

## Exploring Bias and Accountability in Military Artificial Intelligence

Philip Alexander\*

---

### ABSTRACT

*With the evolution of machine learning, predictive risk assessment will establish itself as the standard rather than the exception. However, excessive reliance on computer software can be detrimental to a state's international human rights framework, threatening the fundamentality of customary international law. This letter highlights concerns with algorithmic detention during armed conflict, a procedure where the detainee is attributed a recidivist score by an algorithm based on the degree of threat they pose to national security. A recidivist score predicts an individual's likelihood of reoffending or committing a violent crime in the future. Accordingly, the algorithm determines whom to detain and the duration of their detention. The primary concern with predictive detention is the instability of data collection during hostility, making the entire procedure manifestly arbitrary. Furthermore, algorithms are pre-set with data inputted by humans. As a result, there will always be room for human error or discriminatory biases that affect its decision-making. These concerns require immediate redress before military-owned algorithms for the purpose of security detention enter the mainstream.*

---

\* BA LLB (West Bengal National University of Juridical Sciences, Kolkata) '24. The author warmly thanks Professor Atul Alexander for his motivation and support.

Dear Editor,

The interaction between artificial intelligence and data systems has recently gained momentum across several countries such as the United States, the United Kingdom, India and China, developing software that has completely automated the human decision-making process within the criminal justice system.<sup>1</sup> Predictive policing algorithms, such as ‘PredPol’, ‘HunchLab’ and ‘Patternizer’, tabulate estimates of areas with higher crime rates using existing data, allowing the police to control crime more efficiently.<sup>2</sup> Algorithmic programs also ensure the smoother operation of military activity. Algorithmic detention is the process of employing risk assessment instruments to determine whom to detain and the duration of their detention during armed conflict.<sup>3</sup> Although predictive assessment bridges the procedural roadblocks to justice by assisting in decision-making and efficient resource allocation, the possible reliance on biased data and the lack of transparency make algorithmic detention software a dangerous tool in the wrong hands. The central argument in this letter highlights the arbitrariness of algorithmic detention under international human rights law (IHRL) and further proposes a regulatory framework that shifts accountability to the operators of automated decision-making systems used in international and non-international armed conflict.

### PROCEDURAL CONCERNS

Civilians may be subjected to detention during armed conflict for reasons of threat to security and safety measures. For international armed conflict (IAC), Article 5 of the Fourth Geneva Convention permits the detention of protected

---

<sup>1</sup> Odhran James McCarthy, ‘AI & Global Governance: Turning the Tide on Crime with Predictive Policing’ (*Centre for Policy Research*, 26 February 2019) <<https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html>> accessed 21 October 2021.

<sup>2</sup> Alex Chohlas-Wood, ‘Understanding risk assessment instruments in criminal justice’ (*Brookings*, 19 June 2020) <<https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/>> accessed 28 September 2021.

<sup>3</sup> Hannah Kannegieter, ‘Algorithmic Detention and International Human Rights Law’ (*Harvard Human Rights Journal*, 19 February 2021) <<https://harvardhrj.com/online/page/2/>> accessed 28 September 2021.

individuals who have ‘engaged in activities hostile to the security of the State.’<sup>4</sup> For non-international armed conflict (NIAC), detention is governed by Common Article 3 of the Geneva Convention and Article 5 of Additional Protocol II.<sup>5</sup>

State practice has established the prohibition against the arbitrary deprivation of liberty as customary international law for both international and non-international armed conflict.<sup>6</sup> The International Covenant on Civil and Political Rights (ICCPR) mandates due process, which entitles the defendant to trial before a judge and proceedings before a court.<sup>7</sup> Although the terms ‘judge,’ ‘trial’ and ‘court’ remain undefined under the ICCPR, the detention of an individual on the grounds of a decision rendered by computer software raises questions on its conformity with the procedural requirements of Article 14 of the ICCPR, which states that ‘everyone shall be entitled to a fair and public hearing by a competent, independent and impartial tribunal established by law.’<sup>8</sup> The United Nations Human Rights Committee in *Marques de Morais v Angola* has further clarified that detention must not only be lawful under domestic legislation but also ‘reasonable and necessary in all circumstances.’<sup>9</sup> For an algorithm to determine the need for detention, it must objectively state that the accused would be a threat to national security while satisfying the tests of necessity and reasonableness.

---

<sup>4</sup> Geneva Convention Relative to the Treatment of Prisoners of War (adopted 12 August 1949, entered into force 2 November 1950) 75 UNTS 135 (Geneva Convention) art 5.

<sup>5</sup> Kevin Jon Heller, ‘What Exactly is the ICRC’s Position on Detention in NIAC’ (*OpinioJuris*, 6 February 2015) <<http://opiniojuris.org/2015/02/06/exactly-icrcs-position-detention-niac/>> accessed 27 January 2022.

<sup>6</sup> ICRC, Customary IHL Database <[https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1\\_rul\\_rule99](https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule99)> accessed 27 January 2022.

<sup>7</sup> International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR) art 9.

<sup>8</sup> *ibid* art 14.

<sup>9</sup> UN Human Rights Committee, *Rafael Marques de Morais v Angola* (29 March 2005) UN Doc CCPR/C/83/D/1128/2002.

## DISCRIMINATION IN AI

The element of impartiality in Article 14 is crucial in determining the legality of algorithmic sentencing software. A fair trial precludes the judge from the influence of external bias, prejudice or preconceptions on the case presented before the court.<sup>10</sup> The possibility for the external manipulation of the algorithm's input variables compromises the impartiality of the court because it creates the opportunity for external bias to influence the judge's decision-making. If the input of data were flawed in the algorithm, where for example, the defendant's ethnicity constituted 90 per cent of the decision to determine their threat to national security, the judge would continue to rely on such data to arrive at a decision under the garb of 'objectivity', regardless of its constitutionality. This is because neither the judge nor the defendant would be able to question the mode in which the algorithm rendered such results.<sup>11</sup> The lack of knowledge on the intricacies of the algorithm's functioning creates a strong dependence on the objectivity of artificial intelligence. This is referred to as 'mathwashing', where programs that employ AI are idolised solely because Mathematics is at the core of their operation.<sup>12</sup> This approach is flawed because algorithmic bias is well-documented in the public domain.

There are numerous examples where artificial intelligence has misperceived information based on sex and race. For example, algorithmic bias is rampant in artificial intelligence employed in the healthcare sector.<sup>13</sup> A study published in 2019 concluded that an algorithm developed for patients, Medicare, was less

---

<sup>10</sup> UN Human Rights Committee, 'General Comment No. 32, Article 14: Right to equality before courts and tribunals and to a fair trial' (23 August 2007) UN Doc CCPR/C/GC/32.

<sup>11</sup> Caoimhe Anderson, 'The Impact of Algorithms in Criminal Sentencing on Due process Rights' (JD thesis, Queens University Belfast 2019).

<sup>12</sup> Nanette Byrnes, 'Why We Should Expect Algorithms to Be Biased' (*MIT Technology Review*, 24 June 2016) <<https://www.technologyreview.com/2016/06/24/159118/why-we-should-expect-algorithms-to-be-biased/>> accessed 27 January 2022.

<sup>13</sup> Katherine Igoe, 'Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It' (*Harvard T.H. Chan School of Public Health*, 12 March 2021) <<https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/>> accessed 10 October 2021.

likely to refer an equally sick black patient over a white patient.<sup>14</sup> Another study conducted by MIT found that the margin of error for three commercial facial recognition software programmes was 0.8% for white men but up to 35% for women of colour.<sup>15</sup> Such biases could, either maliciously or unintentionally, be incorporated into algorithm detention software, influencing the security detention of a civilian during wartime. Considering the post-9/11 situation with the illegal wiretapping and surveillance of Muslim citizens under the PATRIOT Act in the United States,<sup>16</sup> among several other examples of arbitrary laws that incorporate discriminatory biases, the military could unlawfully target certain groups of people, making algorithmic detention highly discriminatory.

The central argument for objective scientific reasoning is the elimination of the biases that are inherent in human judgement,<sup>17</sup> thus reducing the margin of error to negligible values. However, an example of the unreliability of computer software can be seen through the conviction of eighteen-year-old American teenager, Brisha Borden. In 2014, Brisha was arrested for burglary and petty theft. The COMPAS software classified Borden as 'high risk' for committing a violent crime in the future. In comparison, Vernon Prater had multiple criminal charges to his name and was arrested for a crime similar to Borden. Prater was classified as 'low risk' of reoffending.<sup>18</sup> As of today, Borden has been released from prison with no pending criminal charges whereas Prater

---

<sup>14</sup> Heidi Ledford, 'Millions of black people affected by racial bias in health-care algorithms' (*Nature*, 24 October 2019) <<https://www.nature.com/articles/d41586-019-03228-6>> accessed 8 October 2021.

<sup>15</sup> Larry Hardesty, 'Study finds gender and skin-type bias in commercial artificial-intelligence systems' (*MIT News*, 11 February 2018) <<https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>> accessed 9 October 2021.

<sup>16</sup> Shrin Sinnar, 'Applying 9/11 laws to domestic terrorism could hurt minorities more than white supremacists' (*USA Today*, 11 September 2019) <<https://www.usatoday.com/story/opinion/2019/09/11/reform-9-11-terrorism-laws-dont-expand-to-white-supremacists-column/2258400001/>> accessed 9 October 2021.

<sup>17</sup> Aleš Završnik 'Algorithmic justice: Algorithms and big data in criminal justice settings' (2019) 18 *European Journal of Criminology* 623, 637.

<sup>18</sup> Giesecke+Devrient, 'Doing the right thing: the ethics of AI' (*Giesecke+Devrient*) <<https://www.gi-de.com/en/spotlight/digital-infrastructures/ethics-of-ai>> accessed 26 January 2022.

has returned to prison and is serving an eight-year sentence.<sup>19</sup> It is relevant that Brisha Borden is a black woman and Vernon Prater is a white male. States have the obligation to ensure that fundamental freedoms and entitlements are enjoyed without distinction on birth, national, ethnic or social origin, language, religion, economic condition, political or other opinion, gender, sexual orientation, disability or other status.<sup>20</sup> Any detention or arrest made on the grounds above constitute an arbitrary deprivation of liberty under the ICCPR.

### DATA INSTABILITY

Another significant concern with algorithmic detention is the instability of data collection.<sup>21</sup> Presently, the development of algorithmic software is predominantly associated with criminal sentencing.<sup>22</sup> If similar algorithms used in criminal sentencing are applied to armed conflicts, it may produce varied results. This is described as the ‘portability trap’, where the application of machine learning designed for one social context will not necessarily translate into a similar application for another social context, resulting in potentially inaccurate or misleading outcomes.<sup>23</sup> For example, data sets developed using factors that determine the ‘dangerousness’ of a criminal will differ in federal and military contexts.<sup>24</sup> State law enforcement recognises what constitutes criminal behaviour using objective criteria such as age, employment and past convictions. The military will likely not have access to such data sets, unlike the judiciary which derives its data from factors that are easily retrievable from government databases,<sup>25</sup> creating structural barriers to the collection of sufficient data.

Additionally, there are several constraints on the collection of quantitative data during armed conflict. Empirical observation in conflict zones is considered to produce accurate results if accompanied by foresight and

---

<sup>19</sup> *ibid.*

<sup>20</sup> ICCPR (n 7) art 26.

<sup>21</sup> Ashley Deeks, ‘Detaining by Algorithm’ (*Humanitarian Law and Policy*, 25 March 2019) <<https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/>> accessed 1 October 2021.

<sup>22</sup> *ibid.*

<sup>23</sup> *ibid.*

<sup>24</sup> *ibid.*

<sup>25</sup> *ibid.*

planning.<sup>26</sup> However, the methodological challenges associated with quantitative data collection during armed conflict threaten its accuracy and reliability. Hostility is linked with increased anxiety, fear and perception of threats, causing individual and household participants to be apprehensive of survey participation.<sup>27</sup> Violence and political instability in conflict zones may further reduce the participants' likelihood of welcoming researchers into their homes,<sup>28</sup> resulting in greater non-response bias, which may severely impede the integrity of data collection. These concerns risk yielding inaccurate results which may not objectively satisfy the grounds of detention laid down in *Marques de Morais v Angola* - necessity and reasonableness. For a decision for algorithmic detention to be legally tenable, the software must thus establish with certainty that the accused will flee, tamper with the evidence or reoffend. When detention is permitted only 'when absolutely necessary or for imperative reasons of security, the reliance on a system that uses predictive data analysis accompanied by the above concerns of unstable data collection could potentially result in arbitrary decisions that endanger an individual's right to liberty.'<sup>29</sup>

### THE WAY FORWARD

The solution to the above concerns is establishing a framework that creates an effective governance structure for detention algorithms during armed conflict. Accountability must be enforced to ensure that a decision for security detention made by computer software is not on arbitrary grounds. This is known as algorithmic accountability.<sup>30</sup> Under this proposition, statutory

---

<sup>26</sup> Roos Haer and Inna Becher, 'A methodological note on quantitative field research in conflict zones: get your hands dirty' (2012) 15 International Journal of Social Research Methodology 1.

<sup>27</sup> William Axinn, Dirgha Ghimire and Nathalie Williams, 'Collecting Survey Data during Armed Conflict' (2012) 28 Journal of Official Statistics 153.

<sup>28</sup> *ibid.*

<sup>29</sup> Tess Bridgeman, 'The viability of data-reliant predictive systems in armed conflict' (*Humanitarian Law and Policy*, 8 April 2019) <<https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention/>> accessed 21 February 2022.

<sup>30</sup> Robyn Caplan and others, 'Algorithmic Accountability: A Primer' (*Data & Society*, 18 April 2018) <<https://datasociety.net/library/algorithmic-accountability-a-primer/>> accessed 13 October 2021.

restrictions shift accountability onto the developers of Artificial Intelligence-operated decision-making systems.<sup>31</sup> For example, the developers of predictive risk assessment software must be required by law to grant the defendant access to the data used for detention as well as providing transparency on how the software operates. Furthermore, the use of predictive risk scores as evidence must be accompanied by expert testimony. Expert testimony provides the defendant the opportunity for cross-examination in contrast to the blanket reliance on the objectivity of the algorithm. Finally, external auditing from government regulators could be implemented to maintain the integrity of the algorithm while ensuring its compliance with international human rights standards.<sup>32</sup>

In 2019, the United States Senate introduced a bill known as the Algorithmic Accountability Act which regulates automated decision-making systems and ensures the ethical application of Artificial Intelligence.<sup>33</sup> The proposed law requires companies of automated systems to study the algorithms developed, and to identify and fix discriminatory outcomes or bias.<sup>34</sup> This could be achieved through the statutory imposition of impact assessments that measure the implications of an algorithm and the significance of the changes it creates on the broader social environment. The Act emphasised the importance of privacy, non-discrimination and compliance with procedural laws.<sup>35</sup> Similar accountability statutes may mandatorily impose statutory restrictions such as impact assessments during the development and deployment stages of an algorithm's life cycle, essentially building a framework for algorithmic

---

<sup>31</sup> Lorna McGregor, Daragh Murray and Vivian Ng, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68 *International and Comparative Law Quarterly* 309.

<sup>32</sup> Anderson (n 11).

<sup>33</sup> Adi Robertson, 'A new bill would force companies to check their algorithms for bias' (*The Verge*, 10 April 2019) <<https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate>> accessed 9 October 2021.

<sup>34</sup> Jones Day, 'Proposed Algorithmic Accountability Act Targets Bias in Artificial Intelligence' (*Jones Day*, June 2019) <<https://www.jonesday.com/en/insights/2019/06/proposed-algorithmic-accountability-act>> accessed 20 February 2022.

<sup>35</sup> McGregor and others (n 31).



governance among manufacturers of detention software. This allows the software developers to incorporate and monitor the protection of human rights standards into the algorithm's life cycle. These statutory restrictions create a system of checks and balances where due process is cemented as an unalienable right under algorithmic detention.

Additionally, IHRL must serve as a framework in developing stricter legal requirements for manufacturers of artificial intelligence by defining and assessing 'harm' comprehensively in order to eliminate any ambiguity with the law. In the present debates surrounding algorithmic accountability, harm is described using abstract terms such as 'unfairness' and 'bias.'<sup>36</sup> Such vague descriptors make it challenging to understand the human rights implications and the extent of legal obligations imposed on the operators of algorithmic software.<sup>37</sup> Bias in algorithmic detention software could be statistical, automated or algorithmic. However, it would be difficult to pinpoint which biases make the algorithm unlawful. IHRL resolves this issue by providing a concrete assessment of harm through stating how and when such biases are unlawful.

For example, Article 1 of the International Convention on the Elimination of All Forms of Racial Discrimination establishes a conclusive and comprehensive definition of racial discrimination.<sup>38</sup> Similarly, concepts such as 'arbitrariness' and 'liberty' could be defined using pre-existing international human rights instruments, which could subsequently be incorporated into domestic legislation governing algorithmic accountability. Although stringent legal compliance may be viewed as an impediment to scientific advancement, it is a necessary precondition for the lawful social integration of Artificial Intelligence when the AI revolution could potentially infringe the rights of millions. Thus, the IHRL approach to harm assessment prevents the unrestricted application of algorithmic detention because clear-cut definitions trigger a pre-existing framework of rights and liabilities.

---

<sup>36</sup> *ibid.*

<sup>37</sup> *ibid.*

<sup>38</sup> International Convention on the Elimination of All Forms of Racial Discrimination (opened for signature 21 December 1965) 660 UNTS 195 (ICERD) art 1.

In conclusion, the resolution of the discrepancies within algorithmic detention requires the development of a legal framework that establishes algorithmic accountability on corporate entities that develop the technology used for detention purposes. The enactment of domestic statutory legislation for algorithmic detention using IHRL will ensure that algorithmic decisions for detention during armed conflict will not exceed the scope of constitutionality. Furthermore, ascribing liability and accountability to the manufacturers of detention software and the military serves to deter from the multitude of biases that creep into algorithmic detention. Unless stricter measures are adopted to restrict its operation, algorithmic detention could potentially be exploited to deny innocent civilians trapped in hostility their right to life and liberty. Thus, it is urged that the intersection between artificial intelligence and the law should be regulated by domestic and IHRL frameworks to prevent its misuse.

Yours faithfully,  
Philip Alexander