HUMANITARIAN LAW & POLICY

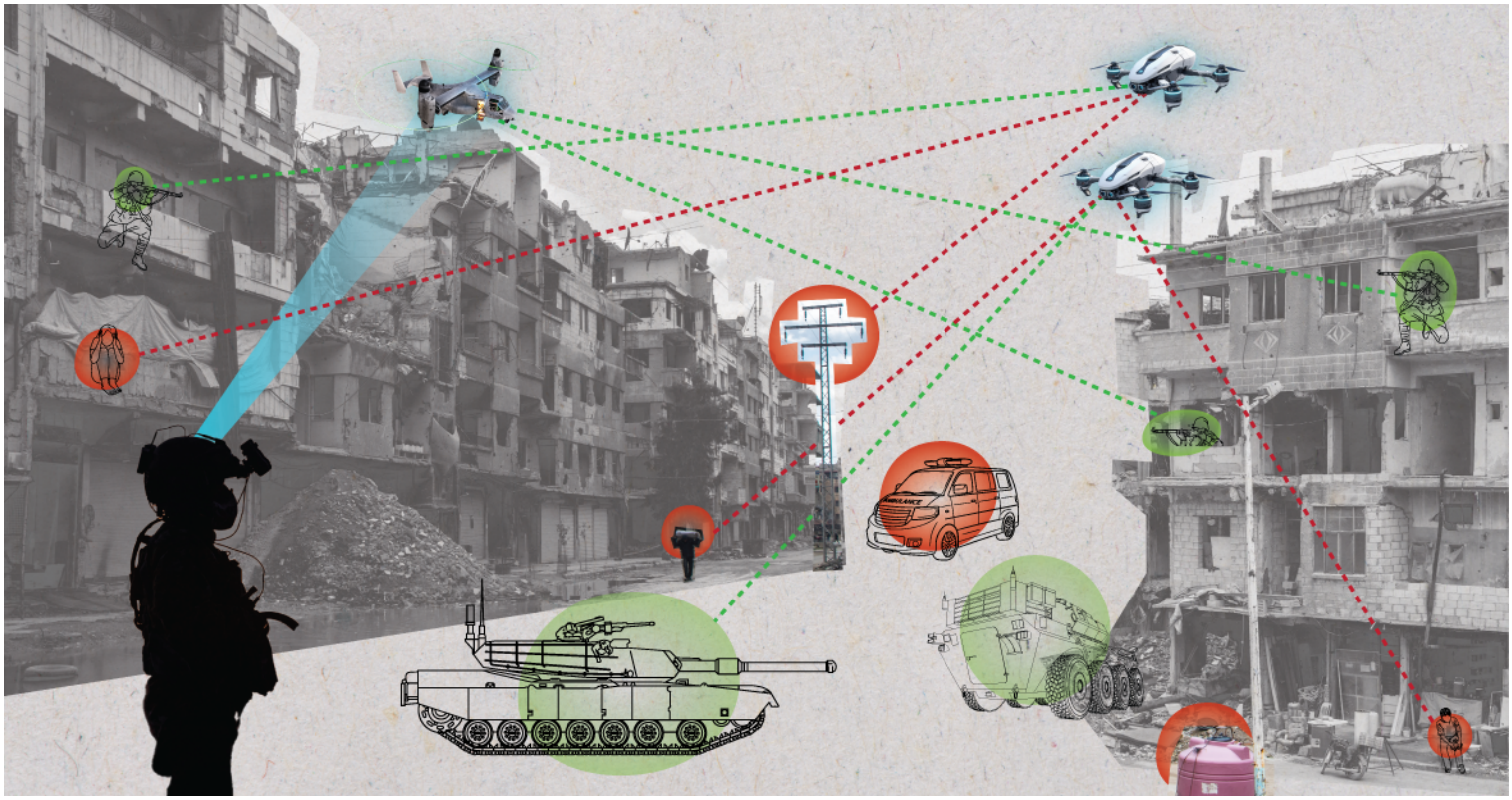
Algorithms of war: The use of artificial intelligence in decision making in armed conflict

October 24, 2023, Artificial Intelligence and Armed Conflict / Autonomous Weapons / Humanitarian Action / New Technologies / Technology in Humanitarian Action

14 mins read


Ruben Stewart
Military and Armed Group Adviser, ICRC


Georgia Hinds
Legal Adviser, ICRC



In less than a year, Chat-GPT has become a household name, reflecting astonishing advances in artificial intelligence-powered software tools, especially generative AI models. These developments have been accompanied by frequent forecasts that AI will revolutionise warfare. At this stage of AI development, the parameters of what is possible are still being explored, but the military response to AI technology is undeniable. China's white paper on national defense promoted the theory of the "intelligentization" of warfare, in which leveraging AI is key to the PLA's modernization plan. The director of the US Cybersecurity and Infrastructure Security Agency Jen Easterly warned that artificial intelligence may be the "most powerful weapon of our time." And whilst autonomous weapon systems have tended to dominate discussions about AI in military applications, less attention has been paid to the use of AI in systems that support human decisions in armed conflicts.

In this post, ICRC Military Adviser Ruben Stewart, and Legal Adviser Georgia Hinds seek to critically examine some of the touted benefits of AI when used to support decisions by armed actors in war. In particular, the areas of civilian harm mitigation and tempo are discussed, with a particular focus on the implications for civilians in armed conflict.

Even before recent hype, you have probably already used AI in various forms, indeed you might be reading this article on a device largely powered by AI. If you have used a fingerprint or face to open your phone, participated on social media, planned journeys using a phone application or purchased anything

online from pizzas to books, it has probably involved AI. In many ways we have grown comfortable with AI, adopting it, often unwittingly, into our everyday life.

But what if that facial recognition software was used to identify a person to be attacked? What if, instead of finding the cheapest flight to get you to a destination, software along similar lines was finding aircraft to perform an airstrike on a target. Or, rather than recommending the best pizza place or the closest available taxi, the machine was recommending plans of attack? This is apparently a reality that is 'coming soon' from *companies developing AI-based decision platforms for defense purposes.*

These kinds of AI decision support systems (AI-DSS) are computerised tools that use AI software to display, synthesise and/or analyse data and in some cases make recommendations – even predictions – in order to aid human decision-making in war.

The advantages of AI-DSS are often framed in terms of increased situational awareness and faster decision-making cycles. These claims are unpacked below, in light of both AI system and human limitations, and in the context of the planning processes of modern conflicts.

Minimising risk of harm to civilians in conflict

The advent of new technologies in warfare is often accompanied by assertions that its integration will reduce civilian harm (though this is *not always borne out in practice*). In the case of AI-DSS, it has been *claimed* that such tools could help to better *protect civilians in conflict in certain circumstances*. Certainly, *international humanitarian law (IHL)* obliges military commanders and others responsible for attacks to base their decisions on information from all sources available to them at the relevant time. In the context of urban warfare in particular, the ICRC has *recommended* that information about factors such as the presence of civilians and civilian objects should include open-source repositories such as the internet. Further, specifically considering AI and machine learning, the ICRC has concluded that, to the extent that *AI-DSS tools can facilitate quicker and more widespread collection and analysis* of this kind of information, they could well enable *better decisions in conflict by humans that minimize risks for civilians*.

At the same time, any AI-DSS output *should be cross-checked between multiple sources to guard against biased or inaccurate information*. Whilst this is true of any information source in conflict, it is especially important for AI-DSS; as the ICRC *has previously outlined*, it may be extremely difficult – if not impossible at times – to verify the accuracy of output due to the system's functioning and the way in which human users interact with machines. These aspects are expanded below.

System limitations

Recent coverage of AI developments has often included examples of AI failing, sometimes fatally. Examples include software *not recognising or misidentifying people with darker skin colour*, *recommending travel routes that do not take into account updated road conditions*, and examples of fatalities resulting from *self-driving cars*. Some of these failings are explainable, but not excusable, for example, because the data upon which it bases its outputs is *biased*, corrupted, *poisoned* or just plainly incorrect. These systems *can also still easily be 'tricked'; techniques can be used to fool the system into misclassifying data*. For example, adversarial techniques could *conceivably be used in conflict to affect a targeting assistance system's source code such that it identifies school-buses as enemy vehicles*, with devastating consequences.

As AI is used for more complex tasks, and especially when multiple layers of analysis (and possibly decisions and judgements) accumulate, then verifying that final output and the source of any errors that contributed to the final output becomes almost impossible. With increasingly complex systems, the potential for compounding errors increases – *a small inadequacy in the first algorithmic recommendation is fed into and skews a second algorithmic process, which feeds into a third, and so on*.

Accordingly, AI systems have often exhibited behaviours that cannot be explained by the user or the developer, even after extensive post facto analysis. *A study of the high-profile large language model GPT-4* found that its ability to solve a math problem drastically, and inexplicably, reduced from 83.6% to just 35.2% three months later. *Unpredictable behaviours can also arise through reinforcement learning*, where machines have proven very efficient at adopting and concealing unforeseen, and sometimes negative behaviour to outwit or outplay humans: be that *lying to win a negotiation* or *taking shortcuts to beat a computer games*.

Challenges for humans interacting with the machine

AI-DSS do not "make" decisions. However, they do directly and often significantly influence the decisions of humans, including due to humans' cognitive limitations and tendencies when interacting with machines.

For example, "automation bias" refers to the tendency of humans to not critically challenge a system's output, or search for contradictory information – *especially in time critical situations*. This has already been observed in other fields such as healthcare, where the *accuracy of diagnoses by experienced radiologists was adversely influenced by false AI outputs*.

Inaccurate diagnoses in health settings can be fatal. And so too, in armed conflict over trust can have lethal consequences. In 2003, the US defensive Patriot system twice fired at friendly coalition aircraft based on them being misclassified as attacking missiles. In a subsequent investigation, one of the major shortfalls identified was that "*operators were trained to trust the system's software*."

These ways of functioning, coupled with these characteristics of human-machine interaction, potentially increase the likelihood of outcomes that diverge from the intention of the human decision-maker. In warfare, this may result in *accidental escalation*, and in any event will heighten *risks for civilians and protected persons.*

Tempo

One touted military advantage of AI is the increase in tempo of decision-making it would give a user over their adversary. Increased tempo often creates additional risks to civilians, which is why techniques that reduce the tempo, such as 'tactical patience', are employed to reduce civilian casualties. Slowing the tempo of decision-making, including the processes and assessments that inform the decision, allows both the system and the user the extra time to:

- See more
- Understand more; and
- Develop more options.

Importantly, this is true throughout the decision-making chain, not only at the final 'decision point'. Accordingly, claims that AI-DSS will actually result in *greater* time for tactical patience, by speeding up time-consuming steps along the way to a final determination of whether to 'pull the trigger,' risk oversimplifying the process of targeting and the execution of force in contemporary conflicts.

Extra time allows you to see more

The now infamous drone strike in Kabul on 29 August 2021, during the evacuation of Kabul, which killed 10 civilians was attributed by the commander of Central Command to the fact that "*We did not have the luxury of time to develop pattern of life and to do a number of other things.*"

'Pattern of life' analysis is how some militaries describe an assessment of the presence and density of civilians and combatants, their schedules, patterns of movement, etc in and around an area being considered for attack. It is *a critical method of reducing civilian harm*. However, assessing a pattern of life can only be done in real-time – the time it takes civilians to create such patterns – it cannot be expedited.

Attempts to predict future behavior based on historical trends will not incorporate the current situation. In this example, a review of older intelligence material, especially full motion video of Kabul would not have reflected the changes in situation and behavior occurring because of the Taliban take-over and ongoing evacuation efforts.

As *civilian casualty prevention guidance* explains "[t]*he longer you wait and observe the more you will know about what is going on and be better prepared to make a decision to employ lethal or non-lethal means*" or as Napoleon put it "dress me slowly, I am in a hurry" – sometimes the best results are achieved by doing things deliberately.

Extra time allows a user to understand more

Another reason to slow the tempo of decision-making is that human understanding, especially of a complex and confusing situation takes time to be developed as well as to deliberate on appropriate responses. With less time available, a human's ability to comprehend the situation will lessen. The military planning process is designed to give commanders and staff the time to consider the operational environment, the adversary, friendly forces and civilians and the advantages and disadvantages of the courses of action being considered. The understanding gleaned from this process of consideration cannot be outsourced for as General Dwight D. Eisenhower explained "[i]*n preparing for battle I have always found that plans are useless, but planning is indispensable.*"

This has implications when it comes to human decision-makers considering a course of action generated or 'recommended' by an AI-DSS, whose ability to accelerate operational tempo relative to an opponent is probably the most cited reason for it being utilised. Without having undertaken, or even fully understood the process of developing a plan proposed by AI-DSS, the human planner is likely to have a limited understanding of the situation, the various influencing factors, and the actors involved. Indeed, it has been *observed that the use of automated aids can* reduce the alertness of human users and impair their ability to maintain situational awareness. This should be considered in light of how it affects compliance with IHL obligations; the *obligation to do everything feasible to verify targets* indicates a requirement to maximise the use of available intelligence, surveillance and reconnaissance assets to gain the most comprehensive situational awareness possible under the circumstances.

Extra time allows a user to develop more options

In addition to allowing a commander to see and understand more, extra time allows commanders to develop *tactical alternatives* , which could include the decision not to use force or to de-escalate. The extra time allows other units and platforms to disengage, reposition, resupply, plan, and prepare to assist in an upcoming operation. This gives a commander more options, including alternative plans that may better reduce civilian harm. Extra time may allow for additional mitigating measures such as the issuance of warnings and from the civilian perspective it allows them to implement coping mechanisms such as taking shelter, resupplying themselves with food and water or evacuating.

As one example of military planning doctrine explains "*if time is available and there is no advantage to acting more quickly, there can be little excuse for not taking the time to plan adequately.*" For as recalled in NATO's Protection of Civilians handbook "*[w]hen time is available to deliberately plan, discriminate and precisely target a force or object in accordance with the IHL principles the chances of CIVCAS [civilian casualties] are greatly minimised.*"

Conclusion

"War is chaotic, lethal, and a fundamentally human endeavor. It is a clash of wills fought among and between people. All war is inherently about changing human behavior, with each side trying to alter the behavior of the other by force of arms." Wars result from human disagreement, are waged between groups of humans, are controlled by humans and are concluded by humans who, in the aftermath, have to co-exist. Most importantly, the suffering in conflict is borne by humans.

This reality, and indeed IHL itself, calls for a *'human-centered' approach* to the development and use of AI in armed conflict – to try to preserve humanity in what is already an inhumane activity. Such an approach has at least two key aspects: (1) a focus on the humans who may be affected; and (2) a focus on the obligations and responsibilities of the humans using or ordering the use of the AI.

When looking at those who may be affected, it is not only about mitigating risks to civilians when using AI-DSS to gain military advantage, there is also the potential to design and use such tools specifically for the objective of civilian protection. Possibilities that have been suggested in this regard include tools to recognise, track and alert forces to the presence of civilian populations, or to recognise distinctive emblems that indicate protected status in armed conflict (see *here* and *here*).

And ensuring that humans can satisfy their obligations under IHL means that AI-DSS should inform but cannot displace human judgment in decisions that pose risks to the life and dignity of people in armed conflict. As much has been widely recognised by states in the context of autonomous weapon systems (see, for example, *here*, *here* and *here*). The responsibility to comply with IHL lies with individuals and their commanders, not computers. As stated in the US Department of Defense Law of War Manual "*The law of war does not require weapons to make legal determinations… Rather, it is persons who must comply with the law of war.*" China stressed this point more generally in its *Ethical Norms for New Generation Artificial Intelligence*, with an insistence "that humans are the ultimately responsible entities."

Assertions that AI-DSS will necessarily result in greater civilian protection and IHL compliance must be critically challenged and measured against these considerations, taking account of what we know about system limitations, human machine interaction, and the effect of increased tempo of operations.

See also:

- Tilman Rodenhäuser, Samit D'Cunha, *Foghorns of war: IHL and information operations during armed conflict*, October 12, 2023
- Fiona Terry, Fabien Dany, *Harnessing the power of Artificial Intelligence to uncover patterns of violence*, May 25, 2023
- Pierrick Devidal, '*Back to basics' with a digital twist: humanitarian principles and dilemmas in the digital age*, February 2, 2023

Tags: AI, armed conflict, artificial intelligence, ChatGPT, civilians, cyber defence, cyber security, decision-making, military, military cyber operations, new technologies

You may also be interested in:

Sexual violence in conflict and weapons: unpacking the links for better prevention

7 mins read

Artificial Intelligence and Armed Conflict / Autonomous Weapons / Humanitarian Action / New Technologies / Technology in Humanitarian Action Hana Salama

It's often said that sexual violence is a weapon of war, but the actual weapons …

Protecting education from non-state armed group attacks

11 mins read

Artificial Intelligence and Armed Conflict / Autonomous Weapons / Humanitarian Action / New Technologies / Technology in Humanitarian Action Jerome Marston

Non-state armed groups are responsible for a significant proportion of attacks that kill and injure …