The Philosophical Journal of Conflict and Violence
Vol. V, Issue 1/2021

Available online at http://trivent-publishing.eu/

Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control

Elke Schwarz

Queen Mary University, London, United Kingdom.

Abstract

In this article, I explore the (im)possibility of human control and question the presupposition that we can be morally adequately or meaningfully in control over AI-supported LAWS. Taking seriously Wiener's warning that "machines can and do transcend some of the limitations of their designers and that in doing so they may be both effective and dangerous," I argue that in the LAWS human-machine complex, technological features and the underlying logic of the AI system progressively close the spaces and limit the capacities required for human moral agency.

Keywords

Artificial Intelligence; Control; Moral Responsibility; War; Weapons.

DOI: 10.22618/TP.PJCV.20215.1.139004

The PJCV Journal is published by Trivent Publishing

Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control

Elke Schwarz

Queen Mary University, London, United Kingdom.

Abstract

In this article, I explore the (im)possibility of human control and question the presupposition that we can be morally adequately or meaningfully in control over AI-supported LAWS. Taking seriously Wiener's warning that "machines can and do transcend some of the limitations of their designers and that in doing so they may be both effective and dangerous," I argue that in the LAWS human-machine complex, technological features and the underlying logic of the AI system progressively close the spaces and limit the capacities required for human moral agency.

Keywords

Artificial Intelligence; Control; Moral Responsibility; War; Weapons.

Introduction

Autonomy in military weapons systems has been advancing at a rapid pace in recent years. A growing number of countries, including the US, the UK, China and Russia either develop, produce or use military systems of varying degrees of autonomy, including lethal ones. With machine learning and computer processing power making great strides, the integration of Artificial Intelligence (AI) into military systems is likely to accelerate the shift toward increased and more complex forms of autonomy in the near future. This desire to develop ever-increasing levels of autonomy in military weapons technology is by no means a recent trend. Indeed, AI and military research and development programmes have traditionally co-evolved, and researchers in cybernetics and AI in the 1950s and 1960s had already raised pertinent questions about how increased levels of machine autonomy might affect human control over military technology.

One of the foremost thinkers of early cybernetic systems, Norbert Wiener, raised this issue early on and warned in no uncertain terms of the risks involved in applying machine rationality to the infinitely complex and plural condition of humanity. If we ask automated machines to do our bidding in human affairs, he suggests, the consequences may well be that we can neither understand their logic nor control their actions upon the world. This twinning of "two agencies essentially foreign to each other," which operate on, and within, vastly different timescales, thus may well yield catastrophic outcomes.[1] For Wiener, this raised pressing questions about moral challenges and responsibility, especially in the context of war.

[1] Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131 (1960): 1355-1358.

"When human atoms are knit into an organisation in which they are used, not in their full rights as responsibly human beings, but as cogs and levers and rods, it matters little that their raw material is flesh and blood," for the human and the human world is increasingly shaped along the lines of a machine logic.[2] And this is not a trajectory one should aspire to.

Although Wiener was responding to emerging military technologies in his own time, his concerns resonate with present-day discussions on contemporary developments and debates around lethal autonomous weapons systems (LAWS) in particular. These debates over whether the development and use of lethal autonomous systems equipped with so-called critical function capabilities (the ability to select and engage or attack a target without human intervention) should be banned outright or, at the very least, regulated at the international level are ongoing, heated, and not likely to be resolved in the near future. The topic has been on the agenda of the UN Convention on Certain Conventional Weapons (CCW) meetings since 2013 and regular meetings of the Group of Governmental Experts (GGE) have been held since 2017. Talks are advancing only slowly and are fraught with delays and obfuscations. However, there is a consensus among the debating parties that, regardless of the degrees of autonomy of a system, a human should retain a meaningful or appropriate level of control over the lethal autonomous system in question. This concept of *meaningful human control* broadly contours the "ability to make timely, informed choices to influence AI-based systems that enable the best possible operational outcome."[3] There have been various efforts to define the requisite core elements. These usually include that an operator, or human 'on the loop,' is in a position to make a conscious and informed decision about the appropriate use of the system; that he or she has adequate information about the target, the system itself and the context within which it will be used; that the system is predicable, transparent and reliable; that it has been adequately tested to operate as intended; that operators have been trained; and that there is the potential for timely human action and intervention, such that control over the system is facilitated to the best possible degree.[4] However, what this might mean, specifically and across contexts, for systems that take on increasing decision-making roles is not clear.

In this article, I am concerned with Lethal Autonomous Systems that employ Artificial Intelligence to achieve autonomy in the kill chain operation, but where the human — in their task to control the system — remains a significant element in the chain toward the lethal action.

Within this context, I explore the (im)possibility of human control and question the presupposition that we can be morally adequately or meaningfully in control over AI-supported LAWS. Taking seriously Wiener's warning that "machines can and do transcend some of the limitations of their designers and that in doing so they may be both effective and dangerous,"[5] I argue that in the LAWS human-machine complex, technological features and the underlying logic of the AI system progressively close the spaces and limit the capacities required for human moral agency. The article begins with some ground-clearing on the theoretical casting of the human in relation to technological systems, wherein I offer a

[2] Norbert Wiener, *The Human Use of Human Beings* (Cambridge: Da Capo Press, 1954), 181.

[3] Michael Boardman and Fiona Butcher, "An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It," MP-IST-178-07 NATO Report. file:///C:/Users/Admin/Downloads/MP-IST-178-07.pdf

[4] Heather Roff and Richard Moyes, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons Systems," *Briefing Paper for Delegates for the CCW GGE Meeting on LAWS* (April, 2016); see also Michael C. Horowitz and Paul Scharre, "Meaningful Human Control in Weapon Systems: A Primer," *Center for a New American Security (CNAS),* Working Paper (March 2015): 4.

[5] Norbert Wiener, "Some Moral and Technical Consequences of Automation."

challenge to the instrumentalist view, which implicitly underpins most of the current discussion on LAWS and human agency. This is followed by a brief overview of the technical characteristics of the systems in question, so as to better understand the intricacies of the technological ecology within which meaningful human control is to be exercised. I then shift my attention to the place of the human within this technological ecology and to three aspects where the logic of AI-enabled lethal systems stands in clear tension with human agency and control (cognitive, epistemic, and temporal). Here, I argue that the human operator, embedded in complex digital ecologies encounters limits that hamper moral agency and decision-making capacities considerably and in turn complicate aspirations of human control. Instead, the human is cast in a diminished role within the human-machine complex, whereby modes of reasoning are shaped along the techno-logics of algorithmic data processing, and the conditions required for ethical decision-making are limited.

I. Agency in the Human-Machine Complex

Much of the public debate on LAWS, and on military technology in general, knowingly or unknowingly embraces an instrumentalist position. This is not surprising. Particularly in matters of government and policy, ethical concerns about technology are traditionally cast in these terms. An instrumental theory of technology presumes that "technologies are 'tools' standing ready to serve the purposes of users,"[6] which implies that technologies are subordinated to the user and their politics, culture, and values. Users are assumed to retain full agency and direction over the instruments they avail themselves of and thus accountability and responsibility are assigned accordingly. With military technologies the relevance of this position is perhaps evident. Different technological means — the tank, the rifle, the bomb disposal robot, and so on — are used for different goals and it is reasonable to assume that an operator has control over the operation of a rifle or a tank and is responsible for its use or misuse. Understood in these terms, there is an implicit and clear division between technology as a utility-generating device and the human as a utility-seeking actor. Neither the question of human agency nor that of the nature of technology is problematised in these accounts. Instead, it is typically assumed that technology is in essence neutral and that distinctly human ideas about values and actions prevail. In other words, the human always remains in charge when the chips are down.[7] This is both common sensical and practical, as David Gunkel points out, because it "locates accountability in a widely accepted and seemingly intuitive subject position."[8] Such a position is also evident in certain strands of analytical just war theorising, in which a hypothetical human moral agent, together with a host of hypothetical technological means of which they might avail themselves — a flame thrower, a trolley, a ray gun — is inserted into morally vexing situations and asked to solve them through rational moral reasoning.

This mode of thinking about technology, however, risks overlooking how the availability and characteristics of weapons technologies can direct and shape actions in morally relevant ways, whereby the means of technological affordances mould the envisioned outcomes and attendant moral justifications in significant ways.[9] In non-instrumental theories of technology,

[6] Andrew Feenberg, *Transforming Technology: A Critical Theory Revisited* (New York: Oxford University Press, 1991), 5.
[7] David J. Gunkel, "Other Things: AI, Robots and Society," in *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, ed. Zizi Papacharissi (New York: Routledge, 2019), 54.
[8] Ibid.
[9] Elke Schwarz, "Technology and moral vacuums in just war theorising," *Journal of International Political Thought* 14/3 (2018).

the technological artefact is instead seen as a significant element in the web of socio-political relations within which the human is embedded, endowed with its own agentic capacity to shape human practices, aims, ideas and frames of reference for decision making, moral and otherwise. More than a merely instrumental relation of utility, "phenomenological interactions of humans with technologies constitute a shared life-world that shapes knowledge, politics and normativity, among others."[10] Quite contrary to the idea that technological artefacts are merely tools, waiting to be bent to our will, they become participants in the shaping of actions, practices and processes. Agency is distributed along the nodes of participant elements, and within complex systems of distributed data infrastructures (as is the case with autonomous systems), the human is increasingly displaced. In other words, there is a co-constitutive relation between the technological systems we are embedded in and our frames and capacities for human moral decision making. This is particularly pronounced with emerging military technologies, where the human is always already embedded within a complex system of technological networks, and even more so with the development of AI, where the prevalence of digital decision systems and interfaces is amplified. Within the structures of digital technology, the human is de-centered and placed within a material-semiotic web of relations which prioritizes a techno-logic geared toward speed, optimization, and efficient decision-making. Moreover, as Maya Indira Ganesh explains, within such dispersed AI data infrastructures "there are subject positions the human might find herself in that she cannot necessarily predict or control, given what big data infrastructures are."[11]

As technological systems become more complex, more ubiquitous, and as humans are more intricately woven into ecologies of technology, the instrumentalist presupposition, then, is no longer as straightforward as it seems. Specifically, where technologies are purposefully designed to display levels of independence and operational autonomy, as is the case with AI systems, no matter how basic, the claim that humans can and will utilise this type of technology at will, and remain in control of autonomous machines, is called into question. This is even more pertinent in the context of recent advancements in neural network machine learning algorithms, which operate beyond the capacity of the engineer to conceptualise the computational process of the neural network technology.[12] With neural network AIs like AlphaGo, for example, the human can stipulate the ideal outcome, but cannot control or necessarily comprehend the pathway to the outcome. In Gunkel's words, "we now have things that are deliberately designed to exceed our control and our ability to respond or answer for them."[13] This produces a chasm in the human-machine complex, which is patched through dialogue protocols, interface design, and other user-focused technological designs, in which higher-level human cognitive processes and problems — "mental workload, communication, decision-making, skilled performance and situational awareness" — are mediated.[14] The more complex and speedy the digital back-end, the more limited are the human capacities, the greater the scope for mediation and the greater the capacity for the technology to direct the human's practices and focus. I will come back to this point shortly.

Rather than a hierarchical-instrumental relationship in which human agency and control is taken for granted, with complex digital technologies, human-technology relations become what Katharine Hayles calls a "cognitive assemblage," in which humans and technological

[10] Maya Indira Ganesh, "The ironies of autonomy," *Nature: Humanities and Social Sciences Communications* 7/157 (2020): 6.
[11] Ganesh, "Ironies of Autonomy": 2.
[12] Hannah Fry, *How to be Human in the Age of the Machine* (London: Transworld Publishers, 2018), 86.
[13] Gunkel, "Other Things": 60.
[14] Pertti Saariluoma, "Four Challenges in Structuring Human-Autonomous Systems Interaction Design Processes," in *NATO Allied Command Transformation* (The Hague: NCI Agency, 2015).

systems are inter-connected, and whereby "the cognitive decision of each affect the others, with interactions occurring across the full range of human cognition, including consciousness, the unconscious, the cognitive nonconscious and the sensory/perceptual systems that send signals to the cortex."[15] In such interactive cognitive networks, mediated through interfaces and interpretive semiotics, instrumental assumptions about a rational, self-reflexive operator in full control of what they do are difficult to maintain. Hayles illustrates this with the example of digital personal assistants which mediate "cognitive abilities in human brains," such as memory and navigational skills, and produce "a certain homogenization of behaviour" in their users.[16] Similarly, Michael Dieter and David Gauthier explain the complexities of digital interface design which, together with the user, form a hybrid cognitive assemblage in which the human user might be nudged and habituated "into patterns of action, even to the point of compulsion."[17] The human-machine cognitive assemblage, as Hayles rightly notes, raises ethical questions about "how agency is distributed … and in what ways actors contribute to systemic dynamics and consequently how responsibilities — technical, social , legal, ethical —should be apportioned."[18]

This bears directly and significantly on debates over AI-enabled LAWS. Contrary to the popular image of a unitary autonomous weapon, purposefully designed for a specific goal or outcome and (mis)used at will by a wilful actor, a careful consideration of what meaningful human control might mean for LAWS must consider autonomous weapons not as discrete devices programmed to produce a pre-determined outcome, but as cognitive assemblages, a complex of sensors, information networks, transmitters, and hardware, which offer affordances, together with various human designers and operators, that shape our very ideas of what it means to exert moral agency. In this setting, the human ought to be considered not simply as a controller of a system, but as an interactive element within the artificial system. To better understand this situatedness, it is useful to consider the system logic of the AI-enabled weapon, and so to this I now turn.

II. The Banality of Killer Robots

The ethical complexities of LAWS, even with a human somewhere on the loop, reside in the mundane AI determined 'selection–decision' backend operations, and not merely in the actual act of killing through the robotic machine aspect. The 'Killer Robot' is in this way better understood as data infrastructure, paired with a weapons platform and payload, in which AI is employed within the sense-decide-act cycle. As such, a broader scope of systems should be taken into consideration in any analysis of LAWS: not just those systems that are designed as an entity to act with full autonomy in their critical functions and where the human is cast as somewhere in this process able to exert control, but also those that have the capacity to do so even though the human is, technically, still designed into the system as an element in the targeting loop.

Autonomous intelligent weapons that take on a significant role in selection and targeting functions — as they are designed and developed today by the US, China, Russia, and other countries — are systems that aim to increase lethality through an acceleration of information

[15] Katherine Hayles, "Cognitive assemblages: Technical agency and human interactions," *Critical Inquiry* 43/1 (2016): 33.
[16] Ibid. 40.
[17] Michael Dieter, David Gauthier, "On the Politics of Chrono-Design: Capture, Time and the Interface," *Theory, Culture & Society* 36/2 (2019): 61-87.
[18] Hayles, "Cognitive assemblages,": 34

transmission in the kill chain.[19] In most cases, this involves an AI systems component that evaluates sensor data and makes a targeting suggestion to the operator in matters of seconds, which can then be followed through "with the click of a mouse."[20] The US Department of Defense programs and progress in this area is indicative of a global trend in the AI arms race, which envisions AI as playing a crucial role in the accelerated identification and tracking of targets by the computer, which then leaves the human with a limited set of possible courses of action for a potentially lethal decision. At present, the US DoD is partnering up with a host of private sector contractors to achieve its objectives set for 2020, which include the use of AI in "joint all-domain command and control, accelerated sensor-to-shooter timelines, autonomous and swarming systems" and "target development."[21] The more work done by the machine, the better, so as to mitigate the limits to human perception and cognition, or so the official story goes. While the DoD continually maintains that there are no plans underway to outsource the act of killing to the machine itself, and that this decision will continue to reside with the human, it is their explicit aim to let the AI make a pre-selection of possible targets at an accelerated pace. A number of specific programmes are currently in place that work toward this goal. DARPA, for example, has teamed up with Lockheed Martin for Project Squad X to deliver improved situational awareness and alleviate the 'fog of war' through human-machine teaming, with "artificial intelligence as a true partner,"[22] providing "advanced sensor fusion, artificial intelligence and autonomy" to support its "human squadmates [with] tactical electronic and kinetic support," thereby helping to "inform human decision making in complex, time-critical combat situations."[23]

A similar human-machine teaming aim underpins the US Army's Advanced Targeting and Lethality Automated System (ATLAS), although with some scope for more increased kinetic autonomy. The ATLAS system's goal is to leverage advancements in machine learning and computer vision for integration into ground vehicles, so that acquisition, identification and engagement of targets can take place three times faster than currently possible with manual processes. For this, "[t]he ATLAS will integrate advanced sensors, processing and fire control capabilities into a weapon system to demonstrate these desired capabilities."[24] The explicit capacity to autonomously acquire, identify and engage targets through AI capabilities has raised some concerns that the US is advancing their LAWS programme despite DoD stipulations not to do so.[25] Since then, the US Army has gone to great lengths to dispel the idea that ATLAS is the flagship programme for lethal autonomous tanks and other ground vehicles; rather, their narrative has shifted to highlight human-machine teaming capabilities, stressing how the machine has no lethal autonomy as such. In line with DoD Directive 3000.09, it cannot pull the trigger. As Army engineer Don Reago puts it, "[e]nvision it as a second set of eyes that's just really fast, [like] an extra soldier in the tank."[26] The programme

[19] Jack Shanahan, "Lt. Gen. Jack Shanahan media briefing on A.I.-related initiatives within the Department of Defense," *US DEPT OF DENSE* (August 30, 2019).
[20] Nathan Strout, "Inside the Army's futuristic test of its battlefield artificial intelligence in the desert," *C4ISRNET* (September 26, 2020).
[21] Ibid.
[22] DARPA, "With Squad X, Dismounted Units Partner with AI to Dominate Battlespace," *DARPA* (2019).
[23] BAE Systems, "BAE Systems selected to provide autonomy capabilities for DARPA's Squad X Program," *Bloomberg* (June 2, 2020).
[24] Department of Defense, "Industry Day for the Advanced Targeting and Lethality Automated System (ATLAS) Program," *Department of Defense* (2019).
[25] Kristin Houser, "US Military: Our 'Lethality Automated System' Definitely isn't a Killer Robot," *The Byte* (2019); and Sidney J. Freedberg Jr, "Fear and Loathing in AI," *Breaking Defense* (March 6, 2019).
[26] Reago, quoted in Sidney J. Freedberg Jr, "ATLAS: Killer Robots? No. Virtual Crewman? Yes," *Breaking Defense* (March 4, 2019).

employs vast data sets and machine learning algorithms to sift through sensory input data and provide recommendations of potential targets. The system itself does not determine whether the object of identification is hostile or not, it merely presents a "list of 'objects of interest' from which the human operator can choose."[27]

The Pentagon's pioneering military AI programme is perhaps also its most well-known to date. The Algorithmic Warfare Cross-Functional Team — better known as Project Maven — has been in operation since 2017. Similar to ATLAS, Project Maven's main aim is to harness the benefits of machine learning and computer vision to accelerate targeting (and other) decisions in challenging conflict environments. Where ATLAS is intended for ground vehicles, Maven draws on drone technology for its visual and other sensory inputs in order to identify and track potential targets in real-time and give the human operator intelligence for kinetic engagement at speed. Here too, it is stressed that the human remains on the loop, specifically by triggering any lethal action, but, as with the ATLAS system above, it is not difficult to imagine how systems such as these could potentially be opening the door toward ever-greater autonomy.

Crucial here is that the AI-enabled autonomous weapons system is always a *system*, an assemblage, comprised of software processes, hardware delivery platforms and, for now, the human. The human is, in this sense, always an element of the techno-logical system, embedded within digital structures that often operate beyond human intelligibility in terms of both speed and computational logic. This has always been the case for operations in technologically advanced militaries, but it is set to become much more pronounced with advances in neural networks, machine learning, and increased processing power. Operating or controlling an AI-enabled weapons system is not a simple case of command and control. The logic of the system matters.

Autonomy in technology is, in simple terms, "the ability of a machine to execute a task, or tasks, without human input, using interactions of computer programming with the environment."[28] Human supervisory roles notwithstanding, it is always constituted "by the integration of the same three fundamental capabilities: sense, decide and act."[29] Unlike an automated system, which follows a set of simple if/then/else rules, an autonomous system works on the basis of probabilistic reason — it makes "guesses about best possible courses of action given sensor data input."[30] This requires the system to gather sensory input about the environment within which, or upon which, the system operates, to form a 'world model,' so that incoming information can be processed through optimisation and verification algorithms and geared toward a decision and subsequent action. The more complex the real-world environment, the more computational capability, sophistication of algorithms, and sheer processing power that is required for the system to map out a world model with acceptable fidelity and make a probabilistic evaluation of what might be the best course of action. The strength of an AI-enabled automated or autonomous system resides in its capacity to process very large amounts of informatic data and to then perform a predictive analysis, far faster than any human could ever do. However, it must also be noted that an activity such as recognising and evaluating an element within a highly dynamic environmental context, such as a face and the space within which it is situated (which is a relatively simple task, in human

[27] Ibid.

[28] Vincent Boulain and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapons Systems* (SIPRI Report: November 2017), 5.

[29] Boulain and Verbruggen, *Mapping the Development*, 7

[30] Mary Cummings, "Artificial Intelligence and the Future of Warfare," *Chatham House Research Paper* (January 2017).

terms), requires an enormously complex set of calculations by a machine and enormous processing power.[31]

This is where advancements in AI and machine learning come to bear on the process. An AI system requires machine learning to adjust to new circumstances, detect and identify patterns, adjust behaviour based on pattern recognition, and evaluate decisions based on successful or unsuccessful outcome of the behaviours. Machine learning is essentially a mathematical mode of learning which uses "algorithms to parse data, learn from it and then make a determination or prediction about something in the world."[32] Greater processing capacities and the availability of large amounts of data has given a boost to a variant of machine learning known as 'deep learning.' Deep learning draws on neural networks that are modelled after biological brain structures. Deep Mind's AlphaGo system, for example, employs artificial deep neural networks, which are trained by supervised learning from human players and reinforcement learning through self-play.[33] Deep learning neural networks are better than other modes of machine learning at perception: pattern recognition of light and dark edges, finding structures and sequences in images and other input data. This makes such modes of learning suited to tasks like image recognition, face recognition, voice recognition, and so on. However, the logic of the deep learning algorithmic operation neither follows the same patterns as human perception and object recognition, nor is readily intelligible to the human observer, or even those that have provided the training data to the AI.

We may, for example, consider the modes by which a team of researchers trained a neural network to identify and distinguish a wolf from a dog. It learns this distinction from a large number of image files which depict dogs and wolves. However, rather than identify the distinguishing features of the animal itself, the neural network classifies 'wolf' and 'dog' based on background features: wolves are more likely to be featured on snow and dogs on grass.[34] Another such example is that of a neural network tasked with identifying camouflaged tanks. The system was trained on images of tanks in trees and trees without tanks which the researchers took, and it was tested on a set of further images of trees and tanks taken by the research team. It seemed to work smoothly until it was tested outside the lab setting, where the identification system failed to recognise any significant number of camouflaged tanks. It transpired that "in the researchers' dataset, photos of camouflaged tanks were taken on cloudy days, while photos of plain forest had been taken on sunny days." What the neural network learned was to differentiate between cloudy and sunny skies, rather than camouflaged tanks from trees.[35]

What is at work here is the technological rendering of the world as a statistical data relationship, a simulation of the world, in algorithmic terms, of those data features that are useful for the model. John Cheney-Lippold explains the implications of this in the context of the fatal 2016 Tesla crash in which the autonomous vehicle failed to correctly identify the

[31] Anyana Ganesh, Andrew McCallum and Emma Strubell, "Energy and Policy Considerations for Deep Learning in NLP," *57th Annual Meeting of the Association for Computational Linguistics (*ACL) (Florence, Italy: July 2019).

[32] Dhruv Shah, "AI, Machine Learning, & Deep Learning Explained in 5 Minutes," *Medium – Becoming Human* (April 3, 2018).

[33] Gunkel, "Other Things": 58.

[34] Marco Ribeiro, Sameer Singh and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (2016). Stroudsburg, Pa.: Association for Computational Linguistics. doi:10.18653/v1/N16-3020.

[35] Ryan Henderson. "Picasso: A free open-source visualizer for Convoluted Neural Networks – Cloudy with a chance of tanks," *Merantix – Medium.com* (May 16, 2017). Available on: https://medium.com/merantix/picasso-a-free-open-source-visualizer-for-cnns-d8ed3a35cfc5.

white side of a tractor trailer against a bright sky, which had crossed into the lane of the Tesla driver.[36] The Tesla crashed against said trailer, killing the driver instantly. The driver had relied on the autonomous features of the car, became distracted and thus lacked sufficient contextual awareness to bridge the momentary lapse in the action loop that would have allowed for the breaks to be applied, and the crash to be averted. An accident, of sorts, but also a condition of the human-technology system. For Cheney-Lippold this is an ontological gap between the real-world occurrence of "elements moving through time and space," as the white truck crosses into the Tesla's lane, and the algorithmic interpretation of events, which constitutes "a probabilistic evaluation of those elements, represented, as best as the algorithm can, as a deviating new world […] where a white truck ceases to be a white truck and becomes a statistical relationship."[37] The AI system represents and operates on the world along a logic that accepts that it cannot comprise and calculate the 'real world' and every one of its iterations as the human would. This always involves abstraction, truncating, and rendering; it involves "approximation, biases, errors, fallacies and vulnerabilities." With AI, "information flows are diffracted, distorted and lost,"[38] casting the world in statistical approximations and acting, or suggesting actions, upon this interpretation of the world.

In short, Machine Learning works through abstraction, through a rendering of the world as a model, through error and failure, and as such is always underspecified.[39] An AI system does not reason the world the same way a human does; it may not even be intuitive to human reasoning. It is their logic to press the infinity of in-the-wild events into calculable processes and it is by design that "machine learning systems make their own rules for how to achieve the goals that are set for them."[40] The AlphaGo AI, for example, made decisions and choices that continue to be opaque and unintelligible to the grand master as much as to the programmers of the system. As one of the AlphaGo programmers explained: "We just create the data sets and the training algorithms. But the moves it then comes up with are out of our hands."[41] In short, the system uses non-human logics to arrive at its conclusion, decision and action, or recommended action. This condition may be harmless, even desirable, in the playful context of a strategic board game, but where the moral stakes are high, such as in warfare, being able to predict and understand the decision system is important. Although steady progress is made in computer science to help patch the lack of understanding between human and machine perception, the disconnect is — in part — not a bug, but a feature of the technology. The more challenging and open an environment, the more complex and sophisticated an AI system might need to be to be able to be employed in such an environment. And the more complex and opaquer a system is, the less predictable, or understandable, are its decision-relevant actions. This "performance-understandability trade-off poses a central paradox of AI" in weapons systems and complicates matters of agency.[42]

Tracing the data-logic of AI systems provides the basis for understanding the challenges they pose to meaningful human control and moral agency in relation to LAWS, but this

---

[36] John Cheney-Lippold, "Accidents Happen," *Social Research: An International Quarterly* 86/2 (2019): 513-535.

[37] Cheney-Lippold, "Accidents Happen": 527.

[38] Matteo Pasquinelli, "How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence," *Spheres* 5 (2019): 5.

[39] Alexander D'Amour et al., "Underspecification Presents Challenges for Credibility in Modern Machine Learning," arXiv:2011.03395 [cs.LG] (November 24, 2020).

[40] Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (Geneva, Switzerland: United Nations Institute for Disarmament Research, 2020), 9.

[41] Cade Metz, "Google's AI Wins Pivotal Second Game in Match With Go Grandmaster," *WIRED* (October 3, 2016).

[42] Holland Michel, *The Black Box, Unlocked*, 10

constitutes only one analytical side of the coin. In complex and accelerated contexts of technological action, and especially within distributed systems, the question of what can constitute meaningful *human* control looms large. The focus of inquiry must therefore go beyond the technological system alone, its capabilities or limitations; what matters most when we consider the meaning of an action is the human element in this technological system — the human, not as a quasi-technological element in a system, but the human *qua* human.

III. Limitations to Meaningful Moral Action

Decisions in warfare are morally complex for the actors involved. This is particularly so in contemporary conflicts, which are often asymmetric, asynchronous, and of unclear geographical scope. In such contexts, moral challenges are most keenly felt with decisions that involve lethal force exerted through technological instruments that facilitate distance. The soldier or military operator confronting moral dilemmas must bear the psychological burden of the decisions they make with respect to the use of lethal force, and as the incidence of PTSD among drone operators clearly indicates, new technologies might further add to their psychological stress. Recent research on the human-machine interface in autonomous military technology suggests that "Autonomous Weapons Systems may also cause anxiety and concerns" among military personnel, as such systems may end up behaving in unwanted and unpredictable ways. We might consider, then, that the role of the human as a moral decision-maker in the technological loop is considerably more complex than the picture of the functional and rational agent 'in' or 'on' the loop might suggest. Literature in moral cognitive psychology clearly links moral judgement to "a variety of […] fine-grained and disparate processes" that are "affective and cognitive" in nature. When considering where humans retain meaningful control over moral decision-making with LAWS, both of these dimensions matter.

There are competing accounts as to what might constitute moral agency in a digitally mediated environment.[43] Some suggest that in order to adjust to the new technological order, we ought to accept our diminished moral agency and instead embrace technological moral agency as a means of securing a better social and political future. Here is not the place to evaluate the merits of such conceptions of moral agency. But when considering moral agency in the context of LAWS, it is worth making a clear distinction between human agency and artificial agency. As Alexander Leveringhaus explains, humans act morally in relation to other humans, based on human experience and all the knowledge this comprises.[44] Unlike technological artefacts, humans possess a "capacity to have mercy with, feel pity for or empathise with other humans," even those that are by all accounts cast as an enemy.[45] As humans, we understand and are able to judge the specificities of human relations and relationality in a range of social contexts in a way that technological artefacts are simply not able to. Or to put it another way, our morality, and thus our moral decision-making, is

[43] See for example Kenneth Einar Himma, "Artificial agency, consciousness and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?," *Ethics and Information Technology* 11/1 (2009): 19-20; John Sullins, "When is a robot a moral agent?," *International Review of Information Ethics* 6/12 (2006); David Gunkel, *The Machine Question: Critical Perspectives on AI, Robots and Ethics* (Cambridge: MIT Press, 2017); Joanna Bryson. "Robots Should be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Yorick Wilks (Amsterdam: John Benjamins, 2010).

[44] Alexander Leveringhaus, "What's So Bad About Killer Robots?," *Journal of Applied Philosophy* 35/2 (2018).

[45] Ibid., 350.

anchored in our history of human social relations. It is this condition that allows for the human to make a decision in warfare that is not an obvious, logical or pre-programmed decision — to forgive, to empathise, to act on pity, to sacrifice oneself, and so on. This is not a task that can be executed by any artificial element in the technological ecology of LAWS.

And so, the task of exercising moral agency in complex and distributed technological systems such as LAWS must fall fully to the human. To give the challenges to this task some context, a brief discussion of the prerequisites for human moral agency is helpful. I draw for this on accounts of moral agency and responsibility as developed by John Martin Fischer and Mark Ravizza, who posit that the human *per se* remains at the heart of having 'guidance control' over moral acts, and with this, advance an understanding of moral responsibility as a fundamentally social practice. Moral agency means the capacity not just to have, but to *take* moral responsibility. This, in turn, requires that the moral agent understands herself as such — as a moral agent within a social setting of values and expectations; that she has adequate knowledge to act as a moral agent and, importantly, that the moral decision-mechanism is hers to own.[46] This means they must not be under duress or undue influence, and they must have reasonable assumptions of knowledge as to what outcomes might be foreseeable. These conditions, at the very least, need to be fulfilled in order for human control over lethal actions to be morally meaningful. Moreover, to exercise moral agency, especially in military contexts, ethical awareness is required — "assessments of the values and interests underlying and the inherent goodness or badness, rightness or wrongness, fairness or unfairness of the situation to be made."[47] For actions and outcomes that are mediated through machine learning and computational processing, this awareness might well be obscured.

As we have already seen above, this is not straightforward in the case of LAWS understood as technological ecologies. Among other factors, cognitive limitations, the opacity of machine-reason, and the prioritisation of speedy action with autonomous weapons systems can severely complicate this moral agency. In what follows I will briefly explain and illustrate each of these three factors.

A. Cognitive Limitations

An extensive body of scholarship in cognitive psychology attests to the fact that as humans, we experience cognitive limitations when interacting with computational systems. Noel Sharkey outlines this in a briefing paper for the UN CCW GGE meeting in April 2018, which is worth summarizing.[48] In short, Sharkey highlights how humans typically make decisions based on two types of reasoning: the first type is deliberative reasoning, a process for which we draw on more extensive memory resources required for decisions of considerable weight and impact, such as morally challenging decisions. The second type of reasoning is automatic reasoning, which we use for routine events in everyday life. This type of reasoning is, as the name suggests, automatic, and therefore takes place much more quickly. As humans, we tend to choose the path of least resistance, so as to be able to cope with the many demands we face in daily life. Automatic reasoning is our first response to most events and occurrences. It can be overridden by deliberative reasoning in novel, challenging, or exceptional situations,

[46] John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998), 207 - 239

[47] Megan M. Thomspon, Tonya Hendriks & Ann-Renee Blais, "Military Ethical Decision Making: The Effects of Option Choice and Perspective Taking on Moral Decision-Making Processes and Intentions," *Ethics & Behavior* 28/7 (2017): 578-596, 579.

[48] Noel Sharkey, "Guidelines for the human control of weapons systems," *ICRAC Briefing Paper* (April 2018).

but, in short, it is the go-to mode for making decisions. And in our interactions with computing machines, this type of reasoning often wins out. The faster and more autonomous the operational mode of the system, the less likely it is that deliberative reasoning will play a role in decision making.

When it comes to decisions of a lethal nature, understanding the properties of automatic reasoning is important. Automatic reasoning tends to cut corners by sidelining ambiguity and doubt, by assimilating fragments of information into a familiar coherent narrative, and by ignoring absent evidence. This is particularly pronounced in our interaction with computational technologies. Studies have consistently shown that there is a tendency for humans to place uncritical trust in computer-based decision systems (automation bias), as we have a tendency to ignore, or not search for, contradictory information in light of a computer-generated solution, especially in "time-critical decision support systems."[49] This applies not only to autonomous or automated systems, but also to 'mixed-mode' systems where the human is in the loop to review the decisions,[50] and is particularly pronounced in systems with high levels of autonomy in decision making, such as AI systems. In other words, the cognitive asymmetry between humans and AI systems produces outcomes that are invariably skewed in favour of machine decisions. As Missy Cummings notes, this automation bias may be benign in situations where technological autonomy and automation is used for mundane and repetitive tasks, but when lethal decisions are at stake, automation bias risks the loss of situational awareness and may promote the degradation of important skills necessary for the ethical conduct of warfare, as human operators are unable to form an appropriate mental model in conditions of accelerated action and are unlikely able to override a potentially unethical decision by the system.[51] The authority of knowing the parameters for the 'right' moral decision is assigned to the technological system, which effectively becomes the authoritative decision expert.

The literature engaging with the complexities of the human-machine relationship and the associated cognitive challenges is growing and helps explain the finer mechanisms at play in this assemblage, albeit often with an emphasis on the technological aspect. Noteworthy for her focus on human capacities is Lucy Suchman's careful study of human-machine interfaces, in which she identifies the challenges specific to human engagements with computational robotic artifacts and an associated tendency to ascribe authority to the technology.[52] With a particular focus on 'expert' systems, she explains that in interaction with computational robotic machines, intentionality and intelligence is attributed to the machine and 'its' decision authority on the basis of partial evidence and the human tendency to consider opaque technological entities as entities with intentionality. As humans, Suchman notes, we are opaque to one another in our mental states — we cannot "see inside each other's heads."[53] This opacity requires, then, an interpretive approach to consider one another's intentions. In other words, we are always underspecified to one another. And as we have seen earlier, so are the decision-capable machines we engage with and to which we attribute some interpreted intentionality. While each digital system is underspecified in its various actions as a system, the human considers not the individual system processes, but the 'behavior' of the overall

[49] M.L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," American Institute of Aeronautics and Astronautics, *AIAA 3rd Intelligent Systems Conference Chicago* (2004).
[50] Kevin Miller, "Total surveillance, big data and predictive crime technology: Privacy's perfect storm," *Journal of Technology, Law and Policy* 19 (2014): 105-146.
[51] M.L. Cummings, "Automation Bias."
[52] Lucy Suchman, *Human-Machine Reconfigurations: Plans and Situated Actions* (Cambridge: Cambridge University Press, 2007), 42
[53] Ibid.

underspecified system and its intentions. Suchman explains: "Once reified as an entity, the inclination to ascribe actions to the entity rather than its parts is irresistible,"[54] thereby personifying or anthropomorphizing the machine and its decision-making capabilities. Just as is the case with a human, we need not know the inner working of "the mechanism, insofar as one need assume only that the design is rational."[55] An interpreted intentionality determines the interaction practically and theoretically: "Practically, it suggests that, like the human actor, the computer should be able to explain itself, or the intent behind its action to the user. Theoretically, it suggests that the computer actually has intent, as demonstrated precisely in this ability to behave in an accountably rational, intelligible way."[56] This, however, is a construct and possibly a fallacy with significant consequences. As we have seen above, with AI-enabled systems, we cannot readily assume that the technology's rationality is coherent with human modes of reasoning, nor do we have adequate information about the factors on which this perceived rationality rests. Considering this intricate relationship, we might better understand how technological artifacts come to gain authority in decision-making. The more sophisticated the technology, the more the machine is seen "as an expert, and the user as a novice or student."[57]

With most AI decision-making systems in use today (recent work on explainable AI notwithstanding), the technological decision framework is highly opaque to the user, usually "embedded at the backend of systems, working at the seams of multiple data sets, with no consumer-facing interface. Their operations are mainly unknown, unseen, and with impacts that take enormous effort to detect."[58] With this opacity, we infer reason and intentionality to an extent that is not warranted for systems that are highly unpredictable as they are set to act on complex environments. The degree to which the human agent has full capacity to act as a rational moral human agent — overriding and guiding the expert system — with sufficient cognitive ability to judge the situated action on which the technology is set to act, is called into question. Intelligibility of process matters for meaningful human control. As Wiener also pointed out, understanding the 'tactics' of the machine is crucial for the strategic planning of any human-machine action. Otherwise, undesired consequences are always the likely outcome.[59]

As highlighted earlier, computational settings and interfaces produce hybrid cognitive assemblages, which prompt the user — often imperceptibly — toward specific ideas of desired behaviour, which may not be morally appropriate for the dynamic and complex challenges of a fast-paced conflict scenario. This attribution of agency and blame to technology has been extensively discussed in the context of drone warfare, where, as Hugh Gusterson explains, an overreliance on the authority of technology, paired with processes of "narrative infilling and remote individualisations" by operators, produce a condition in which, once a technological frame has been put in place, "ambiguous information is being interpreted within that frame, informational gaps are ignored and moral judgements are rendered."[60] Perceptions of agency in this complex human-machine-decision relation shift toward the machine, as levels of autonomy increase. An influential 2017 study on moral agency perceptions and autonomous weapons confirms that military personnel, in particular,

[54] Ibid.
[55] Ibid.
[56] Ibid. 43.
[57] Ibid. 45.
[58] Kate Crawford and Meredith Whittaker, "Artificial Intelligence is hard to see," *Medium* (September 11, 2016).
[59] Wiener, "Some Moral and Technical Consequences of Automation."
[60] Hugh Gusterson, *Drone: Remote Control Warfare*, (Cambridge: MIT Press, 2016), 69.

perceive autonomous weapons to have a 'mind' and agency in terms of acting on morally relevant decisions, thereby obscuring the question of who may have the requisite knowledge to act meaningfully to control the technology or the outcome.[61]

These human cognitive limitations in the human-machine assemblage — the tendency toward automatic reasoning and the affinity toward ascribing agency and authority to the machine — might potentially be overcome with training, expertise, and extensive experience working with specific technological systems. However, the smooth functioning of an AI system is heavily reliant on frequent software updates, algorithmic fine-tuning, and error mitigation,[62] such that the operator 'on' or indeed 'in' the loop may find herself unable to "develop an appropriate mental model which is crucial to overcome system failure."[63] The human operator may not have the knowledge needed for adequate moral awareness, nor for appropriate moral judgement. They may not even have enough time to process the information required to exercise control. Indeed, the vastly divergent times scales in human and machine operations might even rule out "effective control of our machines."[64]

B. Epistemic and Temporal Limitations

With reference to the Observe-Orient-Decide-Act (OODA) loop model of decision-making, Richard Breton and Éloi Bossé consider what it takes for a military operator to make a decision within an environment populated by technologies with varying degrees of automation. For Breton and Bossé, the standard OODA loop is too narrow and abstract to have traction for understanding the mental processes required for meaningful interventions into automated processes. Rather, these processes are marked by a non-linear and cross-referential dynamic of matching new information and representations with established memory. Whether a human operator can meaningfully intervene in such a process depends in no small part on the mental model they are able to establish in relation to the technological device.

Broadly speaking, where the autonomous technology provides a representation of the real world, the human decision maker is able to intervene into the technological process only if he or she is able to identify features that are meaningful and familiar to the decision-maker. Such features are "meaningful and familiar if in the decision-makers long term memory a mental model that matches the situation can be activated."[65] Only then can a decision-maker intervene meaningfully in a morally relevant decision. This means that it takes time for an operator working with and within technology to understand the complexities of its operations on the real world. New mental models can be established for new technologies and new contexts, but this takes time, and it requires an ability to act on newly acquired memory and knowledge. Unlike in AI systems, learning in humans takes considerably longer and involves embodied modes of learning and experiencing the world. This crucial distinction in how we perceive the world matters. As Wiener notes, "the human brain is a far more efficient control apparatus than is the intelligent machine when we come to the higher areas of logic."[66] This

---

[61] Ilse Verdiesen, "Agency Perception and Moral Values Related to Autonomous Weapons: An empirical study using Value-Sensitive-Design approach," Masters Thesis, submitted to Delft University of Technology (August 28, 2017).

[62] Matteo Pasquinelli, "How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence": 1-17.

[63] Richard Breton and Eloi Brosse, "The Cognitive Costs and Benefits of Automation," *NATO, RTO-MP-088* (October 2003).

[64] Wiener, "Some Moral and Technical Consequences of Automation."

[65] Ibid.

[66] Wiener, "Some Moral and Technical Consequences of Automation."

continues to hold true. The brain is a "self-organising system which depends on its capacity to modify itself into a new machine rather than on ironclad accuracy and speed in problem-solving." The 'data' a human has available for decision making is vastly different and more comprehensive than any technological system to date can offer. Humans are able to draw on embodied knowledge to improvise and act flexibility in any given situation and are, in principle, more sensitive to context.[67] Where the human is fitted within the technical logic of an AI weapons systems, this capacity is compromised.

As we have already seen above, AI systems 'learn' and assess data differently, they work through abstractions of the world as a computable model. AI systems rely on the existence and availability of large amounts of data in order to perform an evaluative or predictive analysis of any given situation. AI is trained to capture the present through the lens of past data, to identify patterns and make efficient, future oriented assessments. This carries a number of ramifications. First, it prioritises a datafication of the environment upon which the AI is set to work. This means that the context within which the AI system is tasked to make an assessment or decision needs to be labeled and categorized so that it can be read as data. By default, then, everything that is not easily rendered in a numerical data format falls outside of any algorithmically determinable decision, as it remains invisible to any AI system. Moreover, the quality, origin and quantity of the data available matters. While the process of AI labeling and classification is quite advanced and reliable for fixed categories (a chair, a cat, a bird, and so on), it is less easy to train an AI to 'understand' relational and more fluid dimensions of life (friendship, enmity, identity, culture, social relations).[68] Decisions based on incomplete data or falsely labeled data may lead to biased and unfair outcomes. If an AI builds a world model, based on available data, it is likely to be much more successful in closed systems where parameters can easily be grasped as data. In the context of warfare, where parameters are likely to be more fluid and dynamic, the AI system may suggest a course of action based on an epistemic foundation that may be biased, incomplete, or otherwise not fully appropriate to the situation. There is already ample evidence that state-of-the art AI facial recognition systems, for example, produce biased and potentially erroneous outcomes.[69] Knowing where the data comes from, and how it is operationalized matters for moral awareness, effecting one's ability to identify situations where a morally problematic outcome is at stake. Having sufficient knowledge and understanding of on what grounds a system calculates object identifications and makes recommendations or, indeed acts on them, is relevant for knowing when to intervene and why (recall here our earlier example of the wolves or the tanks).

Consider, for example, recent studies which show that, generally speaking, AI systems in autonomous vehicles are less reliable in identifying pedestrians with darker skin colours than they are with lighter skin shades. The underlying data-bias would endanger people with darker skin tones more substantially than those with lighter shades of skin. This certainly constitutes a moral challenge which is masked within the algorithmic processes at work. An AI system recommending a kill decision is different than a human doing the same to an operator. With regards to the human commander, the operator is likely to have some memory and sufficient knowledge of the commander's background, training, position, preferences, biases and so on. If a commander asks an operator to engage in a kill decision that disadvantages people with

[67] M.L. Cummings, "Automation Bias": 1.

[68] Kate Crawford, "You and AI: Machine Learning, Bias and Implications for Society," Talk held at *The Royal Society* (July 17, 2018).

[69] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi "The ethics of algorithms: Mapping the debate," *Big Data & Society* 3/2 (2016); Tom Simonite, "How Coders are Fighting Bias in Facial Recognition Software," *WIRED* (March 28, 2018).

a darker skin colour, the operator may have adequate knowledge to assess whether this command should be challenged on moral grounds. The opacity of the AI system complicates this agency. Where the human operator is called on to work as a check point or control mechanism for a morally challenging decision by the AI system, he or she needs to have the requisite knowledge to make a moral judgement. Understanding the system's data infrastructure is relevant in order to be able to predict and understand the system. This, in turn, is relevant to be able to foresee the possible outcomes, which, as we have seen above, is challenging with complex autonomous systems.[70] A small blind spot in the system's dynamic set up can have considerable life and death consequences.[71] It is difficult to challenge an unjust outcome where the potential for injustices is contained within the system's design. Consider, for example, the flash crash in 2010, where high-frequency trading algorithms managed to wipe out 600 points on the Dow Jones Index within a matter of minutes. It took years to get to the heart of what caused the costly event. For human decision makers to be able to retain agency over the morally relevant decisions made with AI they would need a clearer insight into the AI black box, to understand the data, its provenance, and the logic of its algorithms, and have time to evaluate contexts and act on them while embedded in a technological setting where information is mediated through screens at high speeds. Simply knowing what outcome has been specified for the system is not enough. As Wiener notes, "to avoid disastrous consequences, it is not enough that some action on our part should be sufficient to change the course of the machine, because it is quite possible that we lack information on which to base consideration of such an action."[72]

Related to the above, and perhaps the most obvious challenge to meaningful human control is the matter of speed. The main allure and logic of AI enabled systems is the ability to outpace and outmaneuver the enemy with superior speed and efficiency. Where speed is a key factor, time horizons for decision making inevitably shrink. This too was of paramount concern to Wiener, who clearly warns that "when a machine constructed by us is capable of operating on its incoming data at a pace which we cannot keep, we may not know until too late, when to turn it off."[73] This is already evident with automated fire-and-forget systems like *Phalanx* or *SeaRAM*, which complete their missions within a few seconds. Where the decision loop consists of technologically gathered and evaluated data, which is deemed fit to be acted on, in taking out a target, in a matter of second, and where speed is prioritized as a core value for an operation, it becomes very difficult for a human operator to exercise control over the system, morally or otherwise. Such collapsed time horizons are likely to be exacerbated with systems that make calculations in nanoseconds, eclipsing any horizon for timely, meaningful intervention. The machine is too fast for the human mind to keep up with. This is what Wiener was most concerned with as he worried that: "[b]y the time we are able to react to information conveyed by our senses and stop the car we are driving, it may already have run head on into a wall."[74]

But there is also another temporal discrepancy at work here. The logic of the AI system operates on a different temporal measure than the human, let alone human relations and human histories. AI-enabled weapons systems are not merely diagnostic or descriptive; they are designed to be predictive and prescriptive, and as such intervene into human affairs and shape or direct human action. In their predictive and prescriptive capacities, they become

[70] Holland Michel, *Black Box Unlocked*, 10-11; Saariluoma, "Four Challenges": 236.
[71] The Physics arXiv Blog, "Google Reveals Major Hidden Weakness in Machine Learning," *Discovery Magazine* (November 30, 2020).
[72] Wiener, "Some Moral and Technical Consequences of Automation."
[73] Ibid.
[74] Ibid.

distinctly future oriented, prompting the human toward action rather than deliberation. In so doing, they produce a sense of urgency, whereby suggested action points need to be taken in good faith on digital memory and information. And within the setting of a lethal autonomous weapons system, the outcome is likely to be more lethal, rather than less. Recall that greater lethality is, of course, the stated aim of the US DoD military AI initiative. Accelerated temporal action horizons stand in contrast to a less speedy process of human ethical reasoning. In other words, the implicit logic of speed, volume, and efficiency of AI systems and processes stands in stark tension with the slow, complex and unwieldy challenges of social and political institutions, including warfare and the ethical deliberations required therein.

Conclusion: The (Im)Possibility of Moral Responsibility

And so, the ultimate question: to what degree are we able to act as moral agents in the use of lethal autonomous intelligent weapons? If we cannot readily understand or predict how intelligent LAWS might interact not only with the contingent, dynamic environment of warfare but also with our human capabilities and limitations, if we are unable to intervene in a timely manner, if we are unable to challenge an algorithmic decision on its technological authority, is it possible to retain the level of human control required for a morally meaningful decision? I am doubtful. This is not to say that the notion of meaningful human control, as a legal concept, is of no use. Indeed, it can function as an important means of safeguarding a minimum of human involvement in the face of an ever-escalating drive for autonomy by advanced militaries, and it is an important and useful pushback to the narratives that full technological autonomy is inevitable and desirable for humanity. With this article, however, I aim to push further and highlight where the limits to human control over such intelligent technologies reside. Conditions that would safeguard or indeed promote human moral agency cannot be ensured with LAWS. On the contrary, the capacity to take responsibility and feel the weight of a morally complex decision becomes more difficult as these decisions are distributed and mediated through technological interfaces, nodes, and various system components. The human is integrated as an .exe file into a technological ecology that is largely invisible, and which operates far beyond human capacities.

We are left therefore with something of a paradox. In trying to reduce or mitigate morally risky decisions through the use of an emerging technology, we risk losing our moral capacities and competencies in the process. The unforeseen consequence of this may well be a de-skilling of moral faculties in decisions over life and death in warfare. Rather than retaining or ensuring meaningful human control over lethal decisions with LAWS, we risk abdicating our responsibility for such decisions altogether, succumbing instead to a requirement for speed and efficiency more in line with the logic of machines than human values, forged over centuries. As warfare becomes increasingly conceived in systematic terms, through digital networks and algorithmic architectures, we should be mindful that these architectures might affect our ethical thinking and acting in ways that move ever-further away from a humanist framework, edging instead toward the purely calculative logic of machines and causing our capacity for moral agency to atrophy as a result. Killing should always remain a troubling and morally challenging act, not an easy technological option or a choice between a number of technologically ascertained recommendations. To conclude, I come back one more time to the originator of what we now know to be AI — Norbert Wiener — whose warning we would do well to heed in the discussions on LAWS: "for man […] to throw the problem of his responsibility on the machine, whether it can learn or not, is to cast his responsibility to the wind and to find it come back seated on the whirlwind."[75]

[75] Wiener, *The Human Use of Human Beings*, 185.

References

BAE Systems. "BAE Systems selected to provide autonomy capabilities for DARPA's Squad X Program." *Bloomberg* (June 2, 2020).

Boulain, Vincent, Verbruggen, Maaike. *Mapping the Development of Autonomy in Weapons Systems*. SIPRI Report: November 2017.

Breton, Richard, Brosse, Eloi. "The Cognitive Costs and Benefits of Automation." *NATO, RTO-MP-088* (October 2003).

Bryson, Joanna. "Robots Should be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks. Amsterdam: John Benjamins, 2010.

Cheney-Lippold, John. "Accidents Happen." *Social Research: An International Quarterly* 86/2 (2019): 513-535.

Crawford, Kate. "You and AI: Machine Learning, Bias and Implications for Society." Talk held at *The Royal Society* (July 17, 2018).

Crawford, Kate, Whittaker, Meredith. "Artificial Intelligence is hard to see." *Medium* (September 11, 2016).

Cummings, Mary. "Artificial Intelligence and the Future of Warfare." *Chatham House Research Paper* (January 2017).

Cummings, M.L. "Automation Bias in Intelligent Time Critical Decision Support Systems." American Institute of Aeronautics and Astronautics, *AIAA 3rd Intelligent Systems Conference Chicago* (2004).

D'Amour, Alexander, et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning." arXiv:2011.03395 [cs.LG] (November 24, 2020).

DARPA. "With Squad X, Dismounted Units Partner with AI to Dominate Battlespace." *DARPA* (2019).

Department of Defense. "Industry Day for the Advanced Targeting and Lethality Automated System (ATLAS) Program." *Department of Defense* (2019).

Dieter, Michael, Gauthier, David. "On the Politics of Chrono-Design: Capture, Time and the Interface." *Theory, Culture & Society* 36/2 (2019): 61-87.

Feenberg, Andrew. *Transforming Technology: A Critical Theory Revisited*. New York: Oxford University Press, 1991.

Fischer, John Martin, Ravizza, Mark. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.

Freedberg Jr, Sidney J. "ATLAS: Killer Robots? No. Virtual Crewman? Yes," *Breaking Defense* (March 4, 2019).

——. "Fear and Loathing in AI." *Breaking Defense* (March 6, 2019).

Fry, Hannah. *How to be Human in the Age of the Machine*. London: Transworld Publishers, 2018.

Ganesh, Anyana, McCallum, Andrew, Strubell, Emma. "Energy and Policy Considerations for Deep Learning in NLP." *57th Annual Meeting of the Association for Computational Linguistics (*ACL) (Florence, Italy: July 2019).

Ganesh, Maya Indira. "The ironies of autonomy." *Nature: Humanities and Social Sciences Communications* 7/157 (2020).

Gunkel, David J. "Other Things: AI, Robots and Society." In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, edited by Zizi Papacharissi. New York: Routledge, 2019.

——. *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. Cambridge: MIT Press, 2017.

Hayles, Katherine. "Cognitive assemblages: Technical agency and human interactions." *Critical Inquiry* 43/1 (2016).

Henderson, Ryan. "Picasso: A free open-source visualizer for Convoluted Neural Networks – Cloudy with a chance of tanks." *Merantix – Medium.com* (May 16, 2017).

Himma, Kenneth Einar. "Artificial agency, consciousness and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11/1 (2009): 19-20.

Holland Michel, Arthur. *The Black Box, Unlocked: Predictability and Understandability in Military AI*. Geneva, Switzerland: United Nations Institute for Disarmament Research, 2020.

Horowitz, Michael C., Scharre, Paul. "Meaningful Human Control in Weapon Systems: A Primer." *Center for a New American Security (CNAS),* Working Paper (March 2015).

Houser, Kristin. "US Military: Our 'Lethality Automated System' Definitely isn't a Killer Robot." *The Byte* (2019).

Gusterson, Hugh. *Drone: Remote Control Warfare*. Cambridge: MIT Press, 2016.

Leveringhaus, Alexander. "What's So Bad About Killer Robots?" *Journal of Applied Philosophy* 35/2 (2018).

Metz, Cade. "Google's AI Wins Pivotal Second Game in Match With Go Grandmaster." *WIRED* (October 3, 2016).

Miller, Kevin. "Total surveillance, big data and predictive crime technology: Privacy's perfect storm." *Journal of Technology, Law and Policy* 19 (2014): 105-146.

Mittelstadt, Brent Daniel, Allo, Patrick, Taddeo, Mariarosaria, Wachter, Sandra, Floridi, Luciano. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3/2 (2016).

Pasquinelli, Matteo. "How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence." *Spheres* 5 (2019): 1-17.

Ribeiro, Marco, Singh, Sameer, Guestrin, Carlos. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (2016). Stroudsburg, Pa.: Association for Computational Linguistics. doi:10.18653/v1/N16-3020.

Roff, Heather, Moyes, Richard. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons Systems." *Briefing Paper for Delegates for the CCW GGE Meeting on LAWS* (April, 2016).

Saariluoma, Pertti. "Four Challenges in Structuring Human-Autonomous Systems Interaction Design Processes." In *NATO Allied Command Transformation*. The Hague: NCI Agency, 2015.

Schwarz, Elke. "Technology and moral vacuums in just war theorising." *Journal of International Political Thought* 14/3 (2018).

Shah, Dhruv. "AI, Machine Learning, & Deep Learning Explained in 5 Minutes." *Medium – Becoming Human* (April 3, 2018).

Shanahan, Jack. "Lt. Gen. Jack Shanahan media briefing on A.I.-related initiatives within the Department of Defense." *US DEPT OF DENSE* (August 30, 2019).

Sharkey, Noel. "Guidelines for the human control of weapons systems." *ICRAC Briefing Paper* (April 2018).

Simonite, Tom. "How Coders are Fighting Bias in Facial Recognition Software." *WIRED* (March 28, 2018).

Strout, Nathan. "Inside the Army's futuristic test of its battlefield artificial intelligence in the desert." *C4ISRNET* (September 26, 2020).

Suchman, Lucy. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge University Press, 2007.

Sullins, John. "When is a robot a moral agent?" *International Review of Information Ethics* 6/12 (2006).

The Physics arXiv Blog. "Google Reveals Major Hidden Weakness in Machine Learning." *Discovery Magazine* (November 30, 2020).

Thomspon, Megan M. , Hendriks, Tonya, Blais, Ann-Renee. "Military Ethical Decision Making: The Effects of Option Choice and Perspective Taking on Moral Decision-Making Processes and Intentions." *Ethics & Behavior* 28/7 (2017): 578-596.

Verdiesen, Ilse. "Agency Perception and Moral Values Related to Autonomous Weapons: An empirical study using Value-Sensitive-Design approach." Masters Thesis. Submitted to Delft University of Technology (August 28, 2017).

Wiener, Norbert. "Some Moral and Technical Consequences of Automation." *Science* 131 (1960): 1355-1358.

——. *The Human Use of Human Beings*. Cambridge: Da Capo Press, 1954.