Centre for International
Governance Innovation

Policy Brief No. 188 — September 2024

The Accountability Surface of Militaries Using Automated Technologies

Arthur Holland Michel

Key Points

→ There are many people, decisions and technologies involved in the process leading to the use of force in military operations. Each of these touch points is a locus for applying — or failing to apply — accountability when things go wrong.

→ The "accountability surface" is a new term of art to characterize the degree to which humans involved in the use of force can be held accountable for undue harm in warfare.

→ As a metaphor, the accountability surface insists upon a crucial question regarding the use of military autonomous and automating technologies: Will the use of this technology expand or contract the degree to which anyone in the organization will be held accountable for any harms resulting from any process in which it plays a role?

→ Automating technologies have the potential to significantly diminish the accountability surface of militaries that use them.

Introduction

Will humans be held accountable when things go wrong? This is the central question of the debate about military autonomous technologies. The question reflects a valid concern. However, it is, for the most part, poorly framed. At the policy level, the debate has often treated lethal autonomy — and the corresponding question of accountability — as a binary condition: Is this system making a decision to kill on its own, yes or no? And if so, can someone be held accountable for its actions when something goes wrong, yes or no?

In reality, automation in the act of killing is not a binary condition. Think of a drone that autonomously alerts its operators to objects of interest in an area, leaving the humans to fly in fighter jets to those targets and drop bombs on them. That would not meet anyone's definition of an autonomous weapon system. But what if the drone identifies objects of interest *and* proposes a plan for how to shoot at them? What if the human's sole decision is to approve or deny the drone's proposed course of action?

As this example shows, the process leading to the use of force consists of many decisions across the chain of command involving many people. The people

About the Author

Arthur Holland Michel writes about emerging technologies. His work has appeared in a range of publications, including *The Economist*, *The Atlantic*, *The Washington Post* and *Wired*. He has conducted research for the International Committee of the Red Cross, Chatham House, Peace Research Institute Oslo and the United Nations, among others. He was a founder of the Center for the Study of the Drone at Bard College and served as its co-director from 2012 to 2020. His first book, *Eyes in the Sky*, was published in 2019.

who sent the drone to a particular area are doing so because of somebody else's decision that that area *might be worth looking at*; that decision was, in turn, informed by intelligence and analysis and decisions from others in the chain, and so on.

All of these people bear some responsibility for ensuring that their decisions do not result in undue harm (International Committee of the Red Cross 2013). Artificial intelligence (AI) and other computerized technologies can automate a wide range of these decisions, from intelligence analysis, to simulation and planning, to target identification and tracking. Many of these technologies are not even embedded directly inside weapon systems. Rather, they are in the computers that humans use when making critical decisions. In that sense, they are not so much automated technologies as they are automating technologies.

Drawing the line between automation, automating and fully autonomous is not easy. Nor is it necessary. For the foreseeable future, military AI will mostly play a supporting role in the act of killing. While such tools fall short of being an autonomous weapon that can, per the Red Cross's definition, "select and apply force to targets without human intervention," their use can have a dramatic effect on the outcomes of human decisions (ibid.). They can also have a dramatic effect on the application of accountability to this process. Indeed, in some instances, apportioning responsibility for decisions that were made jointly by a human and a machine might be harder than doing so in the case of a human's decision to deploy a fully autonomous weapon.

Accountability is also not a binary. Just as there are many ways that accountability can be applied, there are many ways that organizations can fail to apply it. It can be misapplied (holding the wrong person in the chain accountable for a harm), and it can be applied insufficiently (giving a soldier a stern talking-to for a decision that killed multiple civilians). Just because a military has some instruments of accountability vis-à-vis these tools, that does not mean that accountability could, or would, be appropriately applied to their use in every possible case. In short, accountability for the use of autonomous and automating military technologies will not be assured just because there are humans somewhere "in the loop."

A New Metaphor

A new verbiage is necessary to characterize the problem of accountability. What we need is a term that can accommodate both the full spectrum of automating military technologies and the full range of ways by which any human in a military organization that uses these technologies is held accountable for unintended harms.

In cybersecurity, the "attack surface" is a poetic term of art that refers to the extent to which a computer system is vulnerable to hacking.[1] This simple metaphor captures a key point: cyber vulnerability is not a binary condition. Computer systems are not either vulnerable or invulnerable. Rather, they are either more or less vulnerable. In this metaphor, the smaller the attack surface, the more impenetrable the system.

A similar metaphor might be helpful for our purposes — specifically, the accountability surface. Instead of asking whether a technology is or is not autonomous, and only then asking whether a human can be held accountable, the accountability surface insists upon a new question: Will the use of this technology expand or contract the degree to which anyone in the organization will be held accountable for any harms resulting from any process in which it plays a role?

This metaphor puts aside the meaningless question of whether a technology is or is not autonomous. It expands the question of who to blame beyond the single abstract, almost mythological figure of the "human in the loop." As a result, this metaphor helps broaden the scope of the debate to technologies that were previously left out. And it opens the discussion on accountability to consider the many ways that it might be reduced or misapplied as a result, or in spite, of these technologies. It also helps, in the case of autonomous weapons, to put the single decision of whether to launch that weapon within the context of all the other decisions leading up to that decision.[23]

The accountability surface is also helpful because it allows one to make a more probing, more nuanced analysis of a military's plan for ensuring accountability in their use of automating technologies. Militaries often say that responsibility for military actions always rests, ultimately, with those in a position of "command authority," regardless of whether the actions ordered by said commander were executed by a machine, a human or a human-machine team. They also say that they have detailed rules of engagement and that they dispense swift reprisals for those who do not follow them. This is often true. However, just as there is no such thing as a perfectly impenetrable computer system, there is no such thing as a perfectly accountable military organization. The accountability surface is not a measure of all the ways that a military *says* it applies accountability. Rather, it is a measure of all the *gaps* where accountability is applied in practice — of all the ways that the chain of accountability might just abruptly trail off....

The Accountability Surface of Automated Warfare

Let us extend this metaphor: Does the accountability surface expand as a result of automating technologies, or does it shrink? This is, of course, a big question. It is the basis for a great deal of ongoing research, both technical and legal.

[1] See https://cheatsheetseries.owasp.org/cheatsheets/Attack_Surface_Analysis_Cheat_Sheet.html.

[2] Others have suggested that important forms of human input can exist in relation to autonomous systems long before their moment of actual use. The "so-called" "sunrise chart" diagram, which was introduced in the Group of Governmental Experts on LAWS (lethal autonomous weapons systems) in 2018, indicated that "human-machine touchpoints" were spread across every stage of an autonomous weapon's development, testing, deployment and use. This concept mirrors the idea of the accountability surface. However, the implication of the sunrise metaphor is that each of the touchpoints of human control would expand the surface of accountability. That is not necessarily going to be the case. The more touch points, the harder it might be to figure out who is responsible. The sunrise chart also obviated the fact that more than one automating technology will often be used along the chain of actions and decisions leading up to the use of an autonomous weapon.

[3] *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc CCW/GGE.1/2018/3 (2018), online: <https://documents.un.org/doc/undoc/gen/g18/323/29/pdf/g1832329.pdf?token=Iy7nVM2Iw2X83SKfv1&fe=true>.

Box 1: A Note on Harm

It has been widely speculated that the use of automating technologies in warfare could reduce the likelihood of unintended harm. This argument is not a valid counterargument to concerns about accountability. A military that is very good at avoiding unintended harm can still fail to apply accountability in the rare cases when that harm occurs. Likewise, if an automating system only causes harm very occasionally, it is still possible that that system reduces the accountability surface.

But from what is already known, there is reason to believe that automating technologies can have a shrinking effect on how humans are made accountable for what they do. First, things can go wrong. There are many, many factors that can cause an automated or autonomous tool to make an error. The human user's ability to detect and account for these errors in their decision making is naturally limited. Some types of machine errors are practically impossible to detect, understand and account for when making a decision. It is even harder to know when a machine's shortcomings, such as its biases, are relevant to a decision. And if it is impossible for a human to account for an error in a system, and as a result their decision leads to unintended harms (see Box 1), it would be unreasonable to hold them responsible for those harms (Elish 2019). These mistakes therefore become blameless technical errors (Hurst 2018).

Pushing accountability on those who built, tested and validated the system is also not entirely fair. If they knew that a system was capable of exhibiting an error, and if they were acting in good faith, they would have taken steps to address that problem. No one can be blamed for failing to address a problem that no one knew about. This is why engineers often talk about how AI systems just have to be good enough. But when civilians' lives are on the line, is "good enough" really good enough?

Once in use in the real world, the complexity of these systems and their variegated failure modes interact with the complexity of the human systems that contribute to military decisions on the use of force. The result is that the total complexity of assigning responsibility for harms is greatly multiplied. Errors in any single one of the many decisions leading up to the use of force can cascade through the whole chain. A faulty detection at the analysis phase can lead to a faulty assignment of military value to a target at the planning stage. This can then lead to the misallocation of resources to attack that target, all of which can lead to an undesirable effect (for example, killing nearby civilians, or causing a friendly-fire incident). If the person, or people, who made those decisions was unable to detect the AI system's error, and thus could not be held responsible, neither could the person making the decision at the end of the chain.

Given this complexity, the laws and procedures in place today will not necessarily have the capacity to deal with harms that fall in this grey zone between clear human negligence, malintent and purely blameless technical accidents. Tools that support or supplant human decisions are varied. Different tools work in different ways. They have different architectures, varying levels of complexity and different user interfaces. They will be developed in different ways and owned by different branches of service. Different users of these systems will have different training, discipline regimes and cultures of responsibility. This means that no single set of protocols for human control or accountability will work for all of them.

Ideas for Action

AI systems are not alone among technologies in their likely shrinking effect on the accountability surface. All digitalizing technologies raise the possibility of harms for which no one will be held appropriately accountable. Any time a digital interface is placed between a human who makes a decision and the object of that decision, there is a possibility that harm may fall in the murky space between technical error and human failure.

There have been many cases where tools that would not rise to anyone's definition of machine intelligence have contributed to harms that no one was ultimately held accountable for. Consider all the times, for example, when a drone strike hit civilians because the resolution in the cameras was not high enough to show the operators that the figures on the screen were not soldiers. Think of how the precision of their signals intelligence was not good enough to distinguish between the militant in one house and the civilian next door. In the civilian domain, there have also been cases of the opposite effect, such as the two crashes involving Boeing 737 MAX airplanes, in which humans have been blamed (often out of political expediency) for a failure to respond appropriately to a system's unanticipated technical failures.

Some might worry that broadening the scope of the debate on military AI to all of these digitalizing technologies could derail that debate. But many parties refuse to agree on a definition of what they are talking about in the first place (Congressional Research Service 2019). In this regard, focusing on the accountability surface could be helpful, since the metaphor does not hinge on a single set of technological definitions. The dynamics of the accountability surface hold equally true for autonomous weapons as they do for remote-control drones and hand-to-hand combat. If parties do not need to draw rigid lines around what should and should not be regulated, they can focus on finding and implementing generalizable principles, channels and methodologies of accountability.

Here are some ideas for the kinds of questions and frames of thinking that could undergird this process:

→ Never assume that having a human in the loop on an automating system assures accountability: it does not.

→ Consider that a technology can reduce the likelihood of unintended harm, and also reduce the likelihood that anyone will be held appropriately accountable for that harm.

→ For every new tool that supports or supplants a human decision in warfare, ask: Will this reduce the accountability surface? Even if the tool seems simple, ask the question.

→ Assess proactively, and in a detailed manner, how accountability could be applied to the use of the technology. This means mapping all of the ways that the system might fail. It also means mapping all of the ways that a person might fail when using the system. For each of these mapped failures, ask the question: With whom would the buck stop?

→ Have a plan for when the human or the system, or both of them in tandem, fail in an unexpected manner.

→ Study the cognitive limitations and biases that inhibit humans from making good decisions on the basis of bad decision support.

→ Be transparent. Instruments of accountability that are shrouded from public view are unlikely to be effective; in the long term, the organs of justice will atrophy. Therefore, governments should be as open and detailed as possible about what they plan to do when harm arises from a process involving an automated system.

And perhaps most importantly, keep going. Like cybersecurity, the process of ensuring accountability is perpetual. It does not end with the implementation of laws, protocols and procedures; it begins with it. Just as a computer's attack surface continuously grows unless efforts are constantly made to update its defences, the accountability surface of a chain of decisions involving autonomous systems will steadily shrink unless proactive efforts are made to ensure that the accountability instruments and channels remain effective and current. In this regard, the metaphor sets a standard of always striving to be better, to be more accountable: it lends itself to a never-ending process of improvement.

Constant improvement is necessary regardless of whether a military uses AI or not. But it is especially important with AI systems, because their performance in real-world conditions is not a constant and there has been little time to understand the ways in which these systems are likely to operate in battle. An AI system that did not display unpredictable behaviour when it was validated for deployment can gradually become unpredictable over the course of its real-world use. There is still much more that is unknown, when it comes to these systems, than what is known.

A little humility, in other words, can go a long way.

One Last Thought

The discourse on military AI tends to imply that these technologies will disrupt a system of accountability that currently works. That is a faulty assumption, to say the least. Present-day systems of accountability are deeply imperfect, often injuriously so. Attempting to ensure accountability in the use of automating technologies without also striving to improve military accountability in general is a doomed proposition. In that sense, rule-making efforts on military AI and automation should not be seeking merely to hold these technologies and those who use them to an existing standard; they should seek to raise the standard itself for everyone who takes part in the use of force.

Works Cited

Elish, Madeleine Clare. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society 5*: 40–60. https://doi.org/10.17351/ests2019.260.

Hurst, Jules. 2018. "Warbots and Due Care: The Cognitive Limitations of Autonomous and Human Combatants." *Military Review: The Professional Journal of the U.S. Army*: 1–11. https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2018-OLE/Mar/Warbots/.

International Committee of the Red Cross. 2013. *Decision-Making Process in Military Combat Operations.* Geneva, Switzerland: International Committee of the Red Cross. www.icrc.org/en/doc/assets/files/publications/icrc-002-4120.pdf.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

This research paper has been made possible in part by the Mobilizing Insights in Defence and Security (MINDS) program at the Department of National Defence.

MOBILIZING INSIGHTS IN DEFENCE AND SECURITY
MINDS
MOBILISATION DES IDÉES NOUVELLES EN MATIÈRE DE DÉFENSE ET DE SÉCURITÉ

The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Centre for International Governance Innovation or its Board of Directors.

For publications enquiries, please contact publications@cigionline.org.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Program Manager and Research Associate Kailee Hilt
Publications Editor Christine Robertson
Graphic Designer Sami Chouhdary