Pre-publication Version – Accepted for publication in Philosophy Today (forthcoming 2020/21)

Silicon Valley Goes to War: Artificial Intelligence, Weapons Systems and the De-Skilled Moral Agent

Elke Schwarz

Introduction

In 1960, Norbert Wiener, the father of cybernetics, cautioned in the strongest terms against unthinkingly rushing to use computational power as a substitute for human decision-making in the context of war and politics. In particular, he warns that the time scales and reasoning capacities in cybernetic systems and humans are so distinct from one another that the potential for disaster is high. He urges that "we must always exert the full strength of our imagination to examine where the full use of our new modalities may lead us",[1] as our new machines might well be both very effective and highly dangerous. For Wiener, this was not merely a question of scientific practicality, but a moral question of significant weight. Throughout his work on cybernetics, Wiener was interested in the intersection of and interfaces between humans and machines. As these become more tightly enmeshed in our contemporary times, the question of shifting and mediated agencies, ethical and otherwise, looms large, especially so in the conduct of war.

The drive toward greater automation and autonomy in weapons systems has a long history, whereby military aims and cutting-edge research in autonomy and AI have traditionally co-evolved. While presently a lot of AI innovation is driven by the private sector, this was not always the case. In the early years of AI development, AI and military developments were intricately entwined. Turing's conception of the possibility of computing machines evolved from his wartime work on decoding; John von Neuman's work on system design and programming derived from his high-level military work for the Manhattan Project; and some of Wiener's contribution to systems automation stemmed from wartime advances in automatic gun fire control.[2] But not all involved in this twinned development were at ease with the use of increasing automation, if not autonomy, in the context of warfare. At the heart of Wiener's concern was the question of human control, especially as he saw the temporal structures of cybernetic systems outpacing human cognitive abilities. What are the consequences, he asked, if we avail ourselves of a "mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete"?[3] As advances in machine learning and computational processing power continue to progress at

[1] Wiener, "Some Moral and Technical Consequences of Automation", 1960.
[2] Noble, *The Religion of Technology*, 1999, pp. 152-153.
[3] Wiener, "Some Moral and Technical Consequences of Automation", 1960

pace, this distance between machine and human capacities widens even further. The machine is cast as a superior decision-making agent, over the slow and deliberative processes implicit in human thought; a superior decision-making agent which inevitably comes to replace the human in a growing realm of tasks and decisions. As is the case with most of our new digital dependencies, this is born out of a sense of both inferiority and convenience.

The philosopher Günther Anders identified this exact condition already in the 1960s when he diagnosed the modern human as suffering from what he calls "Promethean Shame".[4] This 'shame' reflects the modern experience of finding oneself always already inferior to the industrial products human society produces. These products are perfectible, they can be fixed and, most importantly, they can be and indeed are perpetually improved. Our products and technologies effortlessly and continually follow the logic of material progress, and progress, understood as linear improvement, is what underpins the modern condition as a primary value. As humans, we are not afforded such transformational capacities, and in the face of this constraint, we are ashamed. We are not made, we are born, with limited functionalities. Our form and capacities as humans have not changed significantly over the centuries. We are outperformed and outlasted by our artifacts and, increasingly, our material products. Our bodies decay, inevitably. In comparison to our products, we are always obsolete. As Anders notes,

> The desire of modern man to become self-made, to become a product has to be considered faced with this changed foil: humans want to make themselves, not because they do not tolerate anything that is not human-made, but rather because they do not want to remain un-made. Not because they are ashamed to be made by others (God, gods, Nature), but because they are not made at all and as un-made humans remain inferior to their own fabrications. What is evident here is a variant of a classic confusion: the inversion between *creator and creatum*.[5]

In this condition, modern humans measure their worth, and that of their actions, on the basis of the flawless, efficient, and perfectible functionality and logic of machines. As the technological environment becomes more intricate and pervasive, so too does the realm of actions to which this shameful inferiority applies. This mode of thinking bears a number of consequences and one of these manifests in our contemporary digital environment: a shift into a regressive state of incompetence, impotence, and sense of inconvenience when faced with problems or dilemmas. We are de-skilling ourselves *qua* humans. In this precarious condition, the unsettled human is more likely inclined to seek authority and absolution – in the political and moral sense – in technological systems that seem to offer solutions. As Christian Müller notes in his discussion of Anders' work: "the more perfect machines become, the less perfect humans need to be and the more limited they begin to look; the more capable machines are, the more they have the potential to produce incapacitated humans".[6] Human inferiority, convenience, and incapacity work in concert here and smooth a path

---

[4] Anders, *Die Antiquiertheit des Menschen*, 1961.
[5] Ibid., p. 25, author's translation.
[6] Müller, *Prometheanism*, 2016, p. 128.

toward less, not more, human responsibility. Again, Anders diagnoses this condition presciently when he points toward 'oracle machines' – computational decision systems – as 'electronic conscience automatons',

> because that is precisely what these cybernetic computers – the pinnacle of science (as well as that of progress and, by extension, inevitably morality) – are, which are now tasked to effortlessly take responsibility for decisions, while the human stands aside, washing his hands of the burden, both relieved and triumphant in equal parts.[7]

Writing in the same year as the term Artificial Intelligence was coined, in 1956, Anders offers one of the first focused critical engagements of how AI might mediate our moral agency, in day-to-day tasks, as well as in warfare as he considers the modern temptation to relinquish human conscience and agency "at the altar of the apparatus"[8] which in turn becomes a proxy-conscience for the human.[9]

With this framing in mind, this paper focuses less on the latest specifics of technological developments and AI capabilities in military contexts, instead foregrounding how digital military technology, including AI, mediates and shapes us as humans, specifically our capacity to act and decide as moral agents. My core claim is that as digital infrastructures and interfaces dominate the human (military) landscape, not only is human moral agency diminished, but ethical practices become cast in distinctly technological terms. Following Anders, I argue that as the human becomes digitally co-machinistic (*mitmaschinell*), he is compelled to adopt the logic of speed and optimisation for ethical reasoning and as a result becomes de-skilled as a moral agent. This a concerning development particularly when it comes to the conduct of war and the use of violent means where the moral stakes are high.

I begin by briefly tracing the role of AI in military operations as it is both in place today and envisioned for the future, with a focus on targeting systems in the US context, before then moving on to exploring these developments and visions for human decision making and agency in collaboration with AI systems. In the final section, I focus more specifically on the nexus of moral agency and AI, highlighting some of the key tensions between digital technological systems and ethical practices. I argue against contemporary attempts to think 'ethically' with and through digital machine logics, which are not only doomed to fail but doomed to do so in ways that preclude ethical reasoning altogether.

Military AI: Between Utopia and Dystopia

The US and China are presently leading the way in shaping global ambitions for military AI use, with Russia following closely on their heels and other military powers – France, the UK, and Israel, among others – investing increasing sums in military AI. As one of the key drivers of AI in military operations, the US is both vocal about its ambitions and representative of

---

[7] Anders, *Die Antiquiertheit des Menschen*, 1961, p. 245. [author's translation]
[8] Ibid., p. 60
[9] Schwarz "Günther Anders in Silicon Valley", 2019.

future global visions of AI for warfare. In 2018, the US established the Joint Artificial Intelligence Center (JAIC) to serve as the Department of Defense (DoD) hub for the digitisation of US warfare. The Center's stated aim is to shift the DoD's "mindset and culture from a hardware-centric and industrial-age force to a software-driven and information-age one".[10] The goal is to make the business of warfare lighter, more agile and, importantly, much faster. In general terms, AI for military purposes can be used to undertake tasks that involve *description* (pattern recognition, voice/object identification), *prediction* (the forecasting of future states)[11] or *prescription* (analysis and subsequent decision). Already it is easy to see how AI can be used to streamline infrastructural and logistical processes, such as supply chain management, optimised communication, or the preventive maintenance of fighter jets, to give just one concrete example.[12] In short, for the faster and more efficient undertaking of non-combat activities and processes, AI promises to make unwieldy and logistically complex operations much leaner. But when it comes to combat activities and the use of AI for tactical, operational, and/or strategic gains, the use of AI in its descriptive, predictive, or prescriptive dimensions is not necessarily an advantage, especially when it comes to tasks that require complex knowledge and expertise-based reasoning in highly dynamic environments.[13]

Particularly pertinent here is the question of whether AI systems can or indeed should be taking on a significant role in critical selection and targeting functions, whether they should be making lethal decisions, be involved in "accelerated sensor-to-shooter timelines",[14] or play a crucial role in predictive suspect selection and any other tasks in which the ethical stakes are patently high. In all of these examples, the paramount question is what the implications of this might be for human oversight, control, agency, and responsibility. 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 increasingly let the AI make a pre-selection of possible targets at an accelerated pace to increase the lethality of the systems. A number of programmes are currently in place that work toward this goal, such as the U.S. Army's Advanced Targeting and Lethality Automated System (ATLAS), the Pentagon's military AI pioneering Algorithmic Warfare Cross-Functional Team programme, better known as Project Maven, and other defence industry programmes that seek to pair advancing drone technology with intelligent capabilities, as illustrated by the recent test flight of an MQ-9 Reaper, equipped with an Agile Condor Pod which enables the integrated system to autonomously find, track,

---

[10] For example, see the About section on the website for the DoD's Joint Artificial Intelligence Center (JAID), https://www.ai.mil/about.html

[11] For a focused discussion of predictive AI for counterterrorism operations see McKendrick, "Artificial Intelligence, Prediction and Counterterrorism", 2019.

[12] Consider, for example, the Autonomic Logistic Information System (ALIS) – a 'smart' system designed to help with the maintenance of F-35s, taking into consideration multiple data inputs in order to streamline maintenance tasks. See Tucker, "The F-35 is about to get a lot smarter", 2018.

[13] Cummings, "Artificial Intelligence and the Future of Warfare", 2017.

[14] Shanahan, "Lt. Gen. Jack Shanahan media briefing on A.I.-related initiatives within the Department of Defense", 2019.

and propose a target to the human operator. The stakes are high and the financial rewards potentially enormous for commercial providers. For example, the Pentagon contract for a full cloud infrastructure – the Joint Enterprise Defense Infrastructure (JEDI) – is worth US$10 billion, and with the Pentagon determined to expand its algorithmic warfare capacity, major industry players like IBM, Microsoft, Amazon, and Google are likely to be vying for similarly lucrative AI contracts in the near future. Aggressive newcomers to the military AI market, such as Palantir and Anduril, look set to further strengthen these links between Big Tech and Military Tech.[15]

Although less is known about specific military AI operations in other countries, there are programmes and products that look similar to the US' aims to harness AI for the identification and lethal engagement of potential targets. The Russian weapons manufacturer Kalashnikov, for example, has developed a "fully automated combat module based on neural-network technologies that enable it to identify targets and make decisions".[16] Another major player, China, is forcefully "pursuing AI-enabled systems and autonomous capabilities", which include AI facilitated targeting technologies.[17] In short, there is a global push to develop AI-enabled weapons systems which make it easier to identify potential targets at higher speeds and with more efficiency, as well as a belief, shared among policymakers and industry proponents alike, that such AI functions must be further developed in order to gain an upper hand in the ostensibly inevitable algorithmic context of future wars.[18] Yet this vision of an effortless integration of AI into military operations, especially where lethal decisions are at stake, is marred by the complexities and ethical challenges that any context of conflict produces.

The Artifice of Military Intelligence

At present, the capacities, robustness, and reliability of AI for complex military operations, such as enemy identification and targeting, are still under-developed and reliant on enormous levels of computational processing power. While narrow AI might be used quite effectively in predictive maintenance or logistical operations, where the context is demarcated by a limited number of relatively stable variables, a more extensive use of AI in unpredictable, unruly, and highly dynamic combat environments remains an aspiration rather than a viable reality. The reasons for this are many, but it bears keeping in mind that AI is essentially a software programme premised on algorithms. It is an "entity (or collective set of cooperative entities), able to receive inputs from the environment, interpret and learn from such inputs, and exhibit related and flexible behaviours and actions that might help the entity achieve a particular goal or objective over a period of time."[19] This entails both traditional, symbolic

[15] Simonite, "The Pentagon Doubles Down on AI – and Wants Help from Big Tech", 2019. For a detailed report on the Silicon Valley/Military Industry nexus, see Poulson, "Reports of a Silicon Valley/Military Divide Have Been Greatly Exaggerated", 2020.
[16] Slijper, Beck and Kayser, "The State of AI", 2019, p. 18.
[17] Kania, "AI Weapons in China's Military Innovation", 2020.
[18] Suchman, "Algorithmic Warfare and the Reinvention of Accuracy", 2020.
[19] Faggella, "What is Artificial Intelligence?", 2018.

rule-based AI systems, as well as newer approaches to AI that draw on neural networks and advanced machine learning techniques. For these to work effectively, vast amounts of input data are needed on which algorithms are then trained. The accuracy and comprehensiveness of the input data available is one arena which poses a challenge to military AI. Datasets can be biased from the outset, they can produce biased outcomes, data might be incomplete or insufficient to produce an accurate world model upon which the AI is intended to work. Moreover, computer vision, although improving at a rapid rate, remains still brittle as a tool for reliable recognition and examples where a system can be spoofed by making minute changes to an image are not rare, as was effectively shown by experiment in which hackers fooled an AI algorithm into consistently misidentifying a turtle as a rifle.[20]

This poses not only a security concern, but also raises questions of accuracy in the complex and dynamic world of warfare. Hardware and network provisions in military contexts can be slow and unwieldy in the best of times,[21] yet in many cases new systems are brought into operation before persistent and serious technical failures are fully addressed.[22] Taken together, these observations suggest that triumphant visions of a fast and efficient military AI to lessen the fog of war are optimistic to say the least, yet the chorus for a full-spectrum integration of AI technology into warfighting continues to swell amidst fears of losing the AI arms race. This faith in AI as a panacea overlooks the fact that AI is indeed very likely to be error prone. This is a feature of AI itself and part of the culture of Silicon Valley. Speed is a core value, not only in AI processing, but also in getting a product to market. Facebook's motto, 'Done is better than perfect', is representative of the general ethos. And it is when military doctrine, organisation, and tactics are constructed around the logic of a certain 'master-weapon'[23], as AI might be cast in the accounts of some proponents, that we ought to pay attention to what practices, perspectives, and skills this weapon facilitates or hinders.

Increased computational processing speeds and the availability of vast amounts of data facilitate the illusion that the world can be rendered in all its facets, quickly, so that it can be acted upon. However, 'Artificial' Intelligence, is only ever just that – an 'artificial' way of knowing and acting upon the world. Digital technologies render the world never quite fully as it is, as it appears to us, collectively or individually. Rather, and this too was one of Anders' concerns about modern computational technologies, a quintessential aspect of computation machines is an "'idée fixe', that is to say, they impose determinants that are limited artificially by what they can do".[24] The upshot of such a quantification of morally relevant aspects is that only those elements can be considered that do not resist transformation into utility-relevant data. This in turn means "that (for example) the destruction of human life or the devastation of countries must solely be considered and

[20] Hutson, "A turtle or a rifle?", 2018.
[21] Conger and Metz, "Eric Schmidt's Pentagon Offensive", 2020.
[22] Consider here, for example, the F-35 US Fighter Jet, which continues to suffer technical errors and problems. Insinna, "Inside America's Dysfunctional Trillion Dollar Fighter-Jet Program", 2019.
[23] Coroalles, *Lectures on FSR III Revisited*, 1988.
[24] Anders, "On Promethean Shame", in Müller, *Prometheanism*, 2016, p. 59.

evaluated as profit and loss figures, for reasons of methodological clarity and purity".[25] In other words, the multiple, messy, unruly, and aleatory dimension of humanity must be pressed into a data logic for an ostensibly objective assessment.[26]

The very logic of AI – for military applications or otherwise – rests on this classification and codification of life into computable data; it employs modes of 'thinking' that are often entirely foreign to human deliberative processes, and that necessarily truncate entire spheres of interaction. An AI system cannot, for example, comprehend the world in complex relational dimensions. The vulnerabilities of AI systems pointed out earlier occur because an AI system does not 'think' the same way humans do, nor is the way it perceives the world the same as a human would. An AI system calculates and approximates. The very AI paradigm, as Matteo Pasquinelli highlights, is marked by "limits, approximations, biases, errors, fallacies and vulnerabilities".[27] What AI presents is a model of the world, an approximation of the perceived environment, calculated based on input data. Machine learning, Pasquinelli notes, calculates "not an exact pattern, but the statistical distribution of a pattern"[28], always truncating that which does not fit, pressing data into assorted categories through which an outcome can be achieved that is actionable on the ground. AI is as such a "technique of information compression".[29] The world is rendered through AI, along the limited possible axes of AI. Phenomena that cannot be captured numerically remain invisible to the AI system, data that is too messy is rendered pliable, categorized and calculable, but never without a loss. AI makes the world, as it perceives the world. The logic of AI asks the world to be captured as data, for patterns to be statistically ascertained and acted on, at speed. It is at this point perhaps all too evident that the human is equally pressed into a schema of optimized calculation and functionality in order to better align with machine logics, to become "co-machinistic (*mitmaschinell*)", as Anders put it in his letter to Adolf Eichman's son, Klaus.[30] The core principle of machines, he notes, is maximum performance (*Maximalleistung*), and according to this quintessential logic, their character is expansionist, that is to say they intend to expand their sphere of output to include ever-greater realms of the human world – an imperialistic ambition to compel all under their sway to work to the same tune of perfection and reliability as they do. It is, for Anders, a totalitarian ambition, realized through modern technologies, to bring the world into line with ostensibly objective logics of machine perfection. That which resists such an integrative techno-logic is rendered trivial, including the always inferior human qua human.[31] This is how we might understand the complexities of agency in the military AI human machine today.

Rather than amplifying human cognition through technology, military AI replaces it with a different and faster logical system of information processing and optimisation that privileges 'direct action', which in turn serves the aim of producing knowledge about possible actionable targets at speed, "to the detriment of an informed engagement with

---

[25] Ibid.
[26] Anders, *Die Antiquiertheit des Menschen*, 1961, p. 61
[27] Pasquinelli, "How a Machine Learns and Fails", 2019, p. 2.
[28] Ibid., p. 4.
[29] Ibid., p. 12.
[30] Anders, *Wir Eichmannsöhne*, 1988, p. 50.
[31] Ibid. p. 51.

relevant actors on the ground".[32] In this, the machine co-produces a specific, more violent mode of militarism, one that prioritises lethal action. Putting AI technology "at the edge"[33] of the weapons system is a logical next step in techno-solutionism, amplifying speed and embedding it ever deeper into our social and political world. Especially in irregular warfare, the risk of misidentification of legitimate targets through algorithmic identification is extraordinarily high, no matter how much data is available, and claims of accuracy are informed by a situated and perspectival reading of enmity through the lens of technology.

The prioritisation of machine logic also has other implications for agency and human cognition. Given technology's role in shaping representations of the world as a form of world-making, we cannot readily assume that the military agent operating in a setting of digital information structures and interfaces is able to act with instrumental autonomy over the technology he or she is embedded in.

Agency at the Interface

There is now an extensive and well-developed literature addressing the complex human-machine relationship in modernity, including the cognitive challenges humans face when working with computer technologies.[34] When AI and human reasoning form an ecosystem, the possibility for human control is limited. As humans, we have a strong bias in favour of our computing machines, and we often lack the knowledge needed to reason well enough to assert proper control over an action, particularly when the utility of the action depends so heavily on speed. As is the case with all technologies, or indeed artefacts, "the potential for harm lodges not solely in the inanimate object, but in the myriad of ways that people interact with them",[35] as Sheila Jasanoff explains. And this is particularly relevant against the background of advancements in neural network machine learning algorithms, which are sometimes deliberately designed to operate beyond the capacity of the engineer, let alone the user, to conceive of the computational process at work in the neural network. Within such technological systems, the human is no longer in full control and instead operates from inside a web of relations that prioritise technological parameters such as speed, optimisation, and efficient decision-making, and within which ideas of good or bad are always already fixed.

The work of Missy Cummings explores these limitations of human agency within the setting of intelligent decision support systems. She notes that although such systems can be useful in alleviating human fatigue, boredom, and cognitive shortcomings, they can also impose a "measurable cost to human performance … such as loss of situational awareness,

---

32 Suchman, "Algorithmic Warfare and the Reinvention of Accuracy", 2020, p. 6.

33 Ibid., p. 8.

34 See for example Suchman, *Human-Machine Reconfigurations*, 2007; Sharkey, "Guidelines for the Human Control of Weapons Systems", 2018; Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems", 2004; Miller, "Total Surveillance, Big Data and Predictive Crime Technology", 2014; Breton and Bossé, "The Cognitive Costs and Benefits of Automation", 2003. See also Wiener, *The Human Use of Human Beings*, 1954.

35 Jasanoff, *The Ethics of Invention*, 2016, p. 56.

complacency, skill degradation and automation bias".[36] These are considerable implications in terms of human agency for morally relevant acts, especially in such a morally charged context as warfare. In her study of automation bias in human-computer interactions, Cummings highlights the well-documented inclination of humans to afford the computer system a substantial role in determining a decision. As humans we have "a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct and can be exacerbated in time-critical domains".[37] Here again, we might consider how the logic of computation decisions at speed functions to privilege quick action over more time-intensive deliberation. And although this could potentially be overcome for specific systems with targeted training, the complexity of AI systems makes this much harder to achieve, especially as they are designed with a non-human and time-critical logic at their foundation.

Moreover, the embeddedness of human operators within interfaces of digital information complicates human deliberative processes. As Dieter and Gauthier show, interfaces are often designed to function as mediators to prompt users toward action over non-action, "complicating conventional notions of the rational, self-reflexive subject by operating beyond consciousness at vast environmental dimensions and accelerated micro-speeds"[38], thus conditioning the human to become a part of the wider technological interface infrastructure. In short, interfaces capture and 'conduct' human cognition and, in turn, behaviour. As is the case with any technology, the interface mediates between the human and computational information flows. Interfaces mediate between that which the computer captures as data and that which the human receives as intelligible and therefore actionable, but always with some range of distortion and loss. Rather than representing and reflecting physical experience, digital computation "deliver[s] us disintegrated impressions at speed", by design. This marginalises experience and knowledge gained through experience, privileging instead an affective mode of automatic reasoning. As Dieter and Gauthier explain, "[u]sers are typically taken in existential terms by these regimes, marking a shift from an interpretive episteme of deep subjectivity to the 'surface metaphors' of an affective dispositive".[39] Interfaces then assist in producing hybrid cognitive assemblages, nudging and guiding the user toward specific ideas of desired behaviour. Interface design might, for example, "habituate users into patterns of action, even to the point of compulsion"[40], or, in the case of commercial applications, serve to homogenise actions to the detriment of "cognitive functions associated with spatial navigation".[41] In short, Dieter and Gauthier show that interfaces serve as trans-active media which expand and also shape cognition in important ways. While studies on this have proliferated for commercial interfaces, there is still insufficient work done on this in specific relation to the military sector. It is, however, not a stretch to consider how in an environment in which the military operator is intricately

[36] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems", 2004.
[37] Ibid.
[38] Dieter and Gauthier, "On the Politics of Chrono-Design", 2019.
[39] Ibid., p. 66.
[40] Ibid.
[41] Ibid.

embedded within screens, data flows, and other interfaces, these dynamics might play out with military interfaces as well.

Within complex digital environments, the human functions as a modular element in the operational assemblage of the machine, which mediates and moderates experience to make it operational. The divergent temporal regimes implicit in the prioritisation of operationality over experience, in turn, privilege action – or 'getting things done' – as a preferred practice. Cognitive capacity, or the ability to have full situational awareness, is not only compromised with computer technologies, but also actively shaped towards specific, computational logics of actionable events. An example from a mundane context might serve to illustrate this best. Consider, for instance, the attention economy a smartphone interface produces. The screens are designed to elicit attention which leads to action. I am, for example, more inclined to engage with my phone if my messenger app displays a red alert symbol which indicates a new item or message. With time, I come to anticipate this symbol as a prompt to act and re-act. Being prepared and available for action and re-action becomes a norm. In the context of video gaming, James Ash explores this relationship as one of an ongoing modulation through deliberately constructed interfaces. He notes that in video game and interface design, a technique is used that aims to "modulate players capacities for recollection and anticipation within envelopes of differing length. In this way, interfaces can be understood as generating a continuously modulating now, in which the ability to reflect on the past and anticipate the future are opened and closed in wars that are specific to this or that envelope".[42]

As the human operator then becomes a modular part in the digital system, the question of agency – including moral agency – becomes complicated. Granted, the role of the soldier has always been that of being an element in the wider war-machine, and this has steadily increased with the greater complexity of digital information systems. Antoine Bousquet's study on the role of vision – the human eye – as an essential module for targeting technologies illustrates this relationship well. As digitalisation advances and, with AI, moves into the space for potentially lethal decisions in warfare, there is a risk that the role of human judgment as to where the limits of ethical acts are becomes equally modular and computational, whereby flattening out, streamlining, and problem-solving become increasingly prevalent modes of decision-making. To streamline and speed up moral judgment in this way is a temptation that goes hand-in-hand with Anders' diagnosis of an expansionist machine principle highlighted earlier. Such a "transference of accountability to the object (considered to be 'objective'), and the replacement of 'responsibility' by a mechanical response",[43] in which matters of right and wrong are rendered as calculable units, becomes trivial in a techno-logical environment. Shifting ethics into the realm of engineering might seem an obvious move in a technological ecology, but it brackets out all experiences that fall within the interstices – or through the cracks – of calculable life. The conception of ethics as a technical matter adheres to the principle of optimization through machine logics but it has human consequences of considerable weight: the unthinking shift of responsibility

[42] Ash, *The Interface Envelope*, 2015. p. 9.
[43] Anders, *Die Antiquiertheit des Menschen*, 1961, p. 246.

for our technologically mediated actions into the sphere of functional risk management by the machine.

This raises questions about the locus of moral agency and responsibility. Who is morally responsible for a wrongful act in war that was decided, or recommended, by an AI system? Moreover, who can 'feel' the weight of the responsibility for a wrongful act? And who has enough information about or control over a highly dynamic situation to assume responsibility? To address this vacuum or 'responsibility gap',[44] some suggest that autonomous technology can indeed assist in making more ethical decisions, and for this a form of ethics should already be encoded into autonomous machines. To this, I will just briefly turn my focus next.

Managing Morality with Machines

The question of how to address the ethics of autonomous, AI powered systems is looming large – in the military realm as well as in the commercial sector. As the human becomes a module in the digital world and the logic of engineering and problem-solving takes hold in an expanding realm of human endeavours, including ethics, it is perhaps not surprising that projects aiming to consider what makes machines moral are advancing at speed. One such example is the MIT Moral Machine project. The project was conceived by a team of MIT researchers as a "multilingual online 'serious game'"[45] in which participants access an online platform and indicate their choice as to how they would like an autonomous vehicle to solve moral dilemmas. The moral dilemmas in question were constituted by a binary choice about possible collisions. For this, users are shown a cartoon rendering of an ostensibly unavoidable accident scenario in which only two outcomes are possible, depending on whether the vehicle maintains course, or veers off the pre-determined path. For example, a player might be presented with a scenario in which an autonomous car might have a sudden brake failure and careen through a pedestrian crossing which would kill "1 large man and 1 criminal" or, alternatively, the car would swerve and crash into a concrete barrier and kill the car's occupants, which consist of "1 large man, 2 criminals, 1 elderly woman, 1 male doctor"[46]. The player has a binary kill choice here as the only option available. Either the pedestrians die, or the occupants die. This repeats with different options – two nurses or three criminals, three pregnant ladies, a large man and a dog or two babies, five criminals, and so on. The outlandishness of the choices is part of the attractiveness of the game (although this might possibly invite a player to make choices that are somewhat morally unpalatable as well). The experiment attracted just short of 40 million decisions by users, as it was played by people in 233 countries and territories. It is available to play in 10 languages.

---

[44] For key scholarship on the ongoing debate of the responsibility gap with LAWS, see Sparrow, "Killer Robots", 2017; and Matthias, "The Responsibility Gap", 2004.
[45] Awad et al., "The Moral Machine Experiment", 2018.
[46] Ibid.

The general idea behind the experiment is to "quantify social expectations about the ethical principles that should guide machine behaviour"[47], so as to better inform policymakers down the line. The findings from the experiment present a survey of preferences of outcomes by country and cultural clusters. The authors stress that the experiment should not be used for policymakers to follow the public opinions presented in the experiment, but rather that policymakers should be aware that certain, different preferences might exist among the public when it comes to the ethics of self-driving cars. As such, the authors make no claim to ethical deliberation, just to inquiring into public opinion. However, there is an undeniably slippery slope here to mistake crowdsourced consensus for the preference of an outcome with ethical deliberation, especially for a project with the suggestive title: 'Moral Machine'. Good intentions of the research team notwithstanding, by couching this binary kill choice experiment – IF/THEN – in moral terms, the groundwork is laid for using such a statistical survey as a form of ethical deliberation which then feeds into the ethics of AI for self-driving cars or, indeed, autonomous weapon systems.

There are a number of problems with using an experiment of this nature to ascertain public opinions about ethical principles, from the experiment design, to the interface, to some of the premises held, to existing assumptions about norms. This is not the place to elaborate on all of these; others have done this effectively elsewhere.[48] But for the purpose of the present discussion, the problem I will briefly focus on is the very logic of this approach toward "universal machine ethics"[49] as a way of producing ethical machine outcomes. While the MIT Moral Machine experiment is often cast as a contemporary version of Philippa Foot's notorious Trolley Problem in moral philosophy, reducing the famous thought experiment to a simple question of who to kill is missing the point of ethical deliberation. The Trolley Problem serves as a challenge to think through the intuitions we have about making morally difficult decisions when confronted with a dilemma. It serves as a prompt to think through utilitarian and deontological modes of ethical thinking. It is not intended as a blueprint for data collection over which scenario might be more acceptable to the wider public. While the questions raised by the MIT project might have the semblance of those relevant to the Trolley Problem, the ethos at its heart differs starkly.

The assumption with the MIT project is that a machine outcome can be considered as ethical if it aligns with ethical values that are widely accepted, as ascertained through data (here by crowdsourcing). Such an approach to ethics reflects a very truncated way of thinking about the manifold complexities of ethical decisions and challenges that may arise in human life, and especially when machine logics and human worlds interact. The logic of ethics as an engineering problem, as something that can be solved – ethical principles that can be extracted as fixed data – is one of cost-calculation and optimisation. It is ethics-as-technology, which relies on an abstraction of possible scenarios and factors that fit within a technical framework which purports to ascertain best-possible outcomes through best-case-scenario calculation. Within this kind of framework, a pre-formed 'good' or 'ideal' end is

---

[47] Ibid.
[48] Dewitt et al., "'Moral Machine' Experiment is No Basis for Policymaking", 2019; Jaques, "Why the Moral Machine is a Monster", 2019.
[49] Ibid., p. 60.

always already stipulated, toward which the ethical reasoning apparatus works. In the MIT Moral Machine case, the pre-formed good is ascertained through crowdsourcing: who would you prefer to die. In order to be able to code ethics into a machine, the assumption is that we first need to ascertain what actions people accept in abstract case scenarios so that the machine can then calculate the best outcome, based on stipulated parameters, given certain input criteria. Here ethics, seen as a problem, rather than a demand or a responsibility, is, at worst, conflated with cost-benefit utility, or, at best, confused with prudence. If confused with prudence, ethics becomes little more than a form of identifying and administering ways to avoid costly risk or unpopular outcomes.

Such modes of thinking are increasingly prevalent in discussions about the ethics of autonomous systems, including projects that aim to make 'moral' lethal weapons systems. As I have written elsewhere, this casting of social, political, and ethical matters in calculative and mathematical terms has a distinctly modern historical trajectory, which has shaped our modes of tracing the world predominantly in its economic-mathematical contours. It is a manifestation of our social and political forms in a technological present, and it is important to recognise these prevalent forms for how they shape our thoughts.[50] Where the world is seen primarily in calculable terms, it is not a stretch to consider it a logical next step to make ethics programmable through codes to find solutions to complex, multiple, and conflicting ethical issues within plural contexts. Such is the aspiration of the engineer – to solve problems through data and computation.

Ron Arkin's work is perhaps the most emblematic of this move toward ethics as a programme in the contemporary context of weapons systems. For him and co-author Patrick Ulam, the possibility of building ethical modules into lethal autonomous weapons systems may serve as a safeguard against wrong-doing and as a way to address the responsibility gap. The "ethical governor", a software and interface module, provides an analysis of a given potentially lethal situation and assesses, broadly speaking, whether the potential lethal action falls within the parameters of the Laws of War (LOW) and the Rules of Engagement (ROE) relevant to a mission. It then alerts an operator if there are any possible violations so that "operator responsibility can be maintained while a mission is actively underway",[51] thereby ostensibly closing the responsibility gap. The calculation is presented to the operator through a graphical user interface "that conveys the ethical governor's status to the operator, providing continuous information regarding an armed unmanned system's potential use of lethal force during the conduct of a mission"[52], thereby providing "ongoing ethical situational awareness of potential LOW and ROE violations". The proposed prototype interface is a relatively rudimentary set of online boxes which show the ethical governor's recommendation – permission to fire either denied or granted – for a given situation, and an option to click on a 'More Info' tab for an explanation as to why the governor suggests that permission might be denied ('Damaging cultural property prohibited'). The operator can

---

[50] Schwarz, *Death Machines*, 2019. See also "Prescription Drones", 2016, and "Technology and Moral Vacuums in Just War Theorising", 2018.

[51] Arkin and Ulam, "Overriding Ethical Constraints in Lethal Autonomous Systems", 2012, p. 4.

[52] Ibid.

then, together with a second operator, decide to override the recommendation, by entering a validation key and permitting the system to fire, thereby retaining responsibility over the lethal action. This process is meant to be completed in highly time-critical situations.

Here, then, ethics is understood in very limited parameters, as codable rules. The ethical principles included in the system are those of the laws of armed conflict (law) and the rules of engagement (rules). Both of these are informed by ethics, but ethics, especially in war, can hardly be reduced to these principles. The system's capacity would, for example, not be able to flag an instance in which the system misidentifies a disarming combatant for a legitimate target, mistakes a turtle for a rifle, or works according to parameters that set the acceptable collateral damage estimates too high. In all of this, the 'responsible' operator has very limited situational awareness, let alone ethical situational awareness of what is taking place on the ground in a highly dynamic, messy, and every-changing context of conflict, particularly with a very crude interface as the only guide. Much of the decision-making already takes place through the machine. Recall, as discussed earlier, the human tendency is to trust the system, and more so if time is of the essence. A prompt to lethal action, rather than an opportunity to survey the scene and assess the situation and conditions on the ground are the more likely outcome here. Arkin's entire discussion, beginning with the title of the paper, is cast in the language of engineering – "Overriding Ethical Constraints" – such that ethics itself ultimately appears as a constraint, to be solved as an engineering problem by the smooth functioning of the machine. In this the human performs a rather limited responsibility function, as a human "moral crumple zone", to borrow Madeleine Elish's apt turn of phrase.[53]

Conclusion: Moral Agency and Ethical Practice

It is of course tempting to see ethics as an engineering problem when warfare itself is cast predominantly as a matter of engineering prowess, rather than a complex social challenge with uncertain dynamics. The smooth functioning of machine operation replaces the messy idea of difficult, unsolvable moral decisions.[54] It suggests that if only we apply sufficient engineering ingenuity and if we have enough data, we can solve whatever we face and outpace the opponent. Such is the logic implicit in the AI arms race as well. However, if war's end is to be the goal, neither can be cast in those terms. Ethics is difficult. Ethics asks questions that have no clear answers, often no answers at all, only better or worse decisions for which human actors will be asked to take responsibility. The contexts in which ethics arise are always plural, rarely finite, and always challenging, especially so in warfare. Both ethics and war are relational and social categories. Accepting this allows for different forms of non-discrete understandings of suffering, pain ,and damage that occur in war and that perpetuate grievances that lead to war's protraction. Ethical reasoning requires training. It requires that we learn and practice ethical reasoning, that we cultivate an awareness of plural ethical dilemmas and that this is an ongoing and dynamic process rather than a mere rule or stipulation, that, once learned, can be applied to a wide range of scenarios. That is not ethical

[53] Elish, "Moral Crumple Zones", 2019.
[54] Anders, *Die Antiquiertheit des Menschen*, 1961, p. 246.

thinking or moral agency. That is applying a set of codes, rules, or laws. These codes, rules, or laws constitute a part of ethics in military practice, but ethics cannot be reduced to them alone.[55]

If we expect our military personnel to engage in ethical reasoning and take responsibility for their acts, then ethical reasoning, especially for targeting decisions, needs to be honed. This requires that we take moral reasoning seriously as an action that is antithetical to the logic of optimisation and speed. Ethics takes time and practice. It asks of the moral agent that they feel the weight of the decision, that they risk failing, that they accept not knowing the correct or optimal outcome and still have the courage to take responsibility for the action. And it asks of the agent that they accept that right and wrong may not always be secured, that the moral dilemma cannot be solved.[56] In short, it requires both courage and imagination. To mitigate the modern condition of *Promethean shame* and avoid losing sight of the monstrous acts we inflict on the world with our technological devices, Anders urges the conscientious honing of our (collective and individual) moral imagination.[57] Imagination as a crucial supplement to the limits of human perception (and cognition) with which we are so starkly confronted in a technologically mediated present. Such a conscientious honing of moral imagination and ethical skills is what David Whetham calls for in his insightful discussion on ethics education for military operatives, and it is worth quoting him at length. Whetham explains that

> [e]xperimental philosophy … suggests that one's surrounding environment has an enormous effect on ethical awareness and motivation to act … Therefore, normalising routine engagement with ethical issues is hugely important in shaping behaviour. This is sometimes referred to as 'ethical muscle memory', representing the idea that it is easier to do the right thing if you have engaged with a situation in advance and are already familiar with the ethical landscape of the problem.[58]

And precisely herein lies the crux. A military operator embedded in a network of digital infrastructures, some of which take on the function of making morally relevant decisions, has neither sufficient awareness of, nor exposure to the environment in which ethical awareness is required. Both situational awareness and the ethical muscle memory atrophies when technologies shape our cognitive scope, knowledge base, and capacity for moral action. The skill degradation Cummings alludes to in her study on human-machine teaming includes a degradation of the skill to think ethically about complex and changing environments. Rather than expanding our moral imagination, as Günther Anders calls for, and honing our sensibilities for consequences of high moral order, ethics becomes yet another machine function where it is routed through the logic of the protocol.[59] In the human-machine teaming

---

[55] Schwarz, *Death Machines*, 2019, pp. 118-142.
[56] Ibid., p. 197.
[57] Anders, "Theses for the Atomic Age", 1962. See also Schwarz, "Günther Anders in Silicon Valley", 2019.
[58] Whetham, "An Introduction and Review", 2018), p. 75.
[59] Anders, *Die Antiquiertheit des Menschen*, 1961, p. 273.

sought by most military AI programmes, we should, then, question any easy assumptions about humans retaining meaningful moral agency and control over life and death decisions, and consider instead that the move toward digital protocols in warfare is likely to orient military operations toward kill choices as the primary form of action in conflict.

The digital protocol stands in stark contrast to the complex, slow, and irresolvable character of ethical thinking. Solving ethically challenging tasks, including the identification of potential targets, with AI, even with the most sophisticated machine learning techniques, constitutes an abdication of that uniquely human task – to weigh, and feel the weight, of a morally difficult decision. Morally relevant decision-making cannot and should not be delegated to machines, nor should we allow such difficult decisions to be obscured by the smooth functioning of technology or the moral relief AI systems might seem to provide in conditions of radical uncertainty during armed conflict. This would be to give up on our humanity in the name of supposed innovation. And it would edge ethics ever further into the techno-logical realm, where the notion of improving ethical standards is conflated with providing enforceable ethical principles. And where an ethical failure occurs, it is, as Zygmunt Bauman observes "blamed on the faults of the ethical code or the laxity of the organs of its promotion and enforcement".[60] Rather than shaping the human toward greater ethical sensibilities, ethical responsibility is externalised. Where knowledge and perspective are routed through the logic of digital technology, the ethics of a messy world always appears as mediated, moderated, and compressed.

In this context, the dominance of the private sector in designing AI technologies for military use is especially troubling. It is clear, that today, AI engineering has different priorities and logics which often are not well aligned at all with the cumbersome and unwieldy demands of social, political and ethical human life. Engineering is the primary, often only, paradigm for the field and the Silicon Valley ethos – to move fast and break things – contains little of the caution raised by some of the founders of AI technology. We would, however, would do well to heed the words of Norbert Wiener father of cybernetics, who warned in the 1960s that coupling together "two agencies essentially foreign to each other" – the human being and the technological system – may herald a future not of progress but disaster. 'Move fast and break things' should not become the mode with which we conduct such morally charged endeavours as warfare.

References

Anders, Günther, *Die Antiquiertheit des Menschen: Über die Seele im Zeitalter der zweiten industriellen Revolution*, München: C.H. Beck, 1961.

Anders, Günther, "Theses for the Atomic Age", in *The Massachusetts Review* 3, 3 (1962), 493-505.

Anders, Günther, *Wir Eichmannsöhne*, München: C.H. Beck, 1988.

[60] Bauman, "Ethics of Individuals", 2000, p. 86.

Arkin, Ronald, and Patrick Ulam, "Overriding Ethical Constraints in Lethal Autonomous Systems", Georgia Instititute of Technology Mobile Robot Laboratory Technical Report GIT-MRL-12-01, 2012, 1-8.

Ash, James, *The Interface Envelope: Gaming, Technology, Power*, London: Bloomsbury, 2015.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, "The Moral Machine Experiment", in *Nature* 563, 7729 (2018), 59-64.

Bauman, Zugmunt, "Ethics of Individuals", in *Canadian Journal of Sociology*, 25, 1 (2000), 83-96.

Breton, Richard and Éloi Bossé, "The Cognitive Costs and Benefits of Automation", NATO RTO-MP-088, October 2003.

Conger, Kate, and Cade Metz, "I Could Solve Most of Your Problems: Eric Schmidt's Pentagon Offensive", in *The New York Times*, May 2, 2020, https://www.nytimes.com/2020/05/02/technology/eric-schmidt-pentagon-google.html?referringSource=articleShare

Coroalles, Anthony, *Lectures on FSR III Revisited: The Tactical Thought of J.F.C. Fuller Applied to Future War*, Ft. Leavenworth, KS: U.S. Army Command and Staff General College, 1988.

Cummings, M.L., "Automation Bias in Intelligent Time Critical Decision Support Systems", Conference Proceedings of the American Institute of Aeronautics and Astronautics, AIAA 2004-6313, September 20-22, 2004, https://doi.org/10.2514/6.2004-6313

Cummings, M.L., "Artificial Intelligence and the Future of Warfare", Chatham House Research Paper, January 26, 2017, https://www.chathamhouse.org/publication/artificial-intelligence-and-future-warfare

Dewitt, Barry, Baruch Fischhoff, and Nils-Eric Sahlin, "'Moral Machine' Experiment is No Basis for Policymaking", in *Nature*, March 5, 2019, https://www.nature.com/articles/d41586-019-00766-x

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

Elish, Madeleine Clare, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction", in *Engaging Science, Technology, and Society* 5 (2019), 40-60.

Faggella, Daniel, "What is Artificial Intelligence? An Informed Definition", in *Emerj: AI Research and Insight*, December 21, 2018, https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/

Hutson, Matthew, "A turtle or a rifle? Hackers easily fool AIs into seeing the wrong thing", in *Science*, July 19, 2018, https://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing

Insinna, Valerie, "Inside America's Dysfunctional Trillion Dollar Fighter-Jet Program", in *New York Times Magazine*, 21st August 21, 2019, https://www.nytimes.com/2019/08/21/magazine/f35-joint-strike-fighter-program.html

Jasanoff, Sheila, *The Ethics of Invention: Technology and the Human Future*, New York: W.W. Norton & Co, 2016.

Jaques, Abby Everett, "Why the Moral Machine is a Monster", *University of Miami School of Law* 10 (2019), 1-10.

Kania, Elsa, "AI Weapons in China's Military Innovation", in *Global China: Assessing China's Growing Role in the World*, Brookings Institute Report, April 27, 2020, https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/

Matthias, Andreas, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", in *Ethics and Information Technology* 6, 3 (2004), 175-183.

McKendrick, Kathleen, "Artificial Intelligence, Prediction and Counterterrorism", Chatham House Research Paper, August 9, 2019, https://www.chathamhouse.org/publication/artificial-intelligence-prediction-and-counterterrorism

Miller, Kevin, "Total Surveillance, Big Data and Predictive Crime Technology: Privacy's Perfect Storm", in *Journal of Technology, Law and Policy* 19 (2014), 105-146.

Müller, Christopher John, *Prometheanism: Technology, Digital Culture and Human Obsolescence*, London: Rowman & Littlefield, 2016.

Noble, David, *The Religion of Technology: The Divinity of Man and the Spirit of Invention*, London: Penguin, 1999.

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

Poulson, Jack, "Reports of a Silicon Valley/Military Divide Have Been Greatly Exaggerated", in *Tech Inquiry*, July 7, 2020, https://techinquiry.org/SiliconValley-Military/#embeddings

Schwarz, Elke, "Prescription Drones: On the Techno-biopolitical Regimes of Contemporary 'Ethical Killing'," in *Security Dialogue* 47, 1 (2016), 59-75.

Schwarz, Elke, "Technology and Moral Vacuums in Just War Theorising", in *Journal of International Political Theory* 14, 3 (2018), 280-298.

Schwarz, Elke, *Death Machines: The Ethics of Violent Technologies*, Manchester: Manchester University Press, 2019.

Schwarz, Elke, "Günther Anders in Silicon Valley: Artificial Intelligence and Moral Atrophy", in *Thesis Eleven* 153, 1 (2019), 94-112.

Shanahan, John, "Lt. Gen. Jack Shanahan media briefing on A.I.-related initiatives within the Department of Defense", *U.S. Department of Defense*, August 30, 2019, https://www.defense.gov/Newsroom/Transcripts/Transcript/Article/1949362/lt-gen-jack-shanahan-media-briefing-on-ai-related-initiatives-within-the-depart/

Sharkey, Noel, “Guidelines for the Human Control of Weapons Systems”, *ICRAC Briefing Paper*, April 2018.

Simonite, Tom, “The Pentagon Doubles Down on AI – and Wants Help from Big Tech”, in *Wired*, December 2, 2019, https://www.wired.com/story/pentagon-doubles-down-ai-wants-help-big-tech/

Slijper, Frank, Alice Beck, and Daan Kayser, “The State of AI: Artificial Intelligence, the Military and Increasingly Autonomous Weapons”, in *PAX Report Series*, May 8, 2019, https://www.paxforpeace.nl/publications/all-publications/the-state-of-ai

Sparrow, Rob, “Killer Robots” in *Journal of Applied Philosophy* 24, 1 (2017), 62-77.

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

Tucker, Patrick, “The F-35 is about to get a lot smarter”, in *Defense One*, December 6, 2018, https://www.defenseone.com/technology/2018/12/f-35-about-get-lot-smarter/153338/

Whetham, David, “An Introduction and Review: The King’s College London Centre for Military Ethics”, in *Journal of Military Ethics* 17, 1 (2018), 72-79.

Wiener, Norbert, *The Human Use of Human Beings*, Boston: Houghton Mifflin, 1954.

Wiener, Norbert, “Some Moral and Technical Consequences of Automation,” in *Science* 131, 3 (1960), 1355-1358.