Centre for International
Governance Innovation

Policy Brief No. 197 — February 2025

Responsible AI and Civilian Protection in Armed Conflict

Daniel R. Mahanty and Kailee Hilt

Key Points

- While the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy holds promise, its supporters should place greater emphasis on how the implementation of its principles will lead to better protection of civilians in armed conflict, especially when combined with other measures not limited to the use of artificial intelligence (AI) and autonomy.
- This policy brief argues that the responsible use invoked by the declaration should not result in only marginally better protection of civilians (PoC)[1] outcomes than "irresponsible" use, but should instead achieve markedly better ones.
- Giving meaning to the declaration's implied PoC content depends on whether the expansion of its membership and stewardship of the process raises the ceiling or lowers the floor for responsible use.
- National and multilateral efforts to promote the responsible military use of AI should be connected to a renewed commitment among all states to mitigate harm to civilians resulting from all military operations, not only those that involve the use of AI.

Introduction

In the last annual report (2023) on the Protection of Civilians in Armed Conflict, the UN Secretary-General described the situation for civilians in armed conflict in the previous year as "resoundingly grim." Over the course of 2022, thousands of civilians died and millions more suffered from the impact of wars.[2] By the time of the report's release in the spring of 2023, it appeared that the scope and severity of civilian harm in 2024 would be even worse. The report also arrived amid a growing crisis of public confidence in international humanitarian law, caused by the pervasive and wanton disregard for its principles exhibited by some states and the tepid compliance modelled by others — with a dearth of meaningful accountability on both sides. All the while, disparate levels of concern among Western states for civilians in Gaza, Sudan and Ukraine had led to charges of double standards and hypocrisy.

Against this backdrop (a mere six days after the release of the report), the United States published an updated version of its Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, replete with an expanded list of supporting states (US Department of State 2023). While the political declaration makes only one explicit reference to the goal of protecting civilians (and that made only in the context of the use of AI and autonomy), several of its elements serve that goal. By committing to use AI in adherence with international law, the declaration's signatories agree to abide by those

1 This brief uses the term "protection of civilians" to mean the measures taken by states to prevent, minimize and address harm resulting from their own military operations (including operations involving allies and partners). This use most closely approximates the concepts of combatant PoC or civilian harm mitigation, rather than its use in the context of peacekeeping or atrocities prevention.

2 See https://acleddata.com/data/#/dashboard.

About the Authors

Daniel R. Mahanty specializes in issues at the intersection of US national security and human rights, with a recent focus on civilian protection in armed conflict. His career has included leadership roles at the Center for Civilians in Conflict, 16 years of service at the US Department of State, and teaching positions at Georgetown University Security Studies Program. This policy brief was written prior to his current role in government service.

Kailee Hilt is a program manager and research associate at CIGI. She focuses on public policy issues tied to emerging technology, privacy and cybersecurity. Kailee is presently pursuing her Certified Information Privacy Professional (CIPP/C) designation. She holds a master of arts in sociology and legal studies from the University of Waterloo, where her research centred on violence against women and access to justice. She also earned a master of information science from Western University, where she examined the influence of emerging technologies, such as artificial intelligence, on the information science profession.

obligations under international humanitarian law serving to protect civilians (for example, distinction, proportionality and necessity), along with the relevant tenets of human rights law. With its reference to avoiding unintended bias, the declaration also infers the PoC from harm resulting from the biased interpretation of data (for instance, when distinguishing civilian or civilian objects from legitimate targets in military attacks).

But the declaration is designed not merely to restate existing obligations or reinforce reciprocal commitments among signatories, but to also differentiate states that are responsible enough to use AI from those that are not as part of an effort to "strengthen international rules and norms"; as such, its supporters should aim much higher when it comes to the PoC. To stand any chance of succeeding in this aim, the "responsible use" invoked by the declaration must result in markedly better PoC outcomes than "irresponsible" use. If treated seriously, the declaration could present both a critical opportunity to better protect civilians by raising the standard of civilian protection observed and enforced by its signatories and as part of a meaningful effort to restore some faith in the international rules-based order itself.

While some observers have panned the declaration as shambolic, others have welcomed the initiative for its timeliness and potential, calling for the United States to further expand its list of supporters. This brief neither dismisses nor embraces the declaration, instead arguing that its legitimacy as an international framework depends entirely on whether it serves as a catalyst for improving PoC outcomes both in contexts where AI and autonomy feature in the conduct of hostilities and where they do not. This result depends upon two eventualities: first, whether the process contributes to a higher standard of practice in PoC, rather than merely serving to legitimize the lowest common denominator; and second, whether states meaningfully adapt their AI and autonomy programs to reduce the risk of harm to civilians.

To these ends, states — particularly those signatories that are also political champions of the PoC — can draw from lessons gained through other similar initiatives. This step will require limiting the benefits of multilateralism to deviants, developing meaningful national and collective action plans, and establishing common frameworks for monitoring and measuring success. This approach neither obviates nor assumes the development of

new treaty-based law but could be a useful basis for improving outcomes with or without it.

PoC Risks of AI and Autonomy

For as long as the use of AI for military purposes has been imagined in fiction and, more recently, emerged in practice, observers and experts have identified its risks. The International Committee of the Red Cross and others have helpfully distilled these risks into a few general categories:

- The use of digital AI and machine learning to control military hardware and weapons systems could lead to greater autonomy and unpredictable outcomes for civilians, along with a lack of attribution and accountability (International Committee of the Red Cross 2021).
- The use of AI and machine learning to select targets based on operational importance could create the risk of disruptions to critical infrastructure systems or assets, such as transportation, medical facilities and telecommunications (ibid.).
- Disinformation generated by AI can distort, or interrupt access to, critical sources of lifesaving information for civilians, such as information about nearby military activity, humanitarian services or evacuation options (Spink 2023).
- Decision making that is enabled or even conducted entirely through AI-generated analysis may lend bias to certain factors that create significant risks for civilians or civilian objects, especially if humans place excessive trust in AI-produced analysis (Conventional Arms and Ammunition Programme 2016).
- Training AI to recognize patterns of life or the distinctive characteristics of combatants and civilians may not translate from one environment to another.

In the context of high-intensity conflict between the forces of industrial nations, scenarios involving these risks are easy to imagine if they are not unfolding already. A military might broadly confer combatant status on civilians, subjecting them to lawful attack, based entirely on algorithmic analysis of large data sets. Provided a list of basic military objectives, a machine might rapidly generate a list of thousands of targets that, if destroyed, could lend a significant, concrete and calculable military advantage, but that would also devastate lives and livelihoods. AI-generated disinformation about military operations or movements — or about "safe" areas for civilians distributed through a wide array of online "bots" with the social characteristics of humans — could lead civilians to make decisions that place them in harm's way. A series of AI-enabled cyberattacks on critical nodes of a telecommunications system could disrupt adversary communications, but also interrupt medical services and instigate a social, political and economic crisis for civilians.

Any of these scenarios could involve violations of the laws of war, and the risks involved with AI do nothing to change the existing legal obligations of states (even if, as some have argued, they may invite the development of more law). As will be addressed in a subsequent section, much depends not only on how states use AI but also on how they interpret their legal obligations.

Raising the PoC Standard Through the Declaration

States that champion the political declaration describe its many intended benefits, including shaping international consensus on an emerging issue of broad relevance and avoiding the risk of inadvertent escalation (Tucker 2023). These intended benefits should be expanded to more explicitly aim for a higher standard of practice in the use of AI and autonomy in the military domain for the benefit of human rights and civilian protection. Without further specific elaboration on how it will improve PoC, the declaration will invite criticism for glossing over AI's risks to civilians or be seen as a transparently political manoeuvre to distinguish certain states from others based largely on realpolitik and rhetoric rather than outcomes. Moreover, if signing the declaration becomes a political licence to develop or acquire technology with the acquiescence of other joining states in the absence of meaningful controls, the declaration will not only lose legitimacy but could also do harm.

The declaration's champions might find inspiration in the design of other non-binding

arrangements that have been useful for norm setting in the absence or anticipation of treaty-based law, which provide a valuable basis for involving a range of interested stakeholders and participants. Certain non-binding agreements have helped to generate norms for the benefit of human rights, such as the Universal Declaration of Human Rights, the Inter-American Declaration of Human Rights and the Montreux Document on Private Military and Security Companies. The soft norms that develop in these contexts can raise a standard of practice or policy by setting specific limitations and aspirational targets that go beyond merely restating existing legal obligations and generally sit above the existing baseline, and by establishing a meaningful framework of mutual accountability for meeting them. This approach can work well to not only create a level playing field, but to also raise a collective standard when states have the incentive to both cooperate and compete through a race to the top.

Supporters of the declaration should also look to lessons in contexts where non-binding political declarations have entrenched minimum standards — or even lowered them — through intentional vagueness, broad caveats or carve-outs, and a lack of any meaningful accountability (Linos and Pegram 2016). In the attempt to create a "big tent" that invites broad participation and includes even states with the most clearly problematic conduct or troubling records, norm setting can default to the lowest common denominator, rather than modelling and upholding a higher standard. For example, civil society and research organizations criticized the US-led Joint Declaration for the Export and Subsequent Use of Armed or Strike-Enabled Unmanned Aerial Vehicles (drones) because it established an export regime that was more permissive than the current US standard, which not only entrenched the lower bar, but also worked against the competitive interests of US industry (Mehta 2016). At a minimum, those who support the declaration should question the suitability of inviting states that seem fundamentally unwilling to change patterns of conduct that lead to significant civilian harm, regardless of whether they believe the harm stemmed from "lawful" conduct.

The outcome and effect of political declarations on norms of practice also depend on the form of governance used both to enlist new signatories and to monitor and enforce adherence. Stewart Patrick (2023) of the Carnegie Endowment for International Peace would likely describe the political declaration as an example of the "club" approach to multilateralism, wherein the initial participants cooperate based on the like-mindedness of liberal states. Approaching the declaration through a "club"-based approach without any form of mutual evaluation or accountability would have the benefit of centring participation on a voluntary commitment to abide by existing rules and standards. But signatories may differ wildly in their interpretation of the rules, and states that traditionally refrain from aligning with great powers may skeptically view a club developed by the United States to compete with China. Moreover, new multilateral structures can help bypass the inertia and bureaucracy involved in the formal (read UN) system, but can also be manipulated to shape and develop norms that soften protection standards or even undermine international humanitarian law or human rights. A number of observers, including the former UN special rapporteur on counter-terrorism and human rights, have publicly reflected on the way that the Global Counterterrorism Forum, designed with the best of intentions, generated a bounty of "soft" law and became a convenient forum for states to gain the legitimacy endowed by multilateralism, while bypassing any meaningful accountability for their human rights conduct (Ní Aoláin 2024).

The working groups developed under the political declaration may provide the necessary underlying system of mutual accountability and reinforce the declaration's intent. Civilian protection could be mainstreamed as a topic within most of their workplans. Adding a working group on civilian protection might lend much-needed attention to the topic and its relevance to the declaration, as well as sharpen the focus of the plenary group on the most important civilian protection issues and what to do about them. Moreover, participants in the political declaration have largely deferred transparency and the participation of civil society and outside experts to the parallel Responsible AI in the Military Domain (REAIM) process. Left unclear is how the input of members of civil society and others on the REAIM process will inform the actions of states involved in the political declaration. Even less clear is whether declaration signatories intend to submit their work for any level of public evaluation and how they would go about doing so.

State-Level Implementation

While revisiting the language of the declaration and the process it has created is worthwhile, the best hope for reinforcing the overarching intent of its initiative may fall to individual states. As acting US Assistant Secretary of Defense Madeline Mortelmans told Breaking Defense, "It is about state practice and how we build states' ability to meet those standards that we call committed to" (Freedberg 2024). Well-designed national plans will provide more specifics on how states intend to follow through on their commitments — including how plans will strengthen the PoC — in part by strengthening the impact of adherence to international humanitarian law. This approach would also allow states to innovate and model their approaches for the collective benefit of other declaration supporters. States may be well served to approach the effort with a few principles in mind:

- **Connect the implementation of the declaration's principles to national PoC plans:** In states such as the Netherlands and the United States, commitments should be integrated in new policies and action plans directly focused on the PoC. The US Department of Defense Civilian Harm Mitigation and Response Action Plan and the Dutch civilian harm mitigation "road map" provide ready-made frameworks for including specific concerns relevant to AI. Failing to connect the initiative to broader frameworks and PoC may seem detached from reality, given the extent of harm caused by conventional warfare.

- **Integrate civilian protection as a feature of AI technology research, development, testing, evaluation and acquisition:** States should ensure that their technology research, development, testing and acquisition strategies include features and capabilities that are specifically designed to prevent harm and better protect civilians, not only by mitigating bias, but by also addressing all of the potential AI failures that can lead to civilian harm. AI tools should be tested to identify and address potential risks to civilians, and states should also integrate civilian harm impact assessments as part of the technology evaluation process to understand and mitigate potential risks before deployment.

- **Consider the possible benefits of AI and machine learning for improving protection:** Although the use of AI can create risks for civilians, recent analysis suggests that AI also brings opportunities for preventing and mitigating harm. During the planning phases of an operation, AI can help map critical civilian infrastructure assets and systems, as well as the interdependencies between them, to minimize damage while also ensuring a more efficient application of force. At the operational level, AI can analyze civilian patterns of life more effectively than human analysts, offering insights into civilian behaviour and movement to reduce harm. Tactically, AI may help locate military targets inside buildings, reducing the risk of mistakenly targeting civilians. AI systems also have the potential for post-attack assessments, analyzing data on civilian harm and the effectiveness of mitigation measures in order to improve future military planning and operations. For example, according to a study by Larry Lewis and Andrew Ilachinski (2022) from the Center for Naval Analysis, AI tools could be used to inform forces of changes to assumptions underlying collateral damage estimation or the presence of transient civilians.

With the importance of centring efforts on state practice, the declaration's supporters should be clear-eyed and introspective about challenges from within, which may pose the single greatest threat to realizing the declaration's potential. First, states rarely willingly join agreements that their most powerful bureaucratic agents (often, but not always, defence ministries) are unlikely to tolerate. If the bureaucracy "tolerated" joining the declaration, it may not perceive any costs or pain in doing so, which could present a challenge for protecting civilians that imposes transaction costs as measured in both time and risk. Governments, and especially those at the forefront of AI development and use, may need to voluntarily restrain their own use of military systems if the bar set by practice is to remain high. This notion runs contrary to a culture that seeks to relieve itself of constraints.

Second, the standard of protection for civilians will not improve if national plans merely default to existing permissive interpretations of international law that have already proven conducive to excessive levels of harm. For the last several decades, states have interpreted the law to allow for the overly broad assignment of

combatant status to civilians and the targeting of "war-sustaining" infrastructure, including energy and banking facilities (Chertoff and Manfredi 2017). The political declaration should further galvanize the need for deep introspection among influential states about their responsibility to tighten the legal loopholes they helped to expand, especially during the counterterrorism era.

Finally, limiting the harmfulness of algorithmic or data-based biases will do little to mitigate the harm caused by states willing to use AI to facilitate harmful practices that endure as a matter of operational culture or even policy.

Conclusion

The year 2024 marked the seventy-fifth anniversary of the Geneva Conventions and the twenty-fifth anniversary of the first formal recognition of the PoC as a priority by the UN Security Council. Both initiatives emerged as features of an international rules-based order in recognition of the global consequences of civilian harm in war, and their commemoration takes place at a time when their limitations, exploited by the cynical self-interests of states and non-states alike, are clear and present for all to see. Those individuals who have vested their faith in the declaration's ability to strengthen the same rules-based order must find concrete ways to make it deliver for those who remain in doubt.

Acknowledgements

This policy brief benefited from the input and advice of Daniel R. Mahanty, the former director of research, learning and innovation at the Center for Civilians in Conflict, prior to his government service.

Works Cited

Chertoff, Emily and Zachary Manfredi. 2017. "The Al-Mayadeen Prison Bombing and the Problem of War-Sustaining Targets." *Lawfare* (blog), July 6. www.lawfaremedia.org/article/al-mayadeen-prison-bombing-and-problem-war-sustaining-targets.

Conventional Arms and Ammunition Programme. 2016. "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies." https://unidir.org/wp-content/uploads/2023/05/safety-unintentional-risk-and-accidents-en-668.pdf.

Freedberg Jr., Sydney J. 2024. "US joins Austria, Bahrain, Canada, & Portugal to co-lead global push for safer military AI." Breaking Defense, March 28. https://breakingdefense.com/2024/03/us-joins-austria-bahrain-canada-portugal-to-co-lead-global-push-for-safer-military-ai/.

International Committee of the Red Cross. 2021. "Artificial intelligence and machine learning in armed conflict: A human-centred approach." Position Paper No. 913. March. https://international-review.icrc.org/articles/ai-and-machine-learning-in-armed-conflict-a-human-centred-approach-913.

Lewis, Larry and Andrew Ilachinski. 2022. *Leveraging AI to Mitigate Civilian Harm*. February. Arlington, VA: Center for Naval Analysis. www.cna.org/reports/2022/02/Leveraging-AI-to-Mitigate-Civilian-Harm.pdf.

Linos, Katerina and Tom Pegram. 2016. "The Language of Compromise in International Agreements." *International Organization* 70 (3): 587–621. https://doi.org/10.1017/S0020818316000138.

Mehta, Aaron. 2016. "White House Rolls Out Armed Drone Declaration." *Defense News*, October 5. www.defensenews.com/breaking-news/2016/10/05/white-house-rolls-out-armed-drone-declaration/.

Ní Aoláin, Fionnuala. 2024. "The Rise of Counter-Terrorism and the Demise of Human Rights." Herbert L. Bernstein Memorial Lecture in Comparative Law, Duke University, February 8. https://scholarship.law.duke.edu/bernstein/18/.

Patrick, Stewart. 2023. "Four Contending U.S. Approaches to Multilateralism." Working Paper. January. Washington, DC: Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2023/01/four-contending-us-approaches-to-multilateralism?lang=en.

Spink, Lauren. 2023. *When Words Become Weapons: The Unprecedented Risks to Civilians from the Spread of Disinformation in Ukraine*. October. Washington, DC: Center for Civilians in Conflict. https://civiliansinconflict.org/wp-content/uploads/2023/11/CIVIC_Disinformation_Report.pdf.

Tucker, Patrick. 2023. "US Woos Other Nations for Military-AI Ethics Pact." *Defense One*, February 16. www.defenseone.com/technology/2023/02/us-woos-other-nations-military-ai-ethics-pact/383024/.

US Department of State. 2023. "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy." November 9. www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

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.

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Program Manager and Research Associate Kailee Hilt
Publications Editor Christine Robertson
Publications Editor Susan Bubak
Graphic Designer Sepideh Shomali

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

CIGI