LEVERAGING AI TO MITIGATE CIVILIAN HARM

MINIMIZING ADDITIONAL CIVILIAN HARM FROM AI IN WARFARE IS NOT AMBITIOUS ENOUGH. TAKE STEPS TOWARD USING AI TO LESSEN WARFARE'S SUFFERING AND DESTRUCTION.

Countries around the world have taken early steps to leverage artificial intelligence in military capabilities. While their goal is greater effectiveness and efficiency, the idea of adapting AI to military applications has also created considerable controversy. Many concerns have been voiced, including potential bias and a lack of fairness, as well as the desire to maintain human judgment and responsibility in engagement decisions. But the chief concern in international discussions is whether military applications of AI could be inherently indiscriminate, unable to differentiate between valid military targets and civilians.

One way to answer this question is to examine specific military applications of AI, including autonomous systems, and examine both technical and operational considerations for how risks to civilians may arise and how they can be mitigated. For example, several presentations during the UN Convention on Certain Conventional Weapons meetings on lethal autonomous weapon systems featured examples of autonomous systems that could be used for warfighting tasks in ways that complied with international law and did not present indiscriminate hazards to civilians. A previous CNA report (*AI Safety: An Action Plan*) also considers some warfighting applications of AI and how risks to civilians from those applications could be minimized through both operational and technical mitigation steps.

Those discussions, however, only address one half of the two-fold responsibilities for civilian protection found in International Humanitarian Law — the *negative* responsibility that militaries should not direct attacks on civilians. The *affirmative* responsibility for militaries to take all feasible precautions to protect civilians from harm has been relatively neglected. With regard to AI and autonomy, states should not only be asking how they can meet their negative responsibilities of making sure that AI applications are not indiscriminate in warfare. They should also be asking: **How can we use AI to protect civilians from harm? And how can AI be used to lessen the infliction of suffering, injury, and destruction of war?**

A new CNA report, *Leveraging AI to Mitigate Civilian Harm*, represents a concrete first step toward answering these questions. It begins by framing the problems that lead to civilian harm. If we understand that AI is a tool for solving problems, what are the problems that need to be solved to better protect civilians or otherwise promote International Humanitarian Law's principle of humanity? While the imperative for avoiding civilian harm is universally acknowledged, the specific mechanisms for how such harm occurs have never been characterized in detail. How does civilian harm occur?

A synthesis of work on civilian harm — including an analysis of several thousand real-world incidents of civilian harm from military operations — helps establish a framework illustrating how civilian harm occurs, depicted on the next page. Categorizing civilian harm into three types of collateral damage and two types of misidentification opens the door to identifying how civilian harm can be mitigated.

CIVILIAN HARM
HOW IT HAPPENS

Two types of Civilian Harm

Misidentification: Mistaken belief that civilians/civilian objects are a valid military target

Collateral damage: Proximity to valid military target

Two types of Misidentification

Misperception

Behavior: Appears to be acting in a threatening way/consistent with threat behavior

Appearance: Matches the physical characteristics of a threat

Signature: Other characteristics that seem to match a threat, like electronic emissions, human intelligence, communication device signature, etc

Misassociation

A wrong correlation between location and information correctly identifying a threat

Three types of Collateral Damage

Anticipated and Accepted
(viewed as proportional)

Unanticipated Presence

Pre-planned Operations: Civilian presence is there, but missed

Pre-planned Operations: Civilian comes later and change is missed (assessment not updated)

Dynamic Operations: Civilian presence is missed

Transient Civilian Presence: Civilian presence moves into target area

Unanticipated Effects

Weaponeering and Weapon Performance: Results in missed target

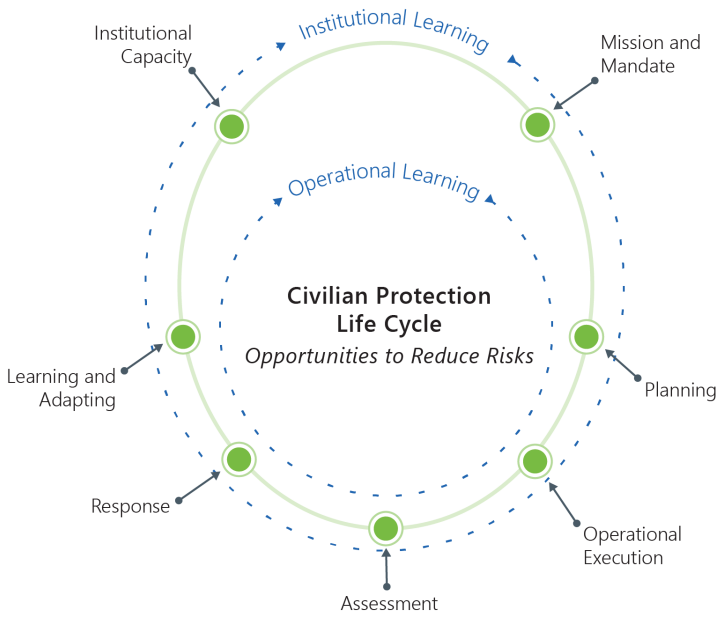
Weaponeering and Weapon Performance: Results in additional unanticipated damage (adjacent structures or individuals)

Secondary Explosion: Results additional damage/casualties

*Accidental or inadvertent direct harm to civilians, excluding intentional harm and where harm is anticipated and accepted as lawful.

The report introduces a "civilian protection life cycle," which demonstrates a comprehensive approach to mitigating harm within each of the following stages: mission and mandate, planning, operational execution, assessment, response, learning and adapting, and institutional capacity. The point is to identify opportunities to protect civilians at each stage in the cycle — not just at the "trigger pull."

Having categorized the stages in the civilian protection life cycle, we are able to see which specific AI applications can assist at each stage. We do this with a matrix that plots 33 types of AI/machine learning applications against the steps within the life cycle. For example, conducting mission rehearsals is a step within the planning stage where civilian protection can be implemented, and one AI application type well-suited to mission rehearsals is an inference engine.



Finding linkages between the risk factors we have observed in real-world operations and specific potential applications of AI brings us a step closer to mitigating harm. Our analysis finds potential to leverage techniques that currently exist, and in many cases have already been applied to other problems. For example, DOD's Project Maven uses machine learning to help identify objects — such as people, buildings and vehicles — in full motion video in order to cue operators to potential targets. Such applications could also employ AI to identify objects to avoid in order to protect civilians. There is no solution that will completely eliminate the problem of civilian harm — military operations will always have a non-zero risk to civilians — but AI can be used to help address patterns of harm and reduce its likelihood.

We suggest that national governments could focus on the following functions as promising starting points:

- **Alerting the presence of transient civilians** — using object identification to automatically monitor for additional individuals around a target area and send an alert if they are detected. This would bring them to the attention of operating forces that might otherwise fixate on the target and miss transient civilian presence.
- **Detecting a change from a collateral damage estimate** — finding differences between the imagery used to originally determine a collateral damage estimate and more recent imagery taken in support of an engagement. This can help bring small details to the surface, details that operating forces might not recognize but could be cues of unanticipated civilian presence, such as additional vehicles near a building.
- **Alerting a potential miscorrelation** — helping to identify that a miscorrelation has taken place. For example, applications could recognize the vehicle being tracked is not the same one that was being tracked previously, showing that a swap has occurred between a threat vehicle and a civilian vehicle.

- **Recognition of protected symbols** — using AI and machine learning methods to identify symbols designating protected objects, such as a red cross or red crescent, and alerting the operator and/or chain of command. This capability would provide a safety net in cases when the protected symbol is present but was missed by operating forces. The Australian Armed Forces have already conducted field experiments showing the utility of this function.

One tragic incident in Afghanistan can illustrate the potential for AI applications to address the root causes of civilian harm. In August 2021, a U.S. drone strike targeted what intelligence suggested was the vehicle of an ISIS-K suicide bomber in Kabul. Instead, 10 civilians were killed. Using another matrix we created that cross-aligns those 33 AI applications against 9 potential mitigations to prevent civilian casualties, we see multiple applications that might have addressed the root causes of this tragedy. For example, an AI application could have a functionality to develop a robust civilian pattern of life that could have picked up signs that the vehicle was involved in humanitarian aid.

This work is merely a first step in exploring a vast space of possibilities, where details matter greatly. **Governments, militaries and academic institutions should be deliberate in developing AI solutions to mitigate harm to civilians, building on this foundation.**

What remains is a matter of will, which we acknowledge is uncertain. While militaries speak of capabilities that help mitigate civilian harm, such as precision-guided munitions, those capabilities were originally acquired to engage military targets more effectively. Militaries may have capabilities that have benefit in mitigating harm in some contexts, but they have neither sought nor even recognized the need to comprehensively develop capabilities to reduce risk to civilians from all the mechanisms we identify here. As a result, the set of current capabilities held by militaries is incomplete. Much more can be done, and existing risks are not always mitigated by capabilities that do exist. For example, a precision-guided munition has no value in mitigating civilian harm when civilians have been misidentified as a military target and are engaged in that mistaken belief.

Despite the potential, we do not see militaries around the world widely seeking to field capabilities on the basis of their value in mitigating civilian harm. We have taken a first step to show how AI-enabled applications for reducing the cost of war on civilians are within the realm of the possible. It remains to be seen whether militaries will choose to pursue them.

ABOUT CNA

CNA is a nonprofit research and analysis organization dedicated to the safety and security of the nation. It operates the Center for Naval Analyses — the only federally funded research and development center (FFRDC) serving the Department of the Navy — as well as the Institute for Public Research. CNA is dedicated to developing actionable solutions to complex problems of national importance. With nearly 700 scientists, analysts and professional staff, CNA takes a real-world approach to gathering data. Its one-of-a-kind Field Program places analysts on carriers and military bases, in squad rooms and crisis centers, working side-by-side with operators and decision-makers around the world. CNA supports naval operations, fleet readiness and strategic competition. Its non-defense research portfolio includes criminal justice, homeland security and data management.

For more information please contact: Dr. Larry Lewis | Lewisl@cna.org | 703.725.3633
Dr. Andy Ilachinski | ilachina@cna.org | 703.824.2045

DMM-2021-U-031086-Final