Ethics and Information Technology (2023) 25:12
https://doi.org/10.1007/s10676-023-09681-2

ORIGINAL PAPER

Artificial intelligence and humanitarian obligations

Daniel Trusilo[1] · David Danks[2]

Published online: 13 February 2023

Abstract
Artificial Intelligence (AI) offers numerous opportunities to improve military Intelligence, Surveillance, and Reconnaissance operations. And, modern militaries recognize the strategic value of reducing civilian harm. Grounded in these two assertions we focus on the transformative potential that AI ISR systems have for improving the respect for and protection of humanitarian relief operations. Specifically, we propose that establishing an interface for humanitarian organizations to military AI ISR systems can improve the current state of ad-hoc humanitarian notification systems, which are notoriously unreliable and ineffective for both parties to conflict and humanitarian organizations. We argue that such an interface can improve military awareness and understanding while also ensuring that states better satisfy their international humanitarian law obligations to respect and protect humanitarian relief personnel.

Keywords Artificial Intelligence · Intelligence, Surveillance, & Reconnaissance · Humanitarian notification · Deconfliction · International Humanitarian Law

International humanitarian law beyond targeting

Much of the debate around Artificial Intelligence (AI) and autonomous systems in military contexts has been on autonomous weapons and targeting systems. There are repeated concerns about whether systems that use AI for targeting and lethal force are consistent with International Humanitarian Law (IHL) (Future of Life Institute, 2021; Russell et al., 2021; Asaro, 2019, 2012), while also noting the value of AI and machine learning (ML) systems for rapid discrimination of valid military targets (DoD, 2022a). In contrast, we focus here on uses of AI technologies for military Intelligence, Surveillance, and Reconnaissance (ISR) operations, particularly as they inform human decision-makers.

✉ Daniel Trusilo
dtrusilo@ucsd.edu

David Danks
ddanks@ucsd.edu

[1] School of Economics and Political Science, St. Gallen, Switzerland

[2] Halicioğlu Data Science Institute & Department of Philosophy, University of California, San Diego, La Jolla, CA, USA

As militaries increasingly deploy sophisticated sensor networks across multiple platforms and modalities that generate a flood of signals, AI/ML systems have become increasingly important to ensure appropriate awareness. A well-known real-world example is the US Department of Defense (DoD) Joint All Domain Command and Control concept, which is intended to process data from numerous sensors using AI (Hoehn, 2022). Reports have also pointed towards Russia's intent to develop reconnaissance and reconnaissance-strike unpiloted aerial vehicles that use neural networks to enhance ISR capabilities (Allen, 2022). Such sociotechnical AI systems for ISR operations will likely increase the speed of targeting processes, and also open new attack vectors for adversary deception, including novel ways to obscure or hide the deception (Danks, 2020). Of course, this same technology could also reduce human error and improve targeting precision (Arkin, 2010, 2018). We are particularly interested in ways that AI ISR systems might impact legitimate humanitarian efforts.

Customary international law holds that "parties to the conflict must allow and facilitate rapid and unimpeded passage of humanitarian relief for civilians in need." (ICRC, 2005, Rule 55). There are thus obligations to respect and protect humanitarian relief personnel for reasons beyond their "mere" status as civilians. We argue that establishing appropriate linkages between AI ISR systems and humanitarian

organizations can enable all parties to a conflict to better satisfy these (and other) IHL obligations, particularly to respect and protect humanitarian relief personnel as per Article 71 of Additional Protocol I to the Geneva Conventions. Humanitarian efforts can also have significant impacts on the conflict itself, potentially contributing to the strategic success of military campaigns. For example, the August 2022 US DoD Civilian Harm Mitigation and Response Action Plan (hereafter Action Plan) states:

> Protecting civilians from harm in connection with military operations is not only a moral imperative, *it is also critical to achieving long-term success on the battlefield*. [emphasis added] (DoD, 2022b, p. 1).

Despite the multiple reasons for parties to conflict to adhere to their IHL obligations, they also face many challenges in efforts to live up to those obligations. AI ISR systems arguably have the potential to improve relevant situational awareness, and thereby enable parties to better honor these critical moral commitments. In the next section, we examine ways that AI ISR could improve target deconfliction processes and humanitarian notification systems. Section three discusses our proposal in more detail. We then conclude by noting some objections and obligations related to deconfliction processes.

The need for an interface

Essentially all modern militaries have processes and systems in place to reduce civilian casualties. For example, DoD Joint Publication 3–60 codifies the use of no-strike and restricted target lists (Joint Chiefs of Staff, 2013) in the DoD's targeting deconfliction process that determines which potential targets are legitimate. Concerns about the accuracy and robustness of this process underlie much of the criticism of DoD's record of civilian harm, with documents pointing to careless targeting or (evidence of) strikes that potentially violate the laws of armed conflict (Yager, 2022). The 2022 DoD Action Plan aims to enhance such processes and systems, including direct calls for "improved knowledge of the civilian environment and civilian harm mitigation capabilities and processes throughout the joint targeting process" (DoD, 2022b, p. 12). The Action Plan acknowledges not only the moral imperative to mitigate civilian harm (including protection of humanitarian relief personnel) but also the strategic benefits of doing so. We suggest that AI ISR could improve the inputs to the deconfliction process as well as the process itself, but we must first provide more details about the current state of deconfliction.

One source of information that currently feeds military no-strike databases is a collection of ad-hoc humanitarian notification systems, often also referred to as humanitarian deconfliction mechanisms. These systems provide location information about humanitarian operations to military actors, but they vary from conflict to conflict and are notoriously unreliable for both the humanitarian organizations that voluntarily provide information and the military entities that receive the information (Lewis, 2022; Ulbricht & Weiner, 2021). At the time of writing, such systems have been or are actively being used in multiple conflicts including Ukraine (OCHA, 2022), Lebanon (OCHA, 2006), Syria (OCHA, 2018), and Yemen (OCHA, 2021).[1]

In all existing humanitarian notification/deconfliction systems, humanitarian relief actors voluntarily report data about their operations. These reports are typically made through email to some central data collection and sharing entity. For example, the UN Office for the Coordination of Humanitarian Affairs (OCHA) has operated what they label a "humanitarian deconfliction mechanism" in Syria since 2014 that shares information with parties to the conflict about static humanitarian locations and humanitarian missions involving movement (OCHA, 2018). Humanitarian organizations voluntarily participate by emailing a completed template to OCHA, which then transmits the exact information received, without attempts to verify or validate the information, to parties to the conflict. The submitted data can contain errors and inconsistencies, which lends credibility to claims that the mechanism cannot be trusted and is vulnerable to misuse (Hill & Hurst, 2019).

Even if OCHA's humanitarian deconfliction mechanism for Syria functions properly as an information dissemination mechanism, the parties to conflict must use that information appropriately in subsequent targeting decisions. In particular, data about participating humanitarian organizations should ideally also be used to ensure that positive identification of military targets actually occurs, particularly since human air strike teams "have misidentified civilians as legitimate targets in case after case after case" (Yager, 2022). To this end, there are multiple opportunities to leverage AI ISR systems to improve awareness of humanitarian relief activities. First, for humanitarian organizations, these

[1] All four of the systems cited are operated by UN OCHA, but there is substantial between-system variation. Little can be gleaned about the system operated in Ukraine, other than the fact that it exists. The system operated in Lebanon in 2006 was specifically focused on notifications to Israeli Defense Forces related to World Food Programme convoys. In contrast, a five-page document outlining the humanitarian deconfliction mechanism used in Syria includes templates for Mission Movement and Static Location information submission, open to all humanitarian organizations. The system operated in Yemen is designed to inform the Saudi-led coalition of humanitarian operations and includes specific guidelines and information submission forms.

technologies could enable a reliable and transparent method of submitting data that can reduce potential sources of error, thereby undercutting a commonly stated reason for parties to conflict to disregard these notifications. Second, if parties to conflict incorporate humanitarian notification data into their AI ISR systems, then they will likely have improved battlespace awareness (assuming the data can be trusted; see the previous sentence). Third, more effective methods of communicating, vetting, and validating data about humanitarian activities though AI ISR systems will allow greater accountability so that humanitarian organizations and the global community as a whole can better identify bad actors and hold them responsible for their actions. Fourth, and most directly, IHL *obligates* parties to conflict to implement feasible civilian risk mitigation measures, and these systems arguably provide such an opportunity.

Current efforts to improve humanitarian notification and our proposal

There are emerging efforts to improve the state of deconfliction processes, the humanitarian notification systems that provide them with information, and overall civilian harm mitigation efforts. Specifically, modern militaries have noted possible ways to leverage AI capabilities when they see their IHL obligations as both a moral and legal imperative as well as a key to strategic success. For example, Actions 4.i., 4.l., and 4.n. of DoD's Action Plan specifically reference the need to enhance civilian harm mitigation efforts using technologies such as AI/ML (DoD, 2022b, pp. 13–14). We suggest here that we can build on this foundation in ways that significantly amplify such efforts.

Specifically, we propose that militaries should establish interfaces for humanitarian organizations into their AI ISR systems. Such interfaces would create possibilities for AI-assisted verification and validation of humanitarian personnel, objects, and activities that extend and improve on ad hoc methods such as emailing data. Currently, human intelligence analysts must spend time analyzing data that is putatively submitted by a humanitarian organization in order to decide if a particular entry in a no-strike database is legitimate. However, an AI ISR system that receives location data from a humanitarian organization can vet the source of the data, whether through military-supplied and -signed credentials to the humanitarian organization or zero-trust architectures (Rose et al., 2020) (or both).

Given confidence in the data source, the AI ISR can then automatically match the submitted location against geospatial imagery and other data sources to validate the reported humanitarian activities. As a concrete example, image recognition techniques applied to geospatial imagery could help to validate information about a surgical hospital for war wounded set up by the ICRC, assuming it followed established marking standards. This vetting and validation can occur at speeds and using information sources that far exceed those of human intelligence analysts.

One key benefit of this proposal is precisely that AI ISR systems can incorporate many different types of data, and so can provide additional confirmation and validation of assertions by humanitarian organizations. For example, Open Source Intelligence (OSINT) AI systems can use web scraping to match publicly available, geotagged images or videos of a humanitarian hospital with submitted data. Signals Intelligence (SIGINT) AI systems can match the location of intercepted calls with the reported site of the humanitarian hospital. Human Intelligence (HUMINT) reports can be analyzed by a natural language processing system for references to the humanitarian hospital. And many other data sources can be used as relevant. Moreover, the humanitarian organizations' data would be provided directly (and presumably securely) to the military AI ISR system, thereby creating possibilities for those organizations to provide data beyond simple text reports. All of these sources of data can then be incorporated into an overall analysis that vets and validates submitted humanitarian data with zero trust required.

Moreover, since the process would occur in a trusted manner, we suggest that the interface need not be unidirectional. In particular, the AI ISR system can provide formal confirmation that data submitted by a humanitarian organization has been vetted and validated, and is being incorporated into the overall assessment (though the AI ISR might conclude that there are reasons to question the input data). Confirmation does not guarantee the safety of humanitarian organizations, but does offer a modicum of transparency by acknowledging that information has been submitted correctly and included in analyses.

Looking beyond situational awareness of parties to conflict, interfaces to AI ISR systems can also have second- and third-order positive effects. First, the bidirectional nature of the interface could be used to inform humanitarian actors about threats from unexploded ordnance. Second, an AI ISR interface for humanitarian organizations offers the potential to establish patterns of IHL violations, which can be used to hold parties to conflict accountable. Third, such an AI ISR interface would allow parties to conflict to leverage AI to rapidly vet and validate submitted humanitarian information on a continuous basis in a way that existing email-based humanitarian notification systems cannot match.

In summary, creating a two-way interface with AI ISR systems for humanitarian organizations not only will improve the communication of quality data for deconfliction processes, but also will support impartial, independent,

and neutral humanitarian operations; create the possibility of accountability for targeting processes; and provide a system that is flexible and fast enough to be appropriate for the dynamic nature of humanitarian operations in conflict zones. Since this technological possibility exists, and since IHL mandates that humanitarian organizations have unimpeded access to civilian populations, one could argue that parties to a conflict have a legal obligation, not a "merely" ethical and moral one, to establish such an interface. Already, the ICRC highlights the development of "environment scanning dashboards" that use AI and ML to assess humanitarian needs in conflict zones (ICRC, 2019). Allowing such dashboards access to broader datasets would allow the international community to better live up to obligations to protect civilians under IHL.

Objections, obligations, and conclusions

We acknowledge that there are numerous reasons for both military actors and humanitarian organizations to object to this proposal. Humanitarian organizations have pressing concerns about security, particularly data security. Although all parties to conflict have an obligation to respect humanitarian operations, there are known examples of nefarious actors who intentionally target humanitarian operations (Mauvais, 2020). All parties to conflict, whether state or non-state actors, are aligned with a side or group, and so their stance is fundamentally different to humanitarian operations, which are grounded in principles of humanity, neutrality, impartiality, and independence. This fundamental difference means that interfaces to AI/ISR systems, and the subsequent uses of those data, must be designed to accord with humanitarian principles, not military desires.

Militaries may object to providing any sort of interface to their AI ISR systems, as it could provide a novel attack vector either directly or via data poisoning. However, we contend that humanitarian interfaces with military AI ISR systems (or something functionally similar) are not merely desirable, but actually *obligatory*. IHL requires that states respect and protect humanitarian relief workers, including ensuring their safety when legitimately delivering humanitarian aid in a conflict zone. Many parties to conflicts are not fully honoring that obligation at present, partly because of their inability to receive vetted and validated location (and other) data about humanitarian efforts. Parties thus have a defeasible obligation to obtain more accurate and usable data, and an interface to their AI ISR systems is a powerful means toward that end.

This position paper is a step toward identifying concrete opportunities for militaries and humanitarian organizations to leverage existing AI technology to reduce human suffering in warfare. There will likely be pushback from both humanitarian organizations who may be skeptical or not have the data/tech literacy, and also militaries who may prefer a status quo that is simpler (and more cynically, that affords them deniability and minimal accountability). We contend that there is a better way that leverages existing AI ISR systems to enhance humanitarian organization safety. This better way would not only improve military awareness and understanding, but also ensure that states better satisfy their obligations. If one has the ability to better understand a conflict zone, and so make more ethical and effective decisions, then one has an obligation to humanity to act.

Acknowledgements Thanks to Shawn Sippin for helpful feedback on an earlier version of this paper. DT was partially supported by the Swiss Drones and Robotics Centre. The views in this paper are solely those of the authors and do not represent the position of any other group or agency.

Funding Open access funding provided by University of St.Gallen
Open access funding provided by University of St.Gallen

References

Allen, G. C. (2022). Russia Probably Has Not Used AI-Enabled Weapons in Ukraine, but That Could Change. Center for Strategic and International Studies. https://www.csis.org/analysis/russia-probably-has-not-used-ai-enabled-weapons-ukraine-could-change

Arkin, R. (2018). Lethal Autonomous Systems and the plight of the non-combatant. In R. Kiggins (Ed.), *The Political Economy of Robots* (pp. 317–326). Switzerland: Palgrave Macmillan.

Arkin, R. (2010). The case for ethical autonomy in Unmanned Systems. *Journal of Military Ethics*, *9*(4), 332–341.

Asaro, P. (2019). Algorithms of violence: critical social perspectives on Autonomous Weapons, Special Issue on Algorithms. *Social Research*, *86*(2), 537–555.

Asaro, P. (2012). On Banning Autonomous Weapon Systems: Human Rights, automation, and the dehumanization of Lethal decision-making. *International Review of the Red Cross*, *94*(886), 687–709.

Danks, D. (2020). How adversarial attacks could destabilize military AI systems.IEEE Spectrum.

DoD (2022a). Artificial Intelligence, Autonomy Will Play Crucial Role in Warfare, General Says. https://www.defense.gov/News/News-Stories/Article/Article/2928194/artificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says/

DoD (2022b).) Civilian Harm Mitigation and Response Action Plan. https://media.defense.gov/2022/Aug/25/2003064740/-1/-1/1/Civilian-harm-mitigation-and-response-action-plan.PDF

Future of Life Institute (2021). Autonomous Weapons: An Open Letter from AI & Robotics Researchers. https://futureoflife.org/%20open-letter-autonomous-weapons/

Hill, E., & Hurst, W. (2019). The U.N. Tried to Save Hospitals in Syria. It Didn't Work.International New York Times. https://www.nytimes.com/2019/12/29/world/middleeast/united-nations-syria-russia.html

Hoehn, J. R. (2022). Joint All-Domain Command and Control (Jadc2). https://sgp.fas.org/crs/natsec/IF11493.pdf

Humanitarian Outcomes (2022). Aid Worker Security Report 2022: *Figures at a Glance*www.humanitarianoutcomes.org/sites/default/files/publications/awsd_figures_2022.pdf

ICRC (2005). *Customary IHL Database*. https://ihl-databases.icrc.org/en/customary-ihl

ICRC (2019). Artificial intelligence and machine learning in armed conflict: A human-centred approach. https://www.icrc.org/en/download/file/96992/ai_and_machine_learning_in_armed_conflict-icrc.pdf

Joint Chiefs of Staff (2013). Joint Publication 3–60 "Joint Targeting." https://www.justsecurity.org/wp-content/uploads/2015/06/Joint_Chiefs-Joint_Targeting_20130131.pdf

Lewis, L. (2022). Improving Protection of Humanitarian Organizations in Armed Conflict. https://www.cna.org/reports/2022/03/Improving-Protection-of-Humanitarian-Organizations-in-Armed-Conflict.pdf

Mauvais, L. (2020). Syria is the deadliest place for aid workers, and there is little hope for change. https://syriadirect.org/syria-is-the-deadliest-place-for-aid-workers-and-there-is-little-hope-for-change/

OCHA (2022). Ukraine: Useful contacts and links. https://reports.unocha.org/en/country/ukraine/card/4HvLutumfJ/

OCHA (2021). The Humanitarian Notification System. Humanitarian Response. https://www.humanitarianresponse.info/en/operations/yemen/deconfliction

OCHA (2006). Lebanon Crisis 2006 Interim Report: Humanitarian Response in Lebanon, 12 Jul to 30 Aug 2006. Relief Web. https://reliefweb.int/report/lebanon/lebanon-crisis-2006-interim-report-humanitarian-response-lebanon-12-jul-30-aug-2006

OCHA (2018). Humanitarian Deconfliction Mechanism: Humanitarian Organisations Operating in Syria (Humanitarian Mission Movements & Static Locations). https://www.humanitarianresponse.info/sites/www.humanitarianresponse.info/files/documents/files/deconfliction_syria_for_static_non_static_feb2018_eng.pdf

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977. https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Article.xsp?action=openDocument&documentId=B67EDFC718BF74E3C12563CD0051DFC4

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture (no. NIST Special publication (SP) 800 – 207)*. National Institute of Standards and Technology.

Royal Australian Air Force (RAAF) (2019). At the Edge Fifth Generation Air Force: Our Human Edge in the Information Age. https://view.publitas.com/jericho/at-the-edge/page/1

Russell, S., Aguirre, A., Javorsky, E., & Tegmark, M. (2021). LethalAutonomous Weapons Exist; They Must Be Banned.IEEE Spectrum. https://spectrum.ieee.org/automaton/robotics/military-robots/lethal-autonomous-weapons-exist-they-must-be-banned

Ulbright, B. R., & Weiner, A. S. (2021). Humanitarian Notification Systems & Intentional Attacks Against Hospitals. https://lieber.westpoint.edu/humanitarian-notification-systems-intentional-attacks-against-hospitals/

UNDSS (2017). Chapter II: United Nations Security Management System (UNSMS), Section F: Saving Lives Together in United Nations Security Management System (UNSMS). https://www.un.org/en/pdfs/undss-unsms_policy_ebook.pdf

Yager, S. (2022). Lost Innocents: The U.S. Military's Shameful Failure to Protect Civilians.Foreign Affairs. https://www.foreignaffairs.com/articles/united-states/2022-01-25/lost-innocents