Program on International Law and Armed Conflict
HARVARD LAW SCHOOL

Exercising Cognitive Agency

A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict

Dustin A. Lewis and Hannah Sweeney

Legal Concept Paper • January 2025

Executive Summary

Most armed forces fighting most wars in most of the parts of the world do not rely on artificial intelligence (AI) to conduct military operations — at least not yet. Nonetheless, it is arguably warranted to consider the use of AI in armed conflict from an international legal perspective at this time. That is in part because some armed forces are already relying on AI-related technologies. Further, these technologies could entail potentially extensive implications not only for how wars are fought and whether people and parties can be held accountable for violations. The use of certain AI systems in war also concerns, more broadly, whether humans should rely on increasingly complex assemblages of sensors, data, algorithms, and machines in decisions that involve mortal endangerment. Unlike corporate codes of conduct, ethics guidelines, and domestic law, international law is the only framework that all States agree is binding in relation to all armed conflicts. Notably, however, there is no specific regime, provision, or rule of international law that expressly pertains to the use of AI in war. Instead, it is necessary to evaluate how the existing legal framework and related responsibility institutions may already regulate AI-related technologies in war.

In this legal concept paper, we seek to provide an analytical framework through which to understand some core issues related to respecting international law concerning the use of AI in armed conflict. Instead of isolating AI, we widen the lens to focus on intelligence and cognitive tasks more broadly. By drawing distinctions and similarities between exercises of natural intelligence by humans, on the one hand, and reliance on artificial intelligence by humans (and the entities they serve), on the other hand, we aim to help uncover part of what the current legal framework expects, assumes, and requires *of humans*.

In short, our understanding is that under the existing law it is assumed that, to administer the performance of obligations binding on States in relation to armed conflict, humans need to exercise what we term cognitive agency. More specifically, our analysis suggests that at least two premises underlie the performance of obligations in the principal field of international law applicable in armed conflict, namely international humanitarian law (IHL)/the law of armed conflict (LOAC). Those premises are that, arguably:

1. Only natural persons — that is, humans — are capable of administering the performance of IHL/LOAC obligations binding on States; and
2. In doing so, the humans concerned must exercise cognitive agency.

By cognitive agency, we mean — with respect to administering the performance of an IHL/LOAC obligation — the undertaking and carrying out of a conscientious and intentional operation of mind by one or more humans vested with State legal capacity, through which that person or those persons implement the execution of the cognitive tasks demanded by the obligation. We reason that these premises arguably reflect a specification or an instantiation of existing conditions of legality, not a new policy approach. We ground that assessment in analyses of assumptions about executing cognitive tasks in connection with war, how IHL/LOAC obligations are performed, and how certain rules of State responsibility operate. If these premises are well founded, they may entail significant consequences with respect to requirements and limits related to the use of AI in armed conflict.

The theoretical groundings form only part of the picture. To implement the identified conditions of legality, it is necessary to ascertain what it means in practice for humans to exercise cognitive agency in relation to each relevant obligation. To help illustrate what the first step in doing so might involve, we briefly explore two obligations under IHL/LOAC — one concerning proportionality in attacks, and another related to detaining civilians — and deduce respective sets of associated cognitive tasks.

Finally, with a view to clarifying what the existing law demands, permits, and prohibits, we formulate a set of guiding questions that States and other relevant stakeholders might consider forming positions on. The foundational questions raised by our inquiry are whether the humans responsible for administering the performance of an IHL/LOAC obligation binding on a State may rely on AI-related technologies in implementing the execution of one or more of the cognitive tasks demanded by the obligation — and, if so, under what circumstances and subject to what conditions can the humans concerned undertake and carry out the requisite conscientious and intentional operation of mind. By reflecting on their positions and publicly articulating their interpretations of how existing obligations may or must be performed, States and other stakeholders can contribute to a more precise and more stable understanding of how international law already regulates the (non-)use of AI-related technologies in armed conflict.

CREDITS

About HLS PILAC

The Harvard Law School Program on International Law and Armed Conflict (HLS PILAC) researches crucial challenges facing the various fields of public international law related to armed conflict. Its mode is critical, independent, and rigorous. HLS PILAC's methodology fuses traditional public international law research with targeted analysis of changing security environments.

About the Authors

Dustin A. Lewis is the Research Director of HLS PILAC. Hannah Sweeney is a graduate researcher at HLS PILAC.

Acknowledgments

The authors gratefully acknowledge Naz K. Modirzadeh, the Founding Director of HLS PILAC and a Professor of Practice at Harvard Law School, for assistance in conceptualizing and editing this paper. The authors also gratefully acknowledge Taha Wiheba, a Program Fellow of HLS PILAC, for providing editorial assistance and critical comments, as well as the following Harvard Law School students, all of whom were HLS PILAC Research Assistants at the time of their contribution to the research underlying this paper: Emma Davies, Eoin Jackson, Elizabeth Peartree, Emma Plankey, Tamar Ruseishvili, Elliot Serbin, Zoe Shamis, and Cecilia Wu. Comments from participants at the following events involving Dustin A. Lewis provided helpful reflections on the core ideas underlying the paper: a roundtable on "The Humanitarian Impact of New and Emerging Technologies of Warfare and the Legal Review of New Weapons" convened in Kyiv, Ukraine by the International Committee of the Red Cross on November 6, 2024; a panel titled "How to Ensure Responsible Use of AI in Military Decision-making—Lessons from the Autonomous Weapon Systems Debate" at the Summit on Responsible Artificial Intelligence in the Military Domain (REAIM Summit) in Seoul, South Korea on September 10, 2024; research briefings titled "New Technologies and Armed Conflict" at the Mission of Switzerland to the United Nations in Geneva on July 18, 2024 and at the Federal Department of Foreign Affairs of Switzerland in Bern on July 19, 2024; and a DILEMA-HILAC lecture titled "Natural and Artificial Intelligence in Armed Conflict: Exploring Settled and Open Legal Questions" at the Asser Institute in The Hague on July 11, 2024. HLS PILAC expresses its gratitude for the generous financial support for independent research provided by the Swiss Federal Department of Foreign Affairs.

Disclaimers

The views and opinions reflected in this paper are those solely of the authors, and the authors alone are responsible for any errors. The views expressed in this paper should not be taken, in any way, to reflect an official opinion of the Swiss Federal Department of Foreign Affairs.

License

Web

This paper is available free of charge at https://pilac.law.harvard.edu.

Correspondence

Correspondence concerning this paper may be sent to pilac@law.harvard.edu.

CONTENTS

1. Introduction

More than 120 armed conflicts are being waged around the world.[1] Reports suggest that, in some of these conflicts, a growing number of armed forces are relying on applications drawn from the science of artificial intelligence (AI).[2] While the vast majority of contemporary conflicts do not involve AI-related technologies, it is nevertheless arguably warranted to focus now on AI in armed conflict and, especially, on foundational international-law issues. That is in part because these applications span an increasingly wide and diverse set of functions and responsibilities that entail significant consequences for people, objects, and the natural environment affected by war. This is perhaps most clearly the case for the use of AI in relation to weapons and methods of warfare. But it also matters for other areas, such as detention, humanitarian services, and legal advice. Further, as with AI tools employed in situations of (relative) peace, the use of these technologies in armed conflict might carry significant implications for such issues as accountability, bias, and moral agency.[3]

It is arguably imperative to ground the growing multilateral discussion on military applications of AI in respect for international law. Unlike domestic law, corporate codes of conduct, or ethics guidelines, international law is the only framework that all States agree is binding in relation to all armed

[1] *See* International Committee of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflicts* (2024), at 6, https://shop.icrc.org/international-humanitarian-law-and-the-challenges-of-contemporary-armed-conflicts-building-a-culture-of-compliance-for-ihl-to-protect-humanity-in-today-s-and-future-conflicts-pdf-en.html. *See also* Geneva Academy, *Today's Armed Conflicts*, https://geneva-academy.ch/galleries/today-s-armed-conflicts (last visited Nov. 23, 2024) (reporting "more than 110" ongoing armed conflicts).

[2] *See, e.g.*, Nathan Strout, *Inside The Army's Futuristic Test Of Its Battlefield Artificial Intelligence In The Desert*, C4ISRNET (Sept. 25, 2020), https://www.c4isrnet.com/artificial-intelligence/2020/09/25/the-army-just-conducted-a-massive-test-of-its-battlefield-artificialintelligence-in-the-desert/; *Israel Claims 200 Attacks Predicted, Prevented With Data Tech*, CBS News (June 12, 2018), https://www.cbsnews.com/news/israel-dataalgorithms-predict-terrorism-palestinians-privacy-civil-liberties/ [hereinafter CBS News, 'Israel Claims Attacks Predicted']; Dustin A. Lewis, *AI and Machine Learning Symposium: Why Detention, Humanitarian Services, Maritime Systems, and Legal Advice Merit Greater Attention*, Opinio Juris (Apr. 28, 2020), http://opiniojuris.org/2020/04/28/ai-and-machine-learning-symposium-ai-in-armedconflict-why-detention-humanitarian-services-maritime-systems-and-legal-advice-merit-greater-attention/; Tess Bridgeman, *The Viability Of Data-Reliant Predictive Systems In Armed Conflict Detention*, ICRC Humanitarian L. & Policy Blog (Apr. 8, 2019), https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictivesystems-armed-conflict-detention/ [hereinafter Bridgeman, 'Viability of Data-Reliant Predictive Systems']; Ashley Deeks, *Detaining by Algorithm*, ICRC Humanitarian L. & Policy Blog (Mar. 25, 2019), https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/ [hereinafter Deeks, 'Detaining by Algorithm'].

[3] On bias, *see generally* Alexander Blanchard & Laura Bruun, *Bias in Military Artificial Intelligence*, SIPRI Background Paper (Stockholm International Peace Research Institute, Dec. 2024), https://www.sipri.org/sites/default/files/2024-12/background_paper_bias_in_military_ai_0.pdf.

conflicts. Yet, notably, there is no specific regime, principle, or rule of international law applicable in relation to armed conflict that pertains expressly to the use of AI. Therefore, it is necessary to evaluate how the existing legal framework and responsibility institutions may already regulate AI-related technologies in war.

In this legal concept paper, we seek to provide an analytical framework through which to understand some foundational issues related to respecting international law concerning the use of AI in armed conflict. Instead of isolating AI, we widen the lens to focus on intelligence and cognitive tasks in war more broadly. Much of the academic and policy attention in this area has addressed issues related to individual criminal responsibility for war crimes. While that is an important field to consider, we cast our attention on the obligations that States must perform and the institution of State responsibility.[4] In particular, we focus on obligations arising in the principal field of international law applicable in relation to armed conflict — often called international humanitarian law (IHL) or the law of armed conflict (LOAC) — and on State responsibility for violations of those obligations. (In short, IHL/LOAC provides a set of obligations that States and certain other subjects are bound to perform in respect of armed conflict. The institution of State responsibility establishes a framework through which to ascertain and implement the international responsibility of a State in case the State fails to perform an obligation binding on it.) Through this lens, we analyze certain conceptual foundations about how the performance of IHL/LOAC obligations and the operation of certain rules of State responsibility relate to cognitive tasks in armed conflict. By drawing distinctions and similarities between exercises of *natural* intelligence by humans, on the one hand, and reliance on *artificial* intelligence by humans (and the entities they serve), on the other hand, we aim to help uncover part of what the current legal framework expects, assumes, and requires of humans in relation to administering the performance of IHL/LOAC obligations and determining responsibility when they fail to do so.

Our core argument is that under the existing law it is assumed that, to administer the performance of IHL/LOAC obligations, humans need to exercise what we term cognitive agency. In particular, our analysis suggests that at least two premises underlie the performance of IHL/LOAC obligations. One

[4] *See generally* JAMES CRAWFORD, STATE RESPONSIBILITY: THE GENERAL PART (Cambridge Univ. Press, 2013). *See also* International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts, U.N. GAOR, 56th Sess., Supp. 10, U.N. Doc. A/56/10 (2001).

of those premises is that, arguably, only natural persons — that is, humans — are capable of administering the performance of IHL/LOAC obligations binding on States. The second premise is that, in doing so, the humans concerned, arguably, must exercise cognitive agency. By cognitive agency, we mean — with respect to administering the performance of an IHL/LOAC obligation — the undertaking and carrying out of a conscientious and intentional operation of mind by one or more humans vested with State legal capacity, through which that person or those persons implement the execution of the cognitive tasks demanded by the obligation.

We reason that these premises arguably reflect a specification or an instantiation of existing conditions of legality, not a new policy approach. We ground that assessment in analyses of assumptions about executing cognitive tasks in connection with war, how IHL/LOAC obligations are performed, and how certain rules of State responsibility operate.

If these premises are well founded, they may entail significant requirements or limits on the use of AI in armed conflict. Indeed, to ensure that one or more responsible humans exercise cognitive agency in administering the performance of a particular IHL/LOAC obligation, States may need to limit — or even prohibit — certain uses of AI in war. Yet while it asks whether those humans are exercising cognitive agency, this conceptual approach does not necessarily stipulate in general (that is, in relation to all IHL/LOAC obligations) the kind and degree of reliance the humans concerned may or may not place on AI-related technologies in administering the performance of the obligations.

These theoretical groundings form only part of the picture. To uphold respect for the law, concerned actors — such as members of the armed forces, legal advisers, data scientists, engineers, or others involved in developing and using AI in war — need to take measures to fully implement these conditions of legality. An important step in doing so is to ascertain what it means in practice for the humans concerned to exercise cognitive agency in relation to each relevant obligation. Such an evaluation would be required to determine whether — and, if so, under what circumstances and subject to what conditions — the humans responsible for administering the performance of an IHL/LOAC obligation binding on a State may rely on AI-related technologies. This paper is thus meant to help lend a conceptual vocabulary through which concerned actors can determine part of what it means to uphold respect for international law in practice.

Following this introduction, we first summarize current expectations and practices surrounding intelligence and cognitive tasks in armed conflict, linking natural and artificial intelligence (Section 2). We then present two premises that, we submit, arguably underlie the performance of IHL/LOAC obligations (Section 3). To implement those conditions of legality, it would be necessary to ascertain what is required for humans to exercise cognitive agency in relation to each relevant IHL/LOAC obligation. To help illustrate what the first step in doing so might involve, we briefly explore two such obligations — one concerning proportionality in attacks, and another related to detaining civilians — and deduce sets of associated cognitive tasks (Section 4). Finally, with a view to clarifying what the existing law demands, permits, and prohibits, we formulate questions that States and other stakeholders might consider forming positions on (Section 5). The foundational questions raised by our inquiry are whether the humans responsible for administering the performance of an IHL/LOAC obligation binding on a State may rely on AI-related technologies in implementing the execution of one or more of the cognitive tasks demanded by the obligation — and, if so, under what circumstances and subject to what conditions the humans concerned can undertake and carry out the requisite conscientious and intentional operation of mind.

Regarding methodology and caveats, we have primarily employed sources and methods of public international law as part of an effort to identify both the existing rules and the responsibility frameworks in which those rules are meant to operate. Further, one of the authors (Lewis) has acted as a participant-observer in multilateral debates on autonomous weapons and the use of AI in military domains, engaging with armed forces, humanitarians, diplomats, human-rights advocates, legal advisers, scientists, and engineers. With a team of research assistants, we have sought to learn about the science of AI, robotics, epistemology, and related fields, yet we are by no means experts in any of those domains. Similarly, also with a team of research assistants, we have examined some of the philosophical underpinnings of the law and legal systems, though, again, as non-experts. The bulk of the research that we have conducted was in English-language sources, while our research assistants conducted research also in French, Spanish, Chinese, and Russian. We do not, and, indeed, could not, claim to have comprehensively surveyed all potentially relevant law, practice, military doctrine, developments in technology, nor humanitarian and human-rights concerns.

2. Intelligence and Armed Conflict

In this section, we set out why we think it is useful to frame certain legal issues concerning the use of AI in armed conflict by addressing both natural intelligence and artificial intelligence.

2.1. Intelligence

Much of the existing literature on AI in military and other security applications focuses on particular applications of AI in isolation. Yet, arguably, there is value in framing certain legal issues in this area in terms of exercises of intelligence — including through executing cognitive tasks — more broadly. The wider frame helps to distinguish between exercises of natural intelligence by humans, on the one hand, and reliance by humans on artificial intelligence, on the other hand. That distinction, in turn, may entail insights regarding what it means to perform international legal obligations and determine responsibility in case of breach. This conceptual framework grounds the performance of IHL/LOAC obligations, first and foremost, in the roles and responsibilities of humans, not artificial or synthetic (in the sense of non-human) agents.[5]

None of these key notions — natural intelligence, artificial intelligence, or intelligence in a broader sense — is expressly defined in an international legal instrument binding in respect of armed conflict. Arguably, there is descriptive, analytical, and normative value in intelligence, in this area, not being construed narrowly in reference only to the collection of militarily valuable information. Rather, exercises of intelligence and their associated cognitive tasks, for the conceptual purposes here, might be said to encompass a diverse array of cognitive-related capabilities and functions. Drawing from cognitive science, intelligence could be said, among other definitions, to include the ability to comprehend, reason, and process information and to decide upon a course of action.[6] Thus, in referring to exercises of intelligence, we refer

[5] Moreover, this approach has been increasingly adopted by certain AI-focused research institutes on the basis that natural and artificial intelligence are often considered to be inherently interconnected in protean combinations. *See, e.g.*, Harvard University's Kempner Institute (last visited Nov. 10, 2024), https://kempnerinstitute.harvard.edu/; Columbia University's NSF AI Institute for Artificial and Natural Intelligence (last visited Nov. 10, 2024), https://arni-institute.org/; UC San Diego's Center for Engineered Natural Intelligence (last visited Nov. 10, 2024), https://ceni.ucsd.edu/; Universitat Pompeu Fabra's Center for Studies in Artificial and Natural Intelligence (last visited Nov. 10, 2024), https://www.upf.edu/web/cesani.

[6] *See generally* José Luis Bermúdez, Cognitive Science: An Introduction to the Science of the Mind (4th ed., Cambridge Univ. Press, 2022) [hereinafter Bermúdez, Cognitive Science].

broadly to the cognitive functions necessary to understand, interpret, and make a decision based upon information. In the rest of this section, we briefly explore definitions of natural intelligence and artificial intelligence, including their applications in relation to armed conflict.

2.2. Natural Intelligence in Armed Conflict

For the purposes of this paper, by natural intelligence, we mean the cognitive and neural capacities of humans, including for abstract thinking, problem-solving, learning, and adapting to changing environments.[7] Characterized in part by its integrative cognitive functions — such as perception, attention, memory, language, and planning — natural intelligence has been viewed by certain scientists as an inherent property of the human mind or brain that distinguishes humans from other species in certain important respects.[8] The study of natural intelligence spans many fields, including, among others, neuroscience, cognitive science, philosophy, linguistics, computer science, psychology, and anthropology.[9] At its core, conceptualizing and evaluating natural intelligence is a fundamentally interdisciplinary activity.

The significance of natural intelligence in the context of armed conflict lies in part in its functional capacities. For this paper, situational awareness and decision making by military actors warrant particular attention. Situational awareness might be said to involve a person's ability to accurately perceive elements of a situation, retrieve relevant memories or scripts for how to react to that situation, and, accordingly, decide on a suitable course of action.[10] In many framings in scientific literature, situational awareness is said to involve three levels: perceiving key features, comprehending those features, and accurately projecting likely outcomes in the near future.[11] The cognitive

[7] *See generally* Ruben Colom, Rex E. Jung, Richard J. Haier, & Sherif Karama, *Human Intelligence and Brain Networks*, 12 DIALOGUES CLIN. NEUROSCI. 489, 489–501 (2010); Natalia A. Goriounova & Huibert D. Mansvelder, *Genes, Cells and Brain Areas of Intelligence*, 13 FRONTIERS HUM. NEUROSCI. 44 (2019).

[8] *See, e.g.*, Jessica F. Cantlon & Steven T. Piantadosi, *Uniquely Human Intelligence Arose from Expanded Information Capacity*, 3 NAT. REVS. PSYCHOL. 275 (2024). *See also* Gilles E. Gignac & Eva T. Szodorai, *Defining Intelligence: Bridging the Gap Between Human and Artificial Perspectives*, 104 INTELLIGENCE 101832 (2024).

[9] *See* Bermúdez, COGNITIVE SCIENCE, *supra* note 6, at 2.

[10] *See* Christopher D. Wickens, *Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement*, 50 HUM. FACTORS 397, 397–403 (2008); Mica R. Endsley, *Measurement of Situation Awareness in Dynamic Systems*, 37 HUM. FACTORS 65, 65–84 (1995); Mica R. Endsley, *Situation Awareness: Operationally Necessary and Scientifically Grounded*, 17 COGNITION, TECH. & WORK 163, 163–67 (2015).

[11] *See* Michael D. Matthews, *Cognitive and Non-Cognitive Factors in Soldier Performance*, in THE OXFORD HANDBOOK OF MILITARY PSYCHOLOGY 30, 198 (Janice H. Laurence & Michael D. Matthews eds., Oxford Univ.
[Footnote continued on next page]

processes that underpin this awareness include attention, sensation, perception, working memory, and long-term memory.[12] Infantry-centric models of situational awareness, which are designed to describe the fundamental cognitive structures and processes that underlie military decision-making for soldiers, also take into account certain human factors. Those factors include, for example, the impact of sleep deprivation, loud noise, extreme physical demands, and the threat of severe bodily injury or death.[13] Such factors impact a person's ability to sense, interpret, and predict events on a battlefield. In other words, they affect one's ability to effectively exercise intelligence and execute cognitive tasks.

The exercise of natural intelligence in armed conflict is perhaps so omnipresent that some may take its existence for granted. Indeed, in situations of armed conflict, individuals, whether members of the armed forces or not, execute cognitive tasks near continuously. For example, the people involved in military chains of command determine which targets to attack, which methods and means of warfare to employ to meet mission objectives, and when it is warranted to retreat. Civilians in conflict zones also exercise intelligence — for example, when they decide whether to flee or seek shelter. Humanitarian actors exercise intelligence when they coordinate relief efforts, source supplies, and determine how to allocate resources. Each of these examples requires the perception of relevant information in one's environment, the cognitive processing of that information, and decision-making based on one's analysis.

2.3. Artificial Intelligence in Armed Conflict

For the purposes of this paper, by artificial intelligence, we mean, in short, the simulation of natural intelligence through constructed systems or (other) machines.[14] Underpinning many AI-related technologies is an effort to model,

Press, 2012); Neil D. Shortland, Laurence J. Alison & Joseph M. Moran, *Situation Awareness*, in CONFLICT: HOW SOLDIERS MAKE IMPOSSIBLE DECISIONS 45, 53–54 (Oxford Univ. Press, 2019); Jonas Lundberg, *Situation Awareness Systems, States and Processes: A Holistic Framework*, 16 THEORETICAL ISSUES IN ERGONOMICS SCI. 447 (2015).

[12] *See* Matthews, *supra* note 11, at 198 (citing John R. Anderson et al., *Theory of Sentence Memory as Part of a General Theory of Memory*, 45 J. MEMORY & LANGUAGE 337, 337–67 (2001); CHRISTOPHER D. WICKENS, ENGINEERING PSYCHOLOGY AND HUMAN PERFORMANCE (Glenview III, 1984)).

[13] *See* Mica R. Endsley, *Situation Models: An Avenue to the Modeling of Mental Models*, 44 PROC. HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING 61, 61–64 (2000).

[14] *See generally* Yongjun Xu et al., *Artificial Intelligence: A Powerful Paradigm for Scientific Research*, 2 [Footnote continued on next page]

simulate, and replicate — in part through computational processes — aspects of the cognitive structures and mechanisms underlying what is considered intelligent behavior.[15] Machine learning, as a subset of AI that a growing number of militaries are reportedly pursuing, has been characterized as involving computers algorithmically training from data and formulating outputs — often in the form of suggestions or predictions — without explicit programming.[16]

In the area of armed conflict, AI-related techniques and methods are increasingly being employed in connection with a wide range of activities. The kinds and degrees of reliance on AI-related technologies vary significantly, and it is infeasible to obtain a comprehensive understanding of the full scope of these technologies currently being utilized by armed forces. That is due in part to a "double black box" in which technical opacity is encased in military secrecy.[17] Despite this secrecy, publicly available and reported sources indicate that armed forces have implemented AI-related technologies to assist with decisions and activities related to reconnaissance and to destroying adversary munitions and installations.[18] AI-assisted technology has also been reportedly developed to help armed forces generate simulations and predictions to assess different courses of action that take into account battlefield conditions, logistical constraints, and

INNOVATION 100179 (2021); John Darzentas et al., *Artificial Intelligence: Theories, Models, and Applications: 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2–4, 2008: Proceedings* (1st ed., Springer, 2008).

[15] *See* Michail E. Klontzas et al., *Introduction to Artificial Intelligence* (1st ed., Springer, 2023).

[16] *See, e.g.*, Kelley M. Sayler, *Artificial Intelligence and National Security,* Congressional Research Service Report No. R45178 (Nov. 21, 2019), https://crsreports.congress.gov/product/pdf/R/R45178/7; International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (Aug. 2019), https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

[17] In contrast to the "black boxes" that function as flight data recorders (making recorded data accessible to authorities when needed), in computing, a "black box" refers to a system wherein the input and output data are known but the process by which the system turns the former into the latter cannot be seen. In this context, a double black box refers to two layers of opacity that limit public understanding of AI-usage in armed conflict. The first layer references general notions of military secrecy, and the second refers to the computing black box that makes it difficult, even for the programmers, to interpret or explain the outputs of systems with high levels of autonomy. On certain issues related to predicting and understanding military applications of artificial intelligence, *see* Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI*, UN INSTITUTE FOR DISARMAMENT RESEARCH (2020), https://unidir.org/publication/black-box-unlocked. Other authors have addressed similar concepts. *See, e.g.*, Ashley Deeks, *The Double Black Box: AI Inside the National Security Ecosystem*, JUST SECURITY (Dec. 7, 2024), https://www.justsecurity.org/98555/the-double-black-box-ai-inside-the-national-security-ecosystem/.

[18] See, e.g., Anastasia Roberts & Adrian Venables, The Role of Artificial Intelligence in Kinetic Targeting from the Perspective of International Humanitarian Law (CCDCOE, 2021); International Committee of the Red Cross, Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach (2019), https://www.icrc.org/sites/default/files/document_new/file_list/ai_and_machine_learning_in_armed_conflict-icrc.pdf [hereinafter ICRC, 'A Human-Centred Approach']; Maggie Gray & Amy Ertan, Artificial Intelligence and Autonomy in the Military: An Overview of NATO Member States' Strategies and Deployment (CCDCOE, 2021); Karel van den Bosch & Adelbert Bronkhorst, Human-AI Cooperation to Benefit Military Decision Making (NATO, 2018).

enemy movements.[19] Reports suggest that armed forces are developing — and, in certain cases, using — AI-related technologies (most often in the form of decision-support systems (DSS)) to assist in processes involving the identification, nomination, or selection of objects of attack.[20] Advanced algorithmic frameworks are reportedly also being considered for use in aspects of detention-related decisions, in particular to help decision-makers predict which actors pose threats to the security of the State.[21] For example, there is speculation that the Chinese and U.S. militaries are each developing criminal-justice-inspired algorithms potentially for use in armed conflicts to help them assess which actors are dangerous, where to allocate patrols, and whom to detain.[22] For its part, Israel has reportedly used large-data algorithms to search social-media posts for predictive information about prospective assailants, which has then been used for detention decisions.[23] Additionally, AI methods are being explored for use in relation to humanitarian services, such as for forecasting needs and allocating resources more efficiently.[24] For example, the Danish Refugee Council has tested a machine-learning model that is reportedly able to predict the number of displaced people, from one to three years into the future.[25]

[19] *See* Anna Nadibaidze, Ingvild Bode, & Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates*, Ctr. for War Stud., Univ. of S. Den. (Nov. 4, 2024), https://usercontent.one/wp/www.autonorms.eu/wp-content/uploads/2024/11/AI-DSS-report-WEB.pdf; Elsa Kania, *AI Weapons in China's Military Innovation*, Brookings Inst. (Apr. 2020), https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf; Merel Ekelhof & Giacomo Persi Paoli, *Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous System*, U.N. Inst. Disarmament Res. (2020), https://unidir.org/sites/default/files/2020-04/UNIDIR_Swarms_SinglePages_web.pdf.

[20] *See*, *e.g.*, Merel A.C. Ekelhof, *AI Is Changing the Battlefield, but Perhaps Not How You Think: An Analysis of the Operationalization of Targeting Law and the Increasing Use of AI in Military Operations*, in Research Handbook on Warfare and Artificial Intelligence 162 (Robin Geiß & Henning Lahmann eds., Edward Elgar, 2024); Yuval Abraham, *'Lavender': The AI machine directing Israel's bombing spree in Gaza*, +972 Magazine (Apr. 3, 2024), https://www.972mag.com/lavender-ai-israeli-army-gaza/; Yuval Abraham, *'A mass assassination factory': Inside Israel's calculated bombing of Gaza,* +972 Magazine (Nov. 30, 2023), https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/; Joshua Hughes, *The Law of Armed Conflict Issues Created by Programming Automatic Target Recognition Systems Using Deep Learning Methods*, 21 Y.B. Int'l Humanitarian L. 99 (2018).

[21] *See* Lorna McGregor, *The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings*, ICRC Humanitarian L. & Policy Blog (Mar. 28, 2019), https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings/; Bridgeman, 'Viability of Data-Reliant Predictive Systems' *supra* note 2; Deeks, 'Detaining by Algorithm', *supra* note 2.

[22] *See* Ashley S. Deeks, *Predicting Enemies*, 104 Va. L. Rev. 1529 (December 2018).

[23] *See* CBS News, 'Israel Claims Attacks Predicted' *supra* note 2; Orr Hirschauge & Hagar Shezaf, *How Israel Jails Palestinians Because They Fit the 'Terrorist Profile,'* Haaretz (May 30, 2017), https://www.haaretz.com/israel-news/2017-05-31/ty-article-magazine/.premium/israel-jails-palestinians-who-fit-terrorist-profile/0000017f-f85f-d044-adff-fbff5c8a0000.

[24] Ana Beduschi, *Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks*, 104 Int'l Rev. Red Cross 1149, 1168 (2022).

[25] Danish Refugee Council, Global Displacement Forecast 2024: Using data modelling to predict displacement crises (Mar. 2024), https://pro.drc.ngo/media/ivvjqetf/240313_global_displacement_forecast_report_2024_final.pdf.

Many of these developments arguably represent a potential transformation in how (aspects of) intelligence in war might be exercised. As we will explain in the next section, the existing IHL/LOAC framework seemingly developed on the assumption that exercises of intelligence and associated cognitive tasks — whether, for example, related to decisions on whom to target, whom to detain, and when and where to provide humanitarian assistance — would be undertaken and carried out by humans, whether acting individually or jointly. Reliance on AI apparently implicates what it means for humans to be involved in the performance of the parts of IHL/LOAC obligations that require executing cognitive tasks.

3. Two Premises Arguably Underlying the Performance of IHL/LOAC Obligations

One key way that international law is respected is when States perform the obligations binding on them. Conversely, a failure to perform a binding obligation constitutes a breach, and legal consequences arise in respect of unexcused breaches. In this section, we set out two premises that, we submit, underlie the performance of IHL/LOAC obligations. In doing so, we seek to link the preceding exploration of exercises of intelligence in war with a structured analysis of what the existing legal framework requires of humans. We submit two conditions of legality under the existing legal framework: (i) arguably, only natural persons are capable of administering the performance of IHL/LOAC obligations binding on States; and (ii), in doing so, the humans concerned must, arguably, exercise cognitive agency. In the rest of this section, we set out the legal and logical foundations concerning each asserted premise.

3.1. Argument #1: Only Natural Persons Are Capable of Administering the Performance of IHL/LOAC Obligations Binding on States

We assert that, arguably, only natural persons are capable of administering the performance of IHL/LOAC obligations binding on States. We reason that this premise arguably reflects a specification or an instantiation of existing conditions of legality, not a new policy approach. We ground that assessment in analyses of assumptions about executing cognitive tasks in connection with war, how IHL/LOAC obligations are performed, and how certain rules of State responsibility operate.

With respect to performing obligations, the logic is that a State vests the authority to exercise legal capacity, ultimately, only in natural persons acting on its behalf. In other words, as far as we are aware, under the existing international law applicable in armed conflict there is arguably no recognized notion of legal capacity that is ultimately ascribable to artificial or synthetic (in the sense of non-human) agents. Rather, under the extant law, legal agency — in the sense of the capacity to administer the performance of an IHL/LOAC obligation binding on a State — is arguably reserved only for humans. This assumption is based in part on a particular understanding of how States act: at the time that the conceptual roots of the existing international legal framework were planted, it was materially impossible for States to undertake and carry out conduct without natural persons materially acting on their behalf.[26] If this legal premise is (still) correct, the State is required to ensure that humans — in particular, those in whom the State has vested legal capacity to act on its behalf and thus those whose conduct is attributable to the State — administer the performance of the relevant IHL/LOAC obligations binding on it. Analytically "reverse-engineering" this approach suggests that it would be impossible for a State to perform a binding IHL/LOAC obligation by relying solely on an artificial or a synthetic (in the sense of non-human) legal agent, such as an AI-related technology. In other words, under this approach, a State cannot perform a binding IHL/LOAC obligation by relying solely on, for example, an AI system that produces behaviors and effects that are not (also) ascribable to one or more humans whose conduct is vested with State legal capacity.

With respect to responsibility, a failure by a State to perform a binding IHL/LOAC obligation (through the acts or omissions of the relevant human(s) vested with State legal capacity) constitutes a breach. Attribution thus serves as the mechanism by which the conduct of certain humans — such as military

[26] This understanding of legal entities is reflected in the Draft Articles on State Responsibility, which details the various methods of attribution for ascribing physical conduct by "agents of the State" to the State as a legal entity. *See* Draft Articles on State Responsibility, *supra* note 4 at commentary to art. 2, ¶ 5 ("For particular conduct to be characterized as an internationally wrongful act, it must first be attributable to the State. The State is a real organized entity, a legal person with full authority to act under international law. But to recognize this is not to deny the elementary fact that the State cannot act of itself. An 'act of the State' must involve some action or omission by a human being or group [of human beings]: 'States can act only by and through their agents and representatives.' The question is which persons should be considered as acting on behalf of the State, i.e. what constitutes an 'act of the State' for the purposes of State responsibility." (internally citing *German Settlers in Poland, Advisory Opinion*, 1923, P.C.I.J., Series B, No. 6, p. 22)).

commanders, political leaders, or others whose conduct is ascribable to the State — may legally entail the responsibility of the State. Further, an implication of analytically “reverse-engineering” this approach is that a use by a State of an AI-related technology in armed conflict involved in the performance of a binding IHL/LOAC obligation would need to be attributable to the State by ascription to one or more humans acting on its behalf. Otherwise, a “responsibility gap” may arise in the sense that conduct meant to be governed by IHL/LOAC would not be capable of being attributed to the incumbent legal subject.[27] In that case, responsibility could not be implemented because a core link in the chain of causality — in particular, attribution to the State — would be missing.

Positions set out by States on related themes and topics appear to support the fundamental logic and normative concerns underlying this approach. Those positions are laid down, for example, in a law-of-war manual,[28] jointly drafted working papers,[29] and political declarations on the responsible use of AI.[30] Alignment with the basic elements may also be detected in guiding

[27] *See* Dustin A. Lewis, *War Crimes Involving Autonomous Weapons: Responsibility, Liability and Accountability*, 21 J. INT’L CRIM. JUSTICE 965, 973 (Nov. 2023).

[28] *See* Office of the Gen. Counsel of the Dep’t of Def., *Department of Defense Law of War Manual* (June 2015, updated July 2023), at § 6.5.9.3. (expressing the position that law-of-war obligations apply to persons rather than to weapons, including that “it is persons who must comply with the law of war”).

[29] *See ‘Towards a “Compliance-Based” Approach to LAWS [Lethal Autonomous Weapons Systems]’ Informal Meeting of Experts on Lethal Autonomous Weapons Systems,* Informal Working Paper by Switzerland to Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (Apr. 11–15 2016), ¶ 16 (expressing the position that “[t]he Geneva Conventions of 1949 and the Additional Protocols of 1977 were undoubtedly conceived with States and individual humans as agents for the exercise and implementation of the resulting rights and obligations in mind.”); *Draft Articles On Autonomous Weapon Systems – prohibitions and other regulatory measures on the basis of international humanitarian law (“IHL”),* Working Paper Submitted by Australia, Canada, Estonia, Japan, Latvia, Lithuania, Poland, the Republic of Korea, the United Kingdom, and the United States to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2024/WP.10 (Aug. 26, 2024), at 8, ¶ 3, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.10.pdf (expressing the view that “IHL imposes obligations on States, parties to armed conflict and individuals, not machines.” (citations omitted)); *Item 5 of the provisional agenda,* Working Paper Submitted by Bulgaria, Denmark, France, Germany, Italy, Luxembourg & Norway to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2024/WP.3 (Mar. 4, 2024), ¶¶ 10, 14, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.3.pdf (expressing the view that “States are responsible at all times for adhering to their obligations under applicable international law, including International Humanitarian Law. As such, States can be held responsible for internationally wrongful acts or violations of IHL resulting from the development or the use of the above-mentioned weapons” and reaffirming that “individual responsibility for violations of international law, specifically IHL, can never be transferred to machines.”).

[30] *See Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* (Nov. 9, 2023), https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/#:~:text=Launched%20in%20February%202023%20at,and%20use%20of%20military%20AI (The U.S. convened a plenary meeting of States endorsing the political declaration in March of 2024; as of November [Footnote continued on next page]

principles on emerging technologies in the area of lethal autonomous weapon systems,[31] a report of a Group of Governmental Experts on such technologies,[32] and statements in related multilateral debates.[33] Comments from the

2024, 52 States have endorsed the political declaration, indicating that AI use in military operations must remain accountable through a responsible human chain of command).

[31] See *Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Final Report*, U.N. Doc. CCW/MSP/2019/9, at Annex III, § (b) (Dec. 13, 2019) (noting that "human responsibility for decisions on the use of weapon systems must be retained, since accountability cannot be transferred to machines.").

[32] *See Report of the 2022 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, U.N. Doc. CCW/GGE.1/2022/2 (2022), ¶ 19 ("For the purposes of its work, the Group recognized that every internationally wrongful act of a State, including those potentially involving weapons systems based on emerging technologies in the area of LAWS entails international responsibility of that State, in accordance with international law. In addition, States must comply with international humanitarian law. Humans responsible for the planning and conducting of attacks must comply with international humanitarian law.").

[33] See *On Agenda Item 5(b) Further consideration of the human element in the use of lethal force; aspects of human machine interaction in the development, deployment and use of emerging technologies in the area of lethal autonomous weapons systems*, Statement by Germany at the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS) (Mar. 26, 2019), ¶ 4, https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/20190326%2BStatement3%2BGermany%2BGGE%2BLAWS.pdf (expressing the position that "[a]ccountability can only be assured as long as humans retain sufficient control over the critical functions of the weapons they operate."); *"Human element" - Agenda Item 5 (d)*, Statement by Brazil at the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS) (Mar. 24, 2019), at 1–2, https://meetings.unoda.org/meeting/29752/statements?f%5B0%5D=author_statements_%3ABrazil (expressing the view that "[t]he human element is what binds autonomous systems to Humanitarian International Law [HIL], since it provides a subject for accountability. In other words, HIL is only applicable to LAWS as long as there is someone to be held accountable. Therefore, the human element is not only the essential concept in understanding the limits and challenges of weapons systems with autonomous functions, but also the element that ensure its compliance to existing norms."); *Agenda item 5(a): an exploration of the potential challenges posed by emerging technologies in the area of Lethal Autonomous Weapons Systems to International Humanitarian Law*, Statement by the United Kingdom at the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS) (Mar. 25-29, 2019), at 1–2, https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/20190318-5%28a%29_IHL_Statement.pdf (stating that "[i]t is the UK[']s view that accountability can never be delegated to a machine or system; should a violation of IHL result from the operation of a weapon or weapon system, processes are already in place to conduct appropriate investigations and, if applicable, apportion responsibility. Legal accountability will always devolve to a human being, never a machine . . . ."); *Approaches of the Russian Federation to the Issue of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Submitted by the Russian Federation at the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS)*, CCW/GGE.1/2024/WP.2 (May 14, 2024), ¶ 19, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.2_English.pdf (stating that "States and individuals (including developers and manufacturers) at any time bear responsibility in accordance with international law for their decisions to develop and use emerging LAWS technologies. We believe that responsibility for the use of such systems rests with the official who assigns a mission for them and orders their use."); *Elements of an international legal instrument on Lethal Autonomous Weapons Systems (LAWS)*, Submitted by Pakistan at the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2024/WP.7 (May 14, 2024), ¶ 9, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.7.pdf (expressing the view that "[h]umans responsible for and in control of LAWS should at all times remain accountable for the [Footnote continued on next page]

International Committee of the Red Cross (ICRC) appear to be consonant with this approach as well.[34]

Counterarguments may include that the asserted premise is incorrect, in part because there is insufficient evidence to establish that extant international law assumes that legal agency — in the sense of the capacity to administer the performance of IHL/LOAC obligations binding on a State — is reserved exclusively to natural persons. Additionally, one could argue that even if the presumption may have been accurate previously, it should not necessarily continue to obtain in an era where, in theory, artificial or synthetic (in the sense of non-human) agents might be capable of materially executing some or all of the cognitive tasks demanded by such obligations, possibly (at least for some cognitive tasks) at a standard of performance higher than humans. That argument, however, collapses two distinct but related elements: the administration of the performance of an IHL/LOAC obligation, and the execution of certain cognitive tasks demanded by the obligation. Our research did not uncover an example of a State expressly subscribing to the view that artificial or synthetic (in the sense of non-human) agents are capable of administering the performance of IHL/LOAC obligations binding on States.[35] Furthermore, we were unable to

consequences of using such weapons, in line with international law and the applicable provisions on the Responsibility of States for Internationally Wrongful Acts."); *Working paper*, Submitted by Japan at the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2024/WP.8 (July 24, 2024), ¶ 2, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.8.pdf (stating that "in the use of weapon systems, human responsibility cannot be transferred to machines, and we must ensure that they are operated under a responsible chain of human command and control in a manner consistent with the obligations of states under IHL, and that responsibility is clearly attributed.").

[34] ICRC, 'A Human-Centred Approach' *supra* note 18, at 7 (expressing the position that "[i]t is humans that comply with and implement the law, and it is humans who will be held accountable for violations. In particular, combatants have a unique obligation to make the judgements required of them by the international humanitarian law rules governing the conduct of hostilities, and this responsibility cannot be transferred to a machine, a piece of software or an algorithm . . . .").

[35] *See* notes 29–33. Beyond the context of IHL/LOAC, our research could not identify any instances where artificial intelligence was afforded legal agency or personhood in a legal proceeding. There are, however, some examples of granting (special) legal status to non-humans, including corporations, rivers, the environment, and animals. *See e.g.*, Kevin Crow, International Corporate Personhood: Business and the Bodyless in International Law (1st ed. 2021); Moe Nakazora, *Environmental Law with Non-Human Features in India: Giving Legal Personhood to the Ganges*, 43 S. Asia Res. 172, 172–91 (2023); Francine Rochford, Environmental Personhood: New Trajectories in Law (2024); *Diego Alberto Ledesma, Chamber 1 - Animal Protection Act, Abuse or Acts of Cruelty*, IPP 149744/2022-0, Court of First Instance in Criminal, Juvenile, Felony, and Misdemeanor Matters No. 3 (2022) (Arg.), English translation available at https://www.nonhumanrights.org/wp-content/uploads/English-Translation-of-Decision-in-Argentine-Puma-Case.pdf (finding that a puma is a legal subject of rights, and thus capable of being a victim of acts of cruelty). These are arguably *sui generis* examples, and our research was unable to uncover any examples where AI-assisted technologies were ascribed legal rights, capacities, or duties. Some AI-assisted technologies have been reportedly employed in the [Footnote continued on next page]

identify any instances in which artificial or synthetic legal agency (in a relevant sense) purportedly arising in other contexts had been transposed into the international legal framework of armed conflict. To be certain, it is important to acknowledge that a lack of such evidence does not necessarily amount to an implied acceptance of this premise, not least in the current era of AI systems with increasingly extensive capabilities. Yet from a legal perspective, the status quo may arguably hold unless and until there is sufficient evidence that States are deliberately moving away from this principle by attempting to modify the law through one or more recognized methods.

3.2. Argument #2: In Doing So, the Humans Concerned Must Exercise Cognitive Agency

Our next argument builds on the preceding premise that legal agency — in the sense of the capacity to administer the performance of an IHL/LOAC obligation binding on a State — is exclusively reserved to humans. For the sake

U.S. domestic context, including with respect, for example, to assessing home-loan eligibility, denying insurance claims, and assisting with hiring decisions, among other things. Thus far, these fields reflect a similar reliance upon human attribution for legal responsibility. *See*, *e.g.*, U.S. Dep't of Justice, Justice Department Files Statement of Interest in Fair Housing Act Case Alleging Unlawful Algorithm-Based Tenant Screening Practices (Jan. 9, 2023), https://www.justice.gov/opa/pr/justice-department-files-statement-interest-fair-housing-act-case-alleging-unlawful-algorithm (stating that "[h]ousing providers and tenant screening companies that use algorithms and data to screen tenants are not absolved from liability when their practices disproportionately deny people of color access to fair housing opportunities"); Consumer Financial Protection Bureau, CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence, (Sept. 19, 2023), https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/ (stating that "[c]reditors must be able to specifically explain their reasons for denial. There is no special exemption for artificial intelligence . . . ."); U.S. Equal Emp. Opportunity Comm'n, Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964, https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial (stating that "if an employer administers a selection procedure [including those assisted by AI], it may be responsible under Title VII if the procedure discriminates on a basis prohibited by Title VII, even if the test was developed by an outside vendor. In addition, employers may be held responsible for the actions of their agents, which may include entities such as software vendors, if the employer has given them authority to act on the employer's behalf."). *See also* Miriam Vogel, Michael Chertoff, Jim Wiley, & Rebecca Kahn, *Is Your Use of AI Violating the Law? An Overview of the Current Legal Landscape*, 26 N.Y.U. J.L. & Pub. Pol'y 1029 (2024). However, close attention should be paid to ongoing litigation against U.S. insurance companies that employ AI to help determine the approval or denial of claims. As of the time of writing, none of the publicly available litigation materials have relied upon AI's potential legal personhood as a possible defense. *See, e.g.*, *Barrows v. Humana, Inc.,* Complaint, No. 3:23-CV-00654, (W.D. Ky. Dec. 12, 2023); *Est. of Lokken v. UnitedHealth Grp. Inc.,* No. 23-CV-3514 (JRT/DJF), 2024 WL 3677896 (D. Minn. Aug. 6, 2024). Nonetheless, some scholars have explored potential pathways for AI to attain (a kind of) legal personhood or (other) special legal status. *See, e.g.*, Nadia Banteka, *Legal Personhood and AI: AI Personhood on a Sliding Scale,* in The Cambridge Handbook of Private Law and Artificial Intelligence 618, 618–35 (2024); Diana Madalina Mocanu, *Gradient Legal Personhood for AI Systems—Painting Continental Legal Shapes Made to Fit Analytical Molds*, 8 Frontiers Robotics & AI (2022).

of argument, we assume that the first premise is well founded. We now further assert that, in administering the performance of an IHL/LOAC obligation binding on a State, the humans concerned must exercise cognitive agency.

By *cognitive agency*, as mentioned above, we mean — with respect to administering the performance of an IHL/LOAC obligation — the undertaking and carrying out of a conscientious and intentional operation of mind by one or more humans vested with State legal capacity, through which that person or those persons implement the execution of the cognitive tasks demanded by the obligation. By *conscientious*, we mean that the humans vested with State legal capacity are required to act with an awareness of the binding IHL/LOAC obligation. By *intentional*, we mean that those humans are required to act with an intention to administer the performance of that obligation, including as regards the execution of the constitutive cognitive tasks, in good faith. And by *implementing the execution of the cognitive tasks demanded by the obligation*, we mean that the humans concerned ensure that each required cognitive task is performed to a sufficient standard, whether the task is performed by a human, by an AI-related technology (if permissible), or by a human relying on an AI-related technology (if permissible).

As with the preceding premise, we argue that this premise arguably reflects a specification or instantiation of an existing condition of legality, not a new policy approach. Here, too, we ground the assessment in analyses of assumptions about executing cognitive tasks in connection with war, how IHL/LOAC obligations are performed, and how certain rules of State responsibility operate.

This second premise is based partly on the assertion that every IHL/LOAC obligation contains within it a requirement to execute certain cognitive tasks and for one or more humans to undertake and carry out a conscientious and intentional operation of mind in connection with those tasks. In other words, in our framework, all exercises of cognitive agency involve the implementation — by one or more humans vested with State legal capacity — of the execution of cognitive tasks entailed in the obligation. For example, with respect to an attack during an armed conflict, the weighing by a military commander of the anticipated concrete and direct military advantage against the potential risk of incidental civilian harm involves cognitive tasks that are demanded by the IHL/LOAC principle (or rule) of proportionality in attack. In addition, as an arguably necessary element to administer the performance of an IHL/LOAC obligation, the human concerned must act with an awareness of that obligation and with an intention to administer the performance of that obligation in good faith, including by ensuring that each required cognitive task is executed to a sufficient standard. Conversely, a failure to do so would contribute to a breach. What conduct suffices to constitute an exercise of cognitive agency depends on the specific IHL/LOAC principle or rule (and the accompanying standard, if any) in respect of which the obligation arises.

Notably, not all cognitive tasks executed in connection with an armed conflict are undertaken in relation to the performance of an IHL/LOAC obligation. In other words, under this conceptual approach, the execution of the required cognitive tasks is a necessary, but not a sufficient, criterion for an exercise of cognitive agency. For example, we draw a legal distinction between an unaffiliated civilian executing cognitive tasks during armed conflict, on the one hand, and a State's military commander exercising cognitive agency in relation to the performance of an IHL/LOAC obligation pertaining to an attack, on the other hand. The unaffiliated civilian deciding whether to flee executes numerous cognitive tasks, but the civilian is not doing so as part of an effort to administer the performance of an IHL/LOAC obligation binding on a State. In contrast, in relation to an attack, a military commander weighing the anticipated military advantage against the potential risk of incidental civilian harm is responsible for implementing the execution of the cognitive tasks required by a principle (or rule) of IHL/LOAC binding on the State. Importantly, in this example, the State has vested the military commander with the legal capacity to administer the performance of (part or all of) that obligation. The commander's undertaking and carrying out of a conscientious and intentional operation of mind here forms an essential element of the performance of (at least part of) the State's IHL/LOAC obligation pertaining to proportionality in attack.

The notion of exercising cognitive agency — as a constitutive element of the performance of an IHL/LOAC obligation binding on a State — should be distinguished from the mental state necessary to apply individual criminal responsibility for an international crime. That is, an exercise of cognitive agency in this sense is not coterminous with *mens rea*. While the two may be related in certain respects, including insofar as they both involve an operation of mind by one or more humans, we distinguish an exercise of cognitive agency as a requirement for administering the performance of an IHL/LOAC obligation binding on a State, on the one hand, and *mens rea* as the mental state (such as concerns the kind of intent or degree of knowledge) that must be established to apply criminal responsibility in case of an international crime, such as a war crime, on the other hand.

The conceptual approach concerning the requirement of an exercise of cognitive agency by humans does not necessarily categorically preclude a State — through the humans it vests with State legal capacity — from relying on AI-related technologies in relation to the execution of cognitive tasks in armed conflict. Rather, it frames part of the legal analysis in terms of the nature and extent of such reliance by those humans; it links that reliance to the administration of the performance of the relevant IHL/LOAC obligation; and it requires a conscientious and intentional operation of mind by those humans, through which those people implement the execution of the cognitive tasks demanded by the obligation. While it asks whether those humans are exercising cognitive agency, this approach does not necessarily stipulate in general (that is, in relation to all IHL/LOAC obligations) the kind and degree of reliance the humans concerned may or may not place on AI-related technologies in implementing the execution of the cognitive tasks demanded by a particular IHL/LOAC obligation. Nor does this approach stipulate — with respect to the humans concerned ensuring that each required cognitive task is performed to a sufficient standard — if it is permissible or not for the relevant cognitive task to be executed, partly or wholly, by an AI-related technology or by a human relying on an AI-related technology. (It is assumed that, under the existing legal framework, there is no question that the relevant cognitive task may be performed by a human who is not relying on an AI-related technology.)

We submit that this notion of exercising cognitive agency provides a theoretically grounded conceptual vocabulary through which to assess whether humans may or may not rely on AI as part of their efforts to administer the performance of an IHL/LOAC obligation. The answer requires ascertaining, among other elements, what specific cognitive tasks the relevant IHL/LOAC obligation requires, as well as determining if it is permissible for an AI-related technology or a human relying on such a technology to execute one or more of those tasks. Depending on the particulars of the applicable situation and assumptions around the use of relevant technologies, it might be argued that, in respect of a particular context and specific IHL/LOAC obligation, the humans concerned ought to rely on certain AI-related methods as part of their efforts to implement the execution of one or more of the cognitive tasks demanded by the obligation. For example, arguably, a human vested with State legal capacity might rely, under certain circumstances and subject to certain conditions, on AI-related technologies that have been sufficiently validated to provide relatively more comprehensive and accurate information concerning the operational environment and presence of civilians. However, reliance on the same or other AI-related technologies in other contexts may be prohibited or subject to different restrictions and requirements. In any event, under this second premise, the mere "rubber stamping" by a human concerned of an AI-related output would arguably not constitute an exercise of cognitive agency because it would lack an adequate conscientious and intentional operation of mind by that person.

We acknowledge that the concept of cognitive agency, as formulated here, does not constitute a recognized international legal term of art. Nor is it, in so many words, recognized in doctrine as a necessary condition underlying the administration of the performance of IHL/LOAC obligations by States. Moreover, we concede that the English term "agency," in this context, may lack a precise equivalent in many other languages.[36] Further, while this theoretical framing may seem to represent a novel conceptual — or, at least, terminological — approach to the performance of IHL/LOAC obligations, we submit that it is more accurate to view it as a specification or instantiation of the established framework. In other words, in setting out this notion, we seek to reflect a standard account, if in admittedly novel terms, of the current presumptions and attributes underlying the performance of IHL/LOAC obligations binding on States. We assert, in short, that this second premise may be deduced from the theoretical underpinnings of the legal framework and assumptions regarding how it is meant to function in practice.

Before proceeding, it may be warranted to address whether the notion of exercising cognitive agency is meant to encompass or be distinct from notions of exercising moral agency or other similar concepts. We observe, for example, that certain IHL/LOAC obligations entail what might be referred to as mandatory evaluative decisions or normative (value) judgements. For example, the IHL/LOAC principle (or rule) of proportionality in attack prohibits a party from launching an attack that may be expected to cause incidental

[36] For example, according to translations provided by a PILAC research assistant, while there is no single equivalent for "agency," in the sense of this paper, in Russian, the concept can still apparently be expressed through the following terms, depending on the context: (1) Способность к самостоятельным действиям (capacity for independent actions); (2) способность к принятию решений (capacity for decision-making); and/or (3) дееспособность (active legal capacity).

civilian harm that is "excessive" relative to the concrete and direct military advantage anticipated.[37] Such a determination clearly involves the consideration of factual information, such as the projected number of civilian casualties, if any, and factual aspects concerning the relative significance of the military objective. Yet the evaluation of "excessive[ness]" may also be described as requiring a normative (value) judgement.[38]

For the purpose of the analysis here, we see two approaches to encapsulating this aspect within the cognitive-agency framework. One approach sees cognitive agency and moral agency in this area as distinct. That approach would recognize a (quasi-)moral agency — separate from cognitive agency — that is arguably reserved by the law (only) to natural persons administering the performance of the part of the obligation demanding a normative (value) judgment. A second approach sees cognitive agency as a composite notion that encompasses moral agency. The composite approach views the exercise of the required normative (value) judgement or evaluative decision as an integral part of the exercise of cognitive agency. Drawing on the example concerning the IHL/LOAC principle (or rule) of proportionality in attack, under the composite approach, the exercise of cognitive agency required to determine the potential excessiveness of an attack entails both the ascertainment of relevant factual elements and an evaluation based on normative (value) judgements. In the rest of this paper, we adopt the composite approach because we think it might better encapsulate the content and structure of legal norms in this area and associated normative concerns.

4. Required Cognitive Tasks: Two Examples

To implement the conditions of legality set out in the preceding section, it is necessary to ascertain what it means in practice for humans to exercise cognitive agency in respect of each relevant IHL/LOAC obligation. That requires, as an initial step, identifying the cognitive tasks involved in performing the

[37] First Additional Protocol to the Geneva Conventions, art. 51(5)(b), June 8, 1977, 1125 U.N.T.S. 3 [hereinafter AP I].

[38] *See, e.g.*, Michael Siegrist, Legal Officer of International Humanitarian Law and International Criminal Justice, Directorate of International Law DIL, A Purpose-Oriented Working Definition for Autonomous Weapons Systems, Statement by Switzerland at the Third Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) (Apr. 13, 2016), https://www.unog.ch/80256EDD006B8954/(httpAssets)/558F0762F97E8064C1257F9B0051970A/$file/2016.04.13+LAWS+Legal+Sessi on+(as+read).pdf (stating that "many pivotal rules of IHL presume the application of evaluative decisions and value judgements. One example would be the assessment of 'excessiveness' of expected incidental harm in relation to anticipated military advantage . . . .").

specific IHL/LOAC obligation concerned. To help illustrate what that step might involve, in this section, we summarize two obligations under IHL/LOAC — one concerning proportionality in attacks, and another related to detaining civilians — and deduce respective sets of associated cognitive tasks. (We do not address other important aspects, such as what action is required, in relation to these specific obligations, in order for the humans concerned to undertake and carry out a conscientious and intentional operation of mind, through which they implement the execution of the cognitive tasks demanded by the relevant obligation. Nor do we discuss whether — with respect to the humans concerned ensuring that each required cognitive task is performed to a sufficient standard — it is permissible or not for the relevant cognitive task to be executed, partly or wholly, by an AI-related technology or by a human relying on an AI-related technology.)

4.1. Example #1: Proportionality in Attack

4.1.1. Anatomy of the Obligation

To respect a core part of the IHL/LOAC principle (or rule) of proportionality in attack, a party to an armed conflict is required to refrain from launching an attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.[39] Commentary by the ICRC has interpreted "concrete and direct" to cover only military advantages that are "substantial and relatively close".[40] The

[39] AP I, *supra* note 37, at art. 51(5)(b). *See also id.* arts. 57(2)(a)(iii), 57(2)(b–c), 85(3)(b); Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 3, art. 8(2)(b)(iv) (classifying such an indiscriminate attack as a war crime, while using the language of "clearly excessive"). The principle (or rule) of proportionality in attack has been acknowledged as customary by States, international tribunals, and the International Committee of the Red Cross. *See* Jean-Marie Henckaerts, Louise Doswald-Beck, & Carolin Alvermann, Proportionality in Attack (Rule 14), in 1 CUSTOMARY INTERNATIONAL HUMANITARIAN LAW 46, 46–50 (Jean-Marie Henckaerts & Louise Doswald-Beck eds., Cambridge Univ. Press 2005). *See also* International Committee of the Red Cross, *Customary International Humanitarian Law*, Rule 14, https://ihl-databases.icrc.org/en/customary-ihl/v1/rule14 (last visited Nov. 23, 2024).

[40] *See* INTERNATIONAL COMMITTEE OF THE RED CROSS, COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977, ¶ 2209 (Yves Sandoz, Christophe Swinarski, & Bruno Zimmermann eds., 1987) [hereinafter COMMENTARY ON THE ADDITIONAL PROTOCOLS] (stating that "[t]he expression 'concrete and direct' was intended to show that the advantage concerned should be substantial and relatively close, and that advantages which are hardly perceptible and those which would only appear in the long term should be disregarded . . . ."). However, various stakeholders differ in their interpretation of the geographical and temporal limits of the relevant military advantage. *See, e.g.*, MATTHEW WAXMAN, INTERNATIONAL LAW AND THE POLITICS OF URBAN AIR OPERATIONS 8, n.14 (2000) ("[The] principle['s] . . . precise meaning remains elusive, in part because of the inherent [Footnote continued on next page]

ICRC has also interpreted the obligation to require an exercise of good faith when comparing the relevant factors.[41] Certain scholars argue that, to prevent systemic errors, these efforts should also include an obligation to assess proportionality in light of prior performance.[42] Jurisprudence from international criminal bodies suggests that the relevant IHL/LOAC standard may be based on that of the "reasonable military commander" or "a reasonably well-informed person in the circumstances of the actual perpetrator, making reasonable use of the information available to him or her".[43] Some military manuals indicate the commander must first obtain the "best possible intelligence," including information on the concentration of civilian persons, protected civilian objects, and the environment.[44] Based on the available information, relevant members of the armed forces are required to determine whether or not the anticipated civilian harm is likely to be excessive relative to the concrete and direct military advantage anticipated. By obliging "those who plan or decide upon an attack" to respect the principle (or rule) of proportionality with regard to certain precautionary measures,[45] the First Additional Protocol (1977) appears to direct the performance of that part of the IHL/LOAC proportionality obligation expressly to that particular set of natural persons.

difficulties in measuring, and then weighing, expected military gain and civilian harm."); Ian Henderson & Kate Reece, *Proportionality Under International Humanitarian Law: The "Reasonable Military Commander" Standard and Reverberating Effects*, 96 INT'L L. STUD. 99 (2020).

[41] *See* COMMENTARY ON THE ADDITIONAL PROTOCOLS, *supra* note 40, ¶ 1978.

[42] Oona A. Hathaway & Azmat Khan, *"Mistakes" in War,* 173 U. PA. L. REV. 1 (2024), at 89.

[43] *See* Office of the Prosecutor, International Criminal Tribunal for the Former Yugoslavia (ICTY): Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia, 39 INT'L LEGAL MATERIALS 1257 (2000), ¶ 58; *Prosecutor v. Galić*, Case No. IT-98-29-T, Judgment, ¶¶ 57–58 (INT'L CRIM. TRIB. FOR THE FORMER YUGOSLAVIA Dec. 5, 2003) (asking whether "a reasonably well-informed person in the circumstances of the actual perpetrator, making reasonable use of the information available to him or her, could have expected excessive civilian casualties to result from the attack . . . .").

[44] *See* Australia, *The Manual of the Law of Armed Conflict*, Australian Defence Doctrine Publication 06.4, §§ 5.53–5.54 (Austl. Def. Headquarters, May 11, 2006). *See also id.* § 2.9 (stating that prior to mounting an attack, "the best possible intelligence is required concerning: concentrations of civilians; civilians who may be in the vicinity of military objectives; the nature of built-up areas such as towns, communities, shelters, etc; the existence and nature of important civilian objects and specifically protected objects; and the environment . . . ."); Sweden, *International Humanitarian Law in Armed Conflict, with Reference to the Swedish Total Defence System*, § 3.2.1.5, at 70–71 (Swed. Ministry of Def., Jan. 1991) (stating that "[a] planning commander must, to be able to decide upon an attack, have access to the best possible information about the objective. The decision should be based upon the information available to the commander at the time of deciding"); France, *Fiche de Synthèse sur les Règles Applicables dans les Conflits Armés*, Note No. 432/DEF/EMA/OL.2/NP, § 5.2 (1992) (signed by Général de Corps d'Armée Voinot for Amiral Lanxade, Chief of Defence Staff) (stating that commanders must "obtain a maximum of information concerning the nature and the location of protected objects").

[45] AP I, *supra* note 37, art. 57(2)(a)(iii).

4.1.2. Associated Cognitive Tasks

The following cognitive tasks are arguably involved in the performance of this IHL/LOAC obligation:

- Perceiving and gathering information regarding civilians in the area and time period;
- Perceiving and gathering information regarding civilian objects in the area and time period;
- Perceiving and gathering information regarding, and making an evaluative assessment of, what may constitute the "direct and concrete military advantage" of the attack;
- Considering if the quality and/or quantity of information gathered is "reasonable" under the circumstances (when applicable, this may include considering if it is the "best possible" information available at the time);
- Analyzing the information to determine the anticipated incidental harm to civilians or civilian objects, or some combination thereof, if any, that would result from the attack;
- Making a normative (value) judgment regarding what would constitute "excessive" incidental civilian harm relative to the concrete and direct military advantage anticipated;
- Comparing the anticipated incidental civilian harm, if any, against the anticipated concrete and direct military advantage in light of the normative (value) judgment regarding "excessive[ness]" in order to determine if the proposed attack would be indiscriminate in IHL/LOAC terms; and
- Refraining from deciding to launch any attack that may be "excessive" in that sense.

4.2. Example #2: Detaining Civilians When Security Makes it Absolutely Necessary

4.2.1. Anatomy of the Obligation

The Fourth Geneva Convention of 1949 provides that — with respect to aliens in the territory of a party to the conflict — "[t]he internment or placing in assigned residence of protected persons may be ordered only if the security of the Detaining Power makes it absolutely necessary."[46] The "security of the Detaining Power" is not expressly defined by the text of the provision. A Commentary by the ICRC interprets this provision to mean that it "is thus left very largely to Governments to decide the measure of activity prejudicial to the internal or external security of the State which justifies internment or assigned residence."[47] The ICRC's Commentary suggests that subversive activities within a party's territory, or actions directly aiding an enemy Power, pose a threat to security in the sense concerned.[48] Consequently, a belligerent may, according to the ICRC's Commentary, intern individuals if there are "serious and legitimate grounds" to believe they belong to organizations aimed at causing disturbances or pose a significant security risk through sabotage or espionage.[49] This standard was also applied by the International Criminal Tribunal for the former Yugoslavia in the *Delalić* case.[50] According to the Fourth Geneva Convention, any protected person who has been interned or placed in assigned residence shall be entitled to have such action reconsidered as soon as possible by an appropriate court or administrative board designated by the Detaining Power for that purpose.[51] If the deprivation of liberty is upheld, the protected person is entitled to have such decision reconsidered periodically (at least twice yearly), with a view to the favorable amendment of the initial decision.[52]

4.2.2. Associated Cognitive Tasks

The following cognitive tasks are arguably involved in the performance of this IHL/LOAC obligation:

- Defining what constitutes a threat to the "security of the Detaining Power";[53]

[46] Geneva Convention Relative to the Protection of Civilian Persons in Time of War, Aug. 12, 1949, 6 U.S.T. 3516, 5 U.N.T.S. 287, art. 42 [hereinafter Fourth Geneva Convention]; see also International Committee of the Red Cross, Customary International Humanitarian Law, Rule 99, https://ihl-databases.icrc.org/en/customary-ihl/v1/rule99 (last visited Nov. 23, 2024).

[47] Jean S. Pictet, ed., *Commentary: IV Geneva Convention Relative to the Protection of Civilian Persons in Time of War* 257-58 (International Committee of the Red Cross, 1958).

[48] *Id.*

[49] *Id.*

[50] *Prosecutor v. Delalić*, Case No. IT-96-21-A, Appeal Judgment, ¶ 323 (Int'l Crim. Trib. for the Former Yugoslavia Feb. 20, 2001).

[51] *See* Fourth Geneva Convention, *supra* note 46, art. 43.

[52] *Id.*

[53] Fulfilling this cognitive task might implicate a series of separate, but related, questions regarding the scope of [Footnote continued on next page]

- Determining who is a protected person under this provision;
- Perceiving and gathering security-related information concerning protected persons in the relevant area and time period;
- Determining if a protected person is a member of a group whose object is to cause disturbances;
- Determining if a protected person is engaging in espionage or sabotage;
- Determining if a protected person is otherwise seriously prejudicing the security of the Detaining Power;
- Based on the collected information, making an evaluative decision as to whether the security of the Detaining Power "makes it absolutely necessary" to order the internment or placing in assigned residence of a particular protected person; and
- If so, reconsidering this decision regularly — at least, twice yearly — to assess if the circumstances have changed so as to no longer warrant the internment or placing in assigned residence of the protected person concerned.

5. Questions for States and Other Stakeholders

In articulating the notion of exercising cognitive agency, we aim to reflect an established account, albeit one expressed in admittedly novel terms, of certain prevailing assumptions and attributes underpinning the performance of IHL/LOAC obligations. As highlighted in the preceding section, framing obligations through the lens of cognitive agency requires identifying the specific cognitive tasks required to perform the relevant obligation. It must then be determined which, if any, of those cognitive tasks may be executed, partly or fully, by an AI-related technology or by humans relying on such technologies. As part of that analysis, it is important to ascertain what is required in terms of the humans concerned undertaking and carrying out a conscientious and intentional operation of mind, through which they implement the execution

what may constitute a threat to the Detaining Power's security. This may include addressing the debate surrounding predictive or preventative detention (or administrative security detention), wherein States use intelligence gathering (whether AI-assisted or otherwise) to predict and prevent threats from emerging in the first place. *See, e.g.*, Deeks, Predicting Enemies, *supra* note 22; Joshua Segev et al., *Detaining Unlawful Enemy Combatants in Israel: A Matter of Misinterpretation?*, in CONSTITUTIONALISM UNDER EXTREME CONDITIONS 121, 121–37 (Richard Albert & Yaniv Roznai eds., 2020); Lawrence Hill-Cawthorne, *Internment in International Armed Conflict under IHL*, in DETENTION IN NON-INTERNATIONAL ARMED CONFLICT 41–45 (Oxford Univ. Press, 2016).

of the required cognitive tasks. While the concept of cognitive agency does not purport to resolve all the growing complexities associated with the possible integration of certain AI-related technologies in armed conflict, it offers a conceptual vocabulary that might assist States and other relevant actors to determine whether — and, if so, under what circumstances and subject to what conditions — humans may rely on AI in administering certain elements of the performance of IHL/LOAC obligations.

With a view to clarifying what the existing law demands, permits, and prohibits, we conclude by formulating a set of core guiding questions that States and other relevant stakeholders might consider forming positions on. In short, the foundational questions for States raised by our inquiry include:

1. Are the humans responsible for administering the performance of a specific IHL/LOAC obligation binding on a State permitted to rely on AI-related technologies in implementing the execution of one or more of the cognitive tasks demanded by the obligation?
2. If so, under what circumstances and subject to what conditions may those humans do so such that the humans concerned can still undertake and carry out the requisite conscientious and intentional operation of mind?

We anticipate a range of potential responses. Certain States might adopt an approach whereby all cognitive tasks integral to fulfilling an IHL/LOAC obligation may be executed only by humans (without reliance on AI-related technologies). At the other end of the spectrum, some States might endorse an approach that permits an extensive reliance — by the humans administering the performance of the particular IHL/LOAC obligation — on AI-related technologies to execute certain required cognitive tasks, subject to general or specific conditions. Between these poles could lie a range of possibilities, from approaches that allocate the majority of relevant cognitive tasks to humans to those that reserve particular kinds of cognitive tasks to humans. Moreover, a State might adapt its approach based on the nature of the specific obligation at issue. For example, IHL/LOAC obligations requiring normative (value) judgements might be said to call for a specific kind or degree of a conscientious and intentional operation of mind by a human vested with State legal capacity. By reflecting upon their positions and publicly articulating their interpretations of existing obligations, States and other stakeholders can contribute to a more precise and more stable understanding of how IHL/LOAC regulates the (non-)use of AI-related technologies in armed conflict.

pilac.law.harvard.edu