant-keep-up-with-open-source/. Updated 2023. Accessed Oct 1, 2024.