UNIDIR

UNITED NATIONS
INSTITUTE FOR
DISARMAMENT
RESEARCH

Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies

A PRIMER

No. 9

UNIDIR RESOURCES

Acknowledgements

This primer builds upon initial discussion at the UNIDIR meeting “Autonomous Weapon Systems: Understanding Bias in Machine Learning and Artificial Intelligence”, held in New York on 5 October 2017. This event was part of Phase III of the UNIDIR project “The Weaponization of Increasingly Autonomous Technologies”, sponsored by the Government of Germany. UNIDIR would like to thank the speakers and participants at that meeting for their thoughtful engagement. UNIDIR would like to thank those who provided input to this paper, in particular David Danks and Emily LaRosa, as well as Kerstin Vignard, John Borrie and Heather Roff. UNIDIR also extends its thanks to those who kindly reviewed the paper at various stages of its production.

About the Project “The Weaponization of Increasingly Autonomous Technologies”

Given that governments have a responsibility to create or affirm sound policies about which uses of autonomy in weapon systems are legitimate—and that advances in relevant technologies are also creating pressure to do so—UNIDIR’s work in this area is focused on what is important for States to consider when establishing policy relating to the weaponization of increasingly autonomous technologies. See http://bit.ly/UNIDIR_Autonomy for Observation Papers, audio files from public events, and other materials.

This is the ninth in a series of UNIDIR papers on the weaponization of increasingly autonomous technologies. UNIDIR has purposefully chosen to use the word “technologies” in order to encompass the broadest relevant categorization. In this paper, this categorization includes machines (inclusive of robots and weapons) and systems of machines (such as weapon systems), as well as the knowledge practices for designing, organizing and operating them.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to a variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and Governments. UNIDIR activities are funded by contributions from Governments and donor foundations.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of UNIDIR. They do not necessary reflect the views or opinions of the United Nations or UNIDIR’s sponsors.

www.unidir.org

Contents

1 Introduction

Algorithms are fundamental to autonomous computational systems. The “brain” of any autonomous system—whether a passenger vehicle or a weapon system—is fundamentally composed of algorithms, whether for sensing, learning, planning, deciding or acting. These systems can exhibit autonomous capabilities to adapt and respond to ill-defined contexts and varying environments only because of the use of various algorithms, whether in development, training or deployment.[1]

As algorithms play a rapidly increasing number of roles in everyday life, there is a growing recognition that algorithms and data are not purely objective and do not always have the impacts or functionality that the designer or user wanted and expected. Such surprises are often viewed as failures of the system, as their unintended consequences might produce harmful effects.

The term “algorithmic bias” is appearing more frequently in newspaper articles, white papers, and reports. Algorithmic biases or failures have been identified and widely reported in non-military applications such as loan processing, teacher evaluation, law enforcement, and numerous other uses. In response, there have been calls for regulation or policy changes in those domains, whether requirements of data transparency, algorithmic auditing, demonstrations of equitable impact, or some other effort to reduce the negative consequences of the use of algorithms.[2]

For many, the term “bias” brings to mind racial or social-economic injustices, and some might question the usefulness of discussing bias in relation to weapon systems. However, there are two important reasons to do so. First, understanding the existence and origins of algorithmic bias helps to counter the narrative—still widespread—that “technology is neutral”. On the contrary, technological systems are designed by humans to meet human needs and objectives and thus they reflect human values and decisions. This awareness reminds us of our ability and responsibility to shape the technologies we design and choose to employ. Secondly, algorithmic bias is a field of intense study that helps illuminate how or why algorithms make particular determinations—and this understanding offers an opportunity to improve the algorithm’s outcome, develop mitigation strategies, or determine whether the bias may result in outcomes that are simply too risky to permit.

When considering the development of Autonomous Weapon Systems (AWS), it is useful to examine what can be learned from other non-military domains where the system’s failure due to algorithmic biases has been well-documented. Algorithms are not perfectly objective and infallible, but they also are not random and capricious. Rather, they exhibit behaviour and biases that result from a variety of decisions and inputs. Using examples and known cases from other fields, this paper offers policymakers an opportunity to consider how biases might be present in

[1] This paper is intended to be an introductory primer for non-technical audiences. Because of the rapidly developing nature of this field, this paper can only provide a snapshot in time. However, many of the underlying concepts about algorithmic bias are likely to remain applicable in the short to medium term.

[2] While there has yet to be much written about bias in weapon systems specifically, there is a vast amount of accessible information on the implications of algorithmic bias and potential responses. See, for example, World Economic Forum Global Future Council on Human Rights, 2018, *How to Prevent Discriminatory Outcomes in Machine Learning*, White Paper; Will Knight, “Forget Killer Robots—Bias is the Real AI Danger”, *MIT Technology Review*, 3 October 2017; Jonathan Vanian, “Unmasking A.I.’s Bias Problem”, *Fortune.com*, 25 June 2018; and O. Osoba, “Keeping Artificial Intelligence Accountable to Humans”, *The RAND Blog*, 20 August 2018.

increasingly autonomous weapon systems, whether such biases can and should be mitigated, and what avenues exist for reducing the potentially harmful consequences of bias.

This primer is divided into three main sections. It begins with a general discussion of algorithmic biases—their nature, types and sources. The second section considers the impacts of algorithmic bias, with both real-world examples and examples of how bias could arise in future weapon systems. The third section considers potential mitigation strategies to address bias determined to be harmful or undesirable.

2 Sources and types of algorithmic bias

In recent years, there has been a growing recognition that an algorithm is not a neutral transformer of data or extractor of information, but rather can exhibit biases or distortions in its operation. These biases can be introduced in many different ways, and in many different forms. Moreover, as systems become more complex, it may become increasingly difficult to understand how an autonomous system arrives at its decision, and so increasingly difficult to determine the nature and extent of algorithmic bias. As a result, significant harms from algorithm-driven systems may go unnoticed or unmitigated until too late.

Throughout this paper, the word "bias" is used in a neutral way to mean "deviation from a standard"; there is not a presumption that the deviation is either negative or positive. The same algorithm can exhibit bias relative to one standard but not relative to another, and so one needs to distinguish the different standards that one might employ for assessing the performance of an algorithm. For example, a ***statistical bias*** occurs when an algorithmic output deviates from a statistical standard, such as the actual frequency of some event in relevant situations. In contrast, a ***moral bias*** arises when the algorithmic judgments deviate from established moral norms. Different types of bias (including regulatory, legal, social, as well as others) can arise when the judgment or result deviates from the relevant established, socially codified norm. Although these standards are distinct from one another, they are sometimes connected. For example, statistical bias in an algorithm may point to moral or normative bias, as when a hiring algorithm *unnecessarily* favours men over women; however, an algorithm might *appear* to be biased towards men, when it is actually biased towards those who can satisfy a specific requirement (which may happen to be imperfectly correlated with being male).

Moreover, algorithmic biases can result from different sources and for different reasons.[3] It is important to determine the different categories and types of biases present in an algorithm, particularly when determining whether a particular bias merits a response and, if so, what mitigation or corrective measures ought to be taken. There are, broadly speaking, three technical or computational sources of bias:

- Training data;
- Algorithm focus; and
- Algorithmic processing.

Beyond these three, there are two forms of bias arising from humans misusing a system or misunderstanding algorithmic output:

- Inappropriate use or deployment (transfer context bias); and

[3] For a more detailed discussion, see D. Danks, and A.J. London, 2017, "Algorithmic bias in autonomous systems" in C. Sierra, (Ed.), *Proceedings of the 26th International Joint Conference on Artificial Intelligence,* pp. 4691–4697.

- Interpretation bias.

The first source of algorithmic bias is the use of ***inappropriate training data***. In general, algorithms are “trained” to perform in particular ways using exemplar data, and so these training data should capture relevant features in the intended context of use. For example, consider a self-driving vehicle intended for use in diverse terrain, traffic conditions, and weather. If its training data comes from only one particular geographic setting or particular lighting conditions, then the vehicle will exhibit biased behaviour in other environments or in different lighting, as it is “trained” to respond in only that one particular setting (i.e. the training data does not possess enough variation or is not sufficiently robust). Algorithmic bias due to inappropriate or insufficient training data can be difficult to detect, as the system will function appropriately in its narrow environment. This bias is revealed only when the system is used more generally. Moreover, inappropriate training data can be a particularly insidious source of algorithmic bias, as most developers do not publicly disclose their training datasets as a matter of proprietary information. As a result, such biases can remain hidden from users or deployers until there has been a substantial accumulation of harms.

In the case of an AWS, the training data could come from the laboratory or the field. In many cases, the data easily available during training will have significant differences from the intended deployment contexts. Developers may not have ready access to people, terrain, or environments that match the intended use cases and these training–deployment mismatches can be quite subtle. For example, perceptual capabilities could be significantly biased or degraded simply due to differences in soil composition. Moreover, since weapon development will frequently occur away from the theatre of operations, it may be quite difficult—perhaps even impossible—to obtain training data that are actually representative of the intended context of use (if there even is a stable “context” in an adversarial environment such as the battlefield[4]). If the type and degree of training–deployment mismatch is known, then developers can use statistical or algorithmic adjustments to minimize the extent of statistical bias due to the training data. That is, one could use one source of algorithmic bias to try to mitigate or compensate for another source. However, developers rarely know the full extent of mismatch, and so typically cannot completely or precisely compensate for bias due to training data.

A second, related source of algorithmic bias is ***inappropriate “focus”***. Focus bias occurs when there is usage of incorrect or inappropriate information in the input or training data. We often believe that an algorithm ought not use some information (even if it is available) in its decision-making processes. An obvious case of focus bias is using morally irrelevant categories, such as whether someone prefers toast with or without jam, to make morally relevant judgments, such as whether that person should be released from prison. A more neutral example of algorithmic bias due to inappropriate focus is in the use of legally protected information in decision-making. An algorithm that uses such information will deviate from the normative legal standard, even though it might exhibit improved statistical performance by using that information. In such instances, there is a case of “forced choice”: either use an algorithm that is biased relative to a legal standard (due to use of protected information), or biased relative to a statistical standard (due to ignoring statistically relevant input data). Of course, there is often no real choice in practice, as the system is required to conform to the legal standards. Thus, much of the technical and mathematical work on algorithmic bias has focused on detecting, measuring, and eliminating this kind of legal or

[4] The forthcoming UNIDIR primer “Learning and Adaptive Weapon Systems in Adversarial Environments” explores questions concerning reliability and predictability of AWS in variable and adversarial environments.

moral bias in which protected information is inappropriately used (perhaps indirectly or even unintentionally).[5]

A third source of algorithmic bias is when the algorithm itself is biased in the way that it transforms data. Known as ***processing bias***, this is most obvious when the algorithm employs a statistically biased estimator in order to learn accurate predictive models from small datasets. In general, algorithmic processing biases are often intentionally used to try to minimize the impacts of other sources of algorithmic bias (i.e. to intentionally skew or counterweight the algorithm's output). For example, one may use algorithmic processing as a bias source in order to mitigate bias due to unrepresentative or inaccurate training data. The resulting algorithm does not accurately capture the training data, and so is biased relative to the usual statistical standards. Nonetheless, this statistical bias is not a negative one, precisely because it can help to counteract other types of bias, such as legal or moral bias. Moreover, this type of algorithmic processing bias is often used to learn more accurate (or stable) models. That is, by intentionally biasing the algorithm, we can reduce the variance in our models so that they are more reliable; this is known as the bias-variance trade-off.[6]

Inappropriate focus and processing biases are areas requiring particular attention in weapon systems. Essentially all learning algorithms make assumptions or have "built-in" information. These might be quite innocuous, such as the assumption that the past is at least somewhat informative about the future (though they could differ in multiple significant ways). In general, there is a trade-off for learning algorithms between (i) size and quality of training data; (ii) number and strength of algorithmic assumptions; and (iii) power and accuracy of the algorithm outputs or learned models. We only have powerful, accurate algorithmic outputs if either the training data are excellent, or the algorithms make a number of (correct) assumptions, or both. For example, if a classification algorithm in an AWS has too few training samples, then it may need to make assumptions about the base rates (i.e. the prior probabilities) of various phenomena. The learning algorithms will thus typically need to make substantive assumptions about features of the environment, the adversaries, or the overall context of use. To the extent that those assumptions are incorrect, the weapon system will likely exhibit significant statistical, moral, or legal biases, even if everything else were done correctly (such as the developers having accurate training data, users being appropriately trained, and the system possessing a robust model capable of adapting to its training data).

Fourth, algorithmic biases can arise from ***inappropriate use or deployment*** of a system. This source of bias is not technical, but instead due to the humans who use the system. Algorithms or systems are intended for particular purposes in particular contexts but are sometimes deployed outside of that narrow window. This "context transfer" can produce algorithmic biases, as one is extending the algorithm beyond its intended capabilities. The system will often behave in surprising or incorrect ways in these novel transfer contexts, depending on the relevant contextual differences. To continue the earlier example of a self-driving car, bias due to transfer context

[5] D. Pedreshi, S. Ruggieri, and F. Turini, 2008, "Discrimination-aware data mining" in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 560–568; F. Kamiran, T. Calders, and M. Pechenizkiy, 2010, "Discrimination-aware decision tree learning" in *Proceedings of the 10th IEEE International Conference on Data Mining,* pp. 869–874; S. Hajian and J. Domingo-Ferrer, 2013, "A methodology for direct and indirect discrimination prevention in data mining", *IEEE Transactions on Knowledge and Data Engineering*, 25(7): pp. 1445–1459; M. Feldman et al, 2015, "Certifying and removing disparate impact" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.

[6] S. Geman, E. Bienenstock and R. Doursat, 1992, "Neural networks and the bias/variance dilemma", *Neural Computation*, 4: pp. 1–58.

would occur if such a vehicle was intended for use in the United States but was deployed instead in the United Kingdom (where traffic differs, most notably by driving on the left-hand side). There are close connections between training data and transfer contexts as sources of bias, as both involve the system not having the "right" information or learned models to behave appropriately. However, inappropriate training data leads to algorithmic biases even in the intended contexts of use; in contrast, bias due to transfer context arises only when the system is used outside of those intended contexts.

This type of bias could arise from the use of AWS in contexts that are substantively different from originally intended use cases. Even if an AWS is trained using appropriate, accurate data from the intended context, the system may act quite differently than expected if used for new purposes. As a concrete example, an AWS that is effective on open terrain could behave in quite biased ways if deployed in an urban, densely populated area. Commanders have a responsibility to ensure that specific weapon systems are used in accordance with rules of engagement and command-and-control authorities. They should know the intended contexts of use for a specific weapon system (and its performance in those contexts), and ensure that it is used only in those contexts. However, there may be circumstances when one is tempted to extend the scope of use of a system. Using weapons in contexts or against targets for which they were not originally designed is a well-documented phenomenon.[7] Thus, to the extent that command-and-control structures are weakened, then there is a risk of an AWS being used in unintended contexts, even though its behaviour can be biased in those situations.

Finally, algorithmic bias can arise from ***interpretation failures*** by the user, as misinterpretation of the algorithm's outputs or function can lead to biased behaviours and outcomes. Interpretation bias arises when the information an algorithm produces does not fit the information requirements of the user (whether that user is a human or some other computational system). There is ample opportunity for this type of "informational mismatch", as developers rarely specify the exact content of the system or algorithm's models in each context. Thus, systems that employ the algorithm's outputs in some way can be misdirected by unreliable features of those outputs. For example, an autonomous monitoring system might employ algorithms that estimate the "surveillance value" of different individuals. The monitoring system could exhibit significant biases if it incorrectly uses the outputs of those algorithms, particularly if it incorrectly understands the meaning (for the algorithm) of "surveillance value", such as treating the output of an algorithm that estimates "uncertainty about an individual's identity" as indicating "probability that an individual is a terrorist".

The behaviour of an AWS could appear biased if ***the user incorrectly understands the information that is used and provided by the AWS***[8]. This source of "algorithmic" bias is perhaps more accurately classified as *a* ***user failure***, since it would arise when the relevant commander misinterprets the behaviour of the system, rather than the system behaving in any intrinsically biased manner. For example, suppose the AWS is designed to identify and fire upon ships that cross a particular boundary. If the commander believes (erroneously) that the system's classification of a ship is based on the categories of "adversary" and "friendly" (rather than "inside" and "outside" perimeter), then the weapon system will likely exhibit quite "biased" (from

[7] See, for example, E. Prokosch, 1995, *The Technology of Killing: A Military and Political History of Antipersonnel Weapons*, Zed Books; E. Prokosch, 1995, "Cluster Weapons" in *Papers in the Theory and Practice of Human Rights,* Human Rights Centre, University of Essex; and John Borrie, 2009, *Unacceptable Harm: A History of How the Treaty to Ban Cluster Munitions Was Won,* UNIDIR, pp. 330–332.

[8] For a more detailed discussion, see UNIDIR, 2016, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, particularly the section "Operators and their limits", p. 14.

the commander's perspective) behaviour, including firing on friendly forces (that cross the perimeter) and not firing on hostile ones (that stay just outside of the perimeter). Although this type of "bias" is due to human error rather than algorithmic failure, it emphasizes the importance of thinking about the larger, complex human-machine-organization system. Failures are rarely the fault of solely the algorithm, but rather are typically the result of multiple sources of bias, whether directly in the AWS algorithms, or in the humans that use them, or in the contexts where they are deployed.

Algorithmic biases are potentially found in any system that uses algorithms in a substantive way. The algorithms in a weapon system are not qualitatively different from those in other, non-weapon systems: they may learn from potentially problematic training data; might use protected or inappropriate information; inevitably have prior knowledge or biases in the algorithm; could be used outside of intended contexts; and provide output information that can be misinterpreted. More generally, weapon developers must make a series of decisions about what the system ought to prioritize or value, often through specification of a "loss function" or "optimization criteria". The developer thus passes along and instils values and biases in the algorithm, even if these choices are usually not described in this way. And if the algorithm has been "taught" to prioritize a particular factor that it ought not to (relative to some standard), then the algorithm will inevitably be biased.

As seen in this section, algorithmic biases do not form a single, monolithic category. In order to assert that a bias is negative or harmful, we must have concrete specifications as to the standards or norms against which to measure the algorithm, and consideration of the source(s) of that bias. We must also consider the role of the algorithm in question within the overall system, and how its biases may affect the overall output decision or information.

3 The real-world impacts of algorithmic bias

Algorithmic biases potentially present themselves wherever algorithms are used to analyse and filter data to extract information or reach a decision. Often, these biases are imperceptible to the layperson or go unnoticed, lending to the conceptualization that the algorithm is, in fact, an objective, impartial black box that takes unbiased data as input and necessarily outputs the correct response. However, there are numerous examples in which algorithmic biases have detrimentally targeted and affected specific populations and demographics because of the deployment context and use. (Hypothetical examples involving AWS are discussed at the end of the section.)

Finance and loan-determining algorithms are an excellent example of how biases impact particular populations and create a cycle of reinforcement. When people apply for loans or lines of credit, banks often will gather a swath of information, including age, profession, and importantly, postal code. These financial institutions no longer hand this data profile to a human who can weigh separate factors and come to a conclusion that prioritizes the individual on a case-by-case basis. Rather, these profiles are now fed into an algorithm that determines whether the candidate is worth the risk of the loan based on a variety of factors, weighted according to the (historical) training data.[9] In practice, the primary goal of these algorithms is to maximize the amount of money loaned, while minimizing the costs of loan defaults. More precisely, the algorithms are

[9] C. O'Neil, 2016, *Weapons of math destruction*, Chapter 8, "Collateral Damage: Landing Credit"; see also: M. Poon, 2007, "Scorecards as devices for consumer credit: The case of Fair, Isaac & Company Incorporated", *Sociological Review*, 55: pp. 284–306.

trained to make ***statistically accurate predictions*** about the likelihood of default for potential loans to particular individuals, perhaps coupled with a decision threshold for granting a loan. These algorithms thus aim to be statistically unbiased: they should accurately capture the patterns in the training data. However, that same aim can lead to moral or legal biases, if the algorithm finds patterns that depend on privileged or protected information. For example, the race of the loan applicant might be a relevant predictor (in the training data) for likelihood of loan default or approval, even though many countries forbid the use of race information in loan decisions.[10] Loan approval algorithms thus frequently provide an example where some type of algorithmic bias is inevitable; developers and users must decide whether to have statistical or moral/legal bias.

The algorithmic biases that have been demonstrated for loan approvals can be quite subtle to the layperson.[11] For example, consider two different individuals: each employed for five years as a school teacher, making the same salary, with the same credit score, and requesting loans of the same size. One might think that these individuals should have roughly the same chance to receive the loan. However, if some other attribute is correlated with loan approval or default in the training data, then the algorithm might give radically different judgments. In particular, even if privileged information (such as race, class or caste) is excluded from the training data, there might be another feature that is ***correlated with*** the privileged information. For example, address is often correlated with race, and so using the applicant's address (not protected information) in the loan decision algorithm can reintroduce the conflict between statistical bias (address is correlated with loan approval) and moral/legal bias (not unjustly harming people from a specific group). Moreover, the potential problems can be exacerbated by the use of an algorithm in modern bureaucratic environments, where the human employee is often discouraged, or even forbidden, to overrule the algorithmic output. There may be no possibility for appeal or correction, which has led to multiple instances in which groups were (unjustly) significantly harmed by the use of morally biased, though statistically unbiased, loan approval algorithms.[12] Importantly, the use of a biased algorithm leads to lack of data about the actual likelihood of loan default in that group (since they are not issued loans in the first place), and thereby perpetuates structural inequalities.

Algorithmic moral biases have also been identified in algorithms that aimed to predict recidivism, or the likelihood that a prisoner will commit another crime after release from prison.[13] The original motivation for these algorithms was admirable, as they aimed to remove the prejudices of judges or parole boards from the process of prisoner release. The use of such predictive algorithms was championed as a more neutral, "objective" or "fairer" way to make such decisions, as they would focus solely on public safety.[14] Unfortunately, the algorithms instead exhibited significant biases; for example, two individuals with essentially the same factors except race could

[10] Of course, race could be a usable predictor because of historical systemic biases, not because it is actually causally relevant to repayment.

[11] C. O'Neil, 2016, Chapter 8; Kochar et al., 2011, "Wealth Gaps Rise to Record Highs Between Whites, Blacks, Hispanics: Twenty-to-one", *Pew Research Center*, 26 July 2011.

[12] FICO scores, which originated in the United States as a "color blind" scoring system, have unfortunately resulted in the gross abuse of mismatching data and feeding it into a pseudoscientific model in order to provide "e-scores." Many companies use these faulty models to determine quality of a person's candidacy for lines of credit, see C. O'Neil, 2016, pp. 104–108.

[13] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, 2016, "Machine bias", 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; A. Chouldechova, 2017, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", Big Data, Special issue on Social and Technical Trade-Offs, arXiv: 1703.00056

[14] Similar arguments have been made that AWS could make less emotional, fairer or better legal determinations about use of force than humans, particularly in stressful circumstances with limited decision time. See, for example, R.C. Arkin, 2010, "The Case for Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics*, vol. 9, no. 4, pp. 332–41; and expert presentation by Professor Mary Cummings (Duke University) at the CCW GGE on LAWS on 12 April 2018, https://www.unog.ch/80256EDD006B8954/(httpAssets)/DF486EE2B556C8A6C125827A00488B9E/$file/Summary+of+the+discussions+during+GGE+on+LAWS+April+2018.pdf.

receive scores at opposite ends of the risk spectrum.[15] These algorithmic biases arose from both inappropriate training data and interpretation errors. First, these algorithms used historical training data, including factors such as the number of past interactions with the police; number of family members who have committed a crime; and whether the individual was subsequently arrested for another crime. Thus, structural social biases in policing practices are "baked into" the training data: some groups are disproportionately surveilled or monitored by the police, so they are disproportionately represented in the training data, and that disparate attention is "learned" by the algorithm. Second, the recidivism algorithms were interpreted as outputting the likelihood of committing another crime, but they actually predicted the likelihood of being arrested. The human users thus misunderstood the algorithm outputs, and so did not include (in their decisions) their human knowledge about structural or systemic biases in arrest practices.

One must be careful, however, not to conclude "moral or legal algorithmic bias" solely from different outcomes in different groups. For example, in hopes of avoiding prejudices in members of management, an employer uses an algorithm to determine whom to hire to work in a warehouse. Such an algorithm cannot be judged simply by observing the ratio of men to women hired and noting a disparity; such an observation does not imply the algorithm was morally biased against women. In particular, if the algorithm focuses only on job-relevant attributes and requirements—for example, the ability to lift 40 kilograms—then it is presumably unbiased (since it uses only appropriate information). If those attributes occur at different rates in the different groups, then the outcome could appear to be biased even though the algorithm is morally unbiased.

In these cases, however, it is crucial that the algorithm use only appropriate or unbiased measures, else one can again have algorithmic bias (due to inappropriate training data that has biased variables). For example, many industries, including colleges and coding firms, employ seemingly standardized and objective measures of skills for hiring, admission, or promotion. Algorithms for these decisions can thus seem analogous to the warehouse hiring case: morally unbiased, even if disparate outcomes result. However, if the measures are themselves biased, then the resulting algorithms will also be biased. For example, in the case of college exams, student applicants from low-income background with less access to learning resources and tutelage will often have a worse score, due to lack of resources rather than lack of abilities. In these cases, one mitigation measure could be to intentionally introduce a statistical bias such as a "bonus score" for individuals from a low-income background in order to minimize the moral algorithmic bias (due to biased inputs). In this case, one type of bias is used to help balance another, more problematic one.

Given what we know from these existing examples, what are the potential challenges and opportunities that may arise in algorithm-dependent weapon systems?

Algorithmic biases present significant technical challenges, and equally major ethical and policy issues. The previous section outlined some of the ways that AWS can behave in surprising or unanticipated ways when one or more sources of bias is present. The system might identify unexpected targets, or find surprising routes through the battlespace. More seriously, this type of surprising behaviour calls into question what sort of human control has been or is being exercised on the system. In the international discussions on Lethal Autonomous Weapon Systems (LAWS) within the Convention on Certain Conventional Weapons (CCW), a growing number of States have acknowledged that humans should retain some form of control over increasingly autonomous weapon systems, whether that requirement is expressed in terms of having a human "in" or "on"

[15] Angwin et al, 2016.

the loop, "meaningful human control", or some other formulation.[16] All of these notions share the idea that the relevant humans should be able to predict the system's behaviour (and so endorse it in advance), or else intervene to alter the behaviour if it proceeds in undesirable or unacceptable ways. The many sources and stages of introduction of algorithmic bias for AWS imply that accurate prediction is unlikely to be possible in all cases.[17] There is a reasonable likelihood that the AWS may sometimes act in surprising ways.[18]

Moreover, these possibilities raise serious challenges for the development of trust in the expected performance of an AWS. Trust is critically necessary on the battlefield, but surprising behaviour by an autonomous system can impair the development of that trust. Most seriously, some unanticipated behaviour could threaten or violate international humanitarian law in particular contexts. This type of behaviour does not simply harm a military's ability to achieve its objectives, but represents potentially serious legal violations, depending on the nature (and possibly outcome) of the weapon's behaviour.

Finally, algorithmic biases imply a distinctive type of regulatory and policy challenge. Traditional arms control policies and regulations have tended to focus on the weapon and its intended uses. In contrast, many of the sources of bias for AWS depend critically on "historical" elements, such as the nature and source of the training data, or the use contexts intended by the development team. Policies that focus on performance in a single, known context are unlikely to provide information about AWS behaviour in novel contexts, or for alternative uses. Training of commanders and deployers is also a potential source of bias, and so is relevant to decisions about the acceptability or usability of particular weapons systems. Traditional systems of prospective and retrospective evaluation, such as Article 36 Reviews under Additional Protocol I of the CCW or internal "After Action Reviews", do not necessarily capture the information that is required to accurately judge the likelihood or origin of biased behaviour in an AWS.[19]

4 Responsibility for mitigation of algorithmic bias in weapon systems

The first step towards potential mitigation is to assess whether a given bias in the algorithm output is even problematic. If a bias is detected but is relatively minor and does not create any problem or harm, then one might decide that no mitigation is required. If the bias is not minor, then the question arises whether it ***should be*** eliminated. There are frequently trade-offs when

[16] For an overview of the concept of Meaningful Human Control, see UNIDIR, 2014, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control could move the discussion forward", UNIDIR.

[17] See section "Predictability and Reliability" in UNIDIR, 2017, "The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches", pp. 12–13.

[18] For a more detailed discussion, see UNIDIR, 2016, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies,* UNIDIR.

[19] Much has been written about the weapon review process and the specific challenges posed by increasingly autonomous technologies. See, for example, V. Boulanin and M. Verbruggen, 2017, "Article 36 Reviews: Dealing with the challenges posed by emerging technologies", *Stockholm International Peace Research Institute*; Article 36, 2016; "Article 26 Reviews and Addressing Lethal Autonomous Weapon Systems", Briefing Paper; and A. Backstrom and I. Henderson, 2012, 'New capabilities in warfare: an overview of contemporary technological developments and the associated legal and engineering issues in Article 36 weapons reviews", 2012, *International Review of the Red Cross*, vol. 94 no. 886.

eliminating biases, as that mitigation can also introduce other biases or problems. One must consider the relative costs and benefits of different options, keeping in mind that diverse societies exhibit significant variation in values. One party might judge an algorithm to be negatively biased, while another would make a different judgment. Moreover, these judgments are based on human values and goals, and so cannot be made using technology alone. As such, in order to assess the potential for problematic bias from an autonomous system, one needs to develop a comprehensive understanding of the system's likely roles in the contexts in which it will be deployed, as well as the relevant ethical, social, legal, political, and other norms. This understanding also decreases the likelihood of algorithmic biases due to transfer contexts or incorrect interpretations, as the human operators or users would then better understand the appropriate uses of the system.

If it is decided that some algorithmic bias—whether statistical, moral, legal, or other—should be mitigated, then there are several different avenues for response. The scope of the system can be reduced or its use altered in some capacity to reduce mismatch of goals and outcomes. Or, in order to maintain the overall functionality, we may attempt to redesign or re-engineer the system. Appropriate balancing and compensation can minimize or mitigate a bias, though only through careful examination of all parts of the system, whether user (human) or algorithm (machine).[20] A mitigation response can involve adjustment to any of these elements, particularly since it may be impossible to eliminate all algorithmic biases. Crucially, we are not limited to only machine-centric compensation: one option is to bring humans in or on the loop to mitigate algorithmic biases, make value-relevant judgment calls, and generally decide how the system should best proceed.[21]

Algorithmic biases cannot be entirely mitigated; arguably, the only way to eliminate all biases would be to use a random number generator with no preferences and no constraints. At the same time, the different sources of algorithmic bias point towards avenues for mitigation or trade-offs. The first question when confronting bias is which biases in a weapon system ***should* be** mitigated? In some cases, identified biases may be desirable, useful parts of an algorithm that improve performance. In other cases, one may be forced to choose between biases, and might conclude that the appropriate trade-off has been made. However, there are cases where the algorithmic biases ought to be mitigated, as they impair the system's ability to support and achieve the mission's goals and intended outcomes.

Responsibility for mitigating unwanted algorithmic biases does not rest with a single actor. A first set of actors are the ***program developers*** designing and creating the system. The developer is intimately familiar with each of the algorithms running in the system. To the extent that an undesirable bias can be mitigated through changes in the underlying algorithms or development process, then developers present a natural locus of intervention. In this way, some potential problems can be avoided before the system is fully built. At the same time, not all algorithmic biases can be addressed purely in the development stage. For example, appropriate training data might not be available, and the developers might have insufficient knowledge of deployment contexts to appropriately adjust their algorithms.

[20] For more information on the challenges of human–machine interaction in weapon systems, see Paul Scharre, 2016, *Autonomous Weapons and Operational Risk*, Center for a New American Security; and UNIDIR, 2016, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies,* UNIDIR.

[21] Of course, humans may also present their own biases; inclusion of humans is not a fail-safe solution, but only a way to lessen the impacts of algorithmic biases. Humans, for example, are often much more flexible than algorithms at absorbing and reacting to all pieces of pertinent information, possess common sense and have a greater understanding of context and salient features, even if they process information less rapidly.

The second set of key actors in potential mitigation of AWS algorithm biases are the ***acquirers*** of the technology. The agency or organization responsible for the purchase of the technology can require that the system have certain features, or meet specific, pre-defined standards. Alternately, the acquirer can require that the developers provide them with precise, detailed information about the training data, intended use contexts, and so forth. In the former case, the acquirer indicates which algorithmic biases are unacceptable, and the developer must find some way of producing such a system. In the latter case, the acquirer gains the knowledge needed to adapt practices (such as rules of engagement) to minimize the harms from the algorithmic biases that remain. In either case, acquisition and procurement teams can minimize the likelihood of algorithmic "failures" or negative biases.

The third set of potential actors in mitigation efforts are ***regulators (including international policymakers) and testers***. Regulators could decide to completely ban the development or use of AWS. Alternatively, they may decide to restrict or regulate some facet of development or use. In this case, they may determine which algorithmic biases are unacceptable, and not allow deployment of systems that exhibit those biases. They could prioritize various conditions, properties, and behaviours of a weapon system, and thereby impose particular ethical, legal, or social norms that the system must follow, though the developers are left with the task of determining how to satisfy those constraints. National or international regulators also have the ability to dictate regulatory constraints and processes that can help guide developers and future testers in their search for these or similar-acting system biases. Lastly, through testing, some algorithmic biases may be identified prior to approval and deployment, allowing for system revisions prior to the negative, real-world or real-life impacts that would impair efficacy or trust in future AWS deployment.

The fourth set of potential actors would be the ***deployers or operators*** of the system. These actors, whether at the strategic or tactical level, would make the final decisions about whether, when and where to use the weapon system, and so have the ability to mitigate algorithmic biases simply by not using the system. Alternately, if a system is used only in settings for which it was designed with appropriate training data (and all of the other conditions), then the system's potentially harmful impacts will be mitigated—though not necessarily completely eliminated.

5 Conclusions and key questions

As algorithms approach ubiquity, there is growing understanding that they are not objective and infallible. Algorithms in all domains, including military applications, can exhibit multiple types of biases that arise from different sources, such as unrepresentative training data or inappropriate transfer of the algorithm to a novel context.

Some degree of algorithmic bias may be inevitable, as it might not be possible to satisfy all relevant norms with a single process, decision, or algorithm. At the same time, algorithmic biases are not mutually exclusive, as some biases feed into one another. Moreover, not all biases are bad, as some biases can be beneficial to achieving the user's end goals. Most pointedly, algorithmic bias can arise at every stage of development and deployment, with each stage bringing its own set of considerations and possibilities for the outcome of bias.

In many cases, mitigation strategies are available, but they require careful engagement with the details of the situation, as one might not want to mitigate; or might be able to mitigate only some biases; or might address problems by changing the users or broader system; and so forth.

Various institutions and organizations are beginning to address these challenges, though policy and technical responses are still in their infancy. As a contribution to the policy response, those participating in the discussion on LAWS within the CCW framework may wish to consider the following questions about algorithmic biases in future systems:

- If governments decide to regulate increasingly autonomous weapon systems, rather than adopt an outright ban, which national or international organizations or instruments would be best placed to offer guidance or assistance to address potential algorithmic biases in AWS, including identifying possible mitigation steps?
- Given the secretive or non-transparent nature of weapon development and weapon review processes, what sorts of "best practices" can provide confidence that key algorithmic biases have been appropriately identified and mitigated?
- Are mitigation steps for algorithmic biases in particular AWS robust against possible loss of communication, interoperability challenges, or reduced human oversight?
- How would training of operators and commanders need to be adapted to ensure that they appropriately understand the algorithmic biases in an AWS, in order to maintain trust in the system and ensure its lawful use?

When recommending that the CCW's High Contracting Parties establish a Group of Governmental Experts on LAWS,[22] governments urged further consideration, *inter alia,* of the topics of responsibility and accountability, ethical and moral questions, and the military value and risks of autonomous weapon systems. A deeper understanding of the issue of algorithmic bias is fundamental to all of these topics and could add nuance—particularly on the topics of reliability and predictability—to the international discussion and proposed responses.

[22] CCW/CONF.V/2 of 10 June 2016, Annex para 4.

UNIDIR
UNITED NATIONS
INSTITUTE FOR
DISARMAMENT
RESEARCH

Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies

A PRIMER

AI-enabled systems depend on algorithms, but those same algorithms are susceptible to bias. Algorithmic biases come in many types, arise for a variety of reasons, and require different standards and techniques for mitigation. This primer characterizes algorithmic biases, explains their potential relevance for decision-making by autonomous weapons systems, and raises key questions about the impacts of and possible responses to these biases.

No. 9

UNIDIR RESOURCES