1 Introduction to Responsible Use of AI in Military Systems

Jan Maarten Schraagen

Technological developments in Artificial Intelligence (AI) continue to add new dimensions and complexities to world security and future conflict scenarios at an increasing pace. While the application of AI holds great potential for progress and economic growth as well as significant opportunities in the fields of security and defense, its potential misuse in international crises and conflicts may undermine the world's security interests and create risks for international peace and stability. The international community is now faced with the central question of how military application of AI can – and should – be dealt with responsibly while at the same time creating an effective deterrent.

This Introduction will set the stage for the chapters that follow, by providing a brief overview of relevant developments in AI, the military, as well as systems engineering practices. This will be followed by a brief introduction to each chapter, providing the reader with an overview of the contents of this volume.

ARTIFICIAL INTELLIGENCE: A BRIEF HISTORY

AI has a long and varied history, with periods of scientific and commercial successes followed by periods of disillusionment, instigated by scientific challenges as well as unrealistically high expectations (Nilsson, 2009). In the early days of AI (1956–1974),

DOI: 10.1201/9781003410379-1

the objective of making machines intelligent was primarily conceived as implementing general search strategies that could reason over symbolic task representations. However, it gradually became apparent that these general search strategies were insufficient for attaining high levels of performance. Researchers subsequently turned to ways of incorporating large amounts of domain knowledge into systems. AI moved from a search paradigm to a knowledge-based paradigm (Goldstein & Papert, 1977), culminating in the heyday of highly domain-specific expert systems in the 1980s (Feigenbaum et al., 1988). However, expert systems were brittle, meaning they only performed well on the limited scope they were designed for, and with the assistance of human experts who were required to close the gap between the designers' intentions and the real-world application (Woods, 2016). In a particular study on fault diagnosis with an expert system, technicians were required to follow underspecified instructions by the expert system, to infer machine intentions, and to recover from errors that led the expert system off-track (Roth et al., 1987). It should come as no surprise that expert systems did not live up to their expectations and rarely made it out of the lab to real-life usage (Leith, 2016). For a long time (roughly from 1990 until 2010), several alternative approaches (e.g., multiagent systems and the Semantic Web) were explored, with little to no success. Then, big data and machine learning entered the scene (Russell & Norvig, 2021). Deep learning turned out to be very successful, leading to unprecedented outcomes such as superhuman performance on image classification tasks, game-playing (Go, chess), and major breakthroughs in voice recognition and automatic language translation. Deep Neural Networks (DNNs) seem to bypass the problem of manual knowledge elicitation and modeling common-sense knowledge that haunted expert systems in the 1980s. However, manual labeling work is still required, for deep learning image classifiers still require labels in order to be able to learn. To obtain a label (for instance, that a certain image qualifies as a 'cat' and another as a 'dog'), a dataset usually requires humans to point out the area and indicate which type of object resides there. As deep learning requires a lot of data, this burden of manual labeling work is often too large or simply not feasible. A second problem with DNNs is that they are no longer understandable by humans. Performing calculations with tens of millions of parameters, the functioning of a deep learning network is inherently incomprehensible to humans (the problem of so-called 'black-box AI models'). Finally, DNNs may turn out to be brittle after all, as small perturbations in the input image may easily fool a neural image classifier (Moosavi-Dezfooli et al., 2016). In conclusion, AI is still very much in development and a future AI era may well go beyond deep learning and evolve into a hybrid of multiple connectionist AI techniques, symbolic approaches, and humans handling unexpected situations that inevitably arise (Peeters et al., 2021).

ARTIFICIAL INTELLIGENCE IN MILITARY SYSTEMS

In the past, AI was funded largely by defense-related funds. This changed around 2010 when AI became a huge commercial success, giving rise to billion-dollar civilian

industries in highly automated driving and data analytics. Still, the recent developments in AI have not gone unnoticed by the defense sector. AI is generally viewed as having large promises in a number of defense areas. AI is expected to speed up and improve decision-making processes, as it is able to process large amounts of data at speeds that are not matched by humans. AI may also be able to select the right information out of large amounts of data, thereby enhancing decision-making processes. AI may also be used to control robots and information agents that can perform dull, dirty, and dangerous tasks without a human operator, thereby freeing up already scarce personnel to focus on more demanding cognitive tasks. Instead of a single tele-operated robot, such as a drone, AI may be used not only to free up personnel, but also to scale up to numerous drones. Also, in communication-denied environments (e.g., underwater or through jamming), where tele-operation is impossible, AI can enable autonomy. The applications of AI lie in several military domains, such as unmanned autonomous systems, decision-making support and intelligence, cyber security, logistics and maintenance, business processes (HR, training, medical, automating work processes), and safety (own personnel as well as civilians).

AI can enhance power on the battlefield, as well as efficiency and effectiveness in the use of unmanned autonomous systems. It can also make work more attractive by delegating particular dull, dangerous, and dirty tasks to AI. If AI takes over certain dangerous tasks, it may make the work of military personnel safer. In military decision support and intelligence, AI can perform automated analysis, combination, and selection of huge amounts of data. This may enhance situation awareness and sensemaking on the battlefield, as well as speed up and qualitatively improve the intelligence-gathering process. AI may also play a role in the automated detection of attacks and vulnerabilities. AI may do this orders of magnitude faster than humans. Also, AI may assist in the automated analysis of the condition of systems, enabling better and faster predictive maintenance and proactive logistics. AI may assist in the automation of work processes, recruitment of personnel, training and education of personnel, as well as in health monitoring and diagnosis.

In conclusion, there are many potential applications of AI in military systems, going beyond merely weapons systems. It is also important to stress that AI will be used to enhance current systems rather than act as a stand-alone ‘AI system’. This implies that AI will be used as an add-on to existing systems in the domains mentioned above.

CHALLENGES OF USING AI IN THE MILITARY DOMAIN

Apart from the perceived benefits, there are also challenges associated with the use of AI. First, if AI-based solutions are to be used, they need to be trusted. This is achieved with sound development and validation methods at different phases of a system’s life cycle. This in turn requires explainability, so that the developers and certification authorities can scrutinize the solution. Explainability is defined here as the capability of an AI agent to “produce details or reasons to make its functioning clear or easy

to understand" (Arrieta et al., 2020, p. 85). Moreover, in some cases also the user or regulator could scrutinize the results of an AI-based solution if it were explainable. As mentioned above, DNNs are no longer understandable by humans and currently have a hard time explaining themselves. The field of 'Explainable AI' is rapidly developing and has grown exponentially over the past few years (Arrieta et al., 2020). Hence, the challenges associated with this topic will remain with us for the foreseeable future. One particular research challenge, to be discussed in more detail below, is what trust repair strategies should be adopted by intelligent teammates working in human-agent teams.

Second, to the extent that large data sets are used by the AI, there is a risk that the data sets are biased. For example, they may work for white males but not for black females, thus leading to discrimination of particular groups in society. There is also the related risk that, as the world constantly changes, there will be 'distributional drift' or 'prediction drift' in the data. In settings with significant changes/distribution shifts, the model based on the past data may not survive contact with the world as it currently is (a state of affairs that has long been recognized in the military, as witnessed by the saying that 'no plan survives first contact with the enemy'). Therefore, the model needs to be monitored and the data need to be as unbiased as possible. This is important not only from an ethical point of view, but also from a performance point of view (biased AI may simply not be effective in particular situations). On the other hand, to the extent that the military is bound by legal obligations on data gathering, as well as dealing with inherently complex situations with a lot of contextual factors, there may in many cases actually be a shortage of data, while the demand for data may be much higher than in civilian settings (e.g., in e-commerce). This may also negatively impact the quality of the models developed in AI.

Third, to the extent that AI takes over certain tasks from humans, there is a fear of humans not being in control anymore over what AI does. This plays a role in the discussion on the use of AI in autonomous weapons systems (AWS). Given the difficulties associated with clearly defining 'meaningful human control', and the fact that 'control' is not a requirement, whereas compliance with the law of war is, the U.S. Department of Defense (2023) prefers the term 'appropriate levels of human judgment' instead of 'meaningful human control'. In response to this, Human Rights Watch (2023) claims that it is not clear what constitutes an "appropriate level" of human judgment. Human Rights Watch also claims that human "control" is an appropriate word to use because it encompasses both the mental judgment and physical act needed to prevent AWS from posing moral, ethical, legal, and other threats. Hence, the debate on the use of the word 'control' is far from over. To make matters more complicated, Ekelhof (2019) has rightfully pointed out that control is distributed over multiple persons at various junctions in the decision-making cycle involved in the target selection and engagement process. Therefore, different forms of control are exercised even before weapons are activated. And even after an AWS has been activated, there may be a human 'in the loop' or 'on the loop', leading to disengagement of the weapon system prior to impact (this is not the case for all AWS; moreover, this discussion largely depends on one's definition of what an AWS is). This leads us, finally, to the issue of the definition of 'autonomous weapons systems' or 'autonomy' in particular. The arguments surrounding this definition are highly contested as well. The UN Convention on Certain Weapons (CCW) established

a Governmental Group of Experts (GGE) to discuss emerging technologies in the area of lethal autonomous weapon systems (LAWS). Over the period 2014–2019, the CCW/GGE has not arrived at a shared definition of AWS. Indeed, in a recent review, Taddeo and Blanchard (2022) identified 12 definitions of AWS proposed by States or key international actors. Clearly, this approach is detrimental in facilitating agreement around conditions of deployment and regulation of their use. However, for the purposes of this article, the discussions surrounding LAWS should not be confused with discussions on the use of AI in military systems. Autonomy in military systems may be enabled by AI, but there are also other technologies to enable autonomy.

RESPONSIBLE USE OF AI

While automation based on AI holds great potential for the military domain, it can also have unintended adverse effects due to various imperfections introduced throughout the life cycle. This can be due to biased data, wrong modeling assumptions, etc. In order to advance the trustworthiness of AI-enabled systems, and hence their ultimate use, an iterative approach to the design, development, deployment, and use of AI in military systems is required. This approach, when incorporating ethical principles such as lawfulness, traceability, reliability, and bias mitigation, is called 'Responsible AI' (U.S. Department of Defense, 2022). This implies that the military use of AI will be conducted in a recognized, responsible fashion across the enterprise, mission support, and operational levels in accordance with international law. The normative statements below constitute a first step toward the responsible use of AI in military systems. It is important to recognize that Responsible AI is not identical to 'explainability' or 'transparency', and therefore should not be confused with the field of Explainable AI. An AI model is considered to be transparent if by itself it is understandable (Arrieta et al., 2020), hence without the need for further explanations. Responsible AI involves other ethical principles besides explainability or transparency, such as lawfulness, bias mitigation, and reliability. In that sense, it encompasses explainable AI but cannot be reduced to it.

In terms of incorporating ethical principles such as data protection and bias mitigation, safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data. Proactive steps should be taken to minimize any unintended bias in the development and use of AI applications. Adequate data protection frameworks and governance mechanisms should be established first within Defense and next with industry at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems.

AI applications should be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. To this end, AI applications should provide meaningful information, appropriate to the context and user, and consistent with the state of art. Transparency and explainability are factors

that can improve human trust in AI systems. The level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety, and security.

An iterative socio-technical systems engineering and risk management approach should be adopted to ensure potential Al risks (including privacy, digital security, safety, and bias) are considered from the outset of an Al project. Efforts should be taken to mitigate or ameliorate such risks and reduce the likelihood of unintended consequences. A robust testing process should be developed, allowing for the assessment of AI applications in explicit, well-defined use cases. This includes continuous identification, evaluation, and mitigation of risks across the entire product lifecycle and well beyond initial deployment.

Appropriate oversight, impact assessment, audit, and due diligence mechanisms should be developed to ensure accountability for AI systems and their impact throughout their life cycle. Both technical and institutional designs should ensure auditability and traceability of (the working of) AI, in particular to address any conflicts with human rights norms and standards and threats to environmental and ecosystem well-being.

AI actors should ensure traceability, including in relation to datasets, processes, and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

CONCEPTUAL DISTINCTIONS AND CLARIFICATIONS

It is important to make a number of conceptual distinctions and clarifications, particularly when talking about the responsible use of AI in military systems.

First, in response to recent fast developments in AI, many organizations, agencies, and companies have published AI ethics principles and guidelines. In a meta-analysis, Jobin, Ienca, and Vayena (2019) included 84 documents containing ethics principles and guidelines. The most frequently mentioned principles were: transparency, justice and fairness, non-maleficence, responsibility, and privacy. The principles and guidelines have been criticized by some for being (i) too abstract to be practical, (ii) reflecting mainly the values of the experts chosen to create them (hence, not being inclusive), and (iii) serving the priorities of the private entities which funded some of this work ('ethics washing') (Hagendorff, 2020; Hickok, 2021). Although some of these criticisms are justified, one should realize that the principles are a starting point. There is great value in all of these documents being publicly accessible (several websites track them and make them available for analysis purposes, e.g., aiethicslab.com and algorithmwatch.org). Some of these principles are useful for structuring the discussion regarding the challenges for human use, for instance, bias mitigation, explainability, traceability, governability, and reliability (taken from the North Atlantic Treaty Organization (NATO) Principles of Responsible Use of AI, 2021).

Second, 'military systems' are much broader than just weapons systems. AI may be of use in a broad array of systems and applications, including business process applications, predictive maintenance, and highly automated responses to cyber-attacks. This does not in any way diminish the importance of discussing the use of AI in (offensive) weapons systems.

Third, 'autonomy' and 'AI' are not identical. AI may be used to achieve the goal of system autonomy, in the general (and, admittedly, vague) sense of achieving tasks with little or no human intervention (Endsley, 2017). However, there are other ways of achieving this goal, including the use of logic-based programming as used in classical automation. An example of the latter would be close-in weapon systems, such as the Goalkeeper or the Phalanx, which are completely automatic weapon systems for short-range defense of ships. These weapon systems may be called 'autonomous' as defined in the U.S. DoD Directive 3000.09 (2023): "A weapon system that, once activated, can select and engage targets without further intervention by an operator". Yet, these close-in AWS do not need AI to function as intended. This is not to deny that data and AI may be key enablers of autonomy.

Fourth, the definition of the concept of 'autonomy' is driven by political and strategic motivations, as briefly discussed above, and is not value-neutral. It is beyond the scope of the current chapter to arrive at a value-neutral definition of 'autonomy' (see Taddeo and Blanchard, 2022, for such an attempt). I will take up the issue of how to define autonomy in the final concluding chapter of this volume.

OVERVIEW OF THE CHAPTERS IN THIS BOOK

The chapters in this book are organized into four major sections. Section I presents models and approaches for implementing military AI responsibly. Section II is an overview of legal aspects regarding the liability and accountability of individuals and states when using AI in the military domain. Section III addresses the shifting role of human control in military teams in which humans and AI have to work together. This section includes both philosophical and human factors contributions. Section IV broadens the scope to include political and economic aspects of using AI in the military domain. Section V contains a concluding chapter in which the issues addressed in the previous sections are critically evaluated. Below, I will briefly summarize the contents of each chapter.

Section I: Implementing Military AI Responsibly: Models and Approaches

This section starts with the chapter by Heijnen et al. who present a Socio-Technical Feedback loop (SOTEF) methodology to establish and maintain the required value alignment at the levels of governance, design, development, and operation of military AI throughout its life cycle. Value alignment is important as the use of military AI

forces us to think about what values are at stake and how we want to ensure these values are accounted for. SOTEF takes an iterative, transdisciplinary, and multistakeholder approach, tailored to the prevailing objectives, context, and AI technology. Ethical, legal, and societal aspects as well as objectives for the human-AI system in high-risk situations are made explicit, commensurable, and auditable (including the attribution of responsibility and accountability). An illustrative scenario and an example set of methods and functions for value alignment exemplify the methodology.

In the second chapter of this section, Koch and Keisinger argue that democracies must be able to defend themselves "at machine speed" if necessary, to protect their common heritage of culture, personal freedom, and the rule of law in an increasingly fragile world. The use of AI in defense in their view comprises responsible weapons engagement as well as military use cases such as logistics, predictive maintenance, intelligence, surveillance, or reconnaissance. Responsibility as a notion poses a timeless question: How to decide 'well' according to what is recognized as 'true'? To arrive at an answer, responsible controllability needs to be turned into three tasks of systems engineering: (i) Design artificially intelligent automation in a way that human beings are mentally and emotionally able to master each situation; (ii) Identify technical design principles to facilitate the responsible use of AI in defense; and (iii) Guarantee that human decision-makers always have full superiority of information, decision-making, and options of action over an opponent. Koch and Keisinger discuss The Ethical AI Demonstrator (E-AID) for air defense as paving the way by letting soldiers experience the use of AI in the targeting cycle along with associated aspects of stress as realistically as possible.

The third chapter by Panwar takes a risk management approach to the responsible use of AI in military systems. Risks posed by different military systems which leverage AI technologies may vary widely and applying common risk-mitigation measures across all systems will likely be suboptimal. Therefore, a risk-based approach holds great promise. Panwar presents a qualitative model for such an approach, termed as the Risk Hierarchy, which could be adopted for evaluating and mitigating risks posed by AI-powered military systems. The model evaluates risks based on parameters that adequately reflect the key apprehensions emerging from AI-empowerment of military applications, namely, violation of International Humanitarian Law (IHL) and unreliable performance on the battlefield. These parameters form the basis for mapping the wide spectrum of military applications to different risk levels. Finally, in order to mitigate the risks, modalities are outlined for evolving a differentiated risk-mitigation mechanism. Factoring in military ethos and analyzing risks against the backdrop of realistic conflict scenarios can meaningfully influence risk evaluation and mitigation mechanisms. The rigor that underpins the Risk Hierarchy would facilitate international consensus by providing a basis for focused discussions. The chapter suggests that mitigating risks in AI-enabled military systems need not always be a zero-sum game, and there are compelling reasons for states and militaries to adopt self-regulatory measures.

Street and Bjelorglic, of the NATO Communications and Information Agency, have written a chapter from the perspective of those developing AI solutions for military users. Their chapter addresses some practical steps to ensure that military AI is developed and deployed responsibly. Specifically, several high-level principles relating to the responsible use of military AI are considered, together with the steps which developers

can take to demonstrate that these areas have been addressed responsibly when developing effective AI solutions for military use. A framework is presented that allows a pragmatic balance between the risks involved in any given AI solution and the tests, checks, and mitigations to be applied during its development.

Section I concludes with a chapter by Gadek, who promotes an existing EU-supported AI application assessment framework, ALTAI, by reviewing three military use cases and highlighting its relevance and shortcomings. Gadek claims that ethics assessments do bring an added value to AI development and that potential solutions such as "explainable AI" or "exhaustive tests", even if desirable, are neither sufficient nor necessary to decide to use AI systems.

Section II: Liability and Accountability of Individuals and States

Cooper, Copeland, and Sanders argue that while AI promises more rapid decision-making, great efficiencies, and enhanced lethality, it also presents a range of risks. States developing new AI capabilities for use in the military domain must establish national processes that allow them to identify and mitigate the risks across the entire life cycle of the AI capability. Their chapter canvases existing military regulatory and governance frameworks designed to address these challenges, particularly during the acquisition and use of highly technical, military capabilities. To mitigate such risks, the chapter identifies and explains the national weapon review process and proposes how such a process may be modified to enable a broader risk-based approach to address legal, ethical, human control, and operational risks associated with the military use of AI technologies.

In his chapter, Mauri argues that the increasing use of AI techniques in the military raises multiple questions, related not only to the ability of AWS to operate within the rules that international law provides for the use of force, but also to issues of international responsibility. In the event that, on the battlefield, AWS (e.g., a drone equipped with systems to select and engage targets without the need for human intervention) are directed to employ force, even lethal force, against an impermissible target (e.g., an unarmed civilian), who is to be held responsible? Numerous authors have begun to speak of possible 'responsibility gaps'. This chapter addresses the issue of the international responsibility of the State and its alleged limitations in regulating AWS.

The chapter by Saxon addresses the use of military AI in unlawful attacks in the ongoing armed conflict in Ukraine and challenges to hold individuals and States accountable for those crimes. The analysis focuses on the more limited context of Russia's 2022–2023 aerial campaign to destroy Ukrainian energy infrastructure. First, the chapter reviews the facts known about these attacks and the technology operating one of the primary weapons used by the Russian armed forces to carry them out – the Iranian-made Shahed drone. Next, it explains the basic principles of IHL, in particular the rules of targeting, which are particularly relevant to the use of military AI. The remainder of the chapter examines how Russia's operation of the Shahed weapon system in the context of repeated targeting of Ukraine energy installations likely constitutes war crimes, and the possibilities of holding persons and States (e.g. Russia and Iran)

accountable for these offenses. It concludes that Russia's use of military AI technology that increases the accuracy of its long-running attacks illustrates the greater likelihood that breaches of IHL occurred, as well as Russia's responsibility for those crimes.

Seixas Nunes, in chapter 10, argues that AWS have thrown into question the traditional framework for assessing accountability in war. Some scholars 'scapegoat' military commanders while others 'scapegoat' AWS for violations of IHL caused by those systems. Seixas Nunes offers a different approach. Specifically, he posits that designers and programmers should be considered as potentially liable for violations of war crimes committed by their systems.

Section III: Human Control in Human-AI Military Teams

The first chapter in this section, by Eggert, examines the normative limits of 'meaningful human control' (MHC). That AWS must, like other weapons, remain under MHC is a popular demand in response to various worries about AWS. These include (i) that AWS may not be able to comply with the laws of war; (ii) that delegating life-and-death decisions to algorithms presents a grave affront to human dignity; and (iii) that it may become impossible to ascribe responsibility for harms caused by AWS. Eggert probes the relationship between the moral significance of human control on the one hand and autonomy in weapon systems, conceived as a certain degree of independence from human agency, on the other. In challenging the justificatory force of MHC in mainstream discussions, Eggert offers a starting point for rethinking what role the notion should play in debates about the ethics of AWS.

Simpson, in his chapter, starts by focusing on a move played by DeepMind's AI programme AlphaGo In a match against Lee Sedol, one of the greatest contemporary Go players. AlphaGo played a move which stunned commentators at the time, who described it as 'unthinkable', 'surprising', 'a big shock', and 'bad'. Move 37 turned out to be key to AlphaGo's victory in that game, and it displays what Simpson describes as the property of 'unpredictable brilliance'. Unpredictable brilliance also poses a challenge for a central use case for AI in the military, namely in AI-enabled decision-support systems. Advanced versions of these systems can be expected to display unpredictable brilliance, while also posing risks, both to the safety of blue force personnel and to a military's likelihood of success in its campaign objectives. This chapter shows how the management of these risks will result in the redistribution of responsibility for performance in combat away from commanders, and toward the institutions that design, build, authorize, and regulate these AI-enabled systems. Surprisingly, this redistribution of responsibility is structurally akin to systems in which humans are 'in the loop' as it is for those in which humans are 'out' of it.

The chapter by Devitt explores moral responsibility for civilian harms by human-AI teams. Devitt argues that although militaries may have some bad apples responsible for war crimes and some mad apples unable to be responsible for their actions during a conflict, increasingly militaries may 'cook' their good apples by putting them in untenable decision-making environments through the processes of replacing human

decision-making with AI determinations in war-making. Responsibility for civilian harm in human-AI military teams may be contested, risking operators becoming detached, being extreme moral witnesses, becoming moral crumple zones, or suffering moral injury from being part of larger human-AI systems authorized by the state. Acknowledging military ethics, human factors, and AI work to date as well as critical case studies, this chapter offers new mechanisms to map out conditions for moral responsibility in human-AI teams. These include: (i) new decision responsibility prompts for critical decision method in a cognitive task analysis, and (ii) applying an AI workplace health and safety framework for identifying cognitive and psychological risks relevant to attributions of moral responsibility in targeting decisions. Mechanisms such as these enable militaries to design human-centered AI systems for responsible deployment.

Miller and Freedman, in the final chapter of this Section, define Responsible Human Delegation as the making of a "responsible" decision (i.e., a technically and ethically sound one) to "delegate" a task or function to automation. Delegation implies that there will be at least periods of no human oversight, after some initial period of the human operator's learning about and perhaps configuring the automation's behavior and performance. Neglect Tolerance is a concept from research on human-robotic interaction which, roughly, uses the amount of time a robot can be "neglected" (i.e., have a function delegated to it for autonomous performance) in context while still maintaining an acceptable level of performance. In this chapter, Miller and Freedman show how Neglect Tolerance can be adapted to a set of moral or ethical hazards and thereby used to provide a quantitative test of whether or not, in a specified set of conditions with a specified set of automation behaviors, a delegation decision can be "responsible". They provide a sample analysis using a hypothetical delegation decision and a Bayesian modeling approach, though alternatives are also discussed.

Section IV: Policy Aspects

Visions of the future of military AI are evergreen, but the reality of military automation is more complicated, Lindsay claims in his chapter. Information system performance is often more about the quality of people and organizations than the sophistication of technology. This is especially true of machine learning, which lowers the costs of prediction but increases the value of data and judgment. For commercial AI, economic institutions help to provide quality data and clear judgment. These enabling complements are likely to be missing or less effective in the contested environment of war. In other words, the economic conditions that enable AI performance are in tension with the political context of violent conflict. This strategic tension is likely to lead to several unintended consequences. These include unmanageable organizational complexity, as militaries and governments struggle to provide quality data and clear judgment, and strategic controversy, as adversaries target the data and judgment that become sources of strength for an AI-enabled organization. The irony, according to Lindsay, is that increasing military automation will make the human dimension of war even more important.

In the final chapter of this Section, Vignard poses the important question of what can be learned from how the international community has approached the development

of norms of responsible State behavior in the absence of appetite for new treaties? Would a similar approach focusing on reaffirming existing international law, agreement on norms, identification of confidence-building measures, and the development of capacity-building initiatives suffice in the field of military applications of AI? Or have these approaches proven too slow to keep pace with the speed of innovation while excluding key stakeholders, such as technologists and the private sector? This chapter identifies key lessons from the UN negotiations on cyber in the context of international security (from 2004 to 2021) and those on lethal AWS (2014-present) applicable to the objectives of developing a shared understanding of Responsible AI (RAI) and accelerating international operationalization of RAI practices.

Section V: Bounded Autonomy

In the final Section, Schraagen critically evaluates the issues addressed in the previous chapters. The aim of this concluding chapter is to reflect on some common themes that run throughout this book, as well as to highlight some issues and research challenges that were not sufficiently highlighted by the contributors. The first issue critically discussed is the debate on 'killer robots'. Three arguments are advanced against the Stop Autonomous Weapons campaign. Secondly, a critical discussion on the various definitions of the concept of 'autonomy' is carried out, resulting in an argument for the concept of 'bounded autonomy'. This concept basically states that the capacity of a system to display autonomous behavior is very limited compared with the variety of the environments in which adaptation is required for objectively autonomous behavior in the real world. This leads to a discussion of the concept of 'meaningful human control'. The argument is that more attention to the testing, evaluation, and certification process of weapon systems is required, rather than to the control exercised by individual commanders or operators. Finally, research challenges in the field of Human Factors and Ergonomics are formulated, in the context of Responsible AI for military systems.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors, 59*(1), 5–27.

Feigenbaum, E. A., McCorduck, P., & Nii, H. P. (1988). *The rise of the expert company*. New York: Times Books.

Goldstein, I., & Papert, S. (1977). Artificial intelligence, language, and the study of knowledge. *Cognitive Science, 1*(1), 84–123.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120.

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics, 1*, 41–47.

Human Rights Watch (2023). Review of the 2023 US policy on autonomy in weapons systems. Review of the 2023 US Policy on Autonomy in Weapons Systems | Human Rights Watch (hrw.org)

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399.

Leith, P. (2016). The rise and fall of the legal expert system. *International Review of Law, Computers & Technology*, *30*(3), 94–106.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574–2582).

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge: Cambridge University Press.

North Atlantic Treaty Organization (NATO) Principles of Responsible Use of AI (2021). Summary of the NATO Artificial Intelligence Strategy.

Peeters, M. M. M., Van Diggelen, J., Van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human-AI society. *AI & Society*, *36*, 217–238.

Roth, E. M., Bennett, K. B., & Woods, D. D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies, 27,* 479–525.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A modern approach* (4th ed.). Pearson Education.

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics, 28*, 37–59.

U.S. Department of Defense (2022). *Responsible artificial intelligence strategy and implementation pathway*. Washington, DC: Department of Defense.

U.S. Department of Defense (2023). *DoD Directive 3000.09. Autonomy in weapon systems*. Washington, DC: Department of Defense.

Woods, D. D. (2016). The risks of autonomy: Doyle's catch. J*ournal of Cognitive Engineering and Decision Making, 10*(2), 131–133.