HUMANITARIAN LAW & POLICY

The risks and inefficacies of AI systems in military targeting support

September 4, 2024, Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / IHL / New Technologies / Special Themes

12 mins read

Jimena Sofía Viveros Álvarez
Lawyer; Member of UNSG High-Level Advisory Body on AI; Commissioner for GC-REAIM; OECD AI Expert

Over the past decade, discussions surrounding artificial intelligence (AI) in the military domain have largely focused on autonomous weapon systems. This is partially due to the ongoing debates of the Group of Governmental Experts on Lethal Autonomous Weapons Systems of the Convention on Certain Conventional Weapons. While autonomous weapon systems are indeed a pressing concern, the critical reality is that AI is hastily deployed to gather intelligence and, even more worrisome, to support militaries to select and engage targets.

As AI-based decision support systems (AI DSS) are increasingly used in contemporary battlefields, Jimena Sofía Viveros Álvarez, member of the United Nations Secretary General's High-Level Advisory Body on AI, REAIM Commissioner and OECD.AI Expert, advocates against the reliance on these technologies in supporting the target identification, selection and engagement cycle as their risks and inefficacies are a permanent fact which cannot be ignored, for they actually risk exacerbating civilian suffering.

AI's military use-cases extend beyond its applications over the use of force or granting autonomy to weapons systems, such as its incorporation into *command and control, information management, logistics and training*. Of particular concern, *AI DSS in targeting* can propose military objectives and give actionable recommendations to its (human) operators, differing from *autonomous weapons systems* which can engage their targets on their own.

In fact, policy makers in the *US*, *UK*, and the *North Atlantic Treaty Organization*, have published defense strategies that *evince that data, not robotics or lethal autonomy, will be the critical enabler in the coming decade*.

They argue that AI DSS could produce actionable intelligence to "*improve*" targeting. Nonetheless, *AI systems will only be as effective as their training data*, as it *provides their underlying specifications*. Moreover, AI's inherent unpredictability gives rise to many issues, challenging international humanitarian law

(IHL) as a whole, and the *black-box* problem of AI makes it impossible for us, humans, to properly understand the decision-making process of these systems. Thus, it is imperative to maintain *data quality and provenance*, as well as preserving *human judgment* in systems capable of selecting and engaging targets, which currently is *very difficult, if not unattainable*.

How does AI DSS affect targeting?

Targeting is broadly defined as *the use of force by warring parties against individuals or objects outside of their control*. Under IHL, target selection and engagement must follow the fundamental principles of *distinction*, *proportionality*, *humanity* and the requirement to take all feasible *precautions*.

However, the "*fog of war*", *i.e.,* the perplexity of maintaining situational awareness in the battlefield, has *increased in the digital age* due to *information operations becoming more common*, alike irregular warfare and combat in *urban areas*. Therefore, in order to protect the civilian population, rules of engagement require *positive identification* (reasonable certainty that the proposed target is legitimate), as part of the necessary precautions in any attack.

Promotors of AI systems for military targeting support, often contend that this technology will "*enhance" military precision*. Albeit, we must be wary of such claims as the incorporation of AI in the military domain *introduces uncertainties on the provenance of data and the reliability of the information provided by the system*, which is particularly perilous when the intelligence it yields becomes actionable for targeting.

Deficient training data will often lead to *biases*, *brittleness*, *hallucinations*, *misalignments*, *privacy risks* and *loss of control*, which paired with the technology's *unpredictability*, pose sundry problems for their use in target identification, selection and engagement.

Lack of diversity in datasets may cause AI systems to single out, for example, members of distinct *ethnic groups* as targets, or even *consider all civilian males as combatants due to encoded gender biases*.

AI's brittleness, *i.e., its incapacity to adapt and perform correctly when presented with data it had never seen before*, may lead to unintended outcomes, *e.g.,* the unlawful use of one school bus or ambulance by enemy combatants could trigger the system to consider *all* school buses and ambulances as legitimate objectives.

"*Hallucinations*" are nonsensical or inaccurate outputs that occur when AI systems perceive patterns or objects that are nonexistent. When paired with any unexpected changes in the battlefield, this may cause the system to perceive patterns where there is none, prompting the targeting of innocent civilians.

Moreover, *misalignments*, which refers to AI hierarchizing a prompt or command over important values or constraints, may be pervasive throughout the targeting process, as it could prioritize the "goal" to eliminate enemy combatants regardless of any incidental and/or disproportionate harm to civilians. It is noteworthy, that *AI's "forecasting" is based on data analytics of past behaviors*, thus lacking context and human logic. This risks rendering them unable to properly scrutinize whether a proposed objective can be legitimately targeted given *the circumstances reigning at the specific time*, as duly required by IHL.

Additionally, it has been widely reported that *states are focusing their resources to implement AI into their defense strategies*. Therefore, the world's major military drivers are *investing hefty amounts of resources* to dominate this field, *igniting a de facto arms race*.

Furthermore, Big Tech leaders are also *actively selling their technologies to states' armed forces*. This raises several concerns, as the models being sold are likely already trained with previously collected *personally identifiable information* (*name, address, telephone, or social security numbers*), or worse, biometrics (*data related to the physical, physiological and/or behavioral characteristics unique to a person*), which could then be (mis)used as intelligence to "optimize" AI-military systems targeting functions. This would detonate a generalized risk to all populations and a massive *breach of individuals' privacy*.

While *many states*, alike the *EU AI Act*, have strict regulations regarding the *processing and use of personal data*, banning the use of real-time remote biometric identification systems in public spaces for law enforcement purposes. However, there is no such regulation during an armed conflict since *the 1949 Geneva Conventions were not developed with data or, even less so, biometrics in mind*, resulting in a *lacuna regarding the lawful use and processing of data during armed conflicts*. Notwithstanding, the aforementioned risks of data processing by military AI-DSS, inevitably extends beyond armed conflicts, as their lifecycle may well initiate even before its detonation, *e.g.* data gathering or system training, and their consequences may linger even after its resolution. Therefore, the author believes that *international human rights law and IHL are not mutually exclusive* regimes in this regard.

In any case, militaries may argue that "benefitting" from live access to biometrics, *could augment reliability on the identification and targeting of enemy combatants*. However, AI systems are immutably fallible due to brittleness, hallucinations and misalignments, and likewise vulnerable to hacking and *adversarial attacks*, *i.e.*, inputs designed to trick AI into submitting incorrect predictions, such as *misclassifying civilian infrastructure as legitimate military objectives*.

Moreover, AI systems are crescively being trained with *synthetic data*, *i.e.,* computer-generated data, as opposed to real-life data, therefore *often lacking accuracy*, since it merely mimics the latter. At the same time, *AI-created synthetic datasets are progressively being used to train new models*. This poses novel and accentuated inexplicability and unpredictability pitfalls, as we become oblivious to what the initial system feeds into the subsequent one, resulting in a *black-box* within a black-box, *i.e.* "black-box$_2$".

AI-supported targeting: enhancing precision or increasing civilian casualties?

Despite all the described risks, AI-DSS are still being deployed in today's battlefields including being used *in target identification, selection and engagement*, which inevitably risks exacerbating civilian suffering, especially when used without the *necessary human judgment*.

To begin with, the use of AI throughout the targeting process relies on *predictions* based on *pattern recognition and classification*, through the *generalization* of data used during the system's training. Hence, its ability to *recommend* targets depends on its capacity to "spot" similarities between the *available and circumstantial* information on the population, and the data it was trained with. This raises a plethora of legal concerns because AI systems *will never be perfect*, and thus will always *be prone to "failure"*, especially when facing *complex real-life* battlefields. This, despite developers' best efforts, simply cannot be pre-designed in a laboratory as there are endless possible scenarios within the "fog of war".

Hypothetically, an AI DSS could be trained and used to identify and locate collaborators of known enemy commanders by monitoring their activities not only through military intelligence surveillance and reconnaissance, but also *via* social media connections, photographs, intercepting communications, or even frequented sites. Thereafter, the system may also provide specific actionable recommendations, such as bombing a building inhabited by their targets.

The problem is that AI systems could misclassify individuals as "targets" if they *perceive* linkage to the adversary combatants, however remote or irrelevant. For example, they could have simply studied in the same school, have a mutual connection or worse, *invent inexistent patterns*, prompting the targeting of innocent civilians.

Although the final determination to use force is made by humans, these AI DSS recommendations will *very likely alter their decision-making process*, as military personnel "*typically privilege action over non-action in a time-sensitive human-machine configuration*" without thoroughly verifying the system's output, which is known as "*automation bias*".

Furthermore, AI's warped *speed and scalability*, enables unprecedented "mass production targeting", heightening the risk of *automation bias* by the human operators, reducing any form of *meaningful human control*, *human-machine teaming* or *human cognitive autonomy*, to *merely pressing a button*.

Considering all of the above, AI usage in targeting decisions *could impact far more lives than even autonomous weapons systems*, especially in urban areas. Yet, militaries often justify civilian casualties as collateral damage, claiming that *they "believed" the system was reliable*, *avoiding accountability*.

Nevertheless, any self-advertised "success or failure rate" could be *deceptive*, due to the exponential velocity with which some AI systems learn based on dynamic data flows that *constantly shift these fluctuating "percentages"*.

Even *Article 36* of Additional Protocol I's requiring State Parties to conduct *legal review of new weapons*, *may not be sufficient to prevent the intricacies of these technologies from violating IHL* because they were not envisaged to validate such auspicating simulations, especially not *ex post facto,* as these systems are being *field-tested ex ante*. In any case, *AI-DSS should absolutely be subjected to these legal reviews* with the same rigor as any other weapon, especially those they intend to "boost", as the core objective of their incorporation into any weapon system is to ultimately "enhance" their autonomous capabilities by enabling in a *novel* way their critical functions of targeting and have an impact on the decisions to apply force on various levels.

Further, armed forces frequently *do not disclose the specifics of their systems*, including the data used to fine-tune and/or train them, arguing national security grounds, rendering futile any monitoring or oversight attempts.

Even if it was possible to corroborate an alleged low margin of error, in terms of reliability and/or predictability, AI's capacity to *provide targets at an unprecedented rate*, would still result in thousands of civilian lives at risk. So, we must ask ourselves – could we condone these killings being the product of a "glitch" or a "hallucination"?

Conclusion

As an intrinsically *dual-use and repurposable technology*, AI systems are frequently employed by different state and non-state actors, which escalates concerns about their impact in the broader peace and security domain, in addition to their applications in the military sector.

AI's revamped targeting capabilities are unchartered waters and despite any alleged advantages, these technologies lack appropriate testing and review standards prior to their deployment. Moreover, being *the result of human activity*, these systems are also fallible since they are inherently unpredictable and unexplainable, the data used to train them will never be perfect, and their inherent risks will only be exacerbated if the increasing use of AI-DSS can lead to *automation bias* in the battlefield. This is why *effective human judgment* must be a hard requirement prior and throughout the entirety of any military operation.

Thus, the unlawful or irresponsible use of AI to support military targeting risks resulting in devastating denouements, led the UN Secretary General to state that "*[n]o part of life and death decisions which impact entire families should be delegated to the cold calculation of algorithms*".

As Pope Francis pronounced earlier this year, "*[w]e need to ensure and safeguard a space for proper human control over the choices made by artificial intelligence programs: human dignity itself depends on it*".

The author steadfastly stands with, and advocates for, both of these calls for action, to the international community at large.

See also:

- Ingvild Bode, Ishmael Bhila, *The problem of algorithmic bias in AI-based military decision support systems*, September 3, 2024
- Wen Zhou, Anna Rosalie Greipl, *Artificial intelligence in military decision-making: supporting humans, not replacing them*, August 29, 2024
- Ingvild Bode, *Falling under the radar: the problem of algorithmic bias and military applications of AI*, March 14, 2024
- Ruben Stewart & Georgia Hinds, *Algorithms of war: The use of artificial intelligence in decision making in armed conflict*, October 24, 2023

Tags: AI, algorithmic bias, armed conflict, artificial intelligence, autonomous weapons, AWS, conduct of hostilities, IHL

You may also be interested in:

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

13 mins read

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

Ingvild Bode & Ishmael Bhila

Algorithmic bias has long been recognized as a key problem affecting decision-making processes that integrate ...

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

13 mins read

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

Wen Zhou & Anna Rosalie Greipl

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