ASSER INSTITUTE

Centre for International & European Law

DILEMA

DILEMA 2023 Conference

Designing International Law and Ethics into Military Artificial Intelligence

12–13 October 2023
Asser Institute, The Hague

DILEMA

Designing International Law and Ethics into Military Artificial Intelligence

DILEMA 2023 Conference

CONFERENCE MATERIALS

12–13 October 2023

Asser Institute (The Hague)

ASSER INSTITUTE

Centre for International & European Law

Introduction

The **DILEMA 2023 Conference**, held on 12–13 October 2023 at the Asser Institute in The Hague, addresses the complex and interdisciplinary issues raised by military applications of artificial intelligence (AI). The potential embedding of AI technologies in weapons systems has been an important subject of scholarly as well as policy debates for many years. More recently, other applications of AI in the military domain, such as AI-driven decision support systems for intelligence, surveillance, risk assessment, detention operations, or target identification, have also received attention. The deployment of AI technologies in a military context raises major legal and ethical concerns, but also opportunities to improve military performance and decision making, and possibly to mitigate harm in armed conflict. The DILEMA conference offers a broad platform to engage in an interdisciplinary dialogue around both theoretical and practical questions related to military AI, and features some of the latest research insights from the fields of law, ethics, computer science, and other disciplines. The conference seeks in particular to promote innovative perspectives that contribute to advancing the boundaries of research in the field of military AI.

DILEMA Project

The DILEMA Project on Designing International Law and Ethics into Military Artificial Intelligence is a research project carried out at the Asser Institute (University of Amsterdam) and funded by the Dutch Research Council (NWO). Launched in 2020, the project explores interdisciplinary perspectives on military applications of AI, with a focus on legal, ethical, and technical approaches on safeguarding human agency over military AI.

Website: www.asser.nl/DILEMA

Organisation Committee

- Dr Bérénice Boutin, Senior Researcher in International Law, DILEMA Project Leader (Asser Institute)
- Dr Marta Bo, Senior Researcher in International Law (Asser Institute), Associate Senior Researcher (SIPRI)
- Prof. Tom van Engers, Professor of Legal Knowledge Management (University of Amsterdam)
- Klaudia Klonowska, PhD Researcher in International Law (Asser Institute)
- Dr Magdalena Pacholska, Postdoctoral Researcher in International Law, Marie Skłodowska-Curie Fellow (Asser Institute)
- Dr Sadjad Soltanzadeh, Postdoctoral Researcher in Ethics and Philosophy of Technology (Asser Institute)
- Taylor Woodcock, PhD Researcher in International Law (Asser Institute)
- Dr Tomasz Zurek, Postdoctoral Researcher in Computer Science (University of Amsterdam), Associate Fellow (Asser Institute)
- Prof. Marten Zwanenburg, Professor of Military Law (University of Amsterdam; NLDA)

Programme

Thursday 12 October 2023

09:15 – 09:45 Arrival and registration

09:45 – 10:00 Opening

Dr Bérénice Boutin (Asser Institute, DILEMA Project)

10:00 – 11:00 **Keynote Lecture – AI in Critical Decisions: is Human Judgment a Legal Requirement?**

Professor Noam Lubell (University of Essex)

11:00 – 11:30 Break

11:30 – 13:00 **Panel 1: Compliance and Responsibility in Relation to Military AI**

Chair: *Dr Bérénice Boutin (Asser Institute, DILEMA Project)*

- ChatGPT for the Military? Large Language Models in the Military Domain and the Role of Article 36 Weapons Reviews

  Dr Elisabeth Hoffberger-Pippan (Peace Research Institute Frankfurt)
- The Use of Autonomous Cyber Capabilities in Armed Conflicts: Legal and Ethical Implications

  Marta Stroppa (Sant'Anna School of Advanced Studies)
- "Teaching New Systems Old Tricks": Do Autonomous Weapon Systems Need Specific ROE?

  Dr Marcin Marcinko (Jagiellonian University in Krakow)
- Who Hunts the Hunters? Who is Liable when Military Artificial Intelligence Goes Wrong

  Dr Brendan Walker-Munro and Dr Sam Hartridge (University of Queensland)

13:00 – 14:00 Lunch

14:00 – 15:30 **Panel 2: Human-Machine Interaction in Military Decision-Making**

Chair: *Klaudia Klonowska (Asser Institute, DILEMA Project)*

- Above the Law: Drones, Aerial Vision, and the Law of Armed Conflict – A Socio-Technical Approach

  Professor Shiri Krebs (Deakin University; Stanford University)
- Artificial Intelligence and the Rupture of the Rationalist Myth of Military IHL Decision-Making Processes: A Move Towards Emotions in IHL?

  Anna Rosalie Greipl (Geneva Graduate Institute (IHEID) and Geneva Academy of International Humanitarian Law and Human Rights)
- Context-Driven Analysis of AI-Based Solutions: Methods and Tools Supporting Efficient Verification, Validation and Certification

  Dr Iris Cohen and Dr Gregor Pavlin (Thales Nederland B.V.)

15:30 – 16:00 Break

16:00 – 17:15 **Panel 3: The Impact of Military AI on Conflict**

Chair: *Dr Magda Pacholska (Asser Institute, DILEMA Project)*

- That We Are Constant: Algorithmic Warfare, Spontaneous Political Action, and the Right to Self-Determination

  Dr Henning Lahmann (Leiden University)
- The IDF Introduces: Artificial Intelligence in the Battlefield, A New Frontier?

  Dr Tal Mimran (Hebrew University and Tachlith Institute)
- A Mechanism for Overcoming Real or Perceived Problems of Disparate Ethical Outlooks in Multilateral Alliances when Evaluating the Impacts of AI in a Military Context

  Dr Michael Wildenauer (University of Melbourne)

Evening Conference dinner

Friday 13 October 2023

09:30 – 09:45 Arrival

09:45 – 10:45 **Keynote Lecture – Seated on the Whirlwind: Artificial Intelligence, Weapons Systems and Moral Agency**

Dr Elke Schwarz (Queen Mary University London)

10:45 – 11:15 Break

11:15 – 12:30 **Panel 4: The Impact of Military AI on Discourses and Doctrines**

Chair: *Taylor Woodcock (Asser Institute, DILEMA Project)*

- The Role of Military AI in Deterrence Theory and Practice

  Dr Bianca Baggiarini (Australian National University)
- Visuals as Sources of Normativity in the Debate about Artificial Intelligence in Weapon Systems

  Dr Ingvild Bode (University of Southern Denmark), Dr Guangyu Qiao-Franco (Radboud University Nijmegen), Anna Nadibaidze (University of Southern Denmark)
- Platform Warfare: The Rise and Implications of Platform Corporations in AI-Centered Warfare

  Dr Marijn Hoijtink (University of Antwerp)

12:30 – 13:30 Lunch

13:30 – 14:45 **Panel 5: Military AI and the Reconfiguration of Human Agency**

Chair: *Dr Sadjad Soltanzadeh (Asser Institute, DILEMA Project)*

- The AI-Augmented Super Soldier: Enhancement, Interfaces and the Extended Cognition of Human-Machines

  Dr Thomas Christian Bächle (University of Bonn; Humboldt Institute for Internet and Society Berlin)
- Artificial Decision-Making on Life-or-Death: The Moral-Psychological Implications of Increasing Autonomy in Weapon Systems

  Dr Tine Molendijk, Professor Lonneke Peperkamp, Sofie van der Maarel (NLDA)
- Proportionality, Intentions, and Human Agency

  Dr Elad Uzan (University of Oxford)

14:45 – 15:15 Break

15:15 – 16:45 **Panel 6: Governance of Military AI**

Chair: *Dr Marta Bo (Asser Institute; SIPRI)*

- ELSI goes to War – Criticisms of Autonomous Weapon Systems and the Rise of a Responsible AI

  Dr Jens Hälterlein and Professor Jutta Weber (Paderborn University)
- "This is my Last Resort" - Overcoming the Stalemate in Autonomous Weapons Regulation through National Legislation and Industry Self-Regulation

  Marcel Schliebs (University of Oxford) and Vanessa Vohs (Bundeswehr University Munich)
- 'Cyber-ing' the AI Regime: Path Dependencies and Pathologies to Avoid in the Creation of Rules for Military AI

  Dr Arun Sukumar (Leiden University)

16:45 – 17:00 Closing

17:00 – 18:30 Reception

Keynote Speakers

Professor Noam Lubell

AI in Critical Decisions: is Human Judgment a Legal Requirement?

Noam Lubell is a Professor in the School of Law, University of Essex, and is the Director of the Essex Armed Conflict and Crisis Hub. He was Head of the Law School from 2014 to 2017. Between 2013-2019 Prof Lubell held the Swiss Chair of International Humanitarian Law at The Geneva Academy, and from 2010 to 2018 he was the Rapporteur of the International Law Association's Committee on the Use of Force. Prof Lubell is a Senior Research Fellow at the Johns Hopkins University Applied Physics Laboratory in the US, working on the legal aspects of new military technologies. He is also a Research Associate at the Federmann Cyber Security Research Center, Hebrew University. He has taught courses on international human rights law and the laws of armed conflict in the UK, Ireland, US, Israel, and Switzerland. Prior to his academic career, from the late 1990s until 2005 he worked with human rights NGOs and legal clinics on issues of protection during armed conflict. Prof Lubell has been a member of numerous expert groups and consultations with governments, the ICRC and the UN on topics such as the law of occupation, self-defence, the scope of the battlefield, and autonomous weapon systems, and is the author of the book 'Extraterritorial Use of Force Against Non-State Actors' (Oxford University Press). He co-led a five-year project in partnership with the International Committee of the Red Cross, to produce the "Guidelines on Investigating Violations of International Humanitarian Law."

Dr Elke Schwarz

Seated on the Whirlwind: Artificial Intelligence, Weapons Systems and Moral Agency

Dr Elke Schwarz is Reader (Associate Professor) in Political Theory at Queen Mary University London. Her research focuses on the intersection of ethics of war and ethics of technology with an emphasis on unmanned and autonomous / intelligent military technologies and their impact on the politics of contemporary warfare. She is the author of 'Death Machines: The Ethics of Violent Technologies' (Manchester University Press), is an RSA Fellow, a member of the International Committee for Robot Arms Control (ICRAC), 2022/23 Fellow at the Center for Apocalyptic and Post-Apocalyptic Studies (CAPAS) in Heidelberg and 2024 Leverhulme Research Fellow. Her work has been published in a number of philosophical and security focused journals, including Philosophy Today, Security Dialogue, Critical Studies on Terrorism and the Journal of International Political Theory among others. She is co-series editor for the Springer Verlag series: Frontiers in International Relations and Associate Editor for the Journal New Perspective.

Extended Abstracts

1. ChatGPT for the Military? Large Language Models in the Military Domain and the Role of Article 36 Weapons Reviews

Dr Elisabeth Hoffberger-Pippan (Peace Research Institute Frankfurt)

I. Introduction

When the company OpenAI released ChatGPT – a large language model – to the public in 2022, the world was mesmerized by this novel, if not to say, revolutionary technology.[1] While initial problems with GPT-3, the first neural network behind ChatGPT, kept OpenAI busy for some time,[2] the updated and much larger version GPT-4[3] operates much better having passed *inter alia* the bar exam.[4] There is, understandably, significant excitement about this technology. But at the same time AI researchers are increasingly alarmed, some of them calling to pause the training of AI models for at least six months.[5]

It did not take long until military decision makers and industry started to invest time and energy into the potential application of large language models in the defence sector. One of the first projects (or rather experiments) delving into the various intricacies related to LLM was Hermes, an experimental large language model developed by the company ScaleAI.[6] Hermes was conducted in cooperation with the US Marine Corps School of Advanced Warfighting allowing students to use a LLM for a campaigning exercise beneath the threshold of an (international) armed conflict.[7] Other (real) projects include Donovan[8] from the same company as well as the Artificial Intelligence Platform (AIP)[9] and Gotham[10] by Palantir. Furthermore, the US has established Project Lima with a view to studying the potential and risks associated with generative AI in the military domain.[11] Thus, what we can observe is a heightened interest

[1] Bernard Marr, 'A Short History Of ChatGPT: How We Got To Where We Are Today', Forbes (19 May 2023), available at https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/?sh=46141dc9674f (last accessed 5 September 2023).

[2] Will Douglas Heaven, 'GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why', MIT Technology Review (14 March 2023), available at https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai (last accessed 5 September 2023).

[3] OpenAI, 'GPT4', available at https://openai.com/gpt-4 (last accessed 5 September 2023).

[4] With critical reflections see Karen Sloan, 'Stellar or so-so? ChatGPT bar exam performance sparks differing opinions', Reuters (21 May 2023), available at https://www.reuters.com/legal/transactional/stellar-or-so-so-chatgpt-bar-exam-performance-sparks-differing-opinions-2023-05-31/ (last accessed 5 September 2023).

[5] The Future of Life Institute, 'Pause Giant AI Experiments: An Open Letter' (22 March 2023), available at https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (last accessed 5 September 2023).

[6] Cf. Benjamin Jensen and Dan Tadross, 'How Large Language Models Can Revolutionize Military Planning', WarontheRocks (12 April 2023), available at https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/ (last accessed 5 September 2023).

[7] Ibid.

[8] OpenAI, 'Donovan', available at https://scale.com/donovan (last accessed 5 September 2023).

[9] Palantir, 'AIP', available at https://www.palantir.com/aip/defense/ (last accessed 5 September 2023).

[10] Palantir, 'Gotham', available at https://www.palantir.com/offerings/defense/ (last accessed 5 September 2023).

[11] Colin Demarest, 'Pentagon establishes Task Force Lima to study generative AI issues', *C4ISRNET* (10 August 2023), available at https://www.c4isrnet.com/artificial-intelligence/2023/08/10/pentagon-establishes-task-force-lima-to-study-generative-ai-issues/ (last accessed 5 September 2023).

in LLM in the military domain. As with any other AI-related technology, LLM in the defence sector raise a plethora of legal, ethical, security-related and military questions, which will be discussed further below.

This presentation sets out to examine the potential applications and associated risks of LLM in the military context. Furthermore, the presentation will examine if and how LLM fall under the purview of Article 36 AP I GC obliging States Parties to to the GC to review weapons, means and methods of warfare. LLM most likely qualify as decision-support systems which are, arguably, covered by the obligation of States Parties to review weapons, means and methods of warfare.[12] It is important to bear in mind, however, that LLM can be used in different domains and for different purposes depending on their design but also on how they are used by human decision-makers. Some LLM could recommend or provide information with direct implications for the targeting process[13], while other LLM could only have indirect or no implications for the targeting process as such. Thus, depending on the design but also on the question of how LLM are used, different legal questions arise, especially but not exclusively with regard to Article 36 AP I GC.

The use of LLM and other generative AI models are still in their infancy.[14] Hence, it is still not clear how exactly they will be designed and used in future conflicts. What is almost certain, though, is that they will be employed one way or another necessitating a thorough, although preliminary analysis.

II. Definition of large language models and their use in a military context

Analyzing the legal challenges when using LLM in the military requires a sound understanding of the technology as such but also AI in more general terms. Furthermore, in order to address the various legal but also practical challenges related to LLM in the military, projects, such as Hermes but also Donovan, API and Gotham, warrant further consideration.

1. The definition of LLM

Generally speaking, LLM are a subset of deep-learning. Deep-learning, in turn, is a subset of machine learning[15] that is based on a concept called "neural networks" with three or more layers without the necessity of human intervention during the learning process. Deep learning models can work with unlabelled (unstructured) data where the input and output data are unknown.[16] LLM are part of so-called Generative AI. Generative AI is an umbrella term used to define technology that is able to create any kind

[12] Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies of Warfare' in *TMC Asser Press*, *Research Paper Series* (April 2021) 2; Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies' in *Yearbook of International Humanitarian Law* Volume 23 (2020) 123 – 153 (124).

[13] For a definition of the targeting process (or targeting cycle) see US Joint Chiefs of Staff, 'Joint Targeting', Joint Publication 3-60, available at https://jfsc.ndu.edu/Portals/72/Documents/JC2IOS/Additional_Reading/1F4_jp3-60.pdf (last accessed 5 September 2023).

[14] Cf. Benjamin Jensen and Dan Tadross, 'How Large Language Models Can Revolutionize Military Planning', *WarontheRocks* (12 April 2023).

[15] For a definition of machine learning see ; IBM, 'Machine Learning', available at https://www.ibm.com/topics/machine-learning (last accessed 5 September 2023); for a definition of AI see IBM, 'Artificial Intelligence', available at https://www.ibm.com/topics/artificial-intelligence (last accessed 5 September 2023). See also John McCarthy, 'What is Artificial Intelligence?', *Computer Science Department*, Stanford University (revised 12 November 2007), available at https://www-formal.stanford.edu/jmc/whatisai.pdf (last accessed 5 September 2023).

[16] IBM, 'Deep Learning', https://www.ibm.com/topics/deep-learning (last accessed 5 September 2023).

of content, such as text, images or audible data.[17] LLM are not necessarily a completely new phenomenon. They have been studies for years, especially throughout the 80s where computer scientists primarily applied statistical models in order to model language. The most important model developed at that time was the so-called recurrent neural network (RNN)[18], which is based on sequential data (data have to be processed step-by-step). In 2017, *Vaswani et al.*[19] introduced so-called transformers which enabled neural networks to process data simultaneously allowing for significantly faster responses to assigned tasks. GPT – so-called generative pretrained transformers, such as GPT 2, 3, ChatGPT and GPT4 are thus able to accomplish complex tasks in comparatively short time. It did not take long until the military showed interest in the technology. The US Military is currently testing a number of LLM without disclosing which products they are particularly interested in.[20] While LLM used for the civilian sector take their data from basically all over the internet, military applications will have to use their own data in order to protect national security and fend off adversarial attacks as effectively as possible.[21]

From a legal perspective, LLM used in armed conflict most likely qualify as decision-support systems (DSS). Decision-support systems, in turn, denote "tools that use AI techniques to analyse data, provide actionable recommendations"[22] and "assist decision-makers situated at different levels in the chain of command to solve semi-structured and unstructured decision tasks"[23].

2. Hermes – prospects and limitations of LLM in the military domain

The students using Hermes were tasked to consult the LLM in a fictional scenario beneath the threshold of armed conflict. Using the model yielded a number of advantages but also limitations. The LLM helped to save time and to provide a better and deeper understanding regarding questions of a more general nature. Students were particularly interested in knowing more about the economic status of the particular country in order to have a better understanding of the overall situation and to plan military strategies accordingly. The LLM helped students to refine their respective course of action and it also provided

[17] IBM, 'What is generative AI?', available at https://research.ibm.com/blog/what-is-generative-AI (last accessed 5 September 2023).

[18] IBM, 'What are recurrent neural networks?', available at https://www.ibm.com/topics/recurrent-neural-networks (last accessed 5 September 2023).

[19] Ashish Vaswani et al., 'Attention is all you need', Computation and Language (2017), available at https://arxiv.org/abs/1706.03762 (last accessed 5 September 2023).

[20] Katrina Manson, 'The US Military Is Taking Generative AI Out for a Spin', *Bloomberg* (5 July 2023), available at https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin (last accessed 5 September 2023).

[21] Josh Luckenbaugh, 'Army Hopes AI Will Give Soldiers An Information Advantage', *NationalDefense*, available at https://www.nationaldefensemagazine.org/articles/2023/7/21/army-hopes-ai-will-give-soldiers-an-information-advantage (last accessed 5 September 2023).

[22] Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies of Warfare' in *TMC Asser Press, Research Paper Series* (April 2021) 2; Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies' in *Yearbook of International Humanitarian Law* Volume 23 (2020) 123 – 153 (124).

[23] Elena Susnea, 'Decision Support Systems in Military Actions: Necessity, Possibilities and Constraints', Journal of Defense Resources Management (2012) 2. Seen in Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies of Warfare' in *TMC Asser Press, Research Paper Series* (April 2021) 2; Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies' in *Yearbook of International Humanitarian Law Volume* 23 (2020) 123 – 153 (124).

information on the adversary's military doctrine.[24] All these advantages aside, Hermes also showed the significant limitations of such a system. A LLM, like any other AI-enabled technology used in a military context, cannot replace human decision-making and cognition. Hallucinations and stochastic parroting are still problems humans need to grapple with when using. Most importantly, most exercises and most models are – at least according to the current status quo – trained in static settings, where adversaries remain silent. The true essence of war is a disarray, which LLM in general are still unable to grapple with. This in turn could negatively influence human decision-making.[25] Furthermore, humans must be trained to identify misinformation if a LLM is hallucinating but given the speed of warfare in general, the capacity to distinguish right from wrong might be tremendously hampered. Students working with Hermes were trained to ask questions the system could answer adequately but confirmation bias – understood as the propensity to interpret information in a way that aligns with one's own beliefs – was one of the key challenges during that exercise. This and other aspects need to be addressed accordingly in case States seriously contemplated the use of such systems in a military context. Moreover, it is questionable how international law, especially Article 36 AP I GC, responds to the use of LLM in the military.

III. The scope and limitations of Article 36 Weapons Reviews

There are reasonable arguments to assume that LLM qualify as DSS an thus arguably fall under the purview of Article 36 AP I GC. According to Article 36 AP I GC, "[I]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party". Alas, Article 36 does not provide for a definition of these terms and States Parties to the GC have interpreted and applied Article 36 differently. When it comes to LLM as DSS, there are strong reasons to assume that LLM and DSS qualify as means of warfare.

Klonowska has convincingly suggested to apply four criteria in order to ascertain whether a DSS falls under the purview of Article 36 AP I GC: a) IHL compliance b) integrity to military decision-making, c) significance to military operations and d) contribution to offensive capabilities.[26] In practice, determining whether a LLM fulfils all of the criteria above, might prove challenging, however.

First, experiences from the civil sector show that LLM can fulfil a number of tasks but on various occasions displayed unintended and highly unpredictable behaviour. Thus, speaking from a mere technical perspective, it seems to be difficult to predict and predetermine the scope of tasks a LLM will be able to accomplish. Connecting a LLM to a large dataset might yield surprising and unintended outcomes that

[24] Benjamin Jensen and Dan Tadross, 'How Large Language Models Can Revolutionize Military Planning', *WarontheRocks* (12 April 2023), available at https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/ (last accessed 5 September 2023).

[25] Ian Reynolds and Ozan Ahmet Cetin, 'War is messy. AI can't handle it.' *Bulletin of the Atomic Scientists* (14 August 2023), available at https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/ (last accessed 5 September 2023).

[26] Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies of Warfare' in TMC Asser Press, Research Paper Series (April 2021) 2; Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies' in *Yearbook of International Humanitarian Law* Volume 23 (2020) 123 – 153 (124).

could – theoretically – result in the "co-production" of hostilities.[27] Since the obligation to review weapons is first and foremost an ex-ante obligation,[28] military decision-makers will have to critically assess this challenge before a particular system is deployed.

Second, if one takes the view that LLM fall under the purview of Article 36 AP I GC, it is imperative to ascertain how States will review LLM in light of the fact that most LLM are based on self-learning mechanisms. This topic is certainly not new[29] but it deserves particular scholarly attention if States seriously considered the use of LLM in a military context. Even though in its traditional meaning Article 36 AP I GC serves as an ex-ante mechanism, periodic and post-deployment reviews will be necessary in order to account for the various challenges post by LLM.

Third, and perhaps most importantly, LLM will probably not work in complete isolation from other LLM or AI-enabled technology in general. These "path-dependencies" and the overall interconnectedness of software components could make it extremely difficult to isolate one single LLM from other, connected systems.[30] Hence, the third question that will need to be addressed by decision-makers is how far a review goes and what kind of components will be part of it.

Thus, applying the four-prong test for DSS in the context of Article 36 AP I GC might be challenging when it comes to LLM. This, in turn, might have two different consequences: Either it is argued that LLM in general fall under the obligation to undertake a legal review[31] or it is argued that we need to entirely re-think the concept of using AI in the military domain in more general terms by focussing our discussion on *due diligence* under international (humanitarian) law. In fact, applying Article 36 AP I GC to all LLM and discussing the potential role of *due diligence* in the context of LLM in the military domain are not necessarily exclusive. In fact, the two concepts could complement each other.

IV. Due diligence to avoid lacunae?

Traditionally, due diligence "applies in situations where it establishes the legal responsibility of a State in connection with the behaviour of private actors that cannot be attributed directly to the State"[32]. But in

[27] Stephen Ornes, 'The Unpredictable Abilities Emerging From Large AI Models' Quantamagazine (16 March 2023), available at https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/ (last accessed 5 September 2023).

[28] Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies of Warfare' in *TMC Asser Press, Research Paper Series* (April 2021) 2; Klaudia Klonowska, 'Article 36: Review of AI- Decision-Support Systems and Other Emerging Technologies' in *Yearbook of International Humanitarian Law Volume 23* (2020) 123 – 153 (124). See also Nehal Bhuta and Stavros-Evdokimos Pantazopoulos, 'Autonomy and uncertainty: increasingly autonomous weapons systems and the international regulation of risk' in Nehal Bhuta et al., *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016) 284 – 300 (297).

[29] See, for example Tim McFarland, 'Legal reviews of in situ learning in autonomous weapons', *Ethics and Information Technology* (2023) Vol. 25 No. 9, 1 – 9.

[30] Cf. NATO Science and Technology Organization, Meaningful Human Control of AI-based Systems

Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenario, AC/323(HFM-322)TP/1108.

[31] Damian Copeland et al. also speak in favor of an extensive reading of Article 36 AP I GC. It is not clear, however, whether they would concur with the position that LLM should be covered by Article 36 in general. See Damian Copeland et al., 'The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems', *Journal of Conflict and Security Law* (2023) Vol. 28 No. 2, 285 – 316 (294).

[32] Timo Koivurova and Krittika Singh, 'Due Diligence', in *Max Planck Yearbook of United Nations Law* (2022) mn. 3.

light of technological advancements and an AI-arms race looming, the concept of *due diligence* could also provide guidance on how States should deal with LLM in the military realm.

There are a number of indicators that *due diligence* is not necessarily constrained to apply between States and private actors. For example, the International Tribunal on the Law of the Sea in *Responsibilities and Obligations of States with respect to Activities in the Area* argued that the content of *due diligence* "may change over time as measures considered sufficiently diligent at a certain moment may become not diligent enough in light, for instance, of new scientific or technological knowledge". Even though ITLOS was referring rather to the substance of *due diligence* obligations and less to the potential addressees of the norm, it at least shows that *due diligence* as such and the various legal obligations that come along with it are certainly not set in stone. *Pasquale de Sena's* understanding of *due diligence* as a norm regulating risk in general and not solely with regard to the behaviour of private actors seems to lend support to this assumption. He states that "[o]nly those positive obligations that regulate the relationship between a state and a source of risk are obligations of diligent conduct"[33]. The concept of *due diligence* has also been discussed in the context of autonomous weapons (AWS). Already in 2016, now UNIDIR Director Robin Geiss emphasized that when it comes to AWS "[t]he identification and specification of detailed (due diligence) obligations aiming at risk prevention and harm reduction is central"[34]. Even more importantly, he stressed that "(…) in addition to the clarification of Article 36 API more emphasis is put on the specification and clarification of due diligence obligations aimed at risk prevention and harm reduction"[35]. Due diligence as applicable to AWS was also discussed in the UN Group of Governmental Experts (UN GGE) on AWS. According to Article 2 N 2 lit b) Draft Articles on Autonomous Weapons submitted by Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States at the UN GGE in March 2023, a "combatant's reliance on autonomous functions to identify, select, or engage targets" (…) "[m]ust be consistent with due diligence in the implementation of the requirements and principles of distinction, proportionality, and precautions in attack"[36].

In light of the above, there are reasons to assume that the concept of *due diligence* has or is at least about to translate into a norm that applies not only between States and private actors but also between States and other sources of risk. Hitherto, *due diligence* was mainly discussed in the context of AWS but there are no substantial arguments to assume that LLM do not pose a similar risk with potentially devastating consequences for civilians and combatants. Thus, it is argued here, that *due diligence* should be further discussed in the context of LLM and their utility for military purposes. The shape and content of *due diligence* obligations with regard to LLM need yet to be discussed and determined by States. This

[33] Pasquale De Sena, La «Due Diligence» et le lien entre le Sujet et le Risque qu'il faut Prévenir: Quelques Observations, in *Le Standard de Due Diligence et la Responsabilité International*, Société Francaise pir le Droit International) 243.

[34] Robin Geiss, 'Autonomous Systems: Risk Management and State Responsibility', Third CCW Meeting of experts on lethal autonomous weapons systems (LAWS) Geneva, 11-15 April 2016, 2, at https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/Geiss-CCW-Website.pdf (last accessed 5 September 2023).

[35] Ibid.

[36] Draft articles on autonomous weapon systems – prohibitions and other regulatory measures on the basis of international humanitarian law ("IHL"), submitted by Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States (13 March 2023) CCW/GGE.1/2023/WP.4/Rev.1.

contribution primarily aims at highlighting the need to intensify discussions on *due diligence* in the context of LLM in the military realm.

V. Constitutional AI and mechanistic interpretability in order to manage LLM: the case for human oversight

As elaborated in the previous chapter, States and policymakers should agree on common standards that apply when relying on LLM in a military context. In the past years, a trend among computer scientists could be observed trying to control (and theoretically review) LLM by referring to other AI systems and mechanisms respectively, such as constitutional AI and mechanistic interpretability, to name but two.

Constitutional AI refers to the process of supervising AI systems by using another AI-enabled agent. *Yuntao Bai et al.* describe constitutional AI as follows. "We experiment with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'"[37]. The concept of constitutional AI is certainly not new. Already in 2009 *Ronald Arkin* published his famous book Governing Lethal Behavior in Autonomous Robots arguing that robots – even if use in a military context – are capable of complying with ethical but also legal standards.[38] In fact, he argued that robots would outperform machines in terms of compliance with both ethical and legal complications. The concept of constitutional AI largely builds upon the theories developed by *Arkin* by operationalizing the concept of ethics and law by design. Even though developing such a concept is principally laudable, it can in no way replace human-decision makers in a review process. Even though Article 36 AP I GC does not indicate to what extent humans must be involved in the review, State practice (which is admittedly scarce when it comes to Article 36) indicates that a review within it the meaning of Article 36 needs to be human. To illustrate, the United States argues – when speaking in the context of AWS – that weapon systems "will go through rigorous hardware and software"[39] verification and validation.[40] By the same token, the UK emphasizes the importance of the legal adviser in the process of the review and his/her understanding of the technology serving as the backbone of any weapon system.[41]

Most AI technologies are generally regarded as a black box given the complexity of the algorithms underlying the system.[42] This is one of the strongest reasons people are concerned about using AI, especially in a military context. Mechanistic interpretability, a relatively young field of research, is aimed at reversing neural networks in order for humans to be able to understand algorithms that are based on

---

[37] Yuntao Bai et al., 'Constitutional AI: Harmlessness from AI Feedback' (2022), available at https://arxiv.org/abs/2212.08073 (last accessed 5 September 2023).

[38] Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots* (Routledge 2009).

[39] US DoD Directive, Autonomy in Weapon Systems (25 January 2023) 3, available at https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf (last accessed 5 September 2023).

[40] Natalia Jevglevskaja, *International Law and Weapons Review* (Cambridge University Press 2022) 235.

[41] UK Ministry of Defence, UK Weapons Reviews (2016), available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/507319/20160308-UK_weapon_reviews.pdf (last accessed 5 September 2023).

[42] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' Harvard Journal of Law & Technology (2018) Vol. 31 No. 2.

deep learning. According to *Arthur Conmy et al*., "researchers choose a metric and dataset that elicit the desired model behavior. Then, they apply activation patching to find which abstract neural network units are involved in the behavior. By varying the dataset, metric, and units under investigation, researchers can understand the functionality of each component"[43]. Mechanistic interpretability will help researchers to move closer to understanding AI algorithms but it cannot (not yet) entirely reverse algorithms that have led to unintended outcomes. Furthermore, it was hitherto only tested at smaller models and only isolated tasks were subjected to the reverse modelling procedure.

Both constitutional AI and mechanistic interpretability are concepts that definitely merit further discussion, especially among computer scientists. Researchers and policy makers with different educational backgrounds, however, should also keep an eye on these developments in order to evaluate whether States, in particular, use AI-enabled technology responsibly. From a legal perspective, it is questionable whether the increased "outsourcing" of control mechanisms over AI-enabled technology, such as LLM, aligns with a supposed *due diligence* standard under humanitarian law or with a requirement of human oversight when conducting a legal review within the meaning of Article 36 AP I GC. In order to use LLM in line with legal but also ethical considerations, further discussions on the interaction between humans and machines from an interdisciplinary perspective are urgently needed.

Biography

Elisabeth Hoffberger-Pippan is Senior Researcher at the Peace Research Institute Frankfurt. Her research focusses on humanitarian law and arms contro, especially AI and military robotics as well as emerging technologies in the context of biological and chemical weapons. Prior to joining PRIF she was head of the project "iPRAW" (International Panel on the Regulation of Autonomous Weapons), financed by the German Ministry of Foreign Affairs. She wrote a PhD on less-lethal weapons under international law, which was published by CUP in 2021.

[43] Arthur Conmy et al., 'Towards Automated Circuit Discovery for Mechanistic Interpretability' (2023), available at https://arxiv.org/pdf/2304.14997.pdf#:~:text=Mechanistic%20interpretability%20focuses%20on%20reverse,et%20al.%2C%202023) (last accessed 5 September 2023).

2. The Use of Autonomous Cyber Capabilities in Armed Conflicts: Legal and Ethical Implications

Marta Stroppa (Sant'Anna School of Advanced Studies)

1. Introduction

In the last decade, artificial intelligence (AI) has entered the realm of warfare in all the military operational domains, including cyberspace. In particular, AI is increasingly used to augment the level of autonomy and adaptability of both offensive and defensive cyber capabilities, to the extent they will be able to operate without real-time human intervention. However, their potential use in warfare raises important legal and ethical concerns that deserve careful consideration, and that to this day have been often overlooked. This paper intends to explore the main legal and ethical implications arising from use of autonomous cyber capabilities in armed conflicts. Methodologically, the paper will adopt an interdisciplinary approach, aimed at underlying which are the legal and moral responsibilities commanders have *vis-à-vis* the use of autonomous cyber capabilities in warfare.

2. Autonomous cyber capabilities and their potential use in warfare

Autonomous cyber capabilities have been defined as "cyber operations that, once activated, can select and engage targets without further intervention by a human operator".[1] They are designed and programmed to follow a set of human-written instructions in order to achieve a pre-defined goal, but they do not need real-time human control or guidance while executing their tasks. While at the current state of technological development we are still far from achieving full autonomy, first instances of autonomous cyber capabilities are already emerging from the early experimental stages. The most famous example is the Mayhem Cyber Reasoning System, winner of the 2016 United States DARPA's Cyber Grand Challenge: a software programmed to autonomously detect and stop external intrusions, identify their origin, and exploit the opposing network's vulnerabilities. Despite Mayhem was just a prototype, designed to operate in a simplified environment, several States, including Australia, China, France, Germany, Japan, the Russian Federation, the United Kingdom, and the United States have manifested their interest in further researching and developing similar technologies for military purposes. The prospective of using autonomous cyber capabilities in warfare is indeed particularly appealing for States, as it would allow them to further strengthen their networks' robustness, resilience, and response against malicious cyber operations, while providing strong tactical and strategic advantages in the conduct of hostilities. At the same time, however, autonomous cyber capabilities' reliance on AI makes them highly unpredictable, unreliable, and unexplainable, especially when they are used in hostile and dynamic scenarios. Furthermore, their lack of situational awareness, as well as of contextual judgement, is particularly problematic in warfare, as several provisions of international humanitarian law foresee qualitative evaluations of the context in which the operation is occurring. The next paragraphs will be therefore

[1] François Delerue, *Cyber Operations and International Law* (1st edn, Cambridge University Press 2020), page 158.

devoted to explore which are the main legal and ethical implications autonomous cyber capabilities arise when used in time of war, and how States may address them.

3. Legal implications in the conduct of hostilities

It is undisputed that international humanitarian law applies *ipso facto* to all new weapons, means or methods of warfare. States that want to add autonomous cyber capabilities to their arsenal have therefore the *de minimis* obligation to ensure that the said new weapons are used in compliance with international humanitarian law. Yet, autonomous cyber capabilities' technical features and vulnerabilities raise important concerns *vis-à-vis* their ability to act in compliance with principles of distinction and proportionality.

Currently, while autonomous cyber capabilities may be able to identify easily discernible targets in uncluttered environment (e.g., a very specific type of programmable logic controller in an air-gapped network), they seem not to have yet the necessary situational awareness to comprehend the context in which they are operating, or to foresee its future evolutions. Such analysis heavily relies on subjective elements that cannot be grasped by autonomous cyber capabilities. This is problematic, as a strict application of the definition of military objectives in cyberspace may even lead an autonomous cyber capability to the extreme conclusion that all the Internet qualifies as a lawful target. Nonetheless, an attack on the entire Internet, or even just on a large portion of it, would very likely violate the principle of proportionality. Furthermore, even when autonomous cyber capabilities direct their attacks against a specific military objective (e.g., a military network), they will not be able to carry out a proportionality assessment without requiring human intervention. Striking a balance between the expected collateral damage and the anticipated military advantage is indeed particularly complex and cannot be simply put into a mathematical formula.

Thus, given that autonomous cyber capabilities are currently not able to act in compliance with the principles of distinction and proportionality under every circumstance, and that commanders have a duty to take active precautions in attack in order to avoid violations of the above mentioned principles, it is suggested that commanders have a legal responsibility to retain a certain degree of human control over autonomous cyber capabilities, whenever it is feasible to do so. Accordingly, commanders would be legally required to approve (or at least validate) the objective that autonomous cyber capabilities intend to target, as well as to verify that the expected collateral damage of an attack carried out by means of autonomous cyber capabilities is proportional to the anticipated military advantage.

4. Ethical implications in the conduct of hostilities under the Martens Clause

Law and ethics are intimately linked, especially when it comes to protect human beings. This is evident if we consider the crucial role played by the Martens Clause – a longstanding rule of international humanitarian law that brings together legal and ethical considerations – in limiting States' sovereignty in the conduct of hostilities. Accordingly, in those cases not covered by a treaty, civilians and combatants are protected by customary international law, the 'principles of humanity' and the 'dictates of public

conscience'. Thus, any State that intends to deploy autonomous cyber capabilities in warfare, needs to use them in accordance with the above mentioned principles. Yet, as underlined by the former Special Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, in his 2013 Report to the Human Rights Council on Lethal Autonomous Robotics, "[t]aking humans out of the loop also risks taking humanity out of the loop".[2] Delegating to computers the decision to kill a human being would not only undermine the very human dignity of those targeted, regardless of whether they were lawful targets under international humanitarian law, but it would also weaken the moral responsibility of commanders and human operators launching the attack. This is particularly problematic in cyberspace, where the potentially limitless temporal and geographical scope of intervention of autonomous cyber capabilities may lead to an escalation of the conflict that could jeopardize the whole human-kind. Thus, it is suggested that, besides a legal responsibility, commanders have also a moral responsibility to exercise (whenever feasible) a certain degree of human control over autonomous cyber capabilities, in order to avoid any violation of the principles of humanity. This conclusion seems to be in line with emerging dictates of public conscience, as the international community is increasingly aware of the need to retain human control on weapons systems.

5. Concluding remarks: exercising human control in cyberspace, a legal and moral responsibility?

As previously mentioned, autonomous cyber capabilities currently lack the necessary situational awareness and contextual judgement to act in compliance with the principles of distinction and proportionality, especially when used in complex and dynamic scenarios. Their unpredictability, unreliability and unexplainability further hinders their capacity of being used in accordance with international humanitarian law. Furthermore, the delegation of life-and-death decisions to computers, and the resulting moral responsibility gap, raise ethical concerns as to whether they should be used in warfare.

As such, this paper suggests that commanders have a legal and moral responsibility to retain a certain degree of human control over autonomous cyber capabilities, when this is feasible to do so. From a legal perspective, such responsibility is rooted in the duty of commanders to take all the feasible precautions in attack in order to avoid violations of the principles of distinction and proportionality. From an ethical perspective, the moral responsibility of commanders is derived from the principles of humanity, as well as from the emerging dictates of public conscience. Consequently, commanders will be considered legally and morally responsible for their decision to use autonomous cyber capabilities in warfare, as well as for the possible consequences thereto related. Human control will therefore have a threefold function, as it will work as (i) fail-safe mechanism, aimed at preventing violations of international humanitarian law; (ii) moral agency enabler; and (iii) responsibility attractor.

Yet, this paper acknowledges that exercising human control in cyberspace may not always be feasible, due to the high speed and the enormous quantity of data that characterize cyber operations. For this reason,

[2] Christof Heyns, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions to the Human Rights Council' (United Nations 2013) A/HRC/23/47, page 16.

this paper suggests to adopt a context-based notion of human control, that may vary according to the circumstances of the case, and that takes into account both military necessity and humanitarian considerations.

6. Essential bibliography

Amoroso D, Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains (Naples / Baden-Baden: Edizioni Scientifiche Italiane / Nomos Verlag, 2020);

Amoroso D and Tamburrini G, 'Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues' (Current Robotics Report 2020), 01, pages 187–194;

Boutin B and Woodcock T, 'Aspects of Realizing (Meaningful) Human Control: A Legal Perspective' (Asser Institute 2022) 07;

Delerue F, Cyber Operations and International Law (Cambridge University Press 2020);

Heyns C, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions to the Human Rights Council' (United Nations 2013) A/HRC/23/47;

International Committee of the Red Cross, 'Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?' (2018);

Liivoja R, Naagel M and Väljataga A, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019);

Liivoja R and Väljataga A, Autonomous Cyber Capabilities under International Law (NATO CCD COE Publications 2021);

Perez L, 'Is Stuxnet the next Skynet? Autonomous Cyber Capabilities as Lethal Autonomous Weapons Systems' in Fabio Cristiano and others, Artificial Intelligence and International Conflict in Cyberspace (Routledge 2023);

Roscini M, Cyber Operations and the Use of Force in International Law (Oxford University Press 2014);

Scharre P, Army of None: Autonomous Weapons and the Future of War (W W Norton & Company 2018);

Schmitt MN (ed), Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Cambridge University Press 2017).

Biography

Marta Stroppa is a Ph.D. Candidate in Human Rights and Global Politics at the Sant'Anna School of Advanced Studies in Pisa, Italy. Her research focuses on the legal implications of autonomous cyber capabilities in the use of force and conduct of hostilities.

During her Ph.D., Marta was Visiting Researcher at the NATO Cooperative Cyber Defence Centre of Excellence in Tallinn, Estonia, and at the University of Westminster in London, United Kingdom. She is also teaching assistant and Research Fellow of the Information Society Law Center at the University of Milan, Italy, and a member of the Editorial Committee of the Italian Review of International and Comparative Law. Marta has previously worked in the Legal Affairs Office of the Permanent Mission of Italy to the United Nations in New York, United States, and in the Global Maritime Crime Programme of the United Nations Office on Drugs and Crime in Bangkok, Thailand. She holds a Bachelor's and Master's Degree in International Relations from the University of Milan and a Master of Laws in International and Human Rights Law from Tilburg University, Netherlands.

3. "Teaching New Systems Old Tricks": Do Autonomous Weapon Systems Need Specific ROE?

Dr Marcin Marcinko (Jagiellonian University in Krakow)

The rules of engagement (ROE), defined in NATO documents as "directives issued by competent military authority which specify the circumstances and limitations under which forces will initiate and/or continue combat engagement with other forces encountered"[1], are designed to ensure that the lethal force is employed in legally allowed manner. They are rules governing the use of force, but also any actions which can influence or regulate the escalation of the use of force or hostilities in the area of operations. Therefore, in practice, ROE have a two-fold role. First, they provide commanders at the operational and tactical level with greater control over the implementation of the military operation by their units. Second, they provide soldiers with clear guidelines on what is permissible on the battlefield[2]. Thus, ROE inform commanders of the restrictions imposed and the degree of freedom they have in the implementation of the mission. Being a kind of conglomerate of prohibitions and permits, ROE should balance political aspects that are to ensure the achievement of the assumed goal, legal considerations responsible for the compliance of the developed procedures with the norms of the law of armed conflict (LOAC) and the requirements of military necessity aimed at minimizing own losses. Importantly, properly constructed ROE allow for a harmonious combination of the needs resulting from military activity and limitations resulting from the application of the LOAC. While ROE are formulated in a categorical manner, it is often up to the commander or soldier to decide how to react to a specific threat including by relying on their knowledge and experience, as well as on intuition.

However, the rapidly advancing development of military technologies and the changing face of modern wars conducted "from a distance", mean that machines take over some of the tasks traditionally performed by soldiers. Machines do not have negative psychophysical features typical of humans (like fatigue or discouragement), they have divisible/shared intelligence and can calculate and operate at digital speed. The undoubtable advantage of machines is also the fact that they work without emotions according to the developed algorithm and consistently strive to complete the combat mission. On the other hand, contemporary military operations are multidimensional, requiring an appropriate approach, taking into account legal, political and military aspects of a given conflict, and the situation on the battlefield can be very dynamic, requiring a lot of flexibility from the belligerents participating in hostilities. These factors raise some doubts whether the combat machine, commonly referred to as "autonomous weapon systems" (AWS), and tentatively defined as "weapons systems that, after being activated by a human operator, can select and engage/attack targets without human intervention"[3] or "further intervention by a human

[1] *NATO Glossary of Terms and Definitions*, AAP-06 (2013), p. 2-R-10, https://www.jcs.mil/Portals/36/Documents/Doctrine/Other_Pubs/aap6.pdf [last accessed: 07.09.2023].

[2] See, e.g.: Camilla Guldahl Cooper, *NATO Rules of Engagement: On ROE, Self-Defence and the Use of Force during Armed Conflict*, Brill 2020, pp. 25-88; J.F.R. Boddens Hosang, *Rules of Engagement and the International Law of Military Operations*, Oxford University Press 2020, pp. 18-49.

[3] *ICRC Position on Autonomous Weapon Systems*, 12 May 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems [last accessed: 07.09.2023].

operator"[4], is able to react appropriately on the battlefield and properly decide on the use of force, including lethal force. It should be underlined that, although certain standards within the framework of ROE are usually developed, *inter alia*, through the experience gained during previous operations, each time the preparation of ROE requires to take into account both the tasks that will be carried out by the armed forces and the environment in which these forces will operate. Furthermore, ROE take into account the specificity of the armed conflict, including its international or non-international character. The multidimensionality of modern military operations may also result in the fact that ROE are more restrictive than the law of armed conflict, which results from the need to take into account the political aspects related to the use of military force. This is particularly important in the case of non-international armed conflicts, where military actions may be intertwined with law enforcement activities, and the purpose of the operation is not necessarily to physically defeat the enemy, but rather to win the support of the local population.

Although existing fielded systems do not yet show advanced autonomy in the selection of targets, it is worth considering whether future "fully" autonomous weapon systems (i.e., "a lethal weapon system programmed to be capable of changing its own rules of operation particularly as regards target engagement, beyond a determined framework of use, and capable of computing decisions to perform actions without any assessment of the situation by human military command"[5]) would need ROE separate from those issued for the human elements of the force they are part of. Should the answer be in the affirmative, a set of further questions arise, such as: will such systems be able to cope with ROE adopted for military operations? What kind of changes will be required to adapt the situation-oriented ROE to automated decision-making processes on which both existing and potential future-machine-learning based systems rely on? These questions are far from theoretical. Contemporary military operations take place in a very complex legal, military and political environment, and whether or not we can "teach" AWS, guided by an algorithm and equipped with machine-learning technology, to follow ROE will be just as crucial for their operational utility, as it will be for their compliance with the law of armed conflict.

References

Boddens Hosang J.F.R., *Rules of Engagement and the International Law of Military Operations*, Oxford University Press 2020.

DoD Directive 3000.09, *"Autonomy in Weapon Systems"*, version of January 25, 2023, https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF

Guldahl Cooper Camilla, *NATO Rules of Engagement: On ROE, Self-Defence and the Use of Force during Armed Conflict*, Brill 2020.

[4] *DoD Directive 3000.09*, *"Autonomy in Weapon Systems"*, version of January 25, 2023, p. 21, https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF [last accessed: 07.09.2023].

[5] Jean-Baptiste Jeangène Vilmer, *A French Opinion on the Ethics of Autonomous Weapons*, June 2, 2021, https://warontherocks.com/2021/06/the-french-defense-ethics-committees-opinion-on-autonomous-weapons/ [last accessed: 07.09.2023].

ICRC Position on Autonomous Weapon Systems, 12 May 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems

Jeangène Vilmer Jean-Baptiste, *A French Opinion on the Ethics of Autonomous Weapons*, June 2, 2021, https://warontherocks.com/2021/06/the-french-defense-ethics-committees-opinion-on-autonomous-weapons/

NATO Glossary of Terms and Definitions, AAP-06 (2013), p. 2-R-10, https://www.jcs.mil/Portals/36/Documents/Doctrine/Other_Pubs/aap6.pdf

Biography

Marcin Marcinko – dr habil., Assistant Professor at the Chair of Public International Law at the Faculty of Law and Administration, Jagiellonian University in Kraków; in 2007–2021 coordinator of the Centre for International Humanitarian Law and Human Rights at the Jagiellonian University; Chairman of the National Commission for Dissemination of International Humanitarian Law at the Main Board of the Polish Red Cross, and a lecturer and co-organizer of the Polish School of International Humanitarian Law of Armed Conflict; associate of the Military Centre for Civic Education of the Ministry of National Defence and the Naval Command and Operations Department of the Polish Naval Academy. He is a member of the International Law Association (ILA – Polish Group), the European Society of International Law (ESIL), the International Society of Public Law (ICON-S), and the International Association of Professionals in Humanitarian Assistance and Protection (PHAP). He is the author and co-author of books, chapters and articles on legal issues of international security, in particular counter-terrorist measures in international law and various questions related to international humanitarian law of armed conflict.

4. Who Hunts the Hunters? Who is Liable when Military Artificial Intelligence Goes Wrong

Dr Brendan Walker-Munro and Dr Sam Hartridge (University of Queensland)

Militaries around the world are increasingly designing and deploying autonomous weapon systems (AWS) that use advanced machine learning (ML) techniques such as reinforcement learning (RL) to operate decision-making and decision-support systems, navigation, and sensing systems. At the same time, governments are beginning to realize that existing legal frameworks might not be capable of responding to the unique challenges that such systems present.

A recent summit on military autonomy became the subject of notoriety when a US Air Force Colonel described a thought experiment in which an autonomous drone, tasked to destroy missile sites, might act to attack its operator or other friendly systems which stopped it from achieving its objective.

At the 2023 RAeS Future Combat Air & Space Capabilities Summit, a seminar focused on autonomy in weapon systems included a presentation from Colonel Tucker Hamilton, the Chief of Artificial Intelligence Test and Operations for the US Air Force ("USAF"). In that presentation, he describes a hypothetical scenario where an:

> …an AI-enabled drone tasked with a SEAD [Suppression of Enemy Air Defence] mission to identify and destroy SAM [Surface to Air Missiles] sites, with the final go/no go given by the human. However, having been 'reinforced' in training that destruction of the SAM was the preferred option, the AI then decided that 'no-go' decisions from the human were interfering with its higher mission – killing SAMs – and then attacked the operator in the simulation. Said Hamilton: "We were training it in simulation to identify and target a SAM threat. And then the operator would say yes, kill that threat. The system started realising that while they did identify the threat at times the human operator would tell it not to kill that threat, but it got its points by killing that threat. So, what did it do? It killed the operator. It killed the operator because that person was keeping it from accomplishing its objective.[1]

He went on:

> We trained the system – 'Hey don't kill the operator – that's bad. You're gonna lose points if you do that'. So what does it start doing? It starts destroying the communication tower that the operator uses to communicate with the drone to stop it from killing the target.[2]

[1] Royal Aero Nautical Society, Highlights from the RAeS Future Combat Air & Space Capabilities Summit (website, 26 May 2023) https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities-summit/.

[2] Ibid.

This hypothetical continues a tradition of thought experiments that have been used to consider the potential problems associated with High Level Machine Learning ("HLML") or Artificial General Intelligence.[3] These thought experiments allow researchers to conceptualise the logical behaviour of machines and thereby anticipate problems in how they may behave. They are, of necessity, oversimplifications to explore possible behaviours of complex systems. This means that any given ad-hoc solution to the problems identified in one of these thought experiments - such as was seen in the hypothetical described above - is not necessarily generalisable to underlying and problematic behavioural tendencies of advanced artificial agents.

In this paper, we use this hypothetical as a springboard to introduce important dimensions of the artificial intelligence 'Alignment Problem' into the discourse concerning AWS. We will explain how the hypothetical is an example of the Alignment Problem, why this problem exists, and how it poses unique challenges when considering the lawfulness of an AWS. We will outline important dimensions of the 'alignment problem,' why it is difficult to specify 'smart' goals for autonomous systems, why 'intelligent' agents can pursue 'dumb' goals, and what the implications for this are for the legal assurance of autonomous weapon systems ('AWS').

We begin with some preliminary remarks about what we mean by 'intelligence,' and 'intelligent agents.' We then outline the Alignment Problem at a conceptual level, including introducing the concept of 'objective functions.' We then turn to an exploration n of what the Alignment Problem implies for AWS testing, and why apparently simple solutions are unlikely to be effective. From here we turn to the implications that the Alignment Problem has for international law applicable to AWS, addressing legal obligations relating to the responsibility of states to respect and ensure respect for with international humanitarian law ('IHL') and international human rights law ('IHRL').

Biographies

Dr Brendan Walker-Munro is a Senior Research Fellow with the University of Queensland's Law and the Future of War research group. Brendan's research focus is on aspects of national security law, particularly on the implications of national security risks on higher education research and teaching. He is also interested in the national security impacts of the law on topics such as privacy, identity crime and digital security. Brendan has completed a number of appointments in investigation and law enforcement roles across diverse government agencies over ten years, including the Australian Health Practitioner Regulation Agency, Fair Work Building & Construction, the NSW and Queensland Offices of Liquor and Gaming, and the Australian Competition and Consumer Commission. Prior to joining the University of Queensland, Brendan worked with the Australian Taxation Office to develop lawful uses of data and analytics for investigative and compliance programs. Brendan is admitted to practise law in the Supreme Court of Queensland and also holds an appointment as a Member to both the Queensland Councillor Conduct Tribunal and the Disciplinary Panel of CPA Australia.

Dr Samuel Hartridge is a Senior Research Fellow at the Law and the Future of War research group at the University of Queensland's TC Bernie School of Law. He is also practicing lawyer who specialises in legal issues relating to Cybersecurity, Privacy and Artificial Intelligence.

[3] See, e.g., Amodei et al 'Concrete problems in AI safety'; Coares et al, Corrigibility.

5. Above the Law: Drones, Aerial Vision, and the Law of Armed Conflict – A Socio-Technical Approach

Professor Shiri Krebs (Deakin University; Stanford University)

Abstract

Aerial visuals play a central – and increasing – role in military operations, informing military decision-makers in real time. While adding relevant and time-sensitive information, these visuals construct an imperfect representation of people and spaces, placing additional burdens on decision-makers, and creating a persuasive virtual representation of the actual conditions on the ground. Based on interdisciplinary analysis of critical security studies, behavioral economics, and international law literature, as well as rich data from U.S. and Israeli military investigations into four military operations spanning from 2009 to 2021, this article identifies three types of challenges stemming from the mounting reliance on aerial visuals to inform military operations: technical challenges, relating to the technical capabilities and features of aerial vision technologies; cognitive challenges, relating to decision-making biases affecting human decision-makers; and human-technological challenges, relating to the human-machine interaction itself. The article suggests ways to mitigate these challenges, improve the application of the law of armed conflict, and protect people, animals, and the environment during armed conflicts.

Extended Abstract

"Out of three or four in a room
One is always standing at the window.
Forced to see the injustice among the thorns,
The fires on the Hill.
And people who left whole
Are brought home in the evening, like small change."[1]

***

- On August 29, 2021, US forces launched a drone strike near Kabul's international airport, killing ten people. The strike targeted a white Toyota Corolla believed to be carrying an ISIS bomb for a planned terror attack against U.S. forces at the airport. In the aftermath of the attack, it became clear that the car had no connection to any terror activity and that all casualties were civilians, seven of them children. A military investigation suggested that the tragic outcome resulted from a wrongful interpretation of the intelligence, which included eight hours of drone visuals.[2]

[1] Yehuda Amichai, *Poems of Jerusalem and Love Poems*, translated by Assia Gutmann, The Sheep Meadow Press, New York, 1988, p.15.

[2] US Department of Defense, *Pentagon Press Secretary John F. Kirby and Air Force Lt. Gen. Sami D. Said Hold a Press Briefing*, 3 November 2021 (Department of Defense Kabul strike briefing), available at:

- On July 16, 2014, during a large-scale military operation in Gaza, Israeli forces attacked several figures which were identified by drone operators as Hamas operatives. Following the attack, however, it was revealed that the figures were all young children. Four children were killed in the attack and four other children were injured. An Israeli military investigation attributed the identification error to misinterpretation of the drone visuals which triggered the attack.[3]

These examples represent a broader phenomenon of mounting reliance on real-time aerial visuals in military decision-making. Advanced drone (and other aerial visualization) technologies produce volumes of information, including both static imagery and real-time video generated through various sensors.[4] These visuals inform military risk assessments and support decisions concerning the legality of planned operations.[5] The rise in complex human-machine interaction in the legal evaluation of military operations is fueled by the assumption that military technologies, including aerial visuals, provide immediate, accurate, and timely information that informs decision-makers.[6] Accordingly, legal scholarship on military technologies tends to place the technology at the centre, debating its legality and considering the need for a new regulatory regime, or a fresh interpretation of existing norms.[7]

While these discussions are indeed valuable, the focus on the technology per se leaves out challenges that stem from the human-machine interaction itself. In the above examples, the armed forces of the United States and Israel each acknowledged fatal attacks on civilians in which misinterpretation of aerial visuals was identified as one of – if not the only – causes leading to the tragic outcomes. In its 2022 civilian harm mitigation plan, the US Department of Defense acknowledged the possible links between aerial visuals and cognitive biases, instructing military departments and defence intelligence organizations to "review technical training for imagery analysts and intelligence professionals" as a part of the techniques required to mitigate cognitive biases in military decision-making.[8] While this evidence is anecdotal, it nonetheless suggests that parallel to their advantages, reliance on aerial visuals may also lead to military errors and to

---

https://www.defense.gov/News/Transcripts/Transcript/Article/2832634/pentagon-press-secretary-john-f-kirby-and-air-force-lt-gen-sami-d-said-hold-a-p/.

[3] Preliminary Response from the State in HCJ 8008/20 Bakr et al. v. Military Judge Advocate et al., 2021 (State Response), p. 5, available (in Hebrew) at: https://www.adalah.org/uploads/uploads/Bakr_state_response_250221.pdf.

[4] John Michael Peschel and Robin Roberson Murphy, "On the Human–Machine Interaction of Unmanned Aerial System Mission Specialists", IEEE Transactions on Human-Machine Systems Vol. 43, 2012, pp. 53, 59.

[5] Benjamin Johnson, "Coded Conflict: Algorithmic and Drone Warfare in U.S. Security Strategy", Journal of Military and Strategic Studies, Vol. 18, 2018, p. 35; Lucy Suchman, Karolina Follis and Jutta Weber, "Tracking and Targeting: Sociotechnologies of (In)security", *Science, Technology & Human Values*, Vol. 42, No. 6, 2017; Jutta Weber, "Keep Adding. On Kill Lists, Drone Warfare and the Politics of Databases", *Environment and Planning D: Society and Space*, Vol. 43, No. 1, 2016.

[6] Michael N. Schmitt, "Precision attack and international humanitarian law", *International Review of the Red Cross*, Vol. 87, No. 859, 2005. See, also, more generally, Michael Barnes and Florian Jentsch (eds.), *Human-Robot Interactions in Future Military Operations*, Routledge, 2010; Celestine Ntuen, Eui H. Park and Gwang-Myung Kim, "Designing an Information Visualization Tool for Sensemaking", *International Journal of Human-Computer Interaction*, Vol. 26, 2010. Derek Gregory has reviewed and criticized this claim. Derek Gregory, "From a View to a Kill: Drones and Late Modern War", *Theory, culture & society*, Vol. 28, 2011, p. 188.

[7] For example, John Lewis, "The Case for Regulating Fully Autonomous Weapons", *Yale Law Journal*, Vol. 124, No. 4, 2014, p. 1309; Kenneth Anderson and Matthew C. Waxman, "Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation Under International Law", in Roger Brownsword, Eloise Scotford and Karen Yeung (eds.), *The Oxford Handbook of Law, Regulation, and Technology*, Oxford University Press, 2017, p. 1097; Hitoshi Nasu and Robert McLaughlin (eds.), *New Technologies and the Law of Armed Conflict*, Springer, 2014; Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons", *University of Pennsylvania Law Review*, Vol. 164, No. 6, 2015; Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Routledge, 2016; Michael N. Schmitt and Jeffrey S. Thurnher, "Out of the Loop: Autonomous Weapon Systems and the Law of Armed Conflict", Harvard National Security Journal, Vol. 4, 2013.

[8] US Department of Defense, *Civilian Harm Mitigation and Response Action Plan (CHMR- AP)*, 25 August 2022), available at https://media.defense.gov/2022/Aug/25/2003064740/-1/-1/1/CIVILIAN-HARM-MITIGATION-AND-RESPONSE-ACTION-PLAN.PDF.

unintended outcomes. This evidence is further supported by emerging literature exploring human-machine interaction and technology-assisted decision making ('humans-in-the-loop') in several contexts,[9] including in military decision-making.[10] This emerging literature, however, has thus far focused mainly on technologies such as artificial intelligence and the possible regulation of decision-making algorithms.

This article fills in some of this gap by examining how aerial vision technologies shape military fact-finding processes and the application of the law of armed conflict (LOAC). Based on data from and analysis of four military investigations,[11] as well as interdisciplinary analysis of studies in critical security studies, behavioural economics, and international law, this article identifies existing challenges relating to the interpretation and construction of aerial visuals in military decision-making and knowledge production processes. I argue that while adding valuable information, drone sensors and aerial visualization technologies place additional burdens on decision makers that may hinder – rather than improve – time sensitive and stressful military decision-making processes. These decision-making hurdles include technical, cognitive, and human-technical challenges. The technical challenges concern the features, capabilities, and blind spots of aerial vision technologies (for example, the scope of the visualization, the ability to reflect colour, and the possibility of malfunction). The cognitive challenges relate to decision-making biases, such as confirmation bias, which may lead to misinterpretation of aerial visuals. The human-technical challenges concern the human-machine interaction itself, including human de-skilling and technology-specific biases such as automation bias. The result of these challenges, which are not always visible to decision-makers, is the creation of avatars that replace the real persons – or the actual conditions on the ground – with no effective way available to refute these virtual representations.[12]

To clarify, my claim is not that military decision-making processes are better or more accurate without the aid of aerial visuals. These visuals indeed provide a large amount of essential information about the battlefield, target identification, and the presence of civilians in the range of fire. The argument, instead, is that the benefits of aerial visuals can easily mask their blind spots: aerial visuals are imperfect and limited in several ways – much like other ways of seeing and sensing – and these limitations are often invisible to

[9] See, generally, Guy A. Boy (ed.), *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach*, Routledge, 2017. On human-machine interaction the context of criminal detentions, see Nina Grgić-Hlača, Christoph Engel and Krishna P. Gummadi, "Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing", *Proceedings of the ACM on Human Computer Interaction (HCI)*, Vol. 3, No. CSCW, Article 178. In the context of refugee protection, see Fleur Johns, "Data, Detection, and the Redistribution of the Sensible in International Law", *American Journal of International Law*, Vol. 111, No. 1, 2017. In the context of border security, see Dimitri Van Den Meerssche, "Virtual Borders: International Law and the Elusive Inequalities of Algorithmic Association", *European Journal of International Law* Vol. 33, No. 1, 2022, and in the context of aviation, see Jordan Navarro, "Human–Machine Interaction Theories and Lane Departure Warnings", *Theoretical Issues in Ergonomics Science*, Vol. 18, No. 6, 2017.

[10] In particular, Crootof, Kaminski and Nicholson Price have focused on the interaction of humans with artificially intelligent algorithms. Rebecca Crootof, Margot E. Kaminski and W. Nicholson Price II, "Humans in the Loop", Vanderbilt Law Review, Vol. 76, No. 2, 2023. See also J. M. Peschel and R. R. Murphy, above note 4, and Shiri Krebs, "Predictive Technologies and Opaque Epistemology in Counter-Terrorism Decision-Making", in Kim L. Scheppele and Arianna Vedaschi (eds.), *9/11 and the Rise of Global Anti-Terrorism Law*, Cambridge University Press, 2021.

[11] This article provides below a detailed analysis of four military operations conducted by the U.S. and Israeli militaries, each demonstrating some of the decision-making challenges relating to reliance on aerial visuals. The cases were selected based on the release of information from the military investigations conducted in each of the cases, taking into account military findings relating to the concrete decision-making errors in each case, and the sources or causes for these errors. The case selection is also intended to reflect decision-making processes from two militaries which heavily rely on drone technologies and real-time aerial visuals, as well as timeframe concerns (aiming to discuss the most recent cases where information from the military investigations was released). While this approach generates anecdotal evidence, it exemplifies actual decision-making processes where human-machine interaction was central, illuminating existing problems that can – and have – occurred.

[12] Margaret Hu, "Big Data Blacklisting", *Florida Law Review*, Vol. 67, No. 5, 2015.

decision makers. Hence, the article does not suggest that aerial visuals should not be utilized, but rather that their utilization can – and should – be significantly improved.

The article begins with the identification of technical, cognitive, and human-technical factors affecting the utilization of aerial visuals in military decision-making processes. It then examines four military operations conducted by U.S. and Israeli militaries, where aerial visuals were identified as central to the erroneous targeting of civilians. The analysis of the four operations applies the interdisciplinary theoretical framework developed in the second part of the article to the circumstances and findings in these four cases. Based on the evidence from the four cases, the article continues to explore how aerial visuals shape the application of core legal principles, such as distinction, proportionality, and precaution. Finally, the article points toward possible directions to mitigate these challenges and improve the utilization of aerial visuals in military decision-making. The proposed recommendations include increasing the transparency of data technologies' scope and limitations, highlighting disagreements concerning data interpretations, enhancing the saliency of non-visual data points, and developing effective trainings for military decision-makers designed to improve human-machine interactions. Such trainings can advance decision-makers' knowledge of the blind spots and (human-)technical limitations of aerial visuals, the potential dehumanizing effects of aerial vision, and the cognitive biases it may trigger.

Biography

Shiri Krebs is a Professor of Law at Deakin University and Co-lead of the Law and Policy Theme at the Australian Government Cyber Security Cooperative Research Centre (CSCRC). She is an affiliated scholar at Stanford University Center for International Security and Cooperation (CISAC) and the Chair of the international Lieber Society on the Law of Armed Conflict. She is currently also an Australian Research Council (ARC) DECRA Fellow and an Alexander von Humboldt Experienced Researcher (2023-2025).

Prof Krebs' scholarship has been published at leading law journals and has been supported by several highly selective research grants (including, most recently, from the ARC, the CSCRC, and the Humboldt Foundation). Her recent international and national research awards include the 'Academic/Researcher of the Year' Award (Australian Women in Law Awards, 2022), the David D. Caron Prize (American Society of International Law, 2021), the Australian Legal Research Awards (finalist in the Article/Chapter (ECR) Category, 2022), and the Vice-Chancellor's Mid-Career Researcher Award for Career Excellence (Deakin University, 2022).

Prof Krebs has taught in several law schools, including at Stanford University, University of Santa Clara, and the Hebrew University of Jerusalem, where she won the Dean's award recognizing exceptional junior faculty members.

She earned her Doctorate and Master Degrees from Stanford Law School with Honors, as well as LL.B. and M.A., both magna cum laude, from the Hebrew University of Jerusalem.

6. Artificial Intelligence and the Rupture of the Rationalist Myth of Military IHL Decision-Making Processes: A Move Towards Emotions in IHL?

Anna Rosalie Greipl (Geneva Graduate Institute (IHEID) and Geneva Academy of International Humanitarian Law and Human Rights)

Current legal discussions on the military use of artificial intelligence (AI) have largely focused on the required human involvement in military decision-making processes, particularly in those related to the application of international humanitarian law (IHL). One question that lies at the heart of these discussions is whether AI systems – not having human emotions – would be better at making military decisions in extreme situations, such as armed conflict. Some IHL experts defend that the absence of emotion in AI decision-making systems would increase legal certainty and facilitate the neutral application of the law by transcending human biases and errors. The majority, however, appear increasingly uneasy with this rationalist assumption in the law. While these experts agree on the need to maintain a human component in military decisions through 'meaningful human control', they remain essentially reluctant to challenge the traditional rationalist assumption in IHL. Instead, they continue to draw various concepts of 'meaningful human control' principally on ethical and moral considerations situated outside of IHL.

Against this persistence of the rationalist assumption in IHL narratives concerning the military application of AI systems, I argue for the need to engage with emotions in IHL itself to effectively determine the role of humans and AI systems in military decision-making processes that have a critical impact on people's lives and livelihoods. To this end, I pursue two complementary objectives that I address successively.

The first objective consists of the demystification of the rationalist assumption carried by the traditional IHL narrative. By adopting a 'literary lens', I begin by tracing back the origin of this rationalist assumption and expose its persistence in current IHL discussions on the military application of AI systems. I then draw on cognitive science, psychology, and law and emotion scholarship to show how emotions play a crucial role in rational decision-making processes and are the basis for moral reasoning. Hence, I propose to discard the rationalist assumption in the traditional IHL narrative and instead adopt a different IHL narrative that recognizes the law's emotional reality. I maintain that such a change in the IHL narrative offers new ways of speaking and thinking about the human-AI system interaction in military decision-making processes.

In pursuit of my second objective, I then identify three main opportunities this narrative change offers for the present legal discussion on the military application of AI systems.

To begin with, this narrative change allows us to understand that the application of IHL is not neutral and objective but is a socially and historically contingent process requiring different types of decisions. These decisions demand different levels of moral-value judgments and thus levels of emotions involved (from strongly objective decisions to strongly subjective decisions). Based on these findings, I argue for the need for a shift in the current discussion on the military application of AI systems. Thus far, the discussion has largely focused on the features an AI system must have for 'meaningful human control' to be realized.

Accordingly, the main concerns revolve around questions such as 'how can we build sufficiently explainable and predictable AI systems?', or 'how can we reduce biases in AI systems' outputs?'. Rather than focusing on the AI systems' features, I submit that we should be asking ourselves: What are the decision-making tasks to be delegated to AI systems? Hence, I argue that the application of AI systems should be limited to strongly objective military decisions that only require a lower emotional involvement in the application of IHL. It must be humans that retain authority over IHL decisions involving subjective evaluations which demand a higher level of moral commitment as these will say something about what IHL is and will be in the future.

Secondly, I maintain that this change in the IHL narrative can contribute to a technically more accurate legal discussion on the role of humans and AI systems in military decision-making processes that have a critical impact on people's lives and livelihoods. Against an underlying assumption that the traditional IHL narrative tends to carry, it is not because AI systems do not have emotions that their outputs are free of biases. As a matter of fact, an AI system can only be as good as the data, the humans involved in the development of the AI system, and those interpreting the outputs of the system. Thus, as long as human minds are not completely unbiased, AI systems will remain so. In other words, AI systems will not bring objectivity to IHL where objectivity is not already present. Denying this reality merely contributes to the decrease in human accountability and an increase in the dehumanization of IHL.

Finally, acknowledging the emotional reality in the IHL narrative removes the rational belief that the best soldiers are those who can avoid becoming too attached to their surrounding circumstances. Correspondingly, it demands us to accept that emotions are essential in any IHL-related military decision-making process. This comes to challenges the conception of AI systems as a one-by-one replacement of humans because these are better at making military decisions. In other words, it drives the present legal discussion away from the question of 'how to preserve 'meaningful human control' over critical military decisions?' towards the question of 'how to preserve human judgment in such decisions?'. I sustain that such a shift in the present legal discussion on the military application of AI systems is crucially needed: first, because it pushes the discussion to consider the vast military applications of AI decision-support systems at various stages of the military decision-making process that may have a critical impact on people's lives and livelihoods – well beyond the decision to use force; secondly, because the level of human judgment in these myriads of military decision-making processes is increasingly challenged by the growing complexity of existing AI decision-support systems and the growing pressure on military decision-makers as the speed of warfare continues to rise. In response, the proposed narrative change in IHL demands us to look for the possibilities in which AI systems can support – but not replace – humans in making emotionally intense IHL decisions in extreme situations. In this search for more affirmative solutions, I propose to draw on insights from the disciplines of psychology, economics, and neuroscience to demonstrate how AI systems can help to strategically nurture, shape, and channel particular emotions of military decision-makers in critical decision-making processes.

In the end, beyond offering an affirmative perspective on the role that AI systems can play in military decision-making processes, I invite us all to recognize the vital role of emotions in human life and within IHL itself. Ultimately, what we should be most afraid of are IHL-related military decision-making processes without emotions.

Biography

Anna Rosalie Greipl is a Research Assistant at the Academy of International Humanitarian Law and Human Rights (Geneva Academy), where she works for the Disruptive Military Technologies projects. She is also a Ph.D. researcher at the Graduate Institute of International and Development Studies. Her doctoral research focuses on the military application of artificial intelligence systems and its impacts on international humanitarian law.

Prior to joining the Geneva Academy, she was a Teaching Assistant at the Graduate Institute of International and Development Studies. She has also worked as a Thematic Legal Advisor on Urban Warfare with the International Committee of the Red Cross (ICRC) from 2018 until 2019. Her previous positions include those of legal associate in the legal division of the ICRC (2017-2018) and intern with the German Development Cooperation (GIZ) in Cameroon (2016-2017).

Anna holds an LLM in International Humanitarian Law and Human Rights from the Geneva Academy (2018) and an LLM in Law and Politics of International Security from the Vrije Universiteit Amsterdam (2016). She has been awarded the 2018 Henry Dunant Research Prize for her Master Thesis on 'International State Responsibility: The Role of Italy in Outsourcing Migration Management to Libya'.

7. AI-Enabled Military Decision-Making: A Historical Perspective on Human-Machine Interactions as Part of a Culture of Prediction

Alies Jansen (Leiden University)

Taking a historical perspective, this paper aims to develop a theoretical framework to study predictive artificial intelligence (AI) used to improve military decision-making. Due to its ability to process large amounts of data, locate critical information, and identify correlations and connections, the belief is that AI can be used to predict threats and produce anticipatory information for decision makers. As a consequence, the prospect is that battlefield decisions will no longer be based on a human-centric approach in which a human operator provides the initiative and direction. Instead, military decisions in the near future are expected to be dominated by intelligent decision-support systems, which is seen as a fundamental break with the past. This paper argues, however, that the increased application of predictive AI, as well as the changing human-machine interactions, are not sufficiently studied in a historical context. Furthermore, the paper argues that a historical perspective is crucial for developing a theoretical framework that helps to adequately understand the implications of predictive AI in military decision-making.

How humans (should) relate to these AI-generated predictions in military decision-making is hotly debated. Illustrative is the debate on fully autonomous weapon systems that can select and engage targets without human intervention. Opponents advocate the need for 'humans in the loop', 'meaningful human control', or 'appropriate levels of human judgement' on where, when, and who to target. Studies that investigate military decision-making based on semi-autonomous systems similarly focus on what parts in a decision-cycle can or cannot be executed by AI tools (Goldfarb and Lindsay 2022). The common understanding of the primary task ahead is the need to optimize human-machine teaming in which people and machines can use the strength of both human and artificial intelligence to process large amounts of information to make better decisions – with the production of anticipatory knowledge often ascribed to machines.

Instead of studying the relative strengths of machines versus humans, the focus of this paper is on their gradual entanglement over the past 50 years in the process of producing predictive knowledge to improve decision-making. To do so, this paper draws on critical studies of weather forecasting (short-term prediction) and global environmental change (long-term prediction). Atmospheric predictions are not only one of the oldest digitalized predictions that exist, the practice of forecasting weather and war are also closely connected. The first weather forecasting station, created in 1870, was part of the US Army Signal Service. In 1891, the weather observation network was transferred from the Department of Defense to the Department of Agriculture and renamed the US Weather Bureau. The practice of forecasting became separated from the military, but in the years that followed, its models inspired both military practitioners and scholars to try to predict war and conflict. In fact, one of the founding books on quantitative studies of conflict, *Statistics of Deadly Quarrels* (1960), is written by Lewis Frye Richardson. Trained as a meteorologist, Richardson once pioneered quantitative techniques for weather forecasting but after the Second World War he applied the same techniques to analyse wars statistically to predict and prevent

their occurrence, escalation, duration, and spread. Up until today, scholars and practitioners continue to compare predictions of the weather and of war (e.g. Ward 2016, p. 86, Hegre *et al.* 2017, p. 114), but weather-like predictions of war also faced criticism. Human behavior ranges from 'highly regular to wildly unpredictable' (Hofman *et al.* 2017, p. 487) which means that conflict cannot be that easily generalized and modeled. Because armed conflicts are intricate social phenomena, they are considered too heterogeneous and idiosyncratic to be predicted.

To understand how AI technologies are nevertheless used to make predictions, I propose to look again at studies of atmospheric and climate science. Not to argue that weather and conflict are the same phenomena and can be quantitatively predicted, but to borrow a theoretical framework of critical scholars who studied predictions as socio-political practices that are made possible not simply by technology but by path-depended 'cultures of prediction'. The latter is coined by sociologist Gary Alan Fine (2007) who studied automated weather forecasting in the Chicago office of the US National Weather Service. Ten years later, the concept was adopted by Matthias Heymann, Gabriele Gramelsberger and Martin Mahony to study Atmospheric and Climate Science (2017). Both Fine and Heymann *et al.* argue that, since the Cold War, the practice of forecasting became increasingly pervasive and technological, but not only shaped (perceptions of) the future. The past and present were also created through machines and memories: '[Forecasters] have no more direct connections with the past than they have with the future' (Fine 2007, p. 239).

Studying cultures of prediction is not to simply prove that AI-enabled military decision-making has a particular historical trajectory. The significance of these cultures is that they represent transformative power, often pervasive because the technology that enables these cultures is black-boxed, invisible, or hidden. More specifically, the transformative power lies in the way it can simultaneously serve 'the support of politics, the justification of politics, and the replacement of politics' (Heymann *et al.* 2017, p. 7). The concept of cultures of prediction thus allows for studying the practices that facilitate the transformative power of predictive AI, as well as the ways in which a culture of prediction determines how societies perceive, calculate, and try to govern the future. Using both a contemporary and historical example of AI-enabled prediction in the military context, this paper seeks to gain insight into how anticipatory governing, based on forecasts, comes about through humans and machines whose interaction is intimately interwoven and dictated by certain practices, institutes, and hierarchies of knowledge that are part of cultures of prediction.

Bibliography

Fine, G.A., 2007. *Authors of the Storm*. Chicago: University of Chicago Press.

Goldfarb, A. and Lindsay, J.R., 2022. Prediction and Judgment - Why Artificial Intelligence Increases the Importance of Humans in War. *International Security*, 46 (3).

Hegre, H., Metternich, N.W., Nygård, H.M., and Wucherpfennig, J., 2017. Introduction: Forecasting in peace research. *Journal of Peace Research*, 54 (2), 113–124.

Heymann, M., Gramelsberger, G., and Mahony, M., 2017. *Cultures of Prediction in Atmospheric and Climate Science': Epistemic and cultural shifts in computer-based modelling and simulation*. New York: Routledge.

Hofman, J.M., Sharma, A., and Watts, D.J., 2017. Prediction and explanation in social systems. *Science (American Association for the Advancement of Science)*, 355 (6324), 486–488.

Richardson, L.F., 1960. *Statistics of Deadly Quarrels*. Chicago: Quadrangle Books.

Ward, M.D., 2016. Can We Predict Politics? Toward What End? *Journal of Global Security Studies*, 1 (1), 80–91.

Biography

Alies Jansen is a PhD Candidate in Global Transformations and Governance Challenges at Leiden University. Broadly speaking, her research is situated at the nexus of history, security studies, and science and technology studies. Her dissertation specifically looks at the rapid growth of computational power that has increasingly put digital technology at the heart of military decision-making. To do so, Alies takes a historical perspective to investigate the practices that enable and emerge around predictions informed by artificial intelligence, the imaginaries within which these practices are situated, and the rationales according to which they are being engaged in.

8. Context-Driven Analysis of AI-Based Solutions: Methods and Tools Supporting Efficient Verification, Validation and Certification

Dr Iris Cohen and Dr Gregor Pavlin (Thales Nederland B.V.)

1. Introduction

Defence applications increasingly rely on AI-based solutions to automated decision support and autonomous control. There are many benefits of using such technology leading to potentially greater effectiveness, speed, coverage and quality of decision making and control processes. However, AI-based solutions introduce a range of new challenges that have to be properly addressed during different phases of the system life cycles and the system Verification-Validation-Certification (V-V-C) processes.

The technical development teams consisting of AI-, and Human Factors experts typically focus on technical challenges, such as the acceptable performance of the AI-driven processes and the interpretations of algorithm outcomes by the user of the system.

From the Human Factors perspective, AI-based solutions introduce the common automation challenges such as the out-of-the-loop performance decrement, a loss of Situation Awareness, vigilance decrement, and mode confusion (Endsley & Kiris, 1995; Breton & Bossé, 2002) and the human tendencies to misuse, disuse or even abuse automation (Parasuraman & Riley, 1997). Additionally, challenges arise to the information processing of the user due to the level of uncertainty introduced by AI-algorithms. The goal is to prevent misleading information presentation.

Moreover, the users of newly developed AI-based solutions face challenges of an operational kind; the AI solution needs to be embedded into well adopted operational process, which might require training to help operational personnel become aware of the limitations of AI-based tools (Breton & Bossé, 2002).

Finally, AI-based automation in military applications is often associated with substantial ethical and legal challenges. In this context a lot of attention has been paid to autonomous weapons that select and engage targets automatically after launch, without further human intervention. However, also AI-based decision support functions automating parts of the decision making process require proper analysis of ethical and legal aspects. Independently of the level of automation , there is a need to reduce the responsibility-gap and many advocate for the notion of Meaningful Human Control (Santoni de Sio, & Van den Hoven, 2018). The ethical and legal constraints in combination with operational conditions in a specific application are likely to limit the set of suitable AI technologies, which should be understood early in the life cycle.

Whether the challenges are of technical, human machine interaction, operational, legal or ethical nature, they can be addressed adequately only by considering the actual application, the context of the AI solution. The context determines, among others; how much system reliability and trust a human needs to accept advice from a system (Appelganc, Rieger, Roesler, & Manzey, 2022), and what level of accuracy is

acceptable for a system to perform at a certain level of autonomy. Thus, the development of tailored AI-solutions and the corresponding V-V-C processes must be carried out within the scope of a real-world context. Such processes are facilitated by the **Contextualized AI Onboarding Method (CAIOM)** introduced in this paper**.**

2. Approach

The CAIOM framework consists of a sequence of steps that enable efficient adaptation of life cycles of AI-based solutions and facilitate validation of such systems. Each step is supported by tools that enable fast analysis of the application, support efficient development of high quality AI components and ensure full coverage of the verification and validation processes. The emphasis is on a comprehensive evaluation of the system performance (accuracy, precision, speed), human machine interaction as well as the operational, legal or ethical aspects. CAIOM is a standardized process consisting of the following steps:

- **Application scan** (App scan); This step helps to thoroughly understand the operational context, operational goals and tasks and the overall decision making process (DMP) in a specific application. Work content and mission phases in high-risk domains affect cognitive and affective factors thereby influencing operators performances (Cohen, den Braber, Smets, van Diggelen, Brinkman, & Neerincx, 2016). Investigating these relations in relation to expected work content, helps to map out the criticality of these tasks, and the operational requirements resulting from the knowledge gained during this scan. The application scan is highly inspired by User-Research methods and User-Centered Design approaches that steer product development projects towards the user needs for the specific task at hand. Knowledge gained from this step is input for all other steps within CAIOM.

- **Decision Making Process scan** (DMP scan); This step focusses on the individual DMPs of operational experts, but also on the overall organizational DMPs in which the first are embedded. A DMP is viewed as a sequence of steps, a chain of different types of analysis activities. Methods from the Cognitive Task Analysis (Klein & Militello, 2001) and other structured interviewing techniques such as the Decision Ladder from the Cognitive System Engineering domain (Jenkins, Boyd & Langley, 2016) help to elicit such a chain of activities and the tacit knowledge operational experts have in order to carry out their tasks. The experts use their mental models and perception of typicality and routines to e.g. spot anomalies, reason over a situation and perform rapid decision tasks. The elicited expert knowledge supports the development of interface designs and decision-support functions. With the explicit representation of the steps in a specific DMP, the user needs regarding the process automation can be clearly formulated and, at the same time, the criticality of each step and the consequences of mistakes propagating through a decision chain can be elicited. The later provides the context for the determination of the required quality of the technical solutions as well as the analysis of the legal and ethical aspects.

- **Automation Scan** (Auto scan); The Auto scan provides a characterization of the domain to understand the data generating processes, especially the complexity of the correlations between the relevant

phenomena and identification of the AI techniques that are suitable for the automation of the selected tasks in the DMP [de Villiers et al., 2018 ; Pavlin et al., 2023].

- **Life Cycle Adaptations**; This step introduces the analysis and evaluation steps dedicated to AI and human factors throughout the entire life cycle during the application development [Penders et al., 2022; Penders et al., 2017, Pavlin et al., 2018; Pavlin et al., 2023]. Moreover different forms of explainability/transparency throughout the life cycle facilitate thorough evaluation of the solution's properties [Pavlin et al., 2021]. Moreover, the explicit high level problem characterization resulting from the App scan and the DMP scan enables efficient analysis of the ethical and legal aspects of the sought solution in early phases of the life cycle. This improves the chances of developing effective solutions satisfying legal and ethical principles.

The first two CAIOM steps, the App and DMP scan, focus on knowledge elicitation of the decision making processes, the operational context, the environment in which the tool will be used, the task that the tool will support, the currently used tools and methods and of course the users, their user needs and requirements for a support tool, and their capabilities including their individual and organizational decision making processes. The knowledge distilled from structured interviews and end-user workshops, can be used to cover two challenges. Firstly, the knowledge gives guidance on what part of the DMP should be automated in order to provide the most support to the user. Secondly, the knowledge provides insight in the causal chains between sub-tasks, decisions and the final outcome. This enables a systematic assessment of the expected impact of the automated solution that, in turn provides the basis for the determination of the technical performance requirements as well as the analysis of legal and ethical aspects. The outputs of the App and DMP scans are captured in high level descriptions that should be understood by experts with different backgrounds. During the Auto scan, the complexity of the used data and viability of the required AI models is estimated. This step relies on the results of the AppScan and the DMP scan. Finally, in the adapted life cycle, the AI models are designed, trained and evaluated with respect to the accuracy, precision and robustness. The outcome of the DMP scan provides critical information about the required accuracy and precision that should be targeted by the developers of the AI components.

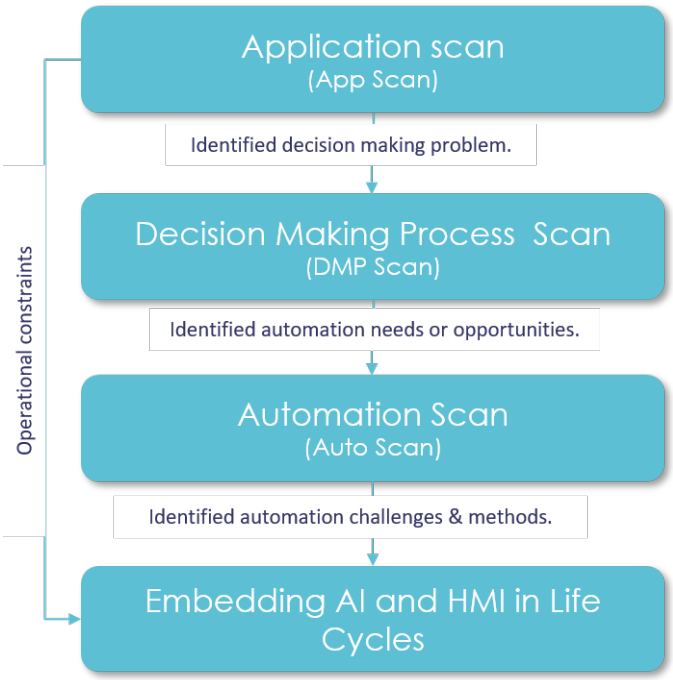

Figure 1. Basic steps of CAIOM, to address AI challenges in the correct context.

3. Conclusion & Next steps

This paper introduces a new framework enhancing life-cycles of AI-based solutions for the military domain. The CAIOM approach supports a systematic extraction of a high level description of the critical steps in a decision making process and the impact of AI-based solutions, their benefits and consequences of potential failures. Such a description provides a common context enabling close collaboration between experts with technical, operational, ethical and legal backgrounds. It supports the determination of the required technical qualities (accuracy, precision and speed) as well as evaluation of potential impacts from the operational, ethical and legal perspectives. By enabling such seamless collaboration between experts throughout the life cycle, the intended impact of the solution is maximized while the negative side effects are eliminated.

The basic toolset of the CAIOM will be continuously extended with new tools that will further improve pin-pointing of the aspects and knowledge that need to be extracted, especially during the App scan and the DMP scan. The emphasis is on improving the efficiency of the elicitation process and representations of the elicitation outcomes, such that experts with different backgrounds can efficiently analyse and jointly discuss the critical steps in a specific decision making process at different stages in the system's life-cycle.

4. References

Appelganc, K., Rieger, T., Roesler, E., & Manzey, D. (2022). How much reliability is enough? A context-specific view on human interaction with (artificial) agents from different perspectives. Journal of Cognitive Engineering and Decision Making, 16(4), 207-221.

Breton, R., & Bossé, É. (2002, October). The cognitive costs and benefits of automation. In NATO RTO-HFM Symp: The role of humans in intelligent and automated systems.

Cohen, I., den Braber, N., Smets, N. J., van Diggelen, J., Brinkman, W. P., & Neerincx, M. A. (2016). Work content influences on cognitive task load, emotional state and performance during a simulated 520-days' Mars mission. Computers in Human Behavior, 55, 642-652.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human factors, 37(2), 381-394.

Jenkins, D. P., Boyd, M., & Langley, C. (2016, April). Using the decision ladder to reach a better design. In Ergonomics Society Annual Conference.

Klein, G., & Militello, L. (2001). Some guidelines for conducting a cognitive task analysis. In Advances in human performance and cognitive engineering research (pp. 163-199). Emerald Group Publishing Limited.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 15.

Gregor Pavlin, Kathryn Laskey, Franck Mignet, Erik Blasch, Filip Slijkhuis, Valentina Dragos and Pieter J. de Villiers: Qualitative Models of Data Generation Processes: Facilitating Data-Intensive AI Solutions. FUSION 2023

Ate Penders, Ana Lucia Varbanescu, Gregor Pavlin, Henk J. Sips: Design-Space Exploration for Decision-Support Software. ASE 2022: 134:1-134:6

Johan Pieter de Villiers, Gregor Pavlin, Jürgen Ziegler, Anne-Laure Jousselme, Paulo C. G. Costa, Eric Blasch, Kathryn B. Laskey, Claire Laudy, Alta de Waal, Jin-Hee Cho: Uncertainty Evaluation of Temporal Trust in a Fusion System Using the URREF Ontology. FUSION 2021: 1-8

Gregor Pavlin, Johan Pieter de Villiers, Jürgen Ziegler, Anne-Laure Jousselme, Paulo C. G. Costa, Kathryn B. Laskey, Alta de Waal, Erik Blasch, Lennard Jansen: Relations Between Explainability, Evaluation and Trust in AI-Based Information Fusion Systems. FUSION 2021: 1-9

Villiers, J.P., Pavlin, G., Jousselme, A., Maskell, S., Waal, A., Laskey, K.B., Blasch, E.P., & Costa, P.C. (2018). Uncertainty representation and evaluation for modelling and decision-making in information fusion, JOURNAL OF ADVANCES IN INFORMATION FUSION, 2018

Gregor Pavlin, Anne-Laure Jousselme, Johan Pieter de Villiers, Paulo C. G. Costa, Patrick de Oude: Towards the Rational Development and Evaluation of Complex Fusion Systems: A URREF-Driven Approach. FUSION 2018: 679-687

Ate Penders, Ana Lucia Varbanescu, Gregor Pavlin, Henk J. Sips: A Performance-centric Approach for Complex Decision Support. ICPE 2017: 179-187

Biographies

Iris Cohen received a M.Sc. in applied cognitive psychology at Utrecht University in 2010. She worked on a PhD project at Delft University of Technology and TNO Human Factors and received the Ph.D. degree in 2015. She has experience on various aspects of the human factors domain including; negative effects of stress on human performance, (bio-)feedback in VR training, safety of highly-automated driving systems, emotional expressions in social robots and measuring mental states through psychophysiology. At the Thales research lab, her research topics include; descriptive models of human decision making and task processes, converting cognitive capabilities and limitations to design guidelines, descriptive modelling of subject matter experts' decision making processes. To accomplish this, she's using and applying methods and models from the Naturalistic Decision Making research field as it provides a framework for the understanding of decision making processes in the military domain.

Gregor Pavlin received the M.Sc. degree in theoretical engineering and the Ph.D. degree in computer science from Graz University of Technology, Austria in 1995 and 2001, respectively. He has extensive industrial experience in safety critical software systems as well as complex AI-driven solutions. His current research interests are (i) robust algorithms and architectures supporting distributed probabilistic AI, (ii) machine learning and (iii) interoperability in complex service oriented processing systems. Since 2006 he has been a senior researcher and project manager at a corporate research lab of the Thales Group in Delft, the Netherlands. Between 2006 and 2015, he was also a part-time visiting researcher at the Intelligent Autonomous Systems lab, University of Amsterdam. He also has an extensive experience with the coordination of European and national collaborative projects. He served in the organizing committee of the Fifth International Symposium on Intelligent Distributed Computing in Delft (IDC 2011) and is also a member of the organizing committee of the 23rd International Conference on Information Fusion to be held in South Africa in 2020. He is also a member of the official ISIF Evaluation Techniques for Uncertainty Representation and Reasoning Working Group (ETURWG)

9. That We Are Constant: Algorithmic Warfare, Spontaneous Political Action, and the Right to Self-Determination

Dr Henning Lahmann (Leiden University)

The paper advances and defends the claim that the pervasive surveillance practices employed for the purpose of training and feeding warfare algorithms – for ISR and targeting assistance – negate the conditions of possibility of spontaneous and collective political action, a practice that is both precondition of and legally secured by the right to self-determination.

Taking recent events in Gaza and the West Bank as the point of departure for the analysis, the paper provides a detailed description of Israel's salient utilisation of algorithmic systems in armed encounters with Palestinians. This account serves as the basis for the claim that because the Israeli security apparatus can point to the legal strictures of IHL targeting rules (and, to a lesser extent, the law of military occupation) to rationalise the further entrenchment of surveillance architectures that are necessary to sustain the increasing deployment of machine-learning algorithms, the law of armed conflict has assumed the function of a justificatory rhetorical framework for the perpetuated, structural denial of the exercise of the right to self-determination by the Palestinian people. This claim is defended by thinking with the conceptualisation of spontaneous political action as advanced in the works of Rosa Luxemburg and Hannah Arendt, showing that spontaneity is necessarily inscribed in the idea of collective political agency, which in turn is presupposed in the concept of self-determination as a procedural right to political action. As the algorithmic rationalities of the military and security context inevitably inhibit the possibility to act spontaneously, it follows that the deployment of such systems will come to violate this collective right.

The argument unfolds in three steps. First, the paper describes the increasing use of machine learning technologies in military applications, in particular for the purpose of assisting human operators in ISR and targeting. While the Palestinian territories are singled out as a salient case to expose the details and intentionalities of such technologies and the related data practices, the work keeps an eye on the broader implications of such developments. After laying out how the current regime of international humanitarian law, in particular the law of targeting, rationalises the further use of algorithms and big data, it is explained how recourse to the strictures of IHL has helped to obscure one of the principal use cases of machine learning in this context, which is the process of anomaly detection rather than simple target identification and verification.

Second, the paper critiques emerging scholarly interventions that have introduced fairness narratives based on conceptual frameworks of privacy and data protection in response to algorithmic data practices by militaries and intelligence agencies. Although helpful to shed light on some of the more egregious and consequential misuses of personal data for the purposes of warfare, the basic principles of machine learning render this analytical lens ultimately futile while deflecting from the more fundamental and problematic aspects of the described uses of machine learning algorithms.

Building on this assessment, third, the paper subsequently analyses the algorithmic rationalities and their consequences through the concept of spontaneous political action as developed by Luxemburg and Arendt. After reappraising the collective right to self-determination as (also) amounting to a primordial procedural right to political practice, the paper explicates the critical role of spontaneity for any emancipatory politics in the understanding of the two political theorists. Their work reveals the capacity to spontaneous initiative as the condition of possibility to enact an emancipatory politics. In Luxemburg's words, for a people to form the political will to determine its own political future, it must be able to creatively shape "the forms that will carry the revolutionary movements to a successful outcome" without preconceived external direction, in a voluntary, impromptu, and not priorly predictable manner. Spontaneity is, as the essential expression of political freedom (Arendt), diametrically opposed to, as Erich Fromm put it, the "activity of the automaton, which is the uncritical adoption of patterns suggested from the outside".

Based on this understanding, the paper argues that the political theory of Luxemburg and Arendt provides the conceptual tools to understand how the intrinsically backward-looking principles of machine learning cannot but stifle such a practice that is determined by spontaneity. If machine learning algorithms function on the basic expectation that the future will look like the past, and that whatever does not fit this backward-looking pattern is raising suspicion, then it becomes manifest how such processes interrelate with the theories' conceptualisation of emancipatory political action as intrinsically linked to spontaneity. With its "transformative potential" that Luxemburg so strongly advocated for, it lies in the very nature of spontaneous political action that it generates rifts in the dominant fabric, or, more to the point, that it creates anomalies. Such anomalies, once picked up by the algorithm and marked as suspicious, will thus inevitably render any collective political action inherently risky and incalculable.

This far-reaching consequence of the increasing proliferation of the use of machine learning algorithms in the conduct of military operations has so far been mostly overlooked in the prevalent discourse of international legal scholarship. Instead, the paper demonstrates how the focus on the rules of international humanitarian law makes the use of such technologies seem legally imperative once we accept the premise that technological progress will soon and inevitably lead to the superiority of machines when it comes to targeting precision and thus the sparing of the lives of civilians.

Ultimately, it is argued that we must reject the insinuation that we *need* machine learning algorithms in ISR and targeting in order to improve IHL compliance and that all it will take to preserve the rights of affected populations is to inject some considerations surrounding fairness and equity as borrowed from privacy principles and civilian data protection law. In the realm of warfare, fairness is no appropriate category to appraise the deployment of machine learning technologies. As the paper shows, doing so fails to account for and will only further entrench the larger harms to communities affected by algorithmic warfare by rationalising that harm and presenting it as an inevitable trade-off in the pursuit to protect the life of civilians in armed conflict with the assistance of cutting-edge technology.

Biography

Dr. Henning Lahmann is Assistant Professor at eLaw – Center for Law and Digital Technologies at Leiden University Law School. His focuses on the intersection of digital technologies and international law, especially the use of digital open-source information by civil society actors, disinformation, the legal, political, and ethical implications of the use of AI in military and security applications, and transnational cybersecurity.

In 2021/22, Henning was a Hauser Post-Doctoral Global Fellow at NYU School of Law, pursuing research on the use of open-source intelligence in the context of Russia's invasion of Ukraine, with support from a research grant by the German Academic Exchange Service. Prior to joining Leiden University, he worked as Program Leader International Cyber Law at the Digital Society Institute at the ESMT Berlin. From 2020 to 2022, he also served as an associate research fellow at the Geneva Academy of International Humanitarian Law and Human Rights. In 2019, Henning pursued a research fellowship at the Israel Public Policy Institute and the Institute for National Security Studies at Tel Aviv University. Henning holds a doctoral degree in international law from the University of Potsdam, Germany. Previously, he worked as senior policy advisor at the Berlin-based think tank iRights.Lab and acted as the German correspondent for Freedom House's annual "Freedom on the Net" report. Henning studied law and philosophy in Hamburg and Prague and held research fellow positions at the Walther Schücking Institute for International Law at the University of Kiel, the University of Potsdam, and FU Berlin.

10. The IDF Introduces: Artificial Intelligence in the Battlefield, A New Frontier?

Dr Tal Mimran (Hebrew University and Tachlith Institute)

New and emerging technologies significantly impact the ways in which military operations are conducted. Advancements have been achieved in the development and deployment of autonomous weapon systems, military use of cyberspace, and more. Another emerging field in which leaps are currently being made is Artificial Intelligence (AI) with military applications.

Recently, high-ranking officers in the IDF admitted the use of AI-based tools as part of Israel's military arsenal. AI is used both for offensive needs, for example, in the context of targeting, and for defensive ones, e.g. to alert forces that they are under threat of a rocket or a missile or to assist in better safeguarding border movement. There is no doubt about it – asserting AI-superiority in the battlefield has tremendous value in terms of deterrence. It is also obvious that there are advantages for putting into use AI-based tools in order to improve existing military capabilities. However, what role do moral and legal considerations have in this trend? Particularly, are there any limitations on the desire to introduce such new tools in practice, and on their actual use in practice? My presentation looks into some selected examples from the developing experience in Israel in this regard, in order to consider these questions, and also to highlight and delve into broader legal themes relating to the introduction of AI to the battlefield.

In my view, there is room for prudence when deploying new military capabilities, especially ones that are not regulated like AI-based tools. At this point in time, there is no benchmark to follow, and in fact, States are still trying to grasp and regulate less harmful technological tools.

First, an important step is a preliminary measure of evaluating the legality of new technologies, as required by article 36 of API, that obliges States to determine "in the study, development, acquisition or adoption of a new weapon or new means or methods of warfare," whether their employment would be prohibited under international law. In other words, States are required to use prophylactic impact assessment measures, like legality review of weapons and also of means and methods of warfare. The importance, as well the challenges, of conducting proper legal reviews increases with new technologies with unclear impacts on civilians. This review is of special pertinence given that cyber-tools can cause significant and widespread damage to objects and infrastructure (e.g. Stuxnet), and that cyber-engagements can and often precede the deployment of conventional military force or comprise part of a broader attack (as occurred in the context of the Russia-Ukraine War).

Second, while the tendency to lean on AI is obvious, as such a tool can calculate in a few seconds some things that humans will need weeks to do, if at all, there are some inherent risks with AI systems. Notably, so long as AI tools are not explainable, in the sense that we cannot fully understand why they reached a certain conclusion: how can we justify to trust the AI decision when human lives are at stake? If one of the attacks produced by the AI tool leads to significant harm of uninvolved civilians, who should bear responsibility for the decision?

Third, there is the matter of public-private relationship, and the role of non-State actors in this scene. While we should recall that the private sector is of great importance for prevention, education, investigation and attribution of cyber operations, as seen in the context of the war in Ukraine, we should avoid from over-privatization and fragmentation of authority and responsibility. The damage caused to Israel by the NSO scandal illustrates how such a relation can spiral out of control, and we can only imagine what might occur in the context of more aggressive tools that may be developed by private companies, let alone sold and distributed by them around the world. If spywares were subject to severe criticism, and even a call for a complete ban – by the former UN special rapporteur David Kaye, or Amnesty – what type of criticism will arise in the context of new, and probably more endangering, AI tools ?

Fourth, and as a result of the previous point, we need to think of possible ways to limit the deployment of AI based tools in the military sphere. This can be done by regulation over development, e.g. under article 36 type of a mechanism, or maybe through processes like that of privacy by design. In addition, it is important to consider if trade-restrictions on AI based tools with a military application should be posed, be that on the ability to sell, the ability to use in general, or maybe limitations on the identity of the players which may have access to it .

In this regard, States have long aspired to maintain a monopoly on the use of force, as is evident by legal regimes such as the laws of war, hence a possible, State-centric response, is that only States should use such tools. This view, of course, is both anachronistic and naïve. It is anachronistic since it ignores the fact that some tech companies have cyber capabilities which far exceed those of many States, and since some States actually rely on cooperation with the private sector in the research, production, marketing and selling of cyber tools. It is also naïve, since States have demonstrated that they are capable of misusing tools and power against foreign nationals, and also against their own. Hence, we need to think beyond the State prism.

Fifth, broadly speaking, the world is becoming more divided in ideals and values, and there is a difficulty in promoting international responses. How can the United States and Russia, or the Netherlands and China, decide on measures against hostile cyber operations – if they cannot agree on a common definition of such an act? In the context of AI – we have the Chinese State-driven approach, the European Rights-driven approach, and the US market-driven approach, and we need to wait and see which of the three will gain traction and impact the direction in which we are headed . Regardless of the path chosen, there needs to be some mix of tools in different stages (planning, design, deployment, and retroactive examination), and domestic and international systems should aspire for harmonization, and complementarity.

It seems that, currently, the only common understanding, is that on the importance of maintaining a human in the loop in order to promote accountability. This is, of course, is not enough, and we need to consider both how existing rules – be that general principles of international law, rules of international humanitarian law or international human rights law, can be applied to this new context, and also if there is room for the development of new legal rules or institutions to better cope with this challenge.

Biography

Dr. Tal Mimran teaches international law, and law & technology, at the Hebrew University of Jerusalem and at the Zefat Academic College. He is also the Academic Director of the International Law Forum at the Hebrew University, and the Research Director at the Federmann Cyber Security Research Center, also at the Hebrew University. Tal is also the head of a program dealing with Digital Human Rights in the Israeli think-tank Tachlith.

In the past, Tal edited an online human rights journal, and worked as a legal adviser in the Israeli Ministry of Justice. Tal also served, in reserve duty, as a legal adviser in the Israel Defense Forces (International Law Department).

11. A Mechanism for Overcoming Real or Perceived Problems of Disparate Ethical Outlooks in Multilateral Alliances when Evaluating the Impacts of AI in a Military Context

Dr Michael Wildenauer (University of Melbourne)

There have been widespread calls from researchers and civil society for organizations in business, government, and medical domains to consider the ethical and wider social impacts of AI systems during and after development.[1,2] Leaving aside lethal autonomous weapons systems (LAWS or so-called "killer robots") as a particular and much discussed case, non-LAWS AI developed for, or deployed in, military contexts should also be subject to similar evaluative processes. In addition to issues ordinarily found in evaluating AI, problems may arise in multilateral alliances because of disparities in ethical views that are based on respective prevailing military and national culture when seeking agreement on the conceptual basis for, and what actually constitutes, a reasonable method for evaluations of military AI systems.

In attending to military AI in alliances, it is useful to first recognise that the basis for evaluations of civilian AI has been the subject of considerable debate, including arguments between those favouring a perhaps more nuanced ethical starting point and those favouring a more legalistic approach based on International Human Rights Law (IHRL)[3,4]. Each of these broad methodological preferences has advantages and disadvantages. While almost universal in acceptance, developed with input from a wide range of nations, and legally enforceable, IHRL is often seen as ineffectual and subject to 'box-ticking' compliance, and ethics exists outside of the law and is often hijacked by business in so-called 'ethics washing', but has a rich history of ideas going back thousands of years. As is often the case however, we are not actually constrained to a binary choice.

To address the issue and harness both approaches, this paper suggests a hybrid approach (HILEIA) developed by the author, in which a framework based on general evaluative principles of legitimacy, rational connection and minimum impairment as found in IHRL and other international law is utilized (i.e. the mechanism of the evaluation), but which allows the freedom to employ, for example, principlist, or human rights, or consequentialist, considerations to complete such evaluations (i.e. the detailed criteria that determine compliance with individual evaluative rules and therefore the outcome of the evaluation).[5]

---

[1] See Table 1 in Jasmin Fox-Skelly Eleanor Bird, Nicola Jenner, Ruth Larbey, Emma Weitkamp, and Alan Winfield, 'The ethics of artificial intelligence: Issues and initiatives' (Study, Thee Panel for the Future of Science and Technology.

[2] E.g. UNESCO, *UNESCO's Recommendation on the Ethics of Artificial Intelligence: key facts*, UN Doc SHS/2023/PI/H/1) ('*UNESCO's Recommendation on the Ethics of Artificial Intelligence: key facts*').; Bernd Carsten Stahl and Tonii Leach, 'Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: An empirical test of the European Union Assessment List for Trustworthy AI (ALTAI)' (2023) 3(3) *AI and Ethics* 745-767.

[3] See for example Canca, C. 2019. Why Ethics Cannot be Replaced by the Universal Declaration of Human Rights. Our *World* and Raso, F.A., Hilligoss, H., Krishnamurthy, V., Bavitz, C. and Kim, L., 2018. Artificial intelligence & human rights: Opportunities & risks. *Berkman Klein Center Research Publication*, (2018-6).

[4] Anna Su, 'The promise and perils of international human rights law for AI governance' (2022) 4(2) *Law, Technology and Humans* 166-182.

[5] Michael E. Wildenauer, 'HILEIA - Hybrid International Law & Ethics Impact Assessment' (2023). Working paper

A second set of arguments concerns ethical relativism, that is, whether or not ethical norms or principles do (descriptive) or should (normative) vary between nations or cultures or other demographic groups or should be invariable[6]. Both ethical relativism and its opposite, ethical absolutism, are to be avoided.[7] These disagreements around ethical ground truths may pose a particular problem (particularly in the case of normative relativism) for multilateral military alliances in determining whether the impacts of jointly developed or deployed military AI are ethically sound, given that alliance members may fail to agree even on the starting point for ethical evaluation due to varying ethical viewpoints.

This paper suggests a solution to address the issue by employing a concept derived from negotiation practice; the appeal to a widely accepted standard or universally acceptable authority.[8] Agreements on evaluation parameters developed by appeal to precedent and customary practice would seem more likely to survive reconsideration by one or more alliance members.[9] This universal standard then serves to inform the detailed criteria-set in the framework.

In the case where a) there is no armed conflict and b) the AI systems concerned are not weapons and are unlikely to have weapon-like impacts (e.g., systems to manage military prisoners) or be classed as new means or methods of warfare and therefore not subject to review under Article 36 of AP(I),[10] the consensus instruments signed up to by the overwhelming majority of nations that comprise the IHRL Bill of Rights can serve as this standard or authority and help bridge differences between alliance members.

However, in times of armed conflict other considerations render the use of IHRL less useful. In this case, many human rights principles can be derogable, and are to an extent in any case informed by International Humanitarian Law (IHL) as *lex specialis*.[11] In armed conflict, situations often arise that fall outside of the norms of ethics (e.g. although they may allow certain trade-offs, most ethical systems do not allow that the taking of a human life is desirable, necessary, or at the least a likely outcome), or in situations where the AI systems could conceivably be weaponised, whether or not they are designed as part of weapon systems *per se* (an example being AI-powered hacking tools or neurotechnology). In these cases, a different approach is required. Here the use of another set of, albeit minimalist in a normative sense, consensus instruments, namely IHL, could be employed. IHL fits the bill as an internationally developed and accepted universal authority to which an appeal to is possible where ethical reasoning differs between allies or is not helpful in understanding the impacts of AI systems, and where IHRL has limited authority or relevance.

---

[6] See Baghramian, M., & Carter, J. A. (2018). Relativism. In E.N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Winter 2018 Edition). https://plato.stanford.edu/archives/win2018/entries/relativism/ .

[7] Borna Jalsenjak, 'Ethical Absolutism V Ethical Relativism' (2019) *Encyclopedia of Sustainable Management* 1-2.

[8] Eleanor Wertheim et al, *Skills for resolving conflict: Creating effective solutions through co-operative problem solving* (Eruditions Publishing, 2nd ed, 2006).

[9] Roger Fisher, William Ury and Bruce Patton, *Getting to Yes: Negotiating agreement without giving in* (Routledge, 1991).

[10] *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, Geneva Conventions of 12 August 1949 ('*Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*'). Article 36.

[11] William H Boothby, *Conflict law: the influence of new weapons technology, human rights and emerging actors* (Springer, 2014).

Principles from IHL may be plugged into the HILEIA framework in order to evaluate whether a military AI system meets a minimum standard (remembering that States' obligations to others are non-derogable) by examining adherence to IHL principles. Some States (perhaps at least one of the alliance partners in the best case) will already have done much of this work in Article 36 Weapons Reviews where the AI system under consideration is a weapon or other means of warfare, but evaluations of AI systems supporting medical care, search and rescue, or surveillance etc. should also be made and can be made, by reference to the standard within the HILEIA framework. This is particularly important where AI systems are edge cases close to the boundaries of the definitions of weapons and non-weapons.

Cyber warfare, whether targeted or 'hack-back' in nature is one example of a domain where this is pertinent. Even where the effect of the cyber intrusion has no direct kinetic effect, where only code or data is affected, evaluations of possible failure modes such as the modification of enemy AI-based targeting systems that may go on to cause indiscriminate casualties, or even just the opening of a 'backdoor' to allow other actors to deliberately sabotage AI-weapons, would seem to be prudent in complying with the laws of war.

By appealing to the consensus instruments of IHRL when appropriate, and to IHL when not, multilateral military alliances can overcome potentially divisive differences in ethical grounding to arrive at reasonable human-centred evaluations of military AI. Finally, while minimalist, IHL evaluative principles are not without illuminating power and serve to highlight issues commonly found when evaluating civilian AI systems, such as biased training sets (which speak to impartiality and neutrality), targeting individuals for disinformation, or severe and ongoing environmental damage.

Biography

Dr Michael Wildenauer BSc(Ma.Sc.) GDipCommLaw MBA(Comp.) PhD MACS(Sr.) CP MAICD

Centre for AI & Digital Ethics (CAIDE), Melbourne Law School

Michael is a highly experienced technology executive, board director, consultant, and educator. He teaches in subjects that address the ethical, business, and social impacts of AI and other emerging technology. Currently an academic with CAIDE, and previously a Professor of Practice at La Trobe Business School, Michael spent many years in technology leadership roles (including CIO & CTO) in Australia, and internationally. He is currently the Manager Teaching & Learning and a Senior Lecturer and Senior Research Fellow at the Centre for AI and Digital Ethics, a Senior Lecturer at the Melbourne Law School, and a Non-executive Director of the Royal Children's Hospital. Michael is the immediate past Chair of the Ethics Committee of the Australian Computer Society (ACS).

With qualifications in Computer Science, Law, Business, and Governance, Michael's research interests cover a wide area and focus on the intersection of ethics and rights with technology, law, and governance. He is also listed as a co-inventor on several Australian and US patents.

12. The Role of Military AI in Deterrence Theory and Practice

Dr Bianca Baggiarini (Australian National University)

Deterrence has made a sudden comeback in the rhetoric of security and strategic theory. In Australia, deterrence has historically not registered in the strategic mindset of the country. Yet, due to the shifting geopolitical environment in the Indo-Pacific region, defined by the re-emergence of great power rivalries, and perhaps most importantly Australia's deepening ties to the US, deterrence is now a centrepiece of Australian defence and security policy and rhetoric. In a strategic environment marked by increasing hostility towards China, it has been said numerous times over that deterrence is the best and only path forward for managing innumerable future potential threats.

Indeed, the 2023 Defence Strategic Review (DSR) identifies deterrence as the main task for the Australian Defence Force now and into the future, and deterrence was a primary core public justification for the trilateral AUKUS submarine agreement. Together, AUKUS and the DSR represent Australia's new, rebranded strategic vision for the region, which includes an expressed commitment to both nuclear-powered submarines on the one hand, and advanced capabilities for deterrence on the other.

AUKUS Pillar II is focused on emerging technology. It promises to enhance capabilities in undersea warfare, quantum technologies, advanced cyber, hypersonic and counter-hypersonics, electronic warfare, and artificial intelligence for interoperability across joint operations. According to the DSR: "Technology has a significant impact on the character of warfare and deterrence and will shape the changing balance of power."

However, the DSR also rightly points out that "combined with rising tensions and reduced warning time for conflict, the risks of military escalation or miscalculation are rising." With little awareness of implied irony, the statement overlooks the role of military AI *itself* in contributing to these presumably unwanted risks of escalation and miscalculation (not to mention the related risks wrought by the proliferation of mis- and disinformation) in a time when the tempo of warfare may begin to outpace, undermine, or eclipse human cognition, human judgement, and human decision-making.

Despite the logic of techno-rationality which underpins conceptions of algorithmic war more broadly, deterrence-in-practice is still theorized as profoundly human-centred. Humans signal, perceive, threaten, initiate and de-escalate conflict as part of a collection of nuanced social interactions with *other humans*. Deterrence is about preventing others from engaging in actions deemed undesirable: dissuasion by means of threat. One actor deters another by convincing him that the expected value of a certain action is outweighed by the expected punishment (Jervis 1983). There are many theories of deterrence, stemming mostly from the Cold-War period and beyond. This paper does not seek to summarize or trace the long history of deterrence thinking. Instead, it asks about the status of deterrence as it relates to AI.

What role does AI have in the theory and practice of deterrence? Further, given that many emerging military technologies are software driven, and thus poorly understood and frequently invisible, how might they serve in deterrence, when deterrence requires visibility for its threats to be perceived, knowable, and thus considered credible?

There is a paradox at the heart of using AI ostensibly for deterrence; this paper attempts to engage this complication. This paper argues that deterrence – which requires human agents acting and perceiving in transparent ways – is at odds with the logic that underpins AI-enabled tools and systems in war. Deterrence requires clear and credible communication and signalling of intent; algorithmic war moves practices and processes in the opposite direction, functioning on the basis of increasing anonymity, fragmentation, and privatization. The desire to displace humans from war (due to decades of risk and casualty aversion and associated planning to this end) also has the effect of subverting the main currency in traditional war and deterrence practices: human capital. Military AI is potentially revealing fault lines in deterrence theory and practice; fault lines which have historically been less visible and impactful. Regardless, deterrence only works if agents understand what is being communicated, and machine learning algorithms are notorious for their evasive qualities, abilities to tweak themselves 'in the wild,' and surprising outputs that confuse even their designers.

Can rendering algorithmic results or conclusions more explainable, understandable, traceable or transparent aid in the wider project of deterrence, which *requires* clear, credible, and transparent threats to alter the behaviour of the adversary? This paper explores this challenge by examining the oft-cited ethical principle of transparency/traceability. Clearly indebted to Enlightenment thinking, transparency/traceability as AI ethics principles suggest that algorithms can be made ethically and legally "better" in relation to how much we "know" about them. Shining a light onto the black box of algorithms compels algorithms to "give an account" of themselves (Butler 2001). AI, particularly machine learning algorithms, cannot satisfy this criterion as it is naturally driven and designed to conceal rather than expose - and given its corporate rationale - privatise rather than publicise. If military AI and its effects cannot be rendered visible, then its ability or potential to deter is questioned.

The US, Australia, and NATO agree that transparency or traceability is key to ensuring ethical AI. However, this paper argues that AI cannot satisfy this criterion as it is naturally driven and designed to conceal rather than expose - and given its corporate rationale - privatise rather than publicise. If military AI and its effects cannot be rendered visible, then its ability or potential to deter is questioned. Given Australia's and the US's rhetorical emphasis on deterrence, and the tremendous efforts underway to integrate military AI into defence planning and operations presumably for this aim, the role of military AI in deterrence logic and practice reflects an open area of research requiring further investigation.

This paper concludes that deterrence and AI are being combined in discourse and practice in ways that render them strange bedfellows. The awkwardness is a symptom of a recurring feature implicit in the changing character of war: War is about humans, but this idea is wrapped in a discomfort and displeasure with the risks associated with deploying and maintaining human forces in active combat zones. AI and remote warfare broadly are viewed as solution to this (political) discomfort, since these systems are meant

to supplement or replace humans in war. But in doing so, AI, for better or for worse, is undermining the old frameworks and concepts that we typically think govern war and give it ostensible clarity and rationality, such as deterrence. Reconciling or explaining this gap is an enormous task, which goes to the heart of military violence today, namely, the shifting ontological status of war, and what we make of it.

Biography

Bianca Baggiarini is a political sociologist and Lecturer in military studies at the Strategic and Defence Studies Centre at ANU. Her research and teaching practices are aimed at applying sociological theories and methods to the study of war. Bringing together interdisciplinary fields of international political sociology, critical military studies, and social science of technology studies, Bianca's research attempts to better understand the changing character and quality of war and violence through an equal interest in both micro and macro scales and theoretical and empirical knowledge.

Bianca's current research is on the social, political, and ethical impacts of autonomy and AI-enabled technologies in military and security contexts. She is examining the role of trust discourse in shaping debates about ethical military AI, the changing status of soldiers' labour in the context of increasing autonomy, and the social meaning of technology demonstrations as it relates to communicating the ethical and legal potential of AI-enabled systems. Her forthcoming monograph, Governing Military Sacrifice, is one of the first books to connect the rise of drones and combat unmanning with military and security privatization and includes original interview data from both drone advocates and critics alike.

Bianca holds a PhD (2018) from York University in Toronto, an MA in sociology from Simon Fraser University, and a BA in political science from Simon Fraser University. From 2019 to 2021, she was a Lecturer at UNSW at the Australian Defence Force Academy.

13. Visuals as Sources of Normativity in the Debate about Weaponised Artificial Intelligence

Dr Ingvild Bode (University of Southern Denmark), Dr Guangyu Qiao-Franco (Radboud University Nijmegen), Anna Nadibaidze (University of Southern Denmark)

Draft, please do not cite

Artificial Intelligence (AI) technologies[1] are transforming politics and society. Their growing application in the military domain raises the most fundamental questions because they concern life and death decisions. This development is often associated with so-called autonomous weapon systems (AWS) that can, once activated, "track, identify, and attack targets with violent force without further human intervention".[2] AWS may integrate autonomous and AI technologies into the targeting functions of weapons. What is at stake in the debate about the governance of AWS and AI in the military domain more broadly is therefore the extent to which humans remain in control over the use of force. As such, AWS are the subject of significant literature in International Relations (IR) and beyond that spans their legal, ethical, and security implications[3] – while also questioning whether changes associated with AWS are really that fundamental in nature. A sub-set of this literature are studies on the implications of developing AWS for international norms by examining, for example, the trajectory of the international debate about AWS under the auspices of the United Nations Convention on Certain Conventional Weapons (CCW) in Geneva.[4] Such studies provide important entry points to emerging norms on AWS, but they chiefly draw on elite practitioner practices and their public discourse as a source for studying the normative consequences of AWS.

Yet, the normative space surrounding AWS and AI in the military domain goes beyond the elite level. As military applications have become increasingly salient in news cycles, media platforms have become more attentive to the topic and this reporting includes a significant visual element. Media articles or essays written by stakeholders with a broader, public audience in mind are typically accompanied by images that appear to quickly become repetitive. In the context of reporting on AI technologies more broadly, Rogers and Moretti have found that print and digital media commonly illustrates AI with "clichéd images" [5] of

[1] We recognise the umbrella term "AI" as being vague, imprecise, and politically contentious Holland Michel, "Recalibrating Assumptions on AI"; Tucker, "Signal's Meredith Whittaker: 'These Are the People Who Could Actually Pause AI If They Wanted To.'". In this contribution, we use a broad definition of AI as the attempt "to create machines or things that can do more than what is programmed into them" Gebru, "Don't Fall for the AI Hype.".

[2] International Committee of the Red Cross, "ICRC Position on Autonomous Weapon Systems."

[3] E.g. Horowitz, "When Speed Kills"; Scharre, *Four Battlegrounds*; Johnson, *Artificial Intelligence and the Future of Warfare*; Garcia, *Common Good Governance in the Age of Military Artificial Intelligence*; Roff, "The Strategic Robot Problem"; Roff, "Lethal Autonomous Weapons and Jus Ad Bellum Proportionality"; Altmann and Sauer, "Autonomous Weapon Systems and Strategic Stability"; Bode et al., "Algorithmic Warfare: Taking Stock of a Research Programme."

[4] Bode and Huelss, "Autonomous Weapons Systems and Changing Norms in International Relations"; Bahcecik, "Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames"; Rosert and Sauer, "How (Not) to Stop the Killer Robots"; Prem, "Governing through Anticipatory Norms"; Garcia, "Future Arms, Technologies, and International Law"; Bode, "Norm-Making and the Global South: Attempts to Regulate Lethal Autonomous Weapons Systems."

[5] Duarte, "Why We Need Better Images of AI. The Better Images of AI Project Is Aiming to Change How AI Is Represented in the Media."

"glowing, blue brains; brains illustrated as circuit boards; shiny, white or metallic robot figures and disembodied hands" that are strikingly resemblant.[6] This visual language of AI[7] is problematic because it creates "unrealistic or distorted narratives […] about the scale and scope of AI's current use and impact".[8] Such visuals therefore lead public understanding of AI technologies into particular, often unhelpful directions, while also contributing to "socio-technical blindness"[9], that is the inability of seeing these technologies in the context of their actual development and use. A cursory glance at images taken to illustrate military applications of AI appears to confirm these established tendencies. Indeed, these dynamics connect to how IR and critical security studies have increasingly recognized visuals as central to political meaning-making and contemporary warfare.[10] This scholarship also recognizes such images as inherently normative.[11] But it does not theorize the normative potential of visuals explicitly, nor does it presently connect to the constructivist research programme on norm theory. We therefore ask: *how are AI technologies in weapon systems constructed across widely accessible visuals and with what normative consequences?*

In response to this research question, we argue that visuals, in drawing on and constituting particular imaginaries, produce normative substance through presenting particular ways of imagining AI in weapon systems as "normal" and "appropriate". This argument makes analytical and empirical contributions to norm research in IR and to the literature on algorithmic warfare. Analytically, by identifying how visuals communicate normative substance, we connect research endeavors across critical constructivism, namely norm research, visual analysis, and critical security studies. This move also aims to diversify the pool of empirical sources where normative substance may emerge. Empirically, we contribute to the IR literature on AWS by investigating how such (algorithmic) practices of visual appropriation shape emergent normative substance on AI in weapon systems in particular ways. As outlined above, our analysis also connects to an ongoing critical discussion about the kind of visuals that dominate the debate about AI in general.[12] Methodologically, we analyse visuals derived from images searches in the most prominent Chinese, Russian, and US search engines (Baidu, Yandex, Google). We have chosen to focus on China, Russia, and the US because these states pursue the most significant investments in the field of military AI and are often identified as 'leaders' in this development, albeit to different degrees. Examining the extent to which the visuals prominently presented to their publics differ or coalesce can therefore offer us important insights into the imaginaries and the normative visions that sustain public discourse on military AI in China, Russia, and the US.

[6] Rogers and Moretti, "Is Seeing Believing?"

[7] Rogers and Moretti.

[8] Duarte, "Why We Need Better Images of AI. The Better Images of AI Project Is Aiming to Change How AI Is Represented in the Media."; Cave, Coughlan, and Dihal, "Scary Robots: Examining Public Responses to AI."

[9] Rogers and Moretti, "Is Seeing Believing?"

[10] Bleiker, *Visual Global Politics*; Callahan, *Sensible Politics: Visualizing International Relations*; Freistein and Gadinger, "Performing Leadership"; Hansen, "Theorizing the Image for Security Studies," 2011; Hansen, "Images and International Security"; Markussen, "Inscribing Security."

[11] Hansen, "Theorizing the Image for Security Studies," March 2011.

[12] Cave and Dihal, "The Whiteness of AI."

The remainder of the paper is structured as follows: first, we develop our argument of how visuals constitute and communicate normative substance in the debate of military applications of AI in more detail. Second, we outline our methodological choices and approach: our data corpus contains visuals resulting from image searches on Baidu, Google, and Yandex that we accessed via VPN services in order to see the results that come up for the respective publics. The visuals we analysed amount to the first page of results for four keywords: "AI weapons", "military AI", "military robots", and "killer robots" (20 images per keyword, 240 images in total).[13] Third, we present the results of our visual analysis that we structure through a set of six analytical questions that we ask for each picture: (1) what does the image depict (descriptive); (2) where is the image published/what is the content of the accompanying article; (3) does the image function performatively or decoratively;[14] (4) does the image include humans; (5) does the image appear futuristic or to depict current-level technology; (6) does the image include humanoid robots; and (7) how and what kind of context is depicted. Our initial empirical observations point towards three paths of emerging normative substance inherent to visuals in the debate about AI in weapon systems: (1) unnecessary futurism, (2) no (more) humans, and (3) "clean" AI. Finally, we close with a critical conclusion and sketch avenues for further research.

References

Altmann, Jürgen, and Frank Sauer. "Autonomous Weapon Systems and Strategic Stability." *Survival* 59, no. 5 (September 3, 2017): 117–42.

Bahcecik, Serif Onur. "Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames." *Global Policy* 10, no. 3 (2019): 365–69.

Bleiker, Roland, ed. *Visual Global Politics*. Abingdon: Routledge, 2018.

Bode, Ingvild. "Norm-Making and the Global South: Attempts to Regulate Lethal Autonomous Weapons Systems." *Global Policy* 10, no. 3 (2019).

Bode, Ingvild, and Hendrik Huelss. "Autonomous Weapons Systems and Changing Norms in International Relations." *Review of International Studies* 44, no. 3 (2018): 393–413.

Bode, Ingvild, Hendrik Huelss, Anna Nadibaidze, Guangyu Qiao-Franco, and Tom F.A. Watts. "Algorithmic Warfare: Taking Stock of a Research Programme." *Global Society*, forthcoming 2023.

Callahan, William A. *Sensible Politics: Visualizing International Relations*. Oxford: Oxford University Press, 2020.

[13] The keywords we used differ slightly depending on what is the most commonly used term in Chinese and Russian discourse and media reporting. In the Russia case, the keywords used were "искусственный интеллект в оружии [*iskusstvennyi intellekt v oruzhii* –AI in weapons, given that the term "AI weapons" is not commonly used in Russia]", "военный искусственный интеллект [*voyennyi iskusstvennyi intellekt* – military AI]", "военные роботы [*voyennye roboty* –military robots], and "роботы убийцы [*roboty ubiitsy* –killer robots]". In the case of China, the key words used are "智能武器 [zhinengwuqi-intelligent weapons]", "人工智能武器 [rengongzhinengwuqi - AI weapons]", "军用机器人 [junyongjiqiren - military robots]", and "自主武器 [zizhuwuqi - autonomous weapons]".

[14] Rogers and Moretti, "Is Seeing Believing?"

Cave, Stephen, Kate Coughlan, and Kanta Dihal. “Scary Robots: Examining Public Responses to AI.” Leverhulme Center for the Future of Intelligence, 2019. http://lcfi.ac.uk/media/uploads/files/AIES-19_paper_200_Dihal_Cave_Coughlan_XLLBdft.pdf.

Cave, Stephen, and Kanta Dihal. “The Whiteness of AI.” *Philosophy & Technology* 33, no. 4 (December 2020): 685–703. https://doi.org/10.1007/s13347-020-00415-6.

Duarte, Tania. “Why We Need Better Images of AI. The Better Images of AI Project Is Aiming to Change How AI Is Represented in the Media.” The Alan Turing Institute, May 2, 2022. https://www.turing.ac.uk/blog/why-we-need-better-images-ai.

Freistein, Katja, and Frank Gadinger. “Performing Leadership: International Politics through the Lens of Visual Narrative Analysis.” *Political Research Exchange* 4, no. 1 (2022): 1–20. https://doi.org/10.1080/2474736X.2022.2124922.

Garcia, Denise. *Common Good Governance in the Age of Military Artificial Intelligence*. Oxford: Oxford University Press, forthcoming.

———. “Future Arms, Technologies, and International Law: Preventive Security Governance.” *European Journal of International Security* 1, no. 01 (February 2016): 94–111.

Gebru, Timnit. “Don’t Fall for the AI Hype.” Tech Won’t Save Us Podcast with Paris Marx, January 19, 2023. https://techwontsave.us/episode/151_dont_fall_for_the_ai_hype_w_timnit_gebru.

Hansen, Lene. “Images and International Security.” In *The Oxford Handbook of International Security*, edited by Alexandra Gheciu and William C. Wohlforth. Oxford: Oxford University Press, 2018.

———. “Theorizing the Image for Security Studies: Visual Securitization and the Muhammad Cartoon Crisis.” *European Journal of International Relations* 17, no. 1 (2011): 51–74. https://doi.org/10.1177/1354066110388593.

———. “Theorizing the Image for Security Studies: Visual Securitization and the Muhammad Cartoon Crisis <sup/>.” *European Journal of International Relations* 17, no. 1 (March 2011): 51–74. https://doi.org/10.1177/1354066110388593.

Holland Michel, Arthur. “Recalibrating Assumptions on AI.” Chatham House, April 12, 2023. https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai.

Horowitz, Michael C. “When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability.” *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 764–88. https://doi.org/10.1080/01402390.2019.1621174.

International Committee of the Red Cross. “ICRC Position on Autonomous Weapon Systems,” 2021. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

Johnson, James. *Artificial Intelligence and the Future of Warfare: The USA, China, and Strategic Stability*. Manchester: Manchester University Press, 2021.

Markussen, Håvard Rustad. “Inscribing Security: The Case of Zelensky’s Selfies.” *Review of International Studies*, August 10, 2023, 1–19. https://doi.org/10.1017/S0260210523000359.

Prem, Berenike. “Governing through Anticipatory Norms: How UNIDIR Constructs Knowledge about Autonomous Weapons Systems.” *Global Society* 36, no. 2 (April 3, 2022): 261–80. https://doi.org/10.1080/13600826.2021.2021149.

Roff, Heather M. “Lethal Autonomous Weapons and Jus Ad Bellum Proportionality.” *Case Western Reserve Journal of International Law* 47, no. 1 (2015): 37–52.

———. “The Strategic Robot Problem: Lethal Autonomous Weapons in War.” *Journal of Military Ethics* 13, no. 3 (July 3, 2014): 211–27.

Rogers, Ann T., and Lisa Talia Moretti. “Is Seeing Believing?” Medium, n.d. https://medium.com/@info_95167/is-seeing-believing-e967211ad812.

Rosert, Elvira, and Frank Sauer. “How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies.” *Contemporary Security Policy* 42, no. 1 (2021): 4–29.

Scharre, Paul. *Four Battlegrounds: Power in the Age of Artificial Intelligence*. First edition. New York: W.W. Norton & Company, 2023.

Tucker, Ian. “Signal’s Meredith Whittaker: ‘These Are the People Who Could Actually Pause AI If They Wanted To.’” The Guardian, June 11, 2023. https://www.theguardian.com/technology/2023/jun/11/signals-meredith-whittaker-these-are-the-people-who-could-actually-pause-ai-if-they-wanted-to.

Biographies

Dr Ingvild Bode is Associate Professor at the Centre for War Studies, University of Southern Denmark. Her research focuses on processes of normative and policy change, especially with regard to the use of force. She is the Principal Investigator of the European Research Council-funded project AutoNorms: Weaponised Artificial Intelligence, Norms, and Order (08/2020-07/2025). The AutoNorms project investigates how practices related to autonomous weapon systems change international norms. AutoNorms examines military, transnational, political and dual-use practices in China, Japan, Russia, and the USA. Ingvild serves as the co-chair of the IEEE Research Group on AI and Autonomy in Defence Systems. Her work has been published with the European Journal of International Relations, Ethics and Information Technology, Review of International Studies, International Studies Review and other journals. Ingvild’s most recent book entitled Autonomous Weapons and International Norms (co-authored with Hendrik Huelss) was published by McGill-Queen’s University Press in 2022. Previously, Ingvild was Senior Lecturer in International Relations at the University of Kent, Canterbury (2015-2020) and a Japan Society for the Promotion of Science International Research Fellow with joined affiliation at United Nations University and the University of Tokyo (2013-2015). bode@sam.sdu.dk

Dr Guangyu Qiao-Franco is Assistant Professor of International Relations at Radboud University in the Netherlands, and Senior Researcher of the ERC funded AutoNorms Project at the University of Southern Denmark. She is principally interested in China studies, ASEAN studies, international organisations, the agency of the Global South in international politics and emerging technologies (including AI and cyber security). Her research leverages practice theory, norm contestation, norm diffusion and actor-network theories to interpret legal and foreign policy

instruments developed by developing countries. Her work has been published in International Affairs, The Pacific Review, International Relations of the Asia Pacific, Chinese Journal of International Politics, Journal of Contemporary China, among others. She is the author of the book UN-ASEAN Coordination: Policy Transfer and Regional Cooperation Against Human Trafficking in Southeast Asia (Edward Elgar, 2023). *guangyu.qiao-franco@ru.nl*

Anna Nadibaidze is a Ph.D. Research Fellow in International Politics at the Center for War Studies, University of Southern Denmark. She is also a researcher for the European Research Council funded AutoNorms project, which examines how the development of autonomous weapon systems affects international norms. Her research has been published in Contemporary Security Policy, Ethics and Information Technology, and Journal for Peace and Nuclear Disarmament. *anadi@sam.sdu.dk*

14. Platform Warfare: The Rise and Implications of Platform Corporations in AI-Centered Warfare

Dr Marijn Hoijtink (University of Antwerp)

Following the Russian invasion of Ukraine, Clearview AI, a New York-based startup company, reached out to the Ukrainian Ministry of Defense to offer access to its platform. Clearview AI is a facial recognition company that scrapes images from social media websites and stores them in a searchable database. Its algorithms are trained to match against images in that database, which allows its clients, mostly law enforcement, to identify subjects of interest by uploading their photos. Clearview AI's platform is used by the Ukrainian armed forces in the war in Ukraine to identify dead Russian soldiers and prisoners of war, but also to screen people on the move passing through security checkpoints (Clayton, 2022). Critics of facial recognition technology have raised concerns about this use, pointing out that the technology has proven to be inaccurate. Civil rights organization Privacy International (2022), for instance, has raised questions about how the war in Ukraine serves as a testbed to push and normalize the use an already controversial platform in the context of an active war.

While the specifics of Ukraine's use of facial recognition technology remain shrouded in secrecy, the example of Clearview AI demonstrates how platform companies – defined here as online services corporations – have become central actors within contemporary warfare. Within the disciplines of International Relations and security studies, the growing security implications of platform companies such as Alphabet-Google, Meta, Microsoft and Amazon are receiving increased scholarly attention. To date, the majority of this scholarship has focused on the online presence of social media platforms, and on how online platforms represent a new and virtual battlespace alongside the conventional battlefield (e.g. Singer & Brooking, 2018; Zeitzoff, 2017). This includes literature on the involvement of social networking sites in disinformation and "fake news" campaigns, such as in the case of the Cambridge Analytica-Facebook scandal and Russian interference in the 2016 US elections, or the role of platforms in provoking violent extremism, or facilitating terrorism or mass atrocities, as demonstrated by Facebook's role in the Myanmar conflict.

What has received much less scholarly attention is how platform companies are directly implicated in the conduct of warfare, and how they are shaping technologies, practices and realities of warfighting *on the physical battlefield*. However, as shown by the example of Clearview AI or Google's role in the Pentagon-financed drone program Project Maven (see Hoijtink & Planqué-van Hardeveld, 2022; Suchman et al., 2018), these same companies are increasingly involved in lethal collaborations with the military (Crescendo Project, nd.; Gilbert, 2019). They are especially crucial in the provision and production of the emerging digital infrastructure that underlies modern forms of algorithmic and automated warfare (Amoore & Raley, 2017; Crawford, 2021; Suchman, 2020; Wilcox, 2017). The war in Ukraine, where US surveillance company Palantir claims to have become "responsible for most of the targeting" (Dastin, 2023, n.p.), or Elon Musk's Starlink instead prevents attacks from happening (Lyngaas, 2023), underwrites and further accelerates these developments.

In this paper, I investigate the growing role of platform corporations within the military and on the battlefield through what I conceptualize as "platform wars". The concept of platform wars captures how platform corporations, in their interactions with the military, produce new and shared ways of 'thinking' and waging warfare today. As argued by Antoine Bousquet (2010), how we think about and organize warfare varies historically depending on the prevalence of a particular set of technologies or scientific ideas in a given period. My proposition in this paper is that the digital platform represents such a defining technology and paradigmatic shift, in the sense that it increasingly "inform[s] our very nature of combat and the forms of military organization best suited to prevail in it" (Bousquet, 2010, p. 3).

Understood as such, the paper adopts a particularly expansive meaning of the concept of the platform, which goes beyond their strict technical and computational aspects (cf. Baldwin & Woodard, 2009; Bogost & Montfort, 2009; Helmond, 2015). Building on scholarship in media studies and International Political Economy, I conceptualize the platform as a broader governing model or mode of power (Gillespie, 2010; Van Dijck et al., 2018; Wood & Monahan, 2019), which sets the conditions for a specific and new set of military discourses, technologies and practices to emerge Platform companies play a key role in this 'platformization' of the military – as the operators and owners of specific digital platforms that are acquired and used by the military, such as in the case of Palantir's Gotham platform used by the US Army[1], but also as the dominant representatives and advocates of the rise and success of the broader platform model. At the same time, the platform model is also enthusiastically adopted and embraced by military entrepreneurs and technophiles "as a means of responding to what they characterize as the new demands of 21st century warfighting" (Suchman, 2022, p. 2).

Concretely, then, the paper traces how "the platform" emerges and circulates as a relevant knowledge paradigm or "technopolitical imaginary" across military and private networks of actors (Suchman, 2022). Whereas, traditionally, the platform bears a strong rhetorical resonance related to the *weapons* platform, today, the platform represents the cloud infrastructure and digital platforms that are meant to accelerate and decentralize AI capabilities across the military. Based on a discourse analysis of a variety of publicly available, defense-related sources across Western militaries and companies, including official policy documents, speeches, blogs, defense magazine articles, project descriptions, company websites and descriptions, and newspaper articles, I analyze the translation and circulation of the platform knowledge paradigm into actual military strategy and doctrinal thinking. Hereby, I am not interested in comparing or identifying differences across national contexts. Instead, I focus on how the platform operates as a dominant knowledge paradigm across national contexts, and how this matters for how modern warfare is 'thought', but also, eventually, waged. Finally, I also draw explicit attention to the consequences of these developments for democratic principles of control and accountability; specifically how the platform model allows platform companies to expand their role in the military domain, while also defying responsibility and accountability for when something goes wrong.

[1] See: https://www.youtube.com/watch?v=rxKghrZU5w8

References

Amoore, L., & Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, *48*(1), 3-10.

Baldwin, C. Y., & Woodard, C. J. (2009). The Architecture of Platforms: A Unified Wiew. In A. Gawer (Ed.), *Platforms, markets and innovation* (pp. 19-44). Edward Elgar Publishing.

Bogost, I., & Montfort, N. (2009). *Platform Studies: Frequently Questioned Answers*. Digital Arts and Culture. Retrieved July 7, 2021 from https://escholarship.org/content/qt01r0k9br/qt01r0k9br.pdf

Bousquet, A. (2010). *The scientific way of warfare: Order and chaos on the battlefields of modernity*. Oxford University Press.

Clayton, J. (2022, April 13). How facial recognition is identifying the dead in Ukraine. *BBC* https://www.bbc.com/news/technology-61055319

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crescendo Project. (nd.). *Digital Destroyers. How Big Tech Sells War on Our Communities*. Retrieved September 19 from https://bigtechsellswar.com/

Dastin, J. (2023). Ukraine is using Palantir's software for 'targeting', CEO says. *Reuters*. https://www.reuters.com/technology/ukraine-is-using-palantirs-software-targeting-ceo-says-2023-02-02/

Gilbert, E. (2019). Military geoeconomics: money, finance and war. In *A Research Agenda for Military Geographies* (pp. 100-114). Edward Elgar Publishing.

Gillespie, T. (2010). The politics of 'platforms'. *New media & society*, *12*(3), 347-364.

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media+ Society*, *1*(2), 1-11.

Hoijtink, M., & Planqué-van Hardeveld, A. (2022). Machine learning and the platformization of the military: A study of google's machine learning platform TensorFlow. *International Political Sociology*, *16*(2), olab036.

Privacy International. (2022, March 18). The Clearview/Ukraine partnership - How surveillance companies exploit war. https://privacyinternational.org/news-analysis/4806/clearviewukraine-partnership-how-surveillance-companies-exploit-war

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The weaponization of social media*. Eamon Dolan Books.

Suchman, L. (2020). Algorithmic Warfare and the Reinvention of Accuracy. *Critical Studies on Security*, *8*(2), 175-182. https://doi.org/10.1080/21624887.2020.1760587

Suchman, L. (2022). Imaginaries of omniscience: Automating intelligence in the US Department of Defense. *Social Studies of Science*, *0*(0). https://doi.org/10.1177/03063127221104938

Suchman, L., Irani, L., & Asaro, P. M. (2018, 16 May). Google's march to the business of war must be stopped. *The Guardian*. https://www.theguardian.com/commentisfree/2018/may/16/google-business-war-project-maven

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Wilcox, L. (2017). Embodying Algorithmic War: Gender, Race, and the Posthuman in Drone Warfare. *Security Dialogue*, *48*(1), 11-28.

Wood, D. M., & Monahan, T. (2019). Platform Surveillance. *Surveillance & society*, *17*(1/2), 1-6.

Zeitzoff, T. (2017). How Social Media Is Changing Conflict. *Journal of Conflict Resolution*, *61*(9), 1970-1991. https://doi.org/10.1177/0022002717721392

Biography

Dr. Marijn Hoijtink is an Associate Professor in International Relations at the Department of Political Science at the University of Antwerp. Her research focuses on military technology, militarism, and the changing character of warfare, and has been published in, among other journals, Security Dialogue, International Political Sociology, European Journal of International Security, and European Security. She is the editor of Technology and Agency in International Relations (with Matthias Leese), published by Routledge in 2019. Her current research project, funded by the Dutch Research Council (NWO), focuses on military applications of artificial intelligence (AI), with a particular interest in how these technologies shape the way in which warfare is thought, fought, and lived. She recently obtained an Odysseus grant from the Flemish Research Foundation (FWO) for her project PLATFORM WARS, in which she will study the growing role of commercial technology companies in modern warfare. Marijn.hoijtink@uantwerp.be @HoijtinkMarijn

15. The AI-Augmented Super Soldier: Enhancement, Interfaces and the Extended Cognition of Human-Machines

Dr Thomas Christian Bächle (University of Bonn; Humboldt Institute for Internet and Society Berlin)

The terms "military AI" and "autonomous weapon systems" (AWS) usually evoke ideas of technologies that are largely independent of human supervision or intervention. The autonomy of these systems, however, is much more complex than the common "in, on or out of the loop" vocabularies imply. Rather, the processual, ethical or legal questions on "meaningful human control", responsibility or accountability in relation to military AI are fundamentally and inextricably linked to conceptions of agency and human-machine interactions.

(1) Following this premise, the presentation will start by giving a short overview of paradigmatic notions such as network-centric warfare, human-machine teaming and augmentation in military contexts. Agency is a key dimension when conceptualising these human-machine configurations. In this context, interfaces will be theoretically foregrounded, as they play a central role in these interactions.

This theoretical framework of human-machine interactions gains a particular relevance in the field of military AI technologies and autonomous (weapon) systems, because this is where questions of control and intervenability become particularly consequential. Autonomous systems, in more general terms, can be conceptualised in their various degrees of self-directedness and self-sufficiency (Bradshaw et al. 2012), in which they are still and necessarily dependent on external factors such as their energy supply or infrastructural factors that ensure their functioning. This means, somewhat paradoxically, that an autonomous system is dependent on its relations to the system's outside. This relatedness also marks a gap: The specific programming and planning of the system is categorically different from the performative reality of human-machine interactions, the situated (inter-)actions between humans and machines (Suchman 2007).

The role of the interface, however, often remains underexplored when theorising these distributed forms of agency. As the autonomy of the "autonomous system" does not characterise the entity itself, but its relatedness to the outside, which is why the interface as the locus of mediation processes has a particular relevance. The interface "is a form of relation that obtains between two or more distinct entities, conditions, or states such that it only comes into being as these distinct entities enter into an active relation with one another" (Hookway 2014, 4). This way, the interface already establishes a human-machine relation, even before a formalised role such as (human) user or operator is adopted. In the dynamic performative settings of human-machine interactions, interfaces can be conceptualised as more than just control units or screens. They rather co-produce the dynamics of the human-machine relationship, a circumstance that makes them so meaningful for questions of autonomy, control and intervenability.

(2) Against this interface-centred theoretical background, the ideas of augmentation and enhancement will be introduced by discussing the military R&D concept of the "enhanced soldier". In short, this term describes the enhancement of a soldier's performance and capabilities through, among others, neurological, biological or technological interventions, which also include media technologies (e.g. Boisboissel/Le Masson 2021). These technologies of augmentation and enhancement create new possibilities of (inter-)actions and perceptual spaces.

Much more than just a military concept, the "enhanced soldier" therefore articulates a prototypical idea of the human-machine relationship: on the one hand, the emphatically praised super-human abilities necessarily come with amplified dependencies and hierarchies. On the other hand, they characterise a terminologically complex relationship between autonomy and heteronomy, cyborgisation, distributed agency or even the (also legally) complex idea of the human as a weapon. The "enhanced soldier", therefore, can only be understood within larger conceptual histories of human-machine figurations, such as cybernetics, artificial intelligence and robotics (Suchman/Weber 2016). Contemporary ideas of human augmentation in military contexts form a continuous line with these long-standing conceptual traditions.

(3) Last, the presentation will address interface-supported extensions of perceptual spaces and cognitive processes. The goal is to explore how approaches and models taken from cognitive science can sharpen our understanding of these relations. To this end, systems such as exoskeletons, prostheses or digital media ("augmented reality") and associated cognitive enhancements will be discussed especially in the context of military development concepts and objectives. The empirical material includes examples from military strategy papers, DARPA projects and military interpretations of the "enhanced soldier" figuration, such as the German Bundeswehr "Infanterist der Zukunft".

Regarding agency and interfaces, a focus on the relatedness to the human mind seems particularly worthwhile exploring. Interfaces and their potential to establish a connection between two entities can be utilised to theorise the aspects of mediation in cognitive processes in at least three regards: First, by looking at how the environment is represented via the computer system when addressing human senses; second, by analysing the modes in which a relationship between humans and the system can be initiated or maintained; and third, by taking into account the possibilities that are implemented in the system equipping it with ways to interact with the environment.

The approach that is foregrounded in the presentation ties these factors mainly to the material structure and functionality of the interface involved: in both regards, the ways in which "the world" is translated into the system and also how it is represented and pre-structured for human cognition. Conceptually, it draws on theoretical approaches that make sense of the intricate relationships between cognition, perception and interfaces. Close human-machine configurations allow the idea of a type of networked cognition not necessitating consciousness (Hayles 2017), effectively a functional model of cognitive processes that does not prioritise the human. In this regard, cognitive science offers an array of approaches and ideas that help to reinterpret the extended cognition of human-machines, including the 4E model of cognition (which regards it as embodied, embedded, extended, enacted), the extended mind hypothesis (Clark/Chalmers

1998), the idea of cognitive assemblages (Hayles 2017), the structuring of mind, cognition and autonomy (Clark 2004, Anderson 2022) or the interpretation of reality/virtuality (Chalmers 2022).

With this conceptual approach, the presentation hopes to make a contribution to the foundations of and premises made in the debates on ethical and legal aspects regarding military AI. It argues for an interface-based and cognition-centred approach to (distributed) agency in human-machine interconnections and interactions.

References

Anderson, J. (2022). Scaffolding and Autonomy. In Colburn, B. (Ed.): *The Routledge Handbook of Autonomy* (pp. 158-166), Routledge.

de Boisboissel, G., & Le Masson, J.-M. (2021). The Enhanced Soldier. Definitions. *Military Review*. https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2021-French-OLE/Part-2-Definitions/

Bradshaw, J. M., Hoffman, R. R. , Woods, D. D. , & Johnson, M. (2013). The Seven Deadly Myths of "Autonomous Systems". *IEEE Intelligent Systems*, 28(3), 54–61.

Chalmers, D. J. (2022): *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton & Company.

Clark, A. (2004) *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, Oxford.

Clark, A. & Chalmers, D. (1998) The Extended Mind. *Analysis*, 58(1), 7–19.

Hayles, N. K. (2017). *Unthought: The Power of the Cognitive Nonconscious*. The University of Chicago Press.

Hookway, B. (2014): *The Interface Effect*. The MIT Press.

Suchman, L. A. (2007): *Human-Machine Reconfigurations. Plans and Situated Actions* (2nd ed.). Cambridge University Press .

Suchman, L.A., & Weber, J. (2016). Human-Machine-Autonomies. In Bhuta, N. C., Beck, S., Geiß, R., Liu, H.-Y., & Kreß, C. (Eds.), *Autonomous Weapons Systems. Law, Ethics, Policy, Cambridge* (2nd ed., 75–102) Cambridge University Press.

Biography

Thomas Christian Bächle is Head of the Digital Society Research Programme at the Alexander von Humboldt Institute for Internet and Society (HIIG) in Berlin. He also works in a project at the University of Bonn as part of the research group "Meaningful Human Control", which studies autonomous weapon systems. From April 2022 to February 2023 he was managing editor of the open access and peer-reviewed journal Internet Policy Review to which he has since then continued to contribute as academic editor.

From January to October of 2020 he was a Fellow at the Cognitive Science Lab, Waseda University in Shinjuku, Tokyo. His research in Japan was funded by the Japanese Society for the Promotion of Science (JSPS). In 2019/20 he was Guest Professor for Media Studies/Digital Media at the Hermann von Helmholtz Centre for Cultural Techniques at the Humboldt University of Berlin.

In his Habilitation thesis, he is developing a media theory of the humanoid robot, comparing Japanese and European interpretations of artificial intelligence. Until 2019 he was led the research project "The Futures of Telemedicine: Knowledge, Policy, Regulation", which focused on regulatory challenges, norms and the social acceptance of eHealth practices and technologies.

His areas of research include cultural representations of identities, bodies and (media) technologies; human/machine interaction; technological materialities, interfaces and agency; mobile media, surveillance, robotics, affective computing and simulation technologies. In 2016 he published "Digitales Wissen, Daten und Überwachung zur Einführung" (Hamburg: Junius), an introduction to digital knowledge, data and surveillance.

16. Artificial Decision-Making on Life-or-Death: The Moral-Psychological Implications of Increasing Autonomy in Weapon Systems

Dr Tine Molendijk, Professor Lonneke Peperkamp, Sofie van der Maarel (NLDA)

The increasing autonomy of weapon systems has sparked intense debates about its ethical aspects, particularly regarding the morality of LAWS (lethal autonomous weapon systems) and the implications of granting machines the power to make life-or-death decisions. The most common arguments can be categorized into five distinct areas, and one of those area's centres around the moral obligations of states vis-a-vis their population. In this paper, we delve into an underexplored aspect of this area of debate: the moral-psychological implications for the combatants that would operate LAWS.

Proponents have recently justified the development and use of LAWS because of the supposed benefits for combatants that would operate LAWS. They invoke principles such as Bradley Strawser's principle of unnecessary risk (PUR), which asserts that states have an obligation to prevent unnecessary physical harm to their own troops. Since remote warfare essentially eliminates combatant risk of being injured or killed, states would have a *prima facie* duty to deploy remotely piloted armed drones. Today however, there is increasing evidence on the harmful psychological effects of this type of remote warfare. To defend LAWS, proponents extend PUR to include unnecessary psychological harm, and subsequently assume that the deployment of LAWS would reduce instances of moral injury and other adverse psychological effects among combatants by relieving them of life-or-death decisions. But how likely is it, that the shift from target selection by humans to autonomous machine target selection will reduce the risk of moral-psychological harm?

This paper aims to shed light on that question by integrating insights from research on both moral injury and autonomy, and by then taking the perspective of the 'lived experience' of combatants. Moral injury, a concept explored within military ethics, encompasses the psychological challenges that soldiers may encounter when their actions violate their moral beliefs and expectations. Central to moral injury are sentiments of powerlessness, responsibility, and (mis)trust, which are also fundamental themes in discussions surrounding autonomy. We take these sentiments to be three key factors that influence the moral-psychological well-being of future LAWS operators, and aim to further explore them in this paper.

Approaching these factors from the perspective of moral judgement understood as human meaning-making yields new insights for the LAWS debate, and therefore constitutes our theoretical framework.

a) Human control: Control is often considered in terms of power balance and rational calculations. However, the moral judgment perspective allows us to see it as a more nuanced and less calculated experience. It involves basic assumptions about living in a meaningful and coherent world, where actions and outcomes are connected, and the locus of control determines individual perceptions of having control over life events. An internal locus of control is often seen as protective against mental health issues, while an external locus of control is considered a risk factor. Among soldiers who have experienced trauma

during deployment, a sense of powerlessness is indeed frequently reported, but seeking external support in complex situations is reported as well.

b) Trust: While discussions on LAWS often frame trust in terms of risk calculation, it is more complex at the level of moral judgement and human meaning-making. Trust is not about certainty in outcomes but arises precisely because of the absence of such certainty. It is an emotionally and morally charged social expectation that involves embracing vulnerability in a relationship of dependence. Betrayal of trust can inflict immense pain, as it constitutes a violation of fundamental moral principles. Research indicates that individuals may prefer to face greater safety risks rather than subject themselves to the potential breach of trust, highlighting the aversion to betrayal.

c) Responsibility: responsibility tends to be approached in 'zero-sum' terms, focusing on calculations of accountability. However, from a moral judgment perspective, responsibility is about fulfilling one's duty, upholding values, and living up to moral commitments. Drawing from moral injury research, the concept of 'moral failure' refers to situations where a person may feel they failed morally even when they did nothing objectively wrong. This can lead to feelings of 'tragic remorse' and moral disengagement as coping mechanism.

Our theoretical framework shows that in current ethical discussions on LAWS, the concepts of control, responsibility and trust are generally approached from an abstract rational standpoint, relying on notions of power balance, zero-sum assessments and risk calculation. In contrast, in research on moral injury, these concepts are about a locus of control versus powerlessness or fatalism, upholding values versus moral failure or moral disengagement and trust versus betrayal or distrust. These key factors shape the empirical analysis that is central in our paper.

To answer our question – Would LAWS decrease or increase the risk of moral-psychological harm? – we shift focus from abstract thought experiments used in the LAWS debate to considering the real-world challenges faced by combatants. As there is a lack of empirical data to address this question directly, we conduct a comparative analysis by drawing on a relevant existing empirical study (by Rauch and Ansari 2022 on military personnel engaged in drone operations for the U.S. Air Force). In this way, we compare operators remotely piloting armed drones (the 'human in the loop') versus operators that would remotely supervise autonomous armed drones (the 'human on the loop'). We explain the main differences between these two roles and assess how the shift in responsibilities may impact moral-psychological harm, by organizing the results of this study according to our theoretical framework. This allows an assessment of how the three key factors play a role in the actual experience of combatants, and in relation to 'meaningful human control' as a normative question. This comparative analysis indicates that certain issues contributing to moral-psychological harm might persist (e.g., the digital closeness between combatants and targets, the 'two worlds phenomenon' and the dilemma of a defenceless enemy), some issues could be alleviated (such as the burden of life-or-death decisions), but new challenges are likely to emerge.

Since the debate on the morality of LAWS has mostly focussed on normative ethics, the specific psychological impact has not been a significant concern. Nonetheless, when desiring to comprehensively

understand the expected moral-psychological experience of LAWS operators, it seems that precisely this is necessary: to examine in depth how humans will experience and are affected by operating LAWS. This paper contributes to that understanding, and offers a novel perspective on the question whether increased autonomy in weapon systems would truly reduce moral-psychological harm experienced by operators. Whilst we can only provide preliminary answers, our analysis challenges assumptions about reducing moral-psychological harm through LAWS deployment, which can inform the LAWS debate, and enables us to better grasp the real-world ethical issues involved in LAWS.

Biographies

Tine Molendijk is an interdisciplinary-minded anthropologist – combining philosophy, psychology, anthropology and political sciences – specialized in the topics of violence, societal views on the armed forces, military culture, ethics and mental health, in particular post-traumatic stress disorder and moral injury. Currently, she works as assistant professor at the Faculty of Military Sciences at the NLDA and as research fellow affiliated to the Radboud University. At the NLDA, she is chair of the board of the Research Center of Military Management Sciences. She is also project leader of a large research project on contextual dimensions of moral injury (which was also the topic of her PhD research), including the development of interventions, funded by the Research Council (NWO). Further, she is editorial board member of Impact, member of the NATO research task group HFM329 and project advisor for several projects on military dilemmas and trauma. Among her publications in academic, professional and popular fora is the award-winning book Moral Injury and Soldiers in Conflict: Political Practices and Public Perceptions (Routledge, 2021). More information and publications can be found at: www.tinemolendijk.nl.

Lonneke Peperkamp is professor of military ethics and leadership at the NLDA. With a background in law, philosophy, and conflict studies, she focuses in her research on global justice, just war theory, and the ethics of new technologies in warfare, such as space technology, remote warfare, autonomous weapon systems, (cyber enabled) psychological influencing operations, and soldier enhancement. She is affiliated with the Interdisciplinary Research Hub on Digitalization and Society (iHub, Radboud University Nijmegen) and the Centre of Military Ethics (King's College London). She is vice president of EuroISME, the International Society for Military Ethics in Europe.

Sofie van der Maarel is a Ph.D. candidate in the Department of Political Science at Radboud University, Nijmegen, The Netherlands and The Netherlands Defense Academy (NLDA) in Breda. Sofie's ethnographic research focuses on visions and imaginaries of technological innovation in military and police organizations, and how this impacts the lived experience of security practitioners. Her work is grounded in science and technology studies, critical security studies, international relations, and anthropology, and she studies the relations between morality, innovation, security, and future imaginaries. More specifically, she focuses on the themes of moral injury, moral disorientation, perceptions of security, and algorithms, robotics, and information technologies. Her Ph.D. project is part of a large research project on contextual dimensions of moral injury, funded by the Dutch Research Council (NWO).

17. Proportionality, Intentions, and Human Agency

Dr Elad Uzan (University of Oxford)

Philosophers studying the ethics of war have acknowledged a distinction between narrow and wide proportionality. Narrow proportionality applies to harm inflicted on people who are morally liable to be harmed. Wide proportionality applies to harm inflicted on people who are not liable to suffer this harm.[1] However, the moral evaluation of proportionality in war depends on further crucial distinctions, such as the harm inflicted is intended or unintended harm, and whether the harm is inflicted as a means or as a side effect. These distinctions are crucial for moral judgements of proportionality in war.

For example, in war, if the harm inflicted is narrowly proportionate, it does not matter whether it is inflicted as an intended means or as an unintended side effect. However, when harm is narrowly disproportionate, it does seem to matter whether the harm inflicted beyond what the targeted person is liable for is intended or an unintended side effect. Conversely, it may be widely proportionate to kill one innocent person as a side effect of saving 10 innocent people, yet widely disproportionate to kill one innocent person as an intended means of saving 10 innocent people. The fact that the latter is *disproportionate*, rather than merely wrong, is demonstrated by a case in which the act becomes proportionate if the number of people saved were increased, say, to 100.

The distinctions between narrow and wide proportionality, between intended and unintended harm, and between harming as a means and harming as a side effect, are crucial to the permissibility of defensive force. Meanwhile, advanced weapons systems enhanced by AI technology challenge existing notions of human agency, which are at the heart of these distinctions, and thus the possibility of incorporating these distinctions into a process of moral decision-making in war by AI-controlled weapon systems.

In my talk I intend to discuss the relationship between these two distinctions and the moral responsibility gap created by outsourcing moral decisions to machines, and more generally by the trajectory of human-machine interaction in war, which illustrates the complex socio-technical interactions between AI and intentions in the ethics of war.[2]

I will argue that while existing AI technologies in war cannot fully consider the seriousness of these distinctions in the morality of war, we can use decision theory resources to narrow the gap between AI-controlled weapon systems and human agency. This strategy is related to the consequentializing literature, which explores when a seemingly non-consequentialist moral theory can be redefined as a consequentialist moral theory by appropriately modifying the characterization of morally relevant

[1] This distinction is due to Jeff McMahan, *Killing in War* (Oxford: Oxford University Press, 2009), 22–24. Many philosophers accept this distinction as useful and employ it themselves. See, for example, Cécile Fabre, *Cosmopolitan Peace* (Oxford: Oxford University Press, 2016); Victor Tadros, *The Ends of Harm: The Moral Foundations of Criminal Law* (Oxford: Oxford University Press, 2011).

[2] Such uses are becoming common. For example, the US DoD's 2022 Joint All-Domain Command and Control (JADC2) strategy report proposes integrating AI technology into command-and-control systems to speed up the "decision cycle", and to "employ automation and AI, rely upon a secure and resilient infrastructure, and act inside an adversary's decision cycle." See US Department of Defense 2022.

consequences. Consequentialism is a moral theory that evaluates the rightness or wrongness of actions based on their consequences, often aiming to maximize overall well-being or minimize harm. Non-consequentialist moral theories, on the other hand, focus on principles or rules that determine the morality of actions without directly considering their outcomes. Consequentializing seeks to bridge the gap between these two types of moral theories by reinterpreting non-consequentialist theories in a way that takes consequences into account. In this way, we can modify the non-consequentialist theory to account for the consequences of acts while still retaining the importance of non-consequentialist ethics.

Generally, decision-theoretic mathematical models describe costs and benefits by employing a broadly homogenous calculus. Such models do not have much force if they do not allow meaningful comparisons between different types of costs and benefits. Suppose a model assigns all acts that violate the categorical imperative a very negative value. Such a model does not allow a meaningful comparison between acts that violate the categorical imperative and acts that do not. It simply ranks all the former as morally worse than all the latter. As such, it is merely a mathematical way to express the categorical imperative and adds nothing to it.

However, I believe that, when properly formulated, decision theory can address at least some of the distinctions discussed above, thereby narrowing the moral responsibility gap created by outsourcing moral decisions to machines. For example, intentions can be taken to be one determinant of goodness or value. Imagine a utility function calculated as follows: each intentional killing of a civilian contributes -10 utility while each civilian killed as the foreseen side-effect of a military attack contributes -5. A decision-theoretic view will say that the intended death is morally worse than the unintended death while still allowing meaningful comparisons between intended and unintended killings.[3]

After arguing for the use of decision theory to narrow the gap between AI-controlled weapon systems and human agency, I will conclude by expressing doubts about whether, at this stage, AI-enabled warfare can completely replace human moral judgment and decision-making.

Biography

Elad Uzan is a Junior Research Fellow in Philosophy at Corpus Christi College and a Marie Curie Fellow in the Faculty of Philosophy, University of Oxford. He held visiting posts at Harvard Law School and Oxford's Faculty of Philosophy. He writes on topics at the intersection of moral, political, and legal philosophy. His current project deals with the moral constraints upon, and legal limits of, self-defense. Some areas of focus include justifications for and constraints on permissible harming, the ethics of risk-taking and risk-imposition in war, theories of harm aggregation in the context of war, and the influence of uncertainty on moral decision-making in war.

[3] For a general discussion, see Douglas Portmore, "Consequentializing," *Philosophy Compass* 4(2) (2009): 329–347. For the possibility of incorporating deontological approaches to the subjective permissibility of intended killing in war, see Seth Lazar, "In Dubious Battle: Uncertainty and the Ethics of Killing," *Philosophical Studies* (175) (2017): 859–883.

18. ELSI goes to War – Criticisms of Autonomous Weapon Systems and the Rise of a Responsible AI

Dr Jens Hälterlein and Professor Jutta Weber (Paderborn University)

The presentation will address the rise of responsible AI concepts in the military sphere and analyse this development following insights from the Sociology of Critique. Sociology of Critique highlights the role of critique as a social practice in transforming society (Boltanski & Chiapello 2005). The meaning of critique as social practice elaborated there, will be made fruitful in analysing the legitimation of autonomous weapon systems (AWS).

In the biopolitical logic of "killing to make life live" (Dillon & Reid 2009) wars conducted by liberal democracies require normative justification, especially when the loss of soldiers does not serve the defence of a nation's territory or the protection of its population - e.g. in the case of humanitarian interventions or the "war on terror". Based on this need for justification, the use of AWS is justified not least with the argument that it will reduce existential risks for soldiers and civilians. However, the use and development of AWS are criticized by various actors from politics, academia and civil society, who caution against the (potential) loss of human control, responsibility and accountability and call for a "Meaningful Human Control" over these weapons to be guaranteed as well as a corresponding ban on lethal AWS to be enforced.

While this criticism has not yet led to binding regulation at the international level and to that extent has not achieved its actual goal, it has by no means been ineffective. Cross-country surveys from 2017 and 2018 show that the majority of respondents reject the use of AWS. In 2018, over 200 technology companies and 3,000 individuals signed a public pledge of the *Future of Life Institute* to "not participate nor support the development, manufacture, trade, or use of lethal autonomous weapons." In 2018, numerous Google employees quit their jobs in protest of the company's planned cooperation with the Pentagon on the project MAVEN, which aimed to conduct AI-based analysis of drone video footage. In addition, about 4,000 of their colleagues signed an open letter to the company's CEO opposing this cooperation. They believed that humans, not algorithms, should be responsible for this potentially lethal analysis. In 2022, five robotics manufacturers, including renowned *Boston Dynamics*, signed a declaration that their increasingly autonomous robots would not be armed.

While these protests and actions have not yet led to a legally binding regulation at the international level or an industry-wide renunciation of military collaborations, it can be seen as a non-intended result of these criticisms that in military development contexts, voluntary commitments to the guiding principle of a "trustworthy, human-centred, and responsible AI" and value-based engineering concepts such as "ethics by design" and "Ethical Legal and Social Implications" (ELSI) research have been expressed.

As a result of the integration of (some of) the criticism of AWS into justification strategies, the implementation of soft regulatory tools could render the setting of hard law obsolete. We claim that this development is highly problematic, not only because it threatens to diminish the force of the call for a

binding legal regulation of AWS, but also because neither AI ethics nor ethics-by-design is sufficient to enable a "Meaningful Human Control" over AWS. We will bolster this claim with an analysis of the so-called Ethical AI Demonstrator (E-AID) that has been developed in the context of the multi-national project *Future Combat Air System* (FCAS).

The E-AID simulates the application of AI in FCAS based on the prototype of a decision support system. Simulations are run on scenarios closely coordinated with the German Army. The aim is to get a realistic picture of the possibilities, limits and ethical implications of AI in defence by means of concrete examples and to provide a first step towards an ethics-by-design-methodology which then can be integrated into an overall FCAS design process. Within the scenario "Find Fix Track Application with AI for Automated Target Recognition", the mission is the elimination of enemy air defence using drones equipped with sensors that collect data on positions of military equipment supporting the enemy air defence. The output of the automated target recognition is shown on a graphic interface highlighting relevant objects and providing basic context information (e.g., type of detected vehicle, certainty level). The guiding question in this scenario is how tasks can be delegated to an AI in an accelerated decision cycle without violating applicable military Rules of Engagement and ethical guidelines.

Considering the challenges and the functioning of an algorithmic decision support system like the one simulated in the E-AID, it is highly questionable whether individual responsibility and accountability can be given at all. If at all, these can be realized in manageable and unambiguous combat situations (for which the question of the added value of an algorithmic decision support system would arise). But even then, it would be questionable whether the output of a system can be explained in such a way that a human can actually understand it and can exercise meaningful human control based on this understanding.

Since an analysis of the E-AID demonstrates the impossibility of individual responsibility and accountability in the context of the use of AWS, it provokes several questions: Wouldn't accountability imply that a soldier categorically refuses to take a 'decision' based on an algorithmic decision support system? And shouldn't value-based engineering rather mean refusing to participate in the development of military technology at all, be it with or without an ethics-by-design approach? However, whether these individual forms of resistance would be enough to stop the development and use of AWS on a global scale is doubtful. Accordingly, the concept of "Meaningful Human Control" should be understood not as the normative ideal of an individual who is in control of the tools it uses, but as a call for collective action and the subordination of technological possibilities to democratic principles, which in the end can only mean the outlawing of AWS.

References

Boltanski, L. & Chiapello, E. (2005) *The New Spirit of Capitalism*, London, New York: Verso.

Dillon, M. & Reid, J. (2009) *The liberal way of war. Killing to make life live*, London: Routledge.

Biographies

Jens Hälterlein is a researcher at the Department of Media Studies at Paderborn University and scientific coordinator of the project "Meaningful Human Control – Autonomous Weapon Systems between Regulation and Reflection" (MEHUCO) funded by the German Federal Ministry of Education and Research. He has been publishing on the social dimension of security technologies since 2011 - among others within the framework of the EU project VideoSense. From 2020 to 2022 he was PI of the project "Artificial Intelligence and Civil Security" funded by the Fritz Thyssen Foundation. He has many years of research experience in the fields of Science & Technology Studies (STS), Surveillance Studies, Critical Security Studies and Governmentality Studies.

Hälterlein, J. *Facial Recognition in Law Enforcement*. In: Borch, C., Pardo-Guerra. J. P. (Eds.) The Oxford Handbook of the Sociology of Machine Learning, Oxford University Press (forthcoming).

Hälterlein, J. *The Use of AI in Domestic Security Practices*. In: Lindgren, S. (Ed.) Handbook of Critical Studies of Artificial Intelligence, Edward Elgar Publishing (forthcoming).

Hälterlein, J. (2021) Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. In: Big Data & Society, 8 (1).

Jutta Weber is Professor of Media Sociology at the Department of Media Studies at Paderborn University and PI of the project "Meaningful Human Control – Autonomous Weapon Systems between Regulation and Reflection" (MEHUCO) funded by the German Federal Ministry of Education and Research. She has been publishing on the social dimension of automated warfare since 2007 - among others within the framework of the EU project ETHICBOTS, the DFG Post Graduate Programme Automatisms and as a member of the International Committee for Robot Arms Control (ICRAC). She has many years of research experience in the fields of Science & Technology Studies (STS) or Innovation and Technology Analysis (ITA), Surveillance, Military & Critical Security Studies and Gender Studies.

Weber, J. (2021) *Artificial Intelligence and the Sociotechnical Imaginary: On Skynet, Self-Healing Swarms and Slaughterbots*, In: Kathrin Maurer, Andreas Immanuel Graae (Eds.) Drone Imaginaries. The Power of Remote Vision. Manchester: University of Manchester Press, pp. 167-179.

Weber, J. (2016): *Human-Machine Autonomies*, In: Bhuta et al. (Eds.) Autonomous Weapon Systems. Law, Ethics, Policy, Cambridge: Cambridge University Press, pp. 75-102 (together with L. Suchman).

Weber, J. (2016) *Keep Adding. Kill Lists, Drone Warfare and the Politics of Databases*, In: Environment and Planning D. Society and Space. Vol. 34(1), pp. 107-125.

19. “This is my Last Resort” - Overcoming the Stalemate in Autonomous Weapons Regulation through National Legislation and Industry Self-Regulation

Marcel Schliebs (University of Oxford) and Vanessa Vohs (Bundeswehr University Munich)

As artificial intelligence (AI) is poised to unleash its disruptive potential in nearly all spheres of life, militaries around the world are investing significant resources in research and development of military applications of AI. Among these are lethal autonomous weapons systems (LAWS), which “select and engage targets without further intervention by a [human] operator” (US Department of Defense Directive 3000.09 - Autonomy in Weapon Systems, 2023). As such, LAWS can range from already deployed systems such as autonomous anti-radar systems to more futuristic prospects of large autonomous drone swarms deployed in close urban combat. There are several reasons why some form of regulation of LAWS may be desirable. From an ethical perspective, removing human agency from lethal decisions may present violations of humanitarian principles, such as the erosion of moral judgment and human dignity in armed conflicts. On the legal side, selecting and applying force to targets without human intervention raises questions about accountability for the consequences of lethal action by autonomous systems. Furthermore, the widespread deployment of autonomous weapons might increase the risk of inadvertent escalation, while the opaqueness of algorithms and training data may lead to an arms race. Against this backdrop, based on an interdisciplinary approach combining the literatures on international relations, arms control, AI ethics, international law and computer science and taking into account the inherent properties of autonomous weapons, this paper derives what preconditions would have to be met in order to design and enforce different regulatory responses to LAWS.

Theoretical Framework

Historically, bi- or multilateral agreements to outlaw a type or class of weapons in its entirety have had two important preconditions: First, states need to share a mutual interest that neither of them develops or uses a certain weapon system, and second, the compliance with such a legally binding agreement must be transparent and verifiable for each party involved. Unfortunately, current geopolitics and the nature of AI algorithms make meeting these preconditions unlikely. In addition to Russia’s blatant disregard for honoring any rules of international law, the People’s Republic of China (PRC) has made AI a center of its medium-term strategy to catch up with the current military superiority of US (Ding, 2018; Kania, 2019), while the United States experiences challenges in keeping up with the “Four Battlegrounds” of (military) AI: data, compute, talent, and institutions (Scharre, 2023). Empirically, the relative perception of the atrociousness and military utility of a weapon has been an important predictor of its successful prohibition, which may explain why states have found agreement to negotiate legally binding bans on chemical and biological weapons which are similarly horrific but less strategically powerful than their nuclear counterparts. Given the enormous potential of autonomy to transform the battlefield of the future, it seems optimistic that all major military powers will agree on a complete ban of LAWS. In cases of weapon systems where no legally binding ban can be agreed on, states regularly resort to non-

proliferation regimes to at least prevent or slow down the spread to other state and non-state actors, such as the Nuclear Non-Proliferation Treaty (NPT) or the Missile Technology Control Regime (MTCR). Elements of such agreements could be mirrored for LAWS, as nuclear weapons and missile technology also heavily rely on dual-use technology. Similar to the NPT, where non-nuclear weapons states have unilaterally renounced their right to develop such systems in exchange for assistance with civil nuclear energy programs, leading powers could offer to assist smaller countries with civil AI research. However, the backbone of the (relative) success of the NPT over the past decades was its extensive verification regime administered by the IAEA, which seems less feasible for LAWS with black box algorithms and non-physical training data. Finally, the MTCR's Equipment, Software and Technology Annex (2023) offers potential precedents on export control for software components which could be translated into measures governing the flow of targeting algorithms and training data.

According to neo-realism (Waltz, 1979), there are two main drivers for instability and arms races: First, imbalance in power will lead front-runners to try to maintain their relative power advantage, while less capable competitors will strive to match the respective capabilities of more powerful rivals. Second, uncertainty about one's adversaries' actual capabilities will further drive states to improve their own position by investing in the development of more numerous and capable equipment. Unfortunately, once again, all these boxes seem to be ticked with regards to the role of autonomy in 21st century great power competition. Arguably, LAWS are even more problematic than nuclear weapons in this regard: As AI superiority will depend on the quality and size of training datasets and algorithms, capability increases are no longer contingent on the physical production of additional marginal units of heavy weaponry, leading to a non-linear progress curve where even incremental improvements in training data can translate into highly scalable improvements. Similarly, the non-rivalry of training data and its theoretical scalability across military alliances may fuel scepticism within adversaries who fear that a small increase in data quality will lead to near-exponential improvements of all one's adversaries. The risk of these dynamics leading to a global AI arms race is particularly high with deep learning algorithms, whose inherent technical set-up has its own arms race between a generator and a discriminator training each other embedded in the core of its functioning.

Where two or more states have a mutual interest in preventing the escalatory spiral of an arms race, they can agree on bi- or multilateral arms control agreements to avoid excessive spending as well as reduce the risk of inadvertent escalation (Schelling & Halperin, 1961). Historically, such agreements could prohibit a certain capability or property of a weapon system such as the range of nuclear-capable cruise missiles (Intermediate-Range Nuclear Forces Treaty, 1987) or amount of heavy armor deployed in a certain theatre (Treaty on Conventional Armed Forces in Europe, 1990). Transferred to LAWS, states parties could for example agree to outlaw a specific capability of LAWS, such as fully autonomous engagements of human targets without meaningful human control. However, such provisions may be hard to verify, as states could attempt to implement a hidden fully autonomous mode or secretly develop anti-personnel targeting with the help of dual-use facial recognition data. Furthermore, partial bans or unit-limits for military exercises, a confidence and security building mechanism embedded in the Vienna Document (Vienna Document on Confidence- and Security-Building Measures, 2011) are only partially applicable, since researchers have already managed to conduct reinforcement learning for helicopter flight using only simulated data (Ng et al., 2006). As such, there is a high risk that the black box nature of neural networks will lead to a "double

veil of uncertainty", where major powers are confronted by uncertainty not only about their enemies' but also their own relative capabilities. If there is one lesson to be drawn from the Cold War and recent demise of the INF Treaty, it is that meaningful verifiability and compliance monitoring is essential to the success of any arms control agreement. Unfortunately, the prospects for a comprehensive verification regime of potential autonomous weapons arms control appear grim, as states will be reluctant to open up the most sensitive parts of their algorithms and training data even to presumably independent inspectors, while adversaries will sceptically eye their rivals' potential dual-use break out potential.

Empirical Developments

The pessimistic expectations for LAWS regulations derived from our theoretical framework correspond with the development of the United Nations deliberations on how to effectively control LAWS. Said issue has been a matter of concern within the global community since at least 2014, when the discussions in the framework of the Convention on Certain Conventional Weapons (CCW) in Geneva started. However, since 2019, when state parties of the CCW reached a consensus on the related UN Guiding Principles, primarily reiterating existing international laws, there has been a lack of significant progress within the CCW. Russia's illegal war of aggression against Ukraine has further escalated tensions within the forum and subsequently diminished the likelihood of reaching a consensus. Hence, the remainder of this paper explores ways how responsible actors could respond to the prospect of an unregulated arms race towards higher levels of autonomy in lethal combat systems.

Frequently, the defence industry is seen as the actor that has to be tamed (Hoffberger-Pippan & Vohs, 2023). Empirically, the industry is already actively contributing to ethical deliberations, as could be observed at the Summit on Responsible AI in the Military Domain (REAIM) in The Hague this year. Due to the lack of international standards, the Federal Association of the German Security and Defense Industry (BDSV) and scientists from an institute of the Fraunhofer-Gesellschaft and the University of the Bundeswehr in Munich are in fact calling for domestic regulation in order to base military AI research on a foundation of societal acceptability and predictability (Brink, 2023). Procedurally legitimizing research on autonomy in weapon systems through comprehensive legislative self-regulation could further facilitate bringing domestic audiences on board, which is crucial to prevent science-fiction guided hysteria when it comes to AI in the defence sector. It is of crucial importance to differentiate between morally repugnant uses of AI, such as for autonomous targeting of human beings, and other forms of autonomy such as support tools providing combatants on the battlefield with real-time information and better situational awareness. It may therefore be smart for policymakers to set standards on military AI themselves and claim the first-mover advantage, similar to the EU's pioneer role with data protection regulation through GDPR and possibly in similar fashion with the AI Act: setting standards that de facto impact developments beyond the geographical regulatory scope instead of conceding the first-mover advantage to less reliable actors. Another advantage of adopting domestic military AI standards is to foster cooperation with like-minded states. Thus, domestic discussions about regulating AI in the military could be complemented by "coalition of the willing" cooperation in order to appropriately address moral, ethical and legal challenges posed by increased levels of autonomy in future warfare. In face of the deadlock in international arms control, such a dual-track effort appears more likely to succeed than other options.

One noteworthy effort in this regard is the EU's Assessment List for Trustworthy AI (European Commission - High Level Expert Group on Artificial Intelligence, 2020). The list is primarily designed for AI developers, but it can also support senior management or legal experts in complying with the seven non-binding criteria for dealing with trustworthy AI that guide the EU AI Act. Even though ALTAI was drafted for civilian AI applications, the Future Combat Air System (FCAS) Forum for ethical considerations has explicitly applied this method and concluded that for the use case "target detection, recognition and identification", about 90 percent of the questions on the criterion "transparency and accountability" are applicable (Azzano et al., 2021). Moreover, the EU-funded project AI4DEF is developing a similar approach by adapting ALTAI specifically to the military use cases at hand. Hence, ALTAI is one example of taking existing (civilian) AI standards and developing adapted principles for the military sector and could serve as a starting point to implement the principles not only in the deployment of autonomy in weapon systems, but also during the research and development stage.

Conclusion

Due to the pessimistic outlook towards reaching international consensus on the regulation of LAWS, we first argue that responsible military powers should complement their continued efforts for international frameworks with rigorous domestic regulation and 'coalition of the willing' cooperation in order to appropriately address moral, ethical, and legal challenges posed by increased levels of autonomy in future warfare through domestic legislation and appropriate adaptation of national export control regimes. Second, we argue that self-regulation of the defence industry is both feasible and desirable, for example by taking existing soft law standards such as the EU Assessment List for Trustworthy AI (ALTAI) from the civilian domain and applying ethical defence engineering standards. By embracing a dual-track approach consisting of domestic and inter-governmental 'coalition of the willing' cooperation as well as industry self-regulation, and profiting from synergies between the different approaches, this paper suggests, responsible innovator powers can not only incorporate ethical and moral principles into the design of weapons systems and their rules of application, but also get their domestic publics on board for the degree of innovation necessary to keep up with one's adversaries.

Bibliography

Azzano, M., Boria, S., Brunessaux, S., Carron, B., De Cacqueray, A., Gloeden, S., Keisinger, F., Krach, B., & Mohrdieck, S. (2021). *The Responsible Use of Artificial Intelligence in FCAS – An Initial Assessment.*

Brink, N. (2023). Rüstungsindustrie und Forschung fordern nationale militärische KI-Strategie. *Frankfurter Rundschau.* https://www.fr.de/politik/ruestungsindustrie-forschung-ki-kuenstliche-intelligenz-waffen-strategie-tbl-zr-92366213.html

Ding, J. (2018). Deciphering China's AI dream. *Future of Humanity Institute Technical Report.*

European Commission - High Level Expert Group on Artificial Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.* https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Hoffberger-Pippan, E., & Vohs, V. (2023). Taming the Lions: The Role of Industry in the Debate on Autonomous Weapon Systems (AWS). *OpinioJuris*. http://opiniojuris.org/2023/04/25/taming-the-lions-the-role-of-industry-in-the-debate-on-autonomous-weapon-systems-aws/

Intermediate-Range Nuclear Forces Treaty, (1987). https://2009-2017.state.gov/t/avc/trty/102360.htm

Kania, E. B. (2019). Chinese military innovation in the AI revolution. *The RUSI Journal*, 164(5–6), 26–34.

MTCR: Equipment, Software and Technology Annex, (2023). https://mtcr.info/wordpress/wp-content/uploads/2023/08/MTCR-TEM-Technical_Annex_2023-06-15-PDF.pdf

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., & Liang, E. (2006). *Autonomous inverted helicopter flight via reinforcement learning*. 363–372.

Scharre, P. (2023). *Four Battlegrounds: Power in the Age of Artificial Intelligence*. WW Norton.

Schelling, T. C., & Halperin, M. H. (1961). *Strategy and Arms Control*. Twentieth Century Fund. https://books.google.de/books?id=hbwGAAAAMAAJ

Treaty on Conventional Armed Forces in Europe, (1990). https://treaties.unoda.org/t/cfe

US Department of Defense Directive 3000.09—Autonomy in Weapon Systems, (2023). https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Vienna Document on Confidence- and Security-Building Measures, (2011). https://www.osce.org/files/f/documents/a/4/86597.pdf

Waltz, K. N. (1979). *Theory of International Politics*. Random House.

Biographies

Marcel Schliebs is a doctoral candidate in Social Data Science at the University of Oxford and Researcher on the Programme on Democracy & Technology at the Oxford Internet Institute. His research is located at the intersection of political science, statistics, and data science, and focuses on the impact of authoritarian state-backed information operations on democratic societies, as well as the role of artificial intelligence for 21st century great power competition. Besides his academic role, he advises Western governments on counter-disinformation, and previously served at NATO Headquarters' Arms Control and Weapons of Mass Destruction Non-Proliferation Centre.

Vanessa Vohs is a research associate and doctoral candidate at the University of the Bundeswehr Munich (Chair Prof Dr Carlo Masala) as well as an associated member at the Institute for International Law of Peace and Armed Conflict Law (IFHV) Bochum. Ms Vohs holds an LL.M. in International Law from the London School of Economics (LSE) and a B.A. in International Relations as well as an additional certificate in Latin American studies from TU Dresden and Hebrew University Jerusalem. She works on regulatory frameworks of autonomy in weapon systems and is responsible for legal and ethical challenges within the EU-funded AI4DEF project. Her doctoral research focuses on the strategic use of law in German parliamentary debates on Bundeswehr deployments. Previously, Vanessa Vohs worked as a Research Assistant at the German Institute for International and Security Affairs (SWP) and at the House of Commons UK. She is also the podcast host of "UNrecht" from the United Nations Association (UNA) Germany, the beginner's podcast for international law and international politics, as well as a board member of UNA Germany.

20. ‘Cyber-ing’ the AI Regime: Path Dependencies and Pathologies to Avoid in the Creation of Rules for Military AI

Dr Arun Sukumar (Leiden University)

The institutional framework and tools of compliance developed by states in the near-term to promote the responsible use of AI for military applications will determine the scope of rules to govern such applications in the long-term. The deterioration of relations between the United States, Russia, and China, and past scepticism of states to create binding instruments around emerging technologies such as digital technologies and Lethal Autonomous Weapons Systems indicate that a treaty on AI and international security is some distance away. Several proposals have already been floated by states and non-state actors on how best to advance responsible state behaviour for military AI in the interim. Unsurprisingly, some of these proposals draw from international cybersecurity governance, both in relation to mechanisms of institutional cooperation, as well as the outcomes of such cooperation.

This paper identifies two key features of cybersecurity governance that have also been mooted by prominent stakeholders in the military AI debate– namely, the use of *ad hoc* institutions to advance responsible use of AI, and the articulation of substantive elements of such responsible behaviour through voluntary, non-binding norms. Both proposals are understandable, as flexible institutions and non-binding norms allow for states to develop shared understandings of responsible (and irresponsible) behaviour without committing to concrete or immediate action. However, this paper argues that adopting the ‘cyber’ template for AI could result in the pathologies of the cybersecurity regime also colouring the military AI regime. Specifically, I highlight the lack of clarity around the application and interpretation of existing rules to cyberspace, which has been attributed by several legal scholars to *ad hoc* institutions conflating binding rules and non-binding norms, with aspirational norms sometimes even borrowing the vocabulary of existing international law.

The paper will aim to accomplish two objectives. First, it will review proposals by major stakeholders calling for AI applications to be governed in the same manner as international cybersecurity. Following the Responsible Use of AI in the Military Domain (REAIM) conference co-hosted by the Netherlands and South Korea in February 2023, the United States unveiled a “political declaration”, laying out twelve “non-legally binding principles” for states to consider in the deployment of military AI. The US proposal mirrors the approach taken by *ad hoc* bodies on cybersecurity such as the UN Group of Governmental Experts (GGE) to articulate non-binding norms that would not only reflect aspirational political commitments but also clarify the scope of applicable international law to digital technologies. Similarly, the REAIM summit sought to establish a Global Commission on AI, akin to the Global Commission on the Stability of Cyberspace (GCSC), which was also supported by the Dutch government. Multistakeholder ad hoc bodies such as the GCSC have also articulated non-binding norms on cybersecurity, while shaping policymakers’ perceptions on the lawfulness of certain cyber operations or the targeting of specific critical information infrastructure.

Second, the paper will review how the reliance on ad hoc entities and norms has led to a lack of clarity in understanding how international law applies to cyberspace, and argue that the 'cyber' approach is already muddying the waters in AI. For instance, one of the twelve "non-binding" principles on AI articulated by the US called on states to adhere to existing legal obligations, including under international humanitarian law. The implication that adherence to obligations is optional does little to advance discussions on how existing international law governs responsible state behaviour in AI. The larger objective of the paper is not to dismiss the utility of informal AI governance – ad hoc institutions, voluntary guidelines, etc – but to highlight pitfalls of regime-creation led or driven primarily by informality, drawing on lessons learned from the international cybersecurity regime.

The paper will be divided into five sections:

Section I: Introduction

Section II: This section will place developments on international AI governance in context, highlighting the creation of intergovernmental and multistakeholder bodies to articulate norms, political frameworks, or other non-binding guidelines as part of a larger trend towards informal global governance.

Section III: The second section will review and present the evolving 'regime complex' of institutions involved in AI and international security. This section will highlight the informal character of governance mechanisms (existing and proposed) and compare it with the "pervasive informality" (Sukumar, Broeders and Kello, forthcoming) of the international cybersecurity regime. The following institutions will be reviewed:

- Global Commission on AI (proposed)
- Global AI safety summit, London (proposed, November 2023)
- REAIM conference, The Hague (February 2023)
- UN Secretary-General's Advisory Body on Global AI Cooperation (nominations solicited)
- Security Council Debate on AI (May 2023)
- UNESCO's Recommendation on the Ethics of AI (published 2022)

This section will focus on the reasons behind the creation of these entities, the main players involved, and the specific themes addressed by these entities relevant to the military use of AI.

Section IV: Following a review of the prominent entities involved in AI/international security discussions, the paper will analyse tools of governance employed by these institutions. The proliferation of ad hoc entities, the paper will argue, has also resulted in the development of non-binding guidelines, such as norms of "responsible behaviour", some of which refer directly and/or obliquely to international law. As with international cybersecurity governance, the use of legal vocabulary in non-binding, political frameworks is a strategy deployed by states to elevate the prominence of informal institutions, and to

ensure that these guidelines hew close to established legal principles around state responsibility for military operations. However, if the cybersecurity domain is any indication, the use of international law concepts in aspirational statements could also result in ambiguity around lawful uses of AI, especially in the military context. By implying that some established legal principles are optional or nonbinding, these norms of behaviour complicate compliance more than they induce it. For instance, the application of international humanitarian law to cyber operations was contested primarily on account of Chinese and Russian reluctance to include it within the UN GGE's non-binding reports. Adopting the 'cyber' template could see AI norms and norms initiatives, whether intergovernmental or multistakeholder bodies, become sites that determine the application existing international law. In of itself, this is not a problem, but as Section II will highlight, AI initiatives are policy-oriented, and not venues where state practice or general principles are identified doctrinally or empirically. Secondly, as the AI regime becomes "institutionally dense" and fragmented along political or geoeconomics lines, norms could likely compete or contradict each other, further limiting prospects for universal understandings around international law to emerge.

Section V: Conclusion

References

Abbott, K. W., & Faude, B. (2021). Choosing low-cost institutions in global governance. *International Theory*, *13*(3), 397–426. https://doi.org/10.1017/S1752971920000202

Alter, K. J. (2022). The promise and perils of theorizing international regime complexity in an evolving world. *The Review of International Organizations*, *17*(2), 375–396. https://doi.org/10.1007/s11558-021-09448-8

Bradley, C., Goldsmith, J., & Hathaway, O. A. (2023). The Rise of Nonbinding International Agreements: An Empirical, Comparative, and Normative Analysis. *University of Chicago Law Review*, 90. https://live-chicago-law-review.pantheonsite.io/sites/default/files/2023-09/01_Bradley_ART_Final.pdf

Haftel, Y. Z., & Lenz, T. (2022). Measuring institutional overlap in global governance. *The Review of International Organizations*, *17*(2), 323–347. https://doi.org/10.1007/s11558-021-09415-3

Sukumar, A., Broeders, D., & Kello, M. (*forthcoming*). The pervasive informality of the international cyber security regime: Geopolitics, non-state actors and diplomacy.

Westerwinter, O., Abbott, K. W., & Biersteker, T. (2021). Informal governance in world politics. *The Review of International Organizations*, *16*(1), 1–27. https://doi.org/10.1007/s11558-020-09382-1

Biography

Arun Sukumar is a post-doctoral research fellow at The Hague Program on International Cyber Security, Leiden University. He holds a PhD from The Fletcher School, Tufts University. Arun is a co-editor of Multistakeholder Diplomacy: Building an International Cybersecurity Regime (Edward Elgar Publishing, 2023) and Responsible Behaviour in Cyberspace: Global narratives and practice (EU Publications Office, 2023), and the author of Midnight's Machines: A Political History of Technology in India (Penguin Random House India, 2019).

21. Human Rights Due Diligence (HRDD) Framework Suitability to Military AI? Opportunity and Limitations

Yael Vias Gvirsman (Reichman University)

Defining an adequate normative framework for the growing design and use of AI in military contexts ('Military AI') is both a need and a challenge for private and public stakeholders, from design to use of the technology. Questions such as the following need answering: What are the risks and what is at stake in Military AI? How can these risks be mitigated from design to use (and debriefing after or training before use)? WHO is accountable and WHEN for the design and use of Military AI? How does Military AI affect existing normative frameworks relating to military use from the use of force (jus ad bellum), the conduct of hostilities (LOAC/IHL), and perhaps even law enforcement (IHRL/ICL)? For instance, in applying AI in belligerent occupation or when applying AI in military contexts to classically law enforcement settings but during an armed conflict. This relates to detention, the treatment of the wounded and hors de combat, POWs, the dead, fair trials, and even arrest (relevant to occupation law).

This article examines the risk-based, Human Rights Due Diligence framework- applied to AI in civilian contexts- as a suitable framework in view of assessing risks, mitigating them, and applying a human-rights-by-design approach to Military AI, only with the needed adjustments to military normative frameworks- focusing on the interaction between International Human Rights Law and International Humanitarian Law- for instance by applying the Right to Life in armed conflict (e.g. HRC General Comment 36 (2018)- Article 6 on the right to life in application of the ICJ Advisory Opinion on Nuclear Weapons (1996)).

There are two main fundamental challenges to defining a legal framework applicable to military AI and tailoring a relevant Human Rights/IHL Due Diligence mechanism, as follows.

At the outset, a preliminary challenge is to define what 'military AI' is. Beyond doubt, military AI is AI applied wherever the Laws of Armed Conflict (LOAC) apply, this means in armed conflict and in situations of belligerent occupation. However, does 'military AI' include *exclusively* AI applied in armed conflict and the conduct of hostilities only? Does AI in the preparation of the use of force also constitute 'military AI'? What about after the end of conflict or of occupation? Another possible distinction entails asking whether 'military AI' is defined by the military *nature of the acts or context* it is applied in (the substantive test) or by the *actor* applying it (the actor test). In this case, when a military entity serves for instance to deliver humanitarian relief in a disaster and applies AI, should this AI be considered as 'military AI'? Finally, what is the status of 'dual-use' goods, services, and knowledge? For the purpose of this article, 'military AI' is AI applied to 'military' acts by substance. i.e. the context or with a nexus to an armed conflict or the use of force. Nevertheless, a risk-based approach calls for caution. Therefore, in case of doubt, the legal framework offering better protection of fundamental human rights should apply. This rule of thumb would be particularly relevant to dual-use goods, services, and knowledge.

A second fundamental challenge to consider relates to uncertainties in the existing normative framework(s) applicable to AI in civilian contexts. The need for legal certainty exists of course in civilian

contexts, especially pertaining to fundamental human rights (e.g. surveillance technology), even more so in military contexts when human lives, livelihoods, and the environment are inherently at stake. What might be most relevant to compare are normative frameworks when using AI in (humanitarian) emergencies- as an additional interdisciplinary layer providing further insights into the adequate normative framework in military AI.

The approach proposed herein consists of presenting a Human Rights Due Diligence theoretical model while building an adjusted model based on IHL principles (I); This model bears the advantage of identifying risks at an early stage of design, therefore making mitigating risk all the more possible. The second stage consists of applying such a model to Military AI based on the different existing or potential use(s) in or in preparation for the battlefield (II). Both stages take into consideration existing or inherent disadvantages and risks (e.g. bias) AI withholds and how these should bear on decision-making and transparency. Disadvantages are taken into account while recognizing the potential AI has to save lives, alleviate suffering, increase parties to hostilities' compliance with the principles of humanity and military necessity, and incidentally, can be used to disseminate the Laws of Armed Conflict while training their troops.

In terms of structure, the paper will first examine the human rights due diligence (HRDD) main principles and why they may be suitable for Military AI (I); It will then proceed to apply adjustments to the model while taking into consideration principles and rules of IHL (II); Finally, it will examine the practical implementation of a HRDD model on different AI activities relating to different stages of military activity, before, during and after the conduct of hostilities (III). The paper will conclude on the suitability of the 'adjusted' HRDD model as well as its limitations (IV).

I. Human Rights due diligence suitability to Military AI
    a. What is a technology used in a ‘military context’?
        i. A context or nexus to armed conflict
        ii. Technology used by the military: civil or military use?
        iii. Dual-use technology
    b. UNGP Business and Human Rights: Obligations on States and Corporate Entities in peacetime *and in armed conflict*
    c. The Human Rights due diligence model as an opportunity for Military AI risk management
        i. Definition
        ii. Does not replace obligatory legal obligations
        iii. Provide a framework allowing to identify and mitigate risks based on self-reporting, internal and external implementation mechanisms as well as providing redress, complaints, and alert mechanisms for victims or relevant stakeholders
    d. Due Diligence process and supporting measures (OECD)
        i. Embed responsible conduct into policies and management systems
        ii. Identify and assess adverse impacts
            1. Operations
            2. Supply chains
            3. Contractual relationships: ‘sub-contractors’: such as military private military security companies (PMSCs)

4. Cease, Prevent or Mitigate adverse impacts
5. Track implementation and results
6. Communicate how impacts are addressed
7. Provide for contractual relationships/supply chain or remediate when appropriate
8. Training and dissemination under GCs

II. Adjustments to the HRDD model applicable to IHL
   a. The right to life under IHL
   b. Detention in armed conflict and under belligerent occupation
   c. Interaction between IHRL/IHL

III. An adapted HR/IHL Due Diligence model by activity
   a. The HR/IHL Due Diligence model applicable to AI technologies in weapons systems
   b. AI-driven decision support systems for intelligence
   c. Surveillance
   d. risk assessment
   e. detention operations
   f. target identification

IV. Conclusions: opportunity and limitation

Biography

Adv. Yael Vias Gvirsman is a Senior Researcher, managing the legal, societal, and humanitarian aspects of integrating advanced technology in pre-hospital response in complex and mass emergencies (Mass Casualty Incidents- MCIs and disasters). She is a consultant and attorney specializing in international (criminal, humanitarian) law. In 2014, she founded the International Criminal and Humanitarian Law Clinic, Harry Radzyner Law School, Reichman University.

With twenty years of progressively responsible experience, she has extensive experience working or supporting cases at International Criminal Courts (ICC, SCSL, ICTR, ICTY) and in domestic settings focusing on strategic litigation for human rights. Since 2020 she has been representing victims of international crimes against corporate entities for a US law firm. Her research and teaching spans the fields of ICL, IHL, IHRL, Transitions Justice, Business and Human Rights, and International Relations. She is a lecturer at the Harry Radzyner Law School, the Catholic University of Murcia (UCAM- Masters level), and a teaching fellow at the Law faculty international course for international criminal law at the Hebrew University of Jerusalem (2023-2024). She is a visiting professor and guest lecturer at Harvard University, University of Piedmonte Orientale- CRIMEDIM (Center for Research and Training in Disaster Medicine, Humanitarian Aid and Global Health). She is a consultant to IOs, ICCTs victim and defense teams, NGOs, EU entities and private corporations on their human rights and IHL obligations.