How might the difference between the accountability of humans and the accountability of highly autonomous weapons in war be qualified?

Thesis submitted to King's College London, August 2019.

Abstract

Objections to the use of highly autonomous weapons in war are manifold. The accountability objection dominates. In order to be effective, this objection must be both comparative and qualified: It must explain how the accountability of autonomous weapons *compares* to the present accountability of humans in war, and it must do so using *appropriately comparative concepts and frameworks*. The first purpose of this thesis is to provide such concepts. It offers a de-mystified notion of autonomy, new cross-cutting distinctions amongst weapons of higher and lower control and certainty, and a maximalist, event-centric notion of accountability. The second and subsidiary purpose of this thesis is to demonstrate the utility of these concepts, across numerous case studies involving both human and artificial actors.

The concepts are found to qualify the discourse effectively. Rather than pointing to a simple, dichotomous distinction in accountability between highly autonomous weapons and humans in war, they emphasise *commonalities* in the accountability obstacles faced by both. In this way, the concepts not only allow for comparative accuracy, but also highlight scope for adapting present approaches to human accountability to the highly autonomous weapons of the future.

Contents

Introduction

Close enough to the epicentre, one finds in the discourse on technological ethics a sinister undertone: an undertone, not of sophistication and readiness, but of frenzy. Uncertain whether the thunderheads of progress will yield welcome rains, or hail down on an unprepared world, analysts are scrambling to determine and quantify the greatest threats.

In March 2019, UN Secretary General António Guterres was resolute in naming one: the weaponization of artificial intelligence in the form of fully autonomous, lethal weapons systems. Such machines, he argued, "are politically unacceptable, morally repugnant, and should be banned by international law" (Guterres, 2019: 1). He was echoed not merely by roaring applause, but by 28 national governments. These states, notably excluding the nine most militarily powerful, called for an outright ban on the sensationally designated "killer robots" (Future of Life Institute, 2019: 1).

There are numerous arguments supporting such a ban. Yet, perhaps the most common among them has been the accountability argument: "It is unclear who, if anyone, could be held responsible for unlawful acts committed by a fully autonomous weapon" (Stop Killer Robots, 2019: 1). Accountability concerns have also been raised for weapons with a lower degree of autonomy than full autonomy (Buchanan and Keohane, 2015: 22-23; Weir, 2015: 1). The arguments hinge on theoretical and practical differences in accountability between the illicit actions of human agents and the adverse actions of fully or highly autonomous military artificial intelligence.

Surprisingly, however, systematic comparisons of accountability between the illicit actions of humans and the adverse actions of present and potential autonomous weapons are rare. In fact, many iterations of the accountability argument are monadic in character. Such arguments give mouth honour to the comparative accountability of human behaviour, without seeking to qualify the difference. Importantly, where comparisons are made, they are not rooted in an appropriately comparative conceptual framework. In other words, the differences described are unqualified. This thesis seeks to resolve this problem, through conceptual and case-driven analysis.

Part I lays the groundwork for this thesis, placing and describing the problem, delineating the

scope, outlining the methodology, and suggesting the significance of this endeavour. Notably, a discussion of terms is absent from this groundwork. Rather, the conceptualisation of terms is a major component of the thesis itself.

Parts II and III deal with the two primary purposes of this thesis. Part II deconstructs the conventional distinction between 'automated' and 'autonomous', introduces new, cross-cutting distinctions amongst weapons, and provides a fully adaptable, event-centric conceptualisation of accountability. These concepts are uniquely suited both to humans and autonomous non-humans. Part III demonstrates the utility of these concepts, comparing accountability across a range of human and artificial actors.

This thesis finds that the concepts provided are useful in qualifying the comparative discourse on accountability in war. Rather than pointing to a simple, dichotomous distinction in accountability between highly autonomous weapons and humans in war, they emphasise commonalities in the accountability obstacles faced by both. In this way, the concepts not only allow for comparative accuracy, but also highlight scope for adapting present approaches to human accountability to the highly autonomous weapons of the future. The concepts also highlight a degree of inconsistency in the modern discourse on highly autonomous weapons: if the accountability concern is truly worrisome, we should be very concerned with the present accountability of human soldiers in war.

PART 1 – Groundwork

Placement

Objections to the development and use of highly autonomous weapons in war appear to fall into four broad categories. The first group of objections is **systemic**. These objections regard shifts in relations between international actors. For example, some argue that the development of these weapons will trigger an AI arms race, which could in turn spur unchecked rivalries, overspending, corner-cutting, and zero-sum logic (Cave and OhEigeartaigh, 2018: 3; Barnes and Chin, 2018: 3). Other systemic objections point to the risk that highly autonomous weapons will render wars too easy to start (Singer, 2009: 44), escalate conflict (Altmann and Sauer, 2017: 128), destabilise power balances by creating a vastly superior hegemon (Bostrom, 2017: 141), or undermine state sovereignty by empowering non-state actors (Altmann and Sauer, 2017: 127).

The second group of objections is **strategic**. These objections regard the disadvantages presented by highly autonomous weapons for the purposes of successful warfare. For example, the more autonomous a weapon, the greater the dangers presented by malfunction, a poor human-AI interface, or disruptions to the command chain (Kovic, 2018: 3). Such weapons also present the hereto improbable risk of 'mass fratricide': weapons could be hacked to target allies (Scharre, 2016: 23).

The third group is **ethical** and includes difficulties presented both in creating an ethically capable AI and in using it to ethical effect. This set of objections is perhaps the most powerful, exploring the theoretical obstacles to developing a value-aligned AI, whether through obedience to deontological or consequentialist rules (Purves et al, 2015: 856; Yudkowsky, 2016: 3) or through virtue ethical machine learning (Vamplew et al, 2018: 28-29; Beberich and Diepold, 2018: 4). One fear is brittleness; that logic-based AI is too crude an instrument for ethics, merely approximating a moral nexus built upon discretion, compassion, intuition, and human experience. Another fear is learned misbehaviour, the mascot for which is perhaps Microsoft's racist twitter bot, whose gleaning from the cyberspace swiftly rendered it obnoxious and discriminatory (Garcia, 2019: 112). Other concerns surround whether or not the use of ethically capable autonomous weapons would be ethical itself. For example, many argue that

autonomous weapons excessively dehumanise the act of killing (Asaro, 2012: 697; Heyns, 2017: 60). Each of these concerns asks whether autonomous weapons would be capable of producing the normative outcomes demanded of it.

The fourth category is **legal**. This category is importantly distinct and is dominated by accountability concerns. It is with this category alone that this thesis is concerned. Accountability objections begin at the point at which an autonomous machine, alone or in conjunction with human actors, has already committed an adverse action. Part of the purpose of this thesis is to fully conceptualise accountability, but for present purposes, accountability objections ordinarily ask the following: Once a mistake has already been made, is there a responsible party? Can it be held to account? If so, how easily? (Sharkey, 2012: 790; Amoroso and Tamburrini, 2017: 16) A summary of this categorisation of arguments against highly autonomous weapons can be found in Figure 1.

Figure 1: Categorising objections to highly autonomous weapons

The Problem

Imperfect Actors

Humans make *mistakes* in war. These are unintentional actions that result in negative, and sometimes lethal, outcomes. Humans also commit *violations* of international humanitarian laws.

Some violations are also mistakes, but most violations involve intentional malevolent behaviour. Grave breaches or serious violations of these laws are *war crimes* (United Nations, 2019: 1). Where autonomous weapons are involved, but humans maintain veto power over final actions, the human agents involved can also here commit mistakes and violations.

Fully autonomous lethal weapons, those with no human veto power over lethal actions, are unlikely to be perfect. The language of their imperfection is, however, more complex. Part of the very problem of accountability is that the separation of 'mistakes' from 'violations' and 'war crimes' in the actions of fully autonomous weapons is especially opaque. To avoid ambiguity, we must phrase the problem in this way: It is likely that fully autonomous weapons will perform some actions with outcomes equivalent to the outcomes of human actions we currently call 'mistakes' and 'violations'. In other words, there will be outcomes in common between human mistakes, human war crimes, and the adverse actions of fully autonomous weapons. Moreover, many of these adverse outcomes will be legally relevant outcomes. By this I mean outcomes which are either covered by international law or ought to be covered by international law, according to a given legal framework.

Imperfect Accountability

Wherever adverse actions are committed, we face questions of accountability. Many scholars and practitioners argue that certain, legally relevant, adverse outcomes of fully autonomous weapons are not accompanied by responsible agents who are subject to accountability. However, it is also the case that many crimes of war committed by humans have lacked some form of accountability (Goldsmith, 2003, 98; Joyner, 1996: 155, 162; Thynne, 2009: 981). This is revealed by the simple fact that not every responsible party in war has been identified and held to account. As shall be discussed, human military acts go unrecorded, unjustified, and unpunished for many reasons.

Imperfect Comparative Concepts

Nevertheless, systematic qualifications – refinements and framings - of this difference in accountability, between human and non-human cases, are rare. Where they do exist, they face several problems. First, the most systematic comparative analyses have often focused on the ethical category: differences in the ability of autonomous weapons to meet demands in the first

place. This is partially true of Ronald Arkin's excellent and influential analysis of weapon autonomy (Arkin, 2009). Second, and most crucially for this thesis, research has lacked the reservoir of appropriately comparative concepts necessary to qualify these differences properly. This dearth of rigorous conceptualisation has muddied the discourse, leading to either monadic or poorly comparative assessments of the accountability objection. Comparisons need qualifications: common terms, common measures, and reasonably comparative case studies which use them.

There are four possible outcomes. First, autonomous weapons and humans could face exactly the same problems in war, resulting in identical rates of unaccountable adverse outcomes (No change). Second, autonomous weapons and humans could face similar rates of unaccountable adverse outcomes, but for differing reasons (Qualitative Difference). Third, autonomous weapons and humans could face numerically different rates of unaccountable, adverse outcomes (Quantitative Difference). Fourth, there could be both qualitative and quantitative differences in accountability between humans and highly autonomous weapons in war. If we can expect no significant qualitative or quantitative difference in accountability between human actors and autonomous weapons, this significantly undermines the accountability objection. Thus, qualifying these comparative assessments is vital.

Thus, at its core, this thesis seeks to equip the discourse in its comparative endeavours; to qualify the content of comparative claims. As a result, a significant portion of this project is conceptual, aiming to provide a coherent conceptualisation of autonomy, accountability, and further terms, which can be reasonably adapted to these comparative assessments of human and weapon accountability. These concepts are subsequently applied to a variety of case studies involving both human and artificial actors. The case section fulfils the subsidiary purpose of this thesis, which is to demonstrate the utility of these comparative concepts in practice, and to suggest their impact upon approaches to the accountability of highly autonomous weapons.

Methodology

Conceptualisation

This thesis is predominately conceptual, seeking to qualify the discourse on comparative accountability with appropriately comparative concepts and frameworks. The aim is to provide a useful and well-defined lexicon for comparative assessments of accountability in humans and autonomous weapons. The curation of this conceptual trove falls into two broad categories: autonomy and accountability. While much of this section is comprised of abstract and original thought, it makes extensive, critical use of literature spanning the legal, technical, philosophical, and social scientific spheres.

Case Studies

Following conceptualisation, this thesis deals with war-time outcomes involving both human and artificial causal actors. These case studies demonstrate the utility of the qualifying concepts developed, rather than seeking to conclusively resolve the debate on the accountability of highly autonomous weapons. The case studies fall into two categories. The first category dives into the foggy realm of humans at war: combatants, their war time outcomes, and the attempts to render those outcomes accountable. The second category engages the world of highly autonomous weapons and the problems they present for accountability, while also speculating as to hypothetical technologies, which have not yet been created or deployed.

Importantly, part of the conclusion drawn by this thesis is that these categories cross-cut the spectrums of control, certainty, and – above all – accountability. In this way, while this thesis uses these categorisations, it also works to disassemble them.

Comparison

This thesis argues that systematically comparative tests of the accountability objection are of particular importance. It suggests that, without clear comparative assessment, advocates risk committing a fallacy which we might call the **Problem-in-Common Fallacy**. I coin this fallacy to denote cases in which a policy is rejected exactly and exclusively because it does not meet criterion [X], in favour of a present policy which also does not meet criterion [X]. In other words, the status quo and the proposal under consideration share a problem in common, which nevertheless forms the basis of rejection for the new proposal.

The fallacy is ubiquitous and differs in important ways from *Appeal to Tradition* and *Whataboutism*. One example is perhaps found in the controversial debate surrounding the

criminalisation of marijuana. While some suggest that certain risks of alcohol far surpass the equivalent risks of THC-controlled marijuana use (Whiteman, 2018: 1), the normative opprobrium against decriminalisation has held firm for decades, classing the drug in the same category as heroin. Arguments against decriminalisation have often centred directly on criteria which are also not met by many legal intoxicants, to an equivalent or even greater degree. If this inconsistency is not explained by an additional, independent criteria - such as the relative difficulty of alcohol prohibition - the fallacy is committed. It is normally caused by a monadic assessment of the new policy in its own right, or a passive, rose-tinted bias towards the status quo.

It is possible that such a fallacy is committed by those who argue that autonomous weapons should be banned exactly because they fail to meet certain criteria of accountability. This would be true if humans regularly failed to meet identical criteria of accountability, to an equivalent degree, and yet proponents continued to believe in the lawful participation of humans in warfare.

To assess whether this is the case, one needs a rigorously comparative study, and that study needs to be equipped with appropriately comparative, qualifying concepts.

Scope

The scope of this thesis requires further clarification. First, this thesis begins at the point at which a fully autonomous weapon has already committed an adverse action. This precludes other objections pertinent to the debate, in particular the objection that it will be prohibitively difficult to prevent such weapons from performing these adverse actions in the first place. While this is a fruitful and fascinating area of discussion, it does not concern the question of what can be done, after the fact, with faulty, erroneous, evil, or misused artificial intelligence in war.

Second, as previously noted, the set of adverse actions is difficult to define without already pressing a point of contention. To call these actions 'war crimes' or 'violations' *a priori* would obscure a major challenge presented by such weapons: it is not clear whether, under existing laws, equivalent actions could rightly be given these names. Moreover, to refer to them as 'mistakes' concedes the absence of intent at the outset. Thus, for our purposes, '**adverse**

actions' are the set of legally relevant actions which produce negative outcomes equivalent to those of human mistakes, violations, or war crimes.

Third, the primary weapons of interest are those which are 'highly' autonomous, in which very complex tasks are permitted to be performed independently, both with and without humans in the decision-making loop. Nevertheless, when assessing the accountability of humans, a wide variety of less autonomous and non-autonomous weapons are drawn into the discussion. This is in line with the conclusion of this thesis that autonomy is not a hard line along which accountability decreases linearly, but that autonomy and accountability cross-cut each other.

Significance

The importance of avoiding fallacies is not the only reason for which this pursuit is an important and contemporary one. The legality of developing and using fully autonomous weapons is directly contested on an international scale, by experts across political, military, and technical spheres. Among those supporting an outright ban, one will find twenty-eight national governments, a majority of European MEPs, the UN Secretary General, and over 4500 researchers in AI and robotics, including the prodigious Demis Hassabis, Stuart Russell, Noah Sharkey, and Nils J. Nilsson (Future of Life Institute, 2019: 1). The list would seem almost overpowering, if not for its counterweight. Those who question the justification of such a ban include the governments of China, Russia, France, the US, and the UK, as well as the strategic and technical experts commissioned by the European Parliament in 2018. Among them were the research director at the Hague Centre for Strategic Studies, Tim Sweijs, and the director of Oxford's Governance of AI programme, Allan Dafoe (Teffer, 2018: 1).

The accountability objection to these weapons has frequently taken centre place. When the Secretary General rose to the stage to call for a ban at the AI Summit for Good, it was accountability that he listed as the monumental challenge (Guterres, 2019: 1). It is crucial that, if autonomous weapons are indeed dangerous, only the *strongest* reasons against them rise to the fore. If the discourse is dominated by phantom objections, opposing advocates are done a considerable favour. Those seeking a ban on killer robots should be encouraged to rally around the best objections, whether systemic, strategic, ethical, or accountability-based.

Finally, the debate is a matter of urgency. Advancements in this area are spurring on rapidly, driven and preceded by non-military tech giants. It is plausible that this software will develop first outside of the military sphere, before being rapidly adapted to military ends: In other words, the developments may come quickly, from otherwise innocuous sectors (Cummings, 2017: 2).

PART 2 - Conceptualisation

The meat of this thesis begins in the conceptual sphere. As will become clear, the concepts proposed in this thesis are partially constitutive of its conclusions regarding the accountability of autonomous weapons. In particular, the conceptual section performs three tasks. First, it deconstructs the dichotomous distinction between 'automated' and 'autonomous' weapons in the social scientific area of research. Second, it proposes two new distinctions amongst weapons of war, which cross-cut the spectrum of autonomy. Third, it provides a coherent and adaptable conceptualisation of accountability, which applies to both man and machine.

Conceptualising Autonomy

'Automated' and 'Autonomous': An Overvalued Distinction

The distinction between 'automated' and 'autonomous' machines is common across technical, legal, and social scientific spheres (Roff, 2019: 129; Asaro 2013: 690; Anderson and Waxman, 2013: 2; Cummings, 2017: 4). According to the roboticist Noel Sharkey, the distinction runs as follows: *Automated machines* are those which "operate with pre-programmed instructions to carry out a specific task" and are "deterministic" or "rule-based" (Red Cross, 2014: 13). Examples include an alarm clock, refrigerator light, or landmine. Conversely, autonomous machines "act dynamically to decide if, when and how to carry out a task" using "stochastic (probability-based) reasoning, which introduces uncertainty" (Red Cross, 2014: 13). Other scholars note distinctions along multiple dimensions: complexity, the significance of the task being delegated, or the degree to which the human is directly involved (Scharre and Horowitz, 2015: 5). At times, scholars draw from these distinctions a dichotomous, or categorical, conceptualisation of automation and autonomy.

The following section contends that this distinction is overvalued and misplaced in the social scientific study of weaponry. This may be demonstrated by example: the 'automated' landmine and the partially 'autonomous' Aegis Combat System. The landmine is evidently rule-based. For example, a 'Bouncing Betty' landmine mechanically follows the rule: "If and only if depressed by a weight of 7kg, launch one metre into the air and detonate". The 'autonomous'

Aegis Combat System identifies, targets, and shoots incoming rockets without requiring the intervention of human agents, adapting to changing environmental factors (Kastan, 2013: 50). In what sense do these weapons differ? Both can perform important tasks; the Aegis can destroy incoming threats and the landmine can kill ground troops. Both can do so without human intervention – the landmine even more so, as no-one possess veto power over its detonation (Grut, 2013: 5). Both perform actions determined by their mechanical make-up and environmental inputs. For both, given identical inputs, the outputs will always be the same.

The material difference between the landmine and the Aegis appears to be, not a dichotomous distinction, but variance merely in degree of complexity. The Aegis can use a much wider set of environmental data, learn from that data to acquire new data, make probabilistic assessments of future outcomes, and perform a wider range of actions. This complexity requires intelligence. As the AI scholar Max Tegmark notes, intelligence is simply the ability to "accomplish complex goals" independently (Tegmark, 2018: 49) and artificial intelligence is this ability procured by man. Intelligence is in potentia: It exists even in the absence of permission to utilise it.

This essay therefore takes a demystified approach to autonomy: a scalar concept, denoting the de facto ability – that is, the ability and the permission – to perform a task independently or 'without human input' (Scharre and Horowitz, 2015: 5). A highly autonomous weapon is simply one with the *de facto* ability to independently perform *very complex* military tasks, although the term is disproportionately applied to machines which perform complex tasks that are normatively considered 'serious', such as killing. Landmines, the Aegis, and the Terminator alike, all sit on this demystified, uni-dimensional spectrum of autonomy. This is true in all cases, except where the machine is postulated to possess mystical autonomy of the kind that means 'undetermined', its decisions being neither random, nor determined by mechanical make-up, nor determined by environmental inputs. This form of autonomy – often called '**free will**' – *would* demand a category of its own in the social sciences, however it is unclear whether even humans possess it, or whether it is even conceptually coherent (Van Inwagen, 2000: 1). A summary of these concepts is found in Figure 2.

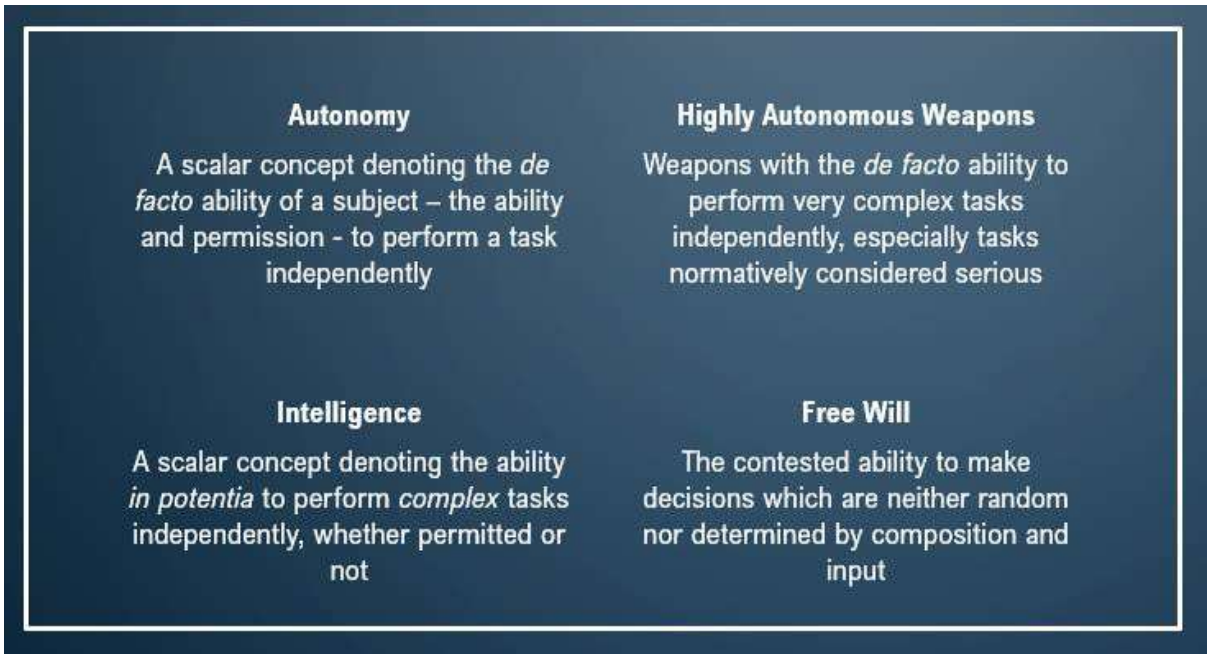

Figure 2: Autonomy, Intelligence, Highly Autonomous Weapons, Free Will. Two differences between autonomy and intelligence are noted: Intelligence is the ability to perform *complex* tasks and it is this ability *in potentia*. Highly autonomous weapons are also necessarily intelligent.

Conceptualising New Distinctions Amongst Weapons

To place the landmine and the killer robots of science fiction along a spectrum of mere complexity may seem a gimmick at best; a grave misunderstanding at worst. As such, there are clarifications to be made. First, the above is not intended to undermine the remarkable increases in capability afforded by machine learning: While I postulate a scale, it is one for which the poles are incredibly far apart. Second, the scale is intentionally de-mystifying. It is designed to highlight that there is nothing magical about highly autonomous weapons: they, like alarm clocks, perform tasks by themselves, only tasks which are far more complex. Third, and most crucially, I disassemble the automated-autonomous dichotomy in order to suggest new, *cross-cutting* distinctions amongst weapons.

The first new distinction is that between weapons for which users have greater or lesser *control* over the probable output of the weapon. Weapons over which there is less control may be called '**Low Control Weapons'**; those over which there is greater control are '**High Control Weapons**'. The second new distinction is that between weapons for which users have greater

or lesser *knowledge* of the probable output of the weapon. This distinction tracks the epistemic distance between the user and the weapon's output. The opposing ends of this certainty spectrum are occupied by '**Low Certainty Weapons**' and '**High Certainty Weapons'**. Probable output refers to the outcomes which the weapon is most likely to produce, both immediately and long-term. Distinctions in weapon control and weapon certainty correlate, but do not always entail each other. A summary of these new concepts is found in Figure 3.

Figure 3: High Control Weapons, Low Control Weapons, High Certainty Weapons, Low Certainty Weapons

Cross-Cutting Distinctions

Crucially, and against the grain, I emphasise that highly autonomous weapons do not monopolise the problems of control and certainty. Landmines are low on the scale of autonomy but are nevertheless Low Control and Low Certainty Weapons. A man who plants a 'Bouncing Betty' knows some things about the likely victims: he knows where they will be located and that they will weigh more than 7kg. He does not have control over the identity of those targets (whether they will be combatants), nor the time at which they will be killed (in battle, ceasefire, or even decades later). This is a serious obstacle to control over - and, adjacently, certainty about - the probable outcome of landmine use.

Consider a second example. In the weeks following the 9/11 attacks, seven letters containing anthrax were posted to numerous media outlets and public offices. Two of these letters were directed to Democrat Senators, Tom Daschle and Patrick Leahy. The first was opened by Daschle's aide, the second was misread and sent to an entirely different State, in which an unsuspecting postal worker subsequently contracted the infection (Foster, 2003: 186). The anthrax letter was unambiguously non-autonomous: being inert, it had no ability to perform any tasks independently. Nevertheless, it belonged definitively to the Low Control and Low Certainty categories; the sender had little control over or knowledge of the immediate and ultimate outcomes of this method of execution.

The distinction between the 'automated' and the 'autonomous' is in many ways an *overvalued proxy* for this more serious concern: for many weapons, the outcome of use is partially or fully beyond human control. Prompted by this concern, one is tempted to ascribe this missing control *to* the weapon under consideration. This, as seen above, leads to considerable conceptual confusion: A landmine performs very serious actions without human oversight, but is considered too simple to accept our ascription of 'control' to the landmine itself. In reality, the Aegis Combat System, however complex, cannot accept our ascription of 'control' any better than the landmine can. For both the landmine and the Aegis, it is merely a question of how many complex or serious actions it independently performs.

The most sensible way of dealing with this problem is to re-centre the discourse: not around the degree of autonomy possessed by the machine, but the degree of control possessed by the user. As here demonstrated, they are far from the same thing.

Machine Autonomy and User Control

Having established that weapons with very little or no autonomy can be Low Control Weapons, it is left to be emphasised that machine autonomy can nevertheless play important roles in reducing human control. Indeed, as we shall see, this concern lies at the foundation of the accountability argument. First, autonomous machines are, by definition, complex. The greater the complexity of the machine, the higher the risk that the machine's processing becomes a 'black box', obscured for the user and even the creator (Bathaee, 2017: 891). This can increase the epistemic distance between the user and the probable output of the weapon, making such weapons Low Certainty. If the user does not *know* what outputs the weapon will produce given

particular inputs, his ability to control the output of the weapon is considerably diminished. Second, where humans do not possess veto power over the autonomous functions of a machine, user control is limited in obvious ways. A further conceptual qualification must be made here: the designation 'fully autonomous', to machines over which humans possess no veto power, is somewhat misleading. Very low-level autonomous machines like landmines can also lack veto power once deployed. Nevertheless, for highly autonomous weapons – which are intelligent – the absence of this veto power has the potential to be more disastrous, with outputs vastly differing from those expected.

Much as expected, then, highly autonomous weapons can be Low Control. Nevertheless, as our cross-cutting distinction is revealing, the relationship between machine autonomy and user control is not so simple. Not only can non-autonomous weapons like anthrax letters be Low Control, but autonomous weapons can be High Control. In fact, autonomy is often developed in machines with the precise intention of *enhancing* human control. This is a far cry from conceptualisations which build the absence of human control into the very notion of autonomy.

Consider a bright rebel who upgrades his landmine with the ability to perform a more complex task independently – the ability to distinguish targets. This increased autonomy significantly increases the rebel's control over and knowledge of the probable output of the weapon. An anthrax letter with the capacity to recognise the faces of those that opened the letters, and release anthrax accordingly, would be a weapon with much higher control, despite being far more autonomous.

Machine Autonomy and User Certainty

In the same way that Low Control has often been conceptualised as something entailed by autonomy, so, too, has Low Certainty. Consider the problematic distinction between the 'automated' and the 'autonomous' previously cited from the Red Cross. The definition strongly linked uncertainty and autonomy at the conceptual level: autonomous weapons were said to use "stochastic" reasoning, thereby "introducing uncertainty" (Red Cross, 2014: 13). This thesis fully contests this point, at a technical and social scientific level.

This thesis fully contests this point. First, the kind of uncertainty present in a highly autonomous weapon is not necessarily 'stochastic', in which the outcome could not be known because it is

produced by processes involving random data selection. In many cases, the uncertainty is merely epistemic or 'subjective': identical inputs will produce exactly the same outputs, but the observer simply does not have cognitive access to what the output will be (Helton, 1997: 4). Indeed, it is disputed whether all stochastic uncertainty collapses into mere epistemic uncertainty, unless dealing with true indeterminacy.[1] The point here is that, much of the time, when we talk about 'uncertainty' in highly autonomous weapons, it is uncertainty of observers, not true indeterminacy of outcome. A summary of this distinction is found in Figure 4.

Second, and less technically, the scope of the epistemic uncertainty matters. It is reasonable that the technicians ascribe uncertainty to autonomous weapons when there is low knowledge of the immediate output of the weapon, for example, whether or not the autonomous weapon will detonate bomb [X] at time [t]. However, the social scientist is concerned with much more than immediate outputs. While he can be certain that a landmine will output 'explosion' given input '7kg depression', it is very significant that he cannot be certain about who will consequently explode. Thus, when the technician says that autonomous weapons increase uncertainty, the social scientist should not only ask what kind of uncertainty, but also the scope of that uncertainty: whether it is uncertainty about the immediate or the ultimate output. Again, the scope of uncertainty can be extremely broad for weapons with very little autonomy.

Finally, weapons with little or no autonomy are capable of being both epistemically uncertain and stochastically uncertain. The latter would be true, for example, if an alarm clock was set to ring at a random time [t] between 9am and 10am. In other words, uncertainty of both kinds cross-cuts the spectrum of autonomy, rather than tracking it, contrary to the Red Cross view. This, like the proposition that control cross-cuts autonomy, is a central proposition of this thesis.

[1]Many contest this 'collapse' of stochastic uncertainty into epistemic uncertainty. One author argues that stochastic uncertainty must be distinct, because if a "mission were to be run several times it would succeed on some occasions and fail on others", leading to an indeterminacy of the outcome itself (Fox and Ülkümen, 2011:1). This point is worth questioning: One might reasonably argue that, given identical inputs into the mission, it would have identical outputs each time; thus, the uncertainty is always merely in the observer, who does not know what those inputs will be.

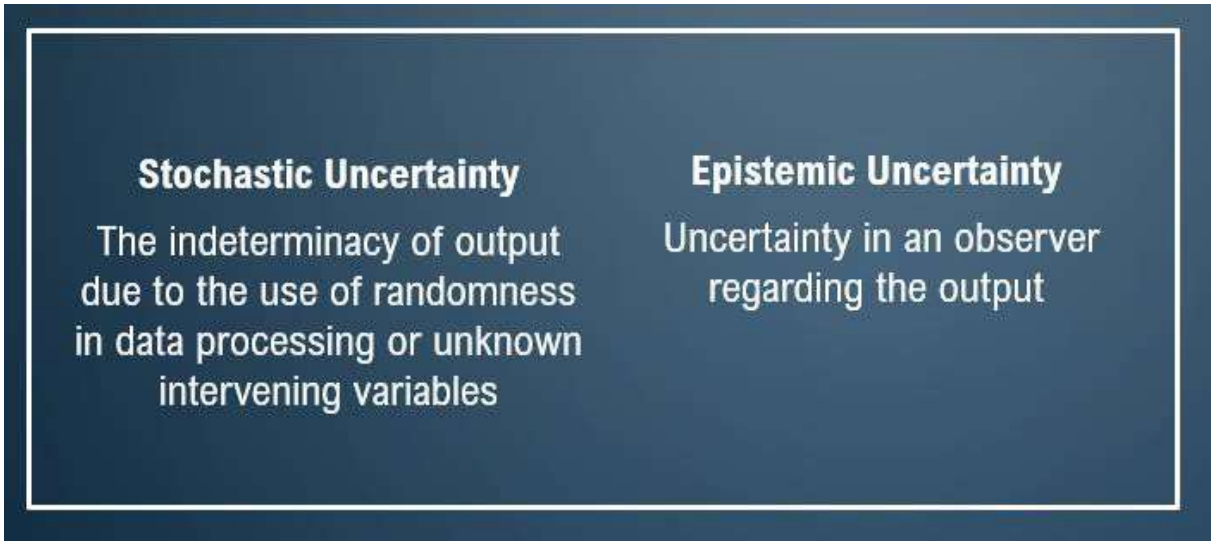

Figure 4: Stochastic Uncertainty and Epistemic Uncertainty: It is disputed whether the former collapses into the latter, but, more importantly, both cross-cut the spectrum of autonomy. Low Certainty, like Low Control, is neither entailed by autonomy nor exclusive to it.

Advantages

The previous section provided several concepts to qualify the discourse on autonomous weapons accountability. In particular, it provided a scalar and de-mystified notion of *autonomy* and further notions of *control* over and *certainty regarding* the probable output of a weapon. There were two notable ways in which these conceptualisations qualify the comparative discourse. First, control and certainty cross-cut the spectrum of autonomy. While it is tempting to define autonomy in terms of a decreased degree of human control, or the unpredictability of the output, these are also features of low-level autonomous weapons and non-autonomous weapons. This conclusion bears crucial implications for comparative assessments of weapon accountability.

Second, these concepts are transferable between artificial and human subjects. They are therefore better suited to comparative assessment. For humans, autonomy – if we once again relinquish the Pandoric subject of free will – similarly concerns the degree of *de facto* ability to perform tasks independently. Autonomy of this sort is materially the same for man, machine and anything else. Moreover, this conceptualisation does not result in a spectrum that runs simplistically from machine to man: A super-intelligent machine, with both the ability and the permission to perform more tasks, and more significant tasks, than a human, would exceed

that human's level of autonomy.

Similarly, High and Low Control and High and Low Certainty are concepts applicable, again, to both man and machine. Human agents of different kinds, in different positions, can offer varying degrees of control and epistemic access to other agents. Consider the infamous Eddie Chapman, or 'Agent Zigzag', the Nazi parachutist who became a double agent for the United Kingdom (MacIntyre, 2009), or the 'Cambridge Five' who relayed considerable intelligence to the Soviet Union (West and Tsarev, 2009). These agents - tools of their respective authorities - were not fully within the control of those authorities, nor did those authorities have knowledge of their most treacherous actions. In other words, like certain autonomous and non-autonomous weapons, these autonomous human subjects were Low Control and Low Certainty.

Conceptualising Accountability

Having conceptualised autonomy, High and Low Control weapons, and High and Low Certainty weapons, we now turn to accountability. As with our earlier conceptualisations, this section aims to equip the notion of accountability with comparability. In this way, it can be used to qualify comparative assessments of accountability for both human and artificial cases. Conceptually, this is perhaps even more challenging: Accountability, whatever it is, is widely understood in human terms. Thus, as in the previous section, this thesis here contests some dominant strains of thought, while nevertheless borrowing heavily from the rich and extensive literature available.

This section has five parts. First, it aims to provide a brief overview of the breadth and multidisciplinary nature of the conceptual discourse on accountability. While necessarily falling short of a comprehensive review, this section acts as a road map: it identifies aspects of accountability that must be considered for incorporation in our base-concept. Second, this thesis contests the agent-centricity of most conceptions of accountability across disciplines. It offers, in its place, a new approach: event-centricity. Third, I suggest that accountability amounts to a set of criteria or 'qualities'. In other words, numerous factors are relevant to its

potential achievement, not merely one. Fourth, I note that the importance of accountability is fundamentally contingent, being determined by normative and practical considerations, about which there can be legitimate disagreement. Building on this contention, I argue that the satisfactory achievement of accountability does not require the meeting of all criteria in all cases. Fifth, and finally, I conceptualise the obstacles to the meeting of accountability criteria. Here, I distinguish between what we may call 'theoretical' and 'practical' obstacles.

Accountability Across Literature

Accountability has been called a 'magical concept'; unsystematically defined, but magnetic in its appeal (Pollitt and Hupe, 2011: 643). While there are common threads, conceptualisations of accountability shift across disciplines. The two most prolific and relevant disciplines are the political and the legal. These differing conceptualisations are often framed as different 'types' of accountability, with some authors listing as many as eight (Stone et al, 1989). Instead of viewing political and legal accountability as different types of accountability, I conceive of these different disciplines as offering applications of the same base-concept.

First, I turn to the political theoretic discourse, which centres on **political accountability.** The discourse provides insight into both the **object** and **content** of political accountability. By 'object', I mean the object to which the discourse ascribes 'accountability' as a descriptor. By 'content', I mean the qualities or criteria which the discourse considers relevant to the essence of accountability.

As to the object of accountability, the political discourse overwhelmingly offers 'agents'. For these theorists, accountability predominately concerns governing agents and their accountability to their subjects – citizens. Accountability, therefore, centrally depends on the relations between 'agents' who possess 'agency'. An **agent** is simply "the actor in an accountability relationship who is to provide an account" (BBC Media Action, 2012: 3). More importantly, however, is the concept of **agency**, which all agents in an accountability relationship must supposedly possess: "An individual's ability to make meaningful choices; that is, the individual is able to envisage options and take action to exercise their right to choose" (BBC Media Action, 2012: 3).

This approach to accountability is affirmed by reams of political theoretic literature. The theorist

Przeworski, for example, frames political accountability "in the context of the principle-agent relationship between elected and administrative officials" (Przeworski et al, 1999: 298). Similarly, Lindberg argues that "an agent or institution who is to give an account" should be "included in the defining characteristics of any form of accountability" (Linberg, 2009: 8). Theories of democratic accountability, in particular, have been dominated by this 'agency model', in which "public officials are in an agency relationship with their electors, and the electorate is limited, for one reason or another, in its capacity to figure out what their agent is doing" (Ferejohn, 1999: 134). In addition to emphasising the agent-centricity of existing conceptualisations of accountability, this latter extract also demonstrates the supposed relationship between accountability and the degree of certainty possessed by agents. This lays important groundwork for the interaction between low-control weapons, autonomy, and accountability: low certainty creates the need for accountability and makes it harder to achieve.

The object of political accountability is clear. As to content, political accountability is ordinarily understood as fulfilling several necessary criteria. Among the most prominent scholars in this sphere is Schedler, who argues that political accountability carries "two basic connotations" (Schedler. 1999: 14). The first is answerability: "the obligation of public officials to inform about and to explain what they are doing" (Schedler, 1999: 14). The second is enforcement: "the capacity of accounting agencies to impose sanctions on power holders who have violated their public duties" (Schedler, 1999: 14). In sum, Schedler's influential conceptualisation of accountability is as follows: "A is accountable to B when A is obliged to inform B about A's (past or future) actions and decisions, to justify them, and to suffer punishment in the case of eventual misconduct" (Schedler, 1999: 17).

Przeworski similarly defines the content of political accountability as a kind of enforceable answerability (Przeworski et al, 1999: 298). In the practitioner's sphere, BBC Media Action offers an equivalent definition of accountability, as "the extent to which people, groups and institutions (principals) are able to hold government and other power holders (agents) responsible for their actions, and the extent to which government and other power holders provide a public account of their decisions and actions" (BBC Media Action, 2012: 3). Other scholars confirm such conceptions (Grant and Keohane, 2005: 29; Finer, 1941: 336). Thus, in the political theoretic discourse, the object of accountability is agents, and the content of accountability is their obligation to inform, their obligation to justify, and their punishability.

Another discourse that refines and applies the concept of accountability is the legal discourse. Here, again, we examine the proposed object and content. Legal notions of accountability are also ordinarily, and intuitively, agent-centric. This is because the content of **legal accountability** is closely related to the notions of culpability and liability, which also centre on agents. In its broadest form, culpability is "more or less synonymous with blameworthiness" (Berman, 2012: 441). In a narrower sense, it refers to "mental states (or in the case of negligence, pseudo- or quasi-mental states) that apply to elements in the actus reus" (Berman, 2012: 441); in other words, "the mental elements of an offense, such as whether the agent intended harm, foresaw harm, recklessly caused harm, or negligently caused harm" (Brink, 2014: 14). Crucially, it is contingent on particular mental states in the party, including willingness to perform the act under consideration and understanding of the act under consideration, although the scope of this contingency has been contested (Sarch, 2017: 719). Beliefs, desires, and intentions are all mental states which feed into whether or not an agent is culpable (Moore, 2010).

While strongly related to culpability, accountability - particularly in international law - has "not acquired a clearly defined legal meaning" (Brunnée, 2005: 22). The US legal dictionary offers a general definition: "that some legal rule(s) exists under which a theory or claim can be made to find one liable in a civil lawsuit or culpable in a criminal matter" (US Legal, 2019). In other words, legal accountability is a theory of the law relating directly to punishable culpability for an illicit action. This notion is similar to political theoretic notions, but simply requires that the behaviour, obligation to answer, and punishment in question are covered by law.

Agent-Centricity and Autonomous Weapons Accountability

The above discourses essentially offer differing applications of the same underlying concept. It is this 'base-concept' I now conceptualise. I begin with the proposition that the base concept of accountability should not be understood as agent-centric. Conceiving of it as such, I argue, is both misleading and inexpedient.

The agent-centricity of accountability means that 'agents' are the necessary object to which 'accountability' is applied. Agents, who possess agency, are the object which 'accountable' modifies. Even a cursory glance at this conceptualisation suggests its implications for the discourse on autonomous weapons. If agents are the sole subject of accountability, it is natural to question the agency of autonomous weapons, and to demand it as a pre-condition of the

satisfactory accountability of these weapons (Brozek and Jakubiec, 2017: 300).

Consequently, the discourse on the accountability of highly autonomous weapons, and AI more broadly, is replete with such queries. Houvast et al. emphasise that:

> Criminal liability as it is traditionally understood within the legal system is predicated on the moral agency and responsibility of the person or entity in question, as well as the need to prosecute guilt and carry out justice. The legal issues that emerge with respect to the criminal liability of robots therefore relate fundamentally to the robot's lack of consciousness and ability to make moral judgements (Houvast et al, 2018: 6).

Davey remarks:

> When a human soldier commits an atrocity and kills innocent civilians, that soldier is held accountable. But when autonomous weapons do the killing, it's difficult to blame them for their mistakes (…) An autonomous weapon's "decision" to murder innocent civilians is like a computer's "decision" to freeze the screen and delete your unsaved project (…) people rarely think the computer intended to complicate their lives (…) So who assumes the blame when autonomous weapons take innocent lives? (Davey, 2016: 1).

These extracts demonstrate that, due to agent-centric conceptualisations of accountability, an immediate agent who possesses particular mental properties is often considered essential to the satisfactory achievement of accountability. A different approach has been to maintain agent-centricity while diminishing its content (Van Genderen, 2018: 20), or to create intermediate categories of 'quasi-agency'. One author, for example, notes:

> As robots begin to approach more sophisticated human-like performances, it seems likely that they might be treated as quasi-agents or quasi-persons by the law, enjoying only partial rights and duties. A closely related concept will be that of diminished responsibility, in which agents are considered as being not fully responsible for their own actions. This will bring us to the more abstract concept of agency itself in the law, and how responsibility is transferred to different agents (Lin, Abney and Bekey, 2008: 55).

Four Problems with Agent-Centricity

Nevertheless, agent-centricity is in many ways an impediment to the discourse. It has four negative consequences.

First, agent-centricity leads to dense, peripheral, and often misguided inquiries into the

prospects for certain mental states in AI, whether robot 'consciousness' or relevant intent. At its worst, this search for agency, as a means to accountability, leads to wishful attributions of agency to even low-level autonomous machines. One famous study found that 83% of subjects "attributed either intentions or decision-making to the computer system – attributes that philosophers generally hold out as prerequisites for moral accountability" (Kahn et al., 2012: 33).

This quest for the mental properties of sentience, consciousness, or intent risks being not only ephemeral, but also needlessly contingent on humans possessing such traits themselves, and in a way that is truly meaningful to the ultimate goals of accountability. One might be forgiven for thinking that a belief in moral determinism would render the entire discourse on accountability null and void, as it would preclude the presence of many of the philosophically ambiguous – even ineffable – traits supposedly demanded of autonomous machines. As we shall see, this need not be true (Shrank, 1978: 801). Crucially, this thesis does not deny that mental states, and indeed the hardest philosophical inquiries regarding those states, are irrelevant to accountability (Himma, 2008: 20). It simply contests that such mental agents do not monopolise the set of qualities relevant to accountability, and that, as such, agents should not be the prime object to which accountability is applied.

Second, agent-centricity obscures the potential for accountability in highly autonomous weapons cases at the outset. If autonomous weapons may not be considered agents, but accountability is fundamentally agent-centric, then the use of autonomous weapons diminishes the prospects for accountability *a priori*. This is because the most immediate factor in the causal chain – the object which most directly caused the act under consideration, is non-agential and so exempt from the descriptor 'accountable'. As a result, authors come to lament an accountability 'gap': a distance between the agents who might be held accountable and the most immediate cause of the act (Bartlett, 2019: 1). This gap is understood as a necessary obstruction to the satisfactory attainment of accountability. The conception of accountability put forward in this thesis does not resolve or heavily bias the debate *a priori.*

Third, agent-centricity is surprisingly non-specific. In the case of political accountability, consider the agent-centric contention that "we must hold policemen accountable". While having a semblance of specificity, the veneer is thin, held together only by a common understanding

of the sorts of things that such a task might entail. The immediate question prompted by such a contention is "for what?" It is certainly not true that policemen either could be or should be 'accountable' for all that they do. Indeed, it is not clear that we are predominately concerned with the policemen themselves at all: we are concerned, rather, with what adverse outcomes – perhaps, the killing of unarmed civilians - might follow from them existing in the positions that they do. As will be elaborated, the coming conceptualisation of accountability provides a more specific, and thus more informative, object.

Fourth, agent-centricity suffers an incompleteness problem. Consider again the agent-centric claim that "we must hold policemen accountable". A further question we might ask is this: "Are the policemen the sole and entire cause and justifier of the adverse events, with which we are most concerned when seeking their accountability?" The answer is likely no. External to the policeman, there are numerous factors and agents contributing to the adverse outcome, capable of justifying that outcome, or providing further information regarding that outcome. To speak about police shootings in terms of "the accountability of policemen" creates a façade of completeness; in reality, it excludes a wealth of data and actors which are fully relevant to the accountability of the outcome concerned.

Accountability as Event-Centric

This thesis proposes that the base-concept of accountability should be conceptualised as event-centric. Accountability should be understood as attaching to events, not to agents, and certainly not to necessarily human agents. Under this conceptualisation, one modifies the discourse not to seek the accountability of policeman, but the accountability of police shootings of non-civilians; not the accountability of war criminals, but of war crimes; not the accountability of soldiers, but the accountability of indiscriminate targeting or disproportionality; not the accountability of autonomous weapons, but the accountability of what we have called their 'adverse outcomes'.

There are multiple advantages to this conceptualisation, paralleling the flaws in agent-centricity highlighted above. First, this conceptualisation offers a degree of specificity not provided by agent-centric notions. It does so by centring the discourse around exactly the phenomena with which we are concerned. If a news anchor remarks upon an accountability problem for hit-and-run accidents, child-soldier recruitment, the sale and distribution of blood diamonds, or the

depletion of coral reefs, we know exactly what we are talking about. Importantly, we know the problem prior to drawing any conclusions regarding which (many) agents feed into the problem and whether there are crucial non-agential components to the causal chain which produced this problem.

Second, this conceptualisation is more inclusive, and thereby more complete. To hold a hit-and-run driver accountable is not necessarily to resolve, nor to fully describe, the accountability problem for a hit-and-run accident. This is because a hit-and-run driver is just *one* of myriad factors that are relevant, both to the degree of accountability necessary for a hit-and-run accident and the degree to which it can be obtained: the adult passenger in the driving seat, the known absence of street cameras, the provision of alcohol to the underage driver, the prevalence of police in the area, the jaywalking of the pedestrian, the appropriate parenting of the driver, the manufacturing of the faulty car, the woman giving birth in the back-seat, and the lightning storm that obscured the driver's vision. Some of these factors involve agents, some involve culpability, some involve liability – others do not. All, however, are extremely relevant to the question of holding the event – a hit-and-run incident – accountable. Even if one were assured that all culpable mental states and actions of all hit-and-run drivers were known, fully explained, and punished where necessary, one could not be satisfied that all hit-and-run incidents were satisfactorily accountable. An event-centric notion of accountability would equip comparative analyses to be both inclusive and complete.

Third, an event-centric conceptualisation of accountability is far more permissive of the demonstrable fact that non-agents can be relevant to the achievement of accountability. This renders it suitable for the comparative discourse on highly autonomous weapons. Under an event-centric concept, there is no *a priori* accountability gap, and it is an open question whether the relevant criteria of accountability for the adverse outcomes of highly autonomous weapons can be satisfied. Ultimately, centring accountability around events, rather than agents, means that a degree of accountability can be maintained even when there are no culpable agents involved at all. This conceptualisation therefore meets our most important criteria: it operates for cases involving both man and machine, because it is applied to outcomes deriving from those subjects, not to the subjects themselves. It is therefore fundamentally comparative, qualifying – framing – how we may understand the differences in accountability between cases involving machines and those involving human agents.

Accountability as Having Numerous Criteria

Having offered its own conceptualisation of accountability's object, this thesis now turns to accountability's content. I begin with the proposition that accountability amounts to a broad set of criteria and should be given a maximalist conceptualisation. This is similar to the way in which some political theorists conceptualise democracy (O'Donnell, 2004; Diamond and Morlino, 2004: 22). For O'Donnell, democracy consists in a set of 'qualities', while for Diamond and Morlino it has multiple 'dimensions'. The critical point, however, is that the net cast for factors relevant to the enhancement of democracy is wide. The individual criteria are scalar, and an increase in the degree to which any criterion is met amounts to an increase in the degree to which democracy can be achieved. Similarly, this conceptualisation of accountability consists in a broad array of 'qualities' or 'criteria', which vary in relevance and attainability from case to case.

First, the accountability of an event may be increased by the existence of agents possessing culpability in the causal chain of that event. In the case of **legal culpability**, the accountability of an event may be enhanced by the existence of agents who are partially or legally culpable under the law. Some events have a high degree of legal culpability present, as where a culpable agent was the direct and immediate cause of an adverse outcome. In other cases, as we note in the next section, culpability is more distantly relevant, such as where a manufacturer or deployer is culpable for his role in the adverse outcomes of a highly autonomous weapon.

Moral responsibility can also enhance the presence of accountability in its base form, where agents are blameworthy under *moral* law (Eshleman, 2001). If a husband cheats on his wife, it is an important contributing factor to his overall accountability that he is indeed blameworthy for doing so; cheating need not be illegal for this to be true.

Countless sub-criteria can aid the fulfilment of the broader culpability and responsibility criteria, including the existence of the relevant moral and legal frameworks, the extent to which those frameworks are known and clear, the existence of certain mental states such as intent in the agents, and the presence of any of a vast rage of wholly and partially 'exculpating' factors. The last of these is important: while culpability is a highly agential contributor to accountability, factors contributing to exculpation, such natural dangers, can be entirely non-agential. Event-centricity is part of what equips us to appreciate this breadth of relevant factors.

A further set of criteria involves **epistemic accessibility.** Epistemic accessibility to the details of an event contribute to the accountability of that event. This is the scalar ability of observing agents or systems to know that the event happened and to understand exactly why. As we saw, this is often proxied by the notion of 'answerability' across literature: that idea that certain agents must be obligated to and able to answer for their actions. There are three things to note about this. First, epistemic accessibility is enhanced by a much wider range of factors a mere obligation on actors to answer for their actions. Smart surveillance technologies, increasingly effective data recording technologies such as blockchain, and AI data processors, can all enhance our epistemic access to events, even where the agents involved are entirely passive or unaware (Susskind, 2018: 134). Epistemic accessibility can also be enhanced by the accidents of an event (such as whether rain washed away the criminal's fingerprints) or by observers who took no culpable part in the event itself. Second, even answerability can be fully met for non-agents. Artificial intelligence regularly explains and justifies its actions to programmers. While AI explainability concerns are legitimate, it is also true that with a sufficiently good AI-user interface, AI can far surpass humans in answerability: Its code can offer the equivalent of every decision tracked out on paper; honest and lacking "the cognitive biases that make human explanations unreliable" (Doshi-Velez et al, 2017: 10).

A further set of criteria involves **deterrability.** The deterrability of an event is the extent to which that event can be prevented from occurring. This is often proxied by 'punishability'. While this thesis is very sympathetic to the limitation of accountability to 'punishability', it also notes that equivalent outcomes can be achieved by deterring non-agents, if they do not have the capacity to experience punishment – as is the case with existing AI. Our aim, when we hold agents to account, is far less to satisfy the ephemeral demands of retributive justice, than it is to simply discourage or prevent the repeat occurrence of the event. In 'Future Politics', Susskind makes an important further point on this subject: As technology advances, we may come to see fewer deterrent laws, and more 'locked doors': cases where, due to technological control, agents will simply be unable to do the illicit actions in question. It is difficult to suggest that, because the measures here concern 'deterrence' rather than 'punishment', then accountability becomes irrelevant. It seems plausible that those subject to technological controls of this kind are, on some level, *highly* accountable: their every action is watched, and if they attempt to do [X] under certain circumstances, they face immediate barriers. While punishability is often vital to

accountability, it is conceivable, in some cases, that deterrence and prevention can play normatively equivalent roles.

It is important to note that these changes are conceptual. Rather than affecting who is punished or responsible, they reframe and qualify the discourse, such that it is both more suited to comparative analyses across man and machine, and more sensitive to the non-agential contributors to accountability.

Accountability as Conditionally Important

Thus far, this thesis has repeatedly referred to the "satisfactory achievement" of accountability. If accountability is event-centric, all events are candidates. However, we do not care about accountability for every event to the same extent. The degree of accountability which we consider satisfactory is determined by both normative and practical considerations: in other words, it is of conditional importance. Moreover, not every criteria of accountability is required for the satisfactory achievement of accountability in every case.

In a 2015 skit, the acclaimed comedian Trevor Noah told the unfortunate tale of 'Death at a Funeral', recounting a series of deaths by lightning on a stormy weekend in the South African province of Eshowe (Smith, 2011: 1). This much was not a point of comedy, but Noah – and a host of commentators - was interested rather in the response of KwaZulu-Natal's MEC for Safety and Security: "We have heard what has happened on this funeral. Let it be known! We will launch a probe. We will not rest, until we know: Where does this lightning come from!" (Noah, 2019). While the quote has been donned with comical flourish, the story is true: Nomsa Dube did indeed call for a comprehensive probe into the deaths by the Department of Science and Technology. The comedy is in the ascription of a demand for accountability where, some might say, there ought to be none.

This case highlights that accountability for adverse outcomes is of conditional importance. However, it also emphasises the benefits of event-centricity. Despite that for some outcomes, like lightning strikes, the immediate and most significant cause is non-agential, *accountability can still matter for them*. For example, one court ruled that a worker's estate was entitled to survivor's benefits from his employer, after he was killed by lightning in an area of increased risk on his employer's property. The court observed: "Although Shelby did not have any control

over the lightening, due to his employment, the worker ended up in an area that increased the likelihood of a lightning strike" (Lumbreras, 2014: 1). The vital thing to note about this case is that *it does not involve a shift in the accountable agent for the lightning, from the lightning itself to the employer*. It is not as though the lightning could not be held accountable and so the employer, as next in line, was subsequently jailed for murder. Rather, the ruling indicates an admission that certain potential contributing factors to accountability – such as certain mental states in an agential, immediate causal factor; the lightening – were simply irrelevant to us in this case. Not all features of the event were subject to accountability, and accountability was satisfactorily obtained through the meeting of *some*, but not all, criteria of accountability.

In this way, our conceptualisation of accountability is again somewhat similar to maximalist conceptions of democracy. The extent to which democracy is desired is normatively and practically determined. Extreme forms of direct democracy, for example, might be undesirable overall, either because we think they produce bad outcomes (normative) or because we think they are too difficult to achieve given existing resources (practical). Equally, particular *criteria* of maximalist democracy are either more or less important depending on the case. Highly free media, for example, may be a luxury desire of polities which are not at high risk of being plummeted into conflict by the spread of extreme ideas.

Obstacles to Meeting Accountability Criteria

As noted, the criteria for the achievement of accountability are both manifold and discretionary. Not every criteria needs meeting in order for the degree of accountability to be satisfactory, and, moreover, the degree of accountability demanded is a normative question. It is now necessary to conceptualise the obstacles to the achievement of accountability. This thesis posits that there are two sorts of obstacles: theoretical and practical.

Theoretical obstacles are obstructions to the satisfaction of a particular criterion, which cannot be addressed because of something fundamental to the event. An obvious example is the non-existence of any law against a particular adverse outcome, which is a theoretical obstacle to the satisfaction of the legal culpability criteria. Certain sadists might be theoretically impossible to punish to an equivalent degree to other convicts. Psychopathy and insanity present theoretical challenges to the criterion of moral responsibility. Non-agential processes such as lightning strikes face theoretical obstructions to culpability and punishability. Highly autonomous

weapons present a theoretical obstacle to the satisfaction of the immediate culpability criteria: the existence of a culpable agent who directly caused the outcome in question. Nevertheless, as with lightening, there can exist agents who are culpable in other ways, whether for the programming, manufacturing, or deployment (Cass, 2014: 1049). Highly autonomous weapons also face, in some cases, the non-existence of the relevant laws and standards of care, a further theoretical obstacle (Kastan, 2012: 47).

For the sake of comprehensiveness, there are further theoretical obstacles to accountability that can be especially elucidated by theological examples. Consider God's execution of the firstborn sons of Egypt in Exodus 11. The event faces the theoretical obstacle of moral indeterminacy, given the infamous Euthyphro Dilemma which renders it difficult to judge the moral wrongness of a deity's act. It faces the theoretical obstacle of event indeterminacy, as it is theoretically impossible to demonstrate to a reasonable standard of proof whether the act took place and exactly how. It also faces theoretical obstacles to punishment, as it is not conceivable that a lone omnipotent being could be meaningfully punished. While this theological point may seem misplaced, the prospects for man-made superintelligence are routinely compared to God, and may face many similar theoretical obstacles to accountability.

Practical obstacles to accountability are simpler to understand: these are simply obstacles which render the satisfaction of a particular accountability criterion more difficult. Practical obstacles can be ameliorated through technological, scientific, political, legal, social, moral, or psychological progress. Examples of practical obstacles to accountability include a lack of street cameras in areas prone to violence or corrupt court systems. For highly autonomous weapons, as for humans, such obstacles are manifold. They include poor AI-user interfaces and explainability, low political will, and ambiguity regarding the distribution of culpability across agents involved in manufacture and deployment.

PART 3 – Case Studies

The aim of this thesis has been to qualify the difference between accountability for humans and highly autonomous weapons in war: creating common terms and a common framework. This qualification has involved the curation of a series of concepts uniquely suited to this comparative goal.

Autonomy has been conceptualised as an ability which can be possessed in equivalent ways by humans and machines alike. It has also been de-mystified as an increase in complexity: a scalar concept denoting the *de facto* ability to perform tasks independently, whether the land-mine, the Iron Dome, the Terminator, or the soldier.

A further distinction between High Control and Low Control has also been noted: the extent to which users possess control over the ultimate outputs of the methods and tools they use. This distinction cross-cuts the distinction between tools which possess a high degree of autonomy and those which do not. It is therefore an important qualification to our comparative analysis between highly autonomous weapons and human cases. These new concepts demonstrate that the lack of control over a weapon's output is not unique to, monopolised, or entailed by, a high degree of weapon autonomy. Moreover, weapons can be either Low Certainty or High Certainty, depending on the degree to which users possess certainty about the ultimate output of the weapon. This distinction, too, cross-cuts the spectrum of autonomy, therefore acting as an important qualifier to the comparative discourse on autonomous weapons. Finally, our conceptualisation of accountability qualifies the discourse in crucial ways. Event-centricity allows us to assess comparative accountability, without biasing the analysis towards agential sources of accountability *a priori*. Moreover, an inclusive, maximalist conception casts a wider net for the sorts of factors which may be thought to contribute to accountability.

To demonstrate the utility of these qualifications, one must apply them. The following applications fall into two categories. The first category examines the accountability of human soldiers and pilots in war, particularly examining the accountability failures and gaps during the Iraq war. It also notes examples of humans and their non-autonomous or low-level autonomous tools and methods of warfare. The second category examines some present highly autonomous weapon systems. It also takes a peak into futures beyond; the world of highly autonomous

weapons that do not yet exist. In many ways, the purpose of this initial categorisation is to later disassemble it. As our concepts assist in elucidating, many factors crucial to accountability cross-cut these distinctions between the human, the automated, and the autonomous. Our case-studies play a subsidiary role, completing the process of qualifying accountability differences across man and machine.

Humans in War

Humans commit mistakes, violations, and crimes against humanity in times of war. Often, the adverse outcomes of human activity in war face serious accountability problems. We must ask: What sorts of accountability problems are these? Given our new concepts, do these problems differ quantitatively and qualitatively from adverse outcomes procured by highly, and 'fully'[2], autonomous weapons? In other words, how do our concepts help to frame or 'qualify' these accountability differences?

The first point to be made is simple: humans frequently and systematically produce unaccountable outcomes in war. In other words, there is a quantitively extensive degree of unaccountability for humans in war. The track record of the US military and, in particular, the most recent Iraq War, is a case in point. The US has rarely meted out penalties for misdemeanours and misdeeds overseas. According to a review of military trials by the Washington Post, while thousands of civilians perished during the Iraq War (Sloboda, 2007: 1), just 20 Iraqi civilian casualties received any criminal investigation for their death: 39 US soldiers were charged in connection with these 20, but only 26 were charged with murder or manslaughter. Less than half of the 26 received any prison time, with the largest sentence spanning 25 years, and most receiving a year or less (White et al, 2006: 1). A former Marine prosecutor noted: "I think there are a number of cases that never make it to the reporting stage, and in some that do make it to the reporting stage, there has been a reluctance to pursue them vigorously" (White et al, 2006: 1). In other words, the criteria for epistemic accessibility and punishability are not met, due to practical obstacles such as low reporting levels, coercive

[2] It has been noted above, and will be further demonstrated, that this distinction is misleading.

failures, and low political will. Similar accountability problems have been infamously noted for peacekeeping abuses (Freedman, 2018: 963), crimes against humanity (Joyner, 1997: 155), and the responsiveness of the International Criminal Court more broadly (Goldsmith, 2003: 98). The latter study argues that the ICC is in fact *systematically* ill-equipped to punish serious abusers of human rights.

The second point to be made is that, while many of the above obstacles are practical, there are also theoretical obstacles to the accountability of humans in war. Evidence for illicit activity can be permanently lost or wiped clean, or, simply, there can be no existing law prohibiting the adverse outcome under the given circumstances. The last of these is not to be underestimated. The central tenets of Just War Theory, the foundation of the laws of war, have the ethical and legal permissibility of highly adverse outcomes built into them. Unarmed mothers and children alike may be permissibly killed in war, as long as such an outcome was proportional and unintended, though foreseen: such is the highly contested doctrine of double effect (Lee, 2004: 234).

A tragic example of this manifested in 2015, when an AC-130 Gunship accidentally attacked the Kunduz Trauma Centre, a hospital run by the well-known NGO Doctors Without Borders, killing over 40 doctors and patients, injuring another 30 (Médecins Sans Frontières, 2015: 1). In response, the US government issued a *mea culpa* and offered a mere $6000 to each of the families of those whose lives were lost (Shear and Sengupta, 2015: 1). No parties were punished for the incident and MSF received nothing in return (Ali, 2016). While the incident has been declared a war crime by some reports (Margulies, 2016), the US Department of Defence declared otherwise. A lack of intention – misidentification - was largely considered a sufficient exculpating factor.

Thus, as emphasised, the accountability of adverse outcomes is only conditionally important. Even though, for any mother, the death of her child by air strike is equally tragic in any case, there are some cases in which, through a normative lens, we simply do not care about holding agents to account. Instead, we have developed a system of norms that determines that the punishment criteria of accountability is, in this case, simply unimportant to us. In response to this problem, Crawford calls for an entirely new framework for responsibility in war, which moves away from individual agents and towards collective organisational responsibility. In this

way, she believes that double effect collateral damage can achieve at least some degree of accountability, despite its exculpation under international law (Crawford, 2013).

To summarise these two lessons, it is clear that the practical and theoretical obstacles to the accountability of humans in war are serious. Moreover, it is evident that the importance of the accountability criteria is conditional on norms and practice. Thus, it remains an option, at least, for highly autonomous weapons, that we draw a similar conclusion: that we understand their adverse outcomes as errors for which punishability is simply unnecessary.

The third lesson that might be learned from accountability failures for humans in war is, perhaps, more speculative. As conceptualised, humans and machines co-exist on the de-mystified scale of autonomy; soldiers simply have greater *de facto* ability to perform complex tasks independently. If this is so, humans may be conceptualised as autonomous weapons themselves: they are used by their superiors to perform lethal military tasks independently. Moreover, in many cases, humans are Low Control, Low Certainty autonomous weapons. This suggestion was foreshadowed by tales of the Cambridge Five and Agent Zigzag, but its applicability is rather more pervasive. From the pilots of the MSF incident to the infamous Blackwater contractors, combatants work for states as autonomous weapons, and the annals of wartime are replete with examples in which they are Low Control, Low Certainty tools of war

– sometimes going rogue, shooting to miss, abusing non-combatants, fleeing combat, or otherwise eschewing the expectations of their superiors.

It is insufficient to object simply that autonomous soldiers can be punished while highly autonomous weapons can be not. This is because we regularly *do not* see fit to punish soldiers for their actions in war, even where they produce highly adverse outcomes. In some cases, as just seen, it is because we make a normative judgement that punishment is not necessary, despite that the outcomes concerned are adverse. In other cases, however, we determine that *other* actors can be held responsible for the adverse actions of these 'autonomous soldiers'. States, in particular, are often the culpable and accountable actors for actions committed by agents on the ground. As a result, some scholars have rightly noted that highly autonomous weapons could be held accountable in very similar ways – through tort law and the responsibility of states (Crootof, 2015: 1347). Here, our functionally comparative conceptions of 'autonomy' and 'accountability' equip us to see the *commonalities* in the sorts of accountability problems

faced by man and machine alike: They therefore highlight the scope for adapting human accountability solutions to the accountability problems of highly autonomous weapons – namely through the culpability of states in tort law.

Highly Autonomous Weapons in War

The previous section addressed the accountability of humans in war and the ways in which our new concepts qualify the discourse. Through this analysis, some indications were made as to the similarities between human and artificial accountability problems. This section aims to take a more concrete look at how a highly autonomous weapon might compare, given the introduced concepts.

Few highly autonomous weapons are used currently, but it is reasonably likely that our future weapons will be predominately autonomous. As mentioned above, one example of such a weapon is the Aegis combat system. Found on battleships, the brainchild of Lockheed Martin, this weapon system is capable of attacking land targets, submarines, and surface ships simultaneously, while at the same time protecting its carrier fleet against any air targets without any human intervention whatsoever (Lockheed Martin, 2019: 1). The Phalanx CIWS is another seaborne autonomous weapon, capable of shooting anti-ship missiles and helicopters, with a firing rate of 4000 rounds per minute (NavWeaps, 2018: 1).

Highly autonomous weapons, however, are not limited to the sea. Sentry guns are rising in popularity, found presently in South Korea and Israel. Because both countries have mandatory military service, they have been the first to try such systems, with the express intent of reducing the need for human service. The most autonomous sentry guns reside on the south side of the DMZ.

Seeming to herald the age of true science fiction, the South Korean SGR-A1 is the most advanced sentry gun on the planet. The system operates based on motion and heat sensors, cameras that operate in low light, and advanced pattern-recognition capabilities. In daylight, it can identify targets at a distance of two miles; by night, one mile (Velez-Green, 2015: 1). The capabilities of the SGR-A1 are profound; the substance of the future. The weapon can not only

warn intruders verbally and identify actions of surrender, but can also engage non-submissive subjects with a hail of gunfire, at an 800 metre distance (Velez-Green, 2015: 1). While South Korea and the manufacturer have consistently denied these capabilities, they have been confirmed repeatedly by watchdogs and researchers (Arkin, 2009: 11; Andersen, 2012: 1; Jeffries, 2014: 1). If this weapon were to misread its inputs – to see defiance where there is actually compliance, and to shoot a person dead mid-surrender – we would face exactly the accountability fears so often exclaimed.

One way in which such autonomous weapons would produce these accountability concerns is through Low Certainty. Due to the extreme complexity and contingency of advanced machine learning processes, artificial intelligence has the capacity to become a black box. As noted, this uncertainty is not reflective of true indeterminacy, but of epistemic uncertainty: users cannot predict outputs – such as the sentry gun killing a surrendered individual – nor can they, at times, even account for the origins of particular outputs when they are produced.

To demonstrate this point, consider a non-military example: DeepMind's AlphaGo, the first AI to beat the world champion at a game of Go; a game vastly more complex than chess and widely understood to require intuition and even creativity to master. In the second game of the grand stand-off between Lee Sedol and AlphaGo, AlphaGo made its 37th move: a move considered so extraordinarily unusual to observers that none could account for the decision.

The move ultimately enabled AlphaGo's victory, despite that even AlphaGo's internal software showed that only one in ten thousand humans would make such a decision (Metz, 2016: 1). Yet, the programmers and Go experts alike, even retrospectively, could not understand *why* AlphaGo made the decision that it did. The black box process of machine learning produced a highly unpredictable output for users. This raises the question: If Move 37 had been a legally relevant, adverse military outcome, how could accountability be achieved? If the decisions of a sentry gun are black-boxed in this way, who is to blame?

The challenge presented here brings many of our concepts together. From AlphaGo to sentry guns, autonomy can contribute to Low Control and Low Certainty. This, in turn, appears to contribute to a lack of accountability: If users cannot predict the outputs of the tools they use, their culpability becomes more ambiguous. The conceptual qualifications provided by this thesis offer important insights.

To demonstrate these insights, let us return momentarily to human cases. Humans in war, using entirely non-autonomous weapons, often suffer the malfunctioning of those weapons, at times to lethal effect. In the above section, we saw the example of the Kunduz Hospital incident, in which technical failings played some role in contributing to the misidentification of the target. A simpler example occurred in 2018, when the US army found that 881 M4A1s had been affected by a functioning problem, which led to unintentional firing when switching from "semi" to "auto" (South, 2018: 1). This provides an important comparative point. The nature of the incidents in which a non-autonomous weapon malfunctions does not seem qualitatively different from the nature of incidents in which a highly autonomous weapon 'malfunctions' (or, more ephemerally, 'makes the wrong decision').

When such accidents take place in non-autonomous weapons, the tools used by humans are still entirely Low Control – in exactly the same sense as would be true if a fully autonomous weapon performed an adverse action. In both cases, an *adverse outcome* is produced, it is produced *beyond the control* of the human user, and the user himself is *non-culpable* for this outcome. In other words, there is little qualitative difference between this kind of malfunctioning in non-autonomous weapons and the faulty decisions of a highly autonomous weapon.

What is to be said, then, for the highly autonomous weapons of the future? Building on our qualifying concepts, this thesis suggests that the discourse has exaggerated the distinction between current autonomous weapons and the supposedly novel 'fully' autonomous weapons of the future. There are three reasons for this.

First, as has already been demonstrated, the notion that there are 'fully' autonomous weapons on the horizon, and that they must be banned before arrival, is more than misleading. The absence of human veto power over the function of machine killing already exists, whether in landmines, sentry guns, or faulty equipment.

Second, as our concepts show, weapon autonomy is not the only thing that leads to humans having low control over their weapons. Sometimes this is because of the nature of the method, as with anthrax letters. Other times, humans have low control because of errors and malfunctions – even non-autonomous weapons make lethal mistakes. This means that the centrality of autonomy – the idea that we are about to step over the line and relinquish our control – is misplaced: humans already face incredibly Low Control across a wide range of

autonomous and non-autonomous weapons.

Third, strong claims regarding the coming shift to 'full' autonomy overlook a vital factor. Imagine two kinds of weapons: Prescriptive Military AI (PMAI) and Active Military AI (AMAI).[3] PMAI is a highly intelligent piece of *advisory* military software, which instructs users regarding optimal outcomes, but possesses no *de facto* autonomy of its own, and cannot enact its own prescriptions. AMAI, conversely, is the software of PMAI attached to the mechanics necessary to perform advanced military tasks. In PMAI, humans possess total control over the ultimate output. In AMAI, they possess no control at all over the output. According to the current discourse, PMAI is not an autonomous weapon, while AMAI is the fully autonomous weapon we fear. Is this leap from PMAI to AMAI that is thought to be so monumentally dangerous. Yet, not only does AMAI already exist in many forms (as discussed), but PMAI has the potential to be *equivalently problematic*. This is simply because the advice of highly intelligent PMAI will become extremely difficult to assess, weigh, and reasonably reject. Once the reliability of the PMAI has been repeatedly demonstrated, users will come to trust it, signing off on its decisions every time. If PMAI identifies an individual as a combatant, the user will accept this ascription. For all intents and purposes, humans will act as the active arm of the PMAI: where the PMAI errs, so too will the human. This, too, critically undermines the notion that 'full' autonomy represents a distinct and novel shift.

Once again, this thesis illustrates that the spectrums of control, certainty, autonomy, and accountability cross-cut each other. These conceptual qualifications give nuance to dichotomous distinctions across the board: from the automated and the autonomous, to present autonomy and 'full' autonomy, to human accountability and artificial accountability. In truth, much of what we fear is already here, in quantitatively and qualitatively similar forms.

Conclusion

It is a strange thesis that invokes lightning, God, comedy, and killer robots. Yet its purpose has been a precise one: to qualify the differences in accountability between highly autonomous weapons and humans in war. The thesis provided a set of concepts uniquely suited to accurate,

[3] This distinction is created in a previous paper by the author.

nuanced, comparative analysis across human and artificial cases. These concepts included a de-mystified spectrum of autonomy, which accommodated the landmine, the human, and the Terminator alike; two new cross-cutting distinctions amongst weapons which proffer higher and lower degrees of control and certainty to users; and an event-centric, maximalist notion of accountability. Accountability was also conceptualised as having only conditional importance and as facing both theoretical and practical obstacles. The concepts were put to use in a series of case studies, involving both man and machine.

These concepts demonstrated that autonomy neither entails nor monopolises a low degree of control or a low degree of uncertainty. Rather, control and certainty cross-cut the spectrum of autonomy: at times, autonomy enhances certainty and control, at other times, the opposite. Thus, accountability problems do not track the spectrum of autonomy either: qualitatively and quantitatively similar accountability problems are faced across human, non-autonomous, and highly autonomous cases. Qualifying the conceptual discourse to accommodate these commonalities allows for more accurate and more appropriately comparative analyses.

There are two directions in which such conclusions can be taken. On one level, as we have seen, these commonalities highlight the scope for adapting present approaches to human accountability to the highly autonomous weapons of the future. On the other hand, these conclusions carry tremendous normative freight: If we worry enough about the accountability problems presented by autonomous weapons to ban them entirely, then the status quo of modern warfare should fill us with dread. Qualitatively similar accountability problems are rife amongst humans in war. To ban a weapon on account of such problems is fallacious, unless accompanied by an admittance that such problems are prohibitive of the status quo; prohibitive of modern warfare.

In summary, the leap from the status quo to a uniquely dangerous, unaccountable and fully autonomous future is often prophesised. Yet there is strong conceptual reason to see this prophesy as a poorly qualified one. In reality, the world already exhibits many of the things it fears, in the form of existing full autonomy, a plethora of Low Control and Low Certainty non-autonomous weapons, and, crucially, the most radically unaccountable autonomous weapons on earth: humans.

References

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival*, *59*(5), 117-142.

Amoroso, D., & Tamburrini, G. (2017). The ethical and legal case against autonomy in weapons systems. *Global Jurist*, *18*(1).

Andersen, R. (2012, Mar 22). Cyber and Drone Attacks May Change Warfare More Than the Machine Gun. *The Atlantic.* Retrieved from: https://www.theatlantic.com/technology/archive/2012/03/cyber-and- drone-attacks-may-change-warfare-more-than-the-machine-gun/254540/

Anderson, K., & Waxman, M. C. (2017). Debating Autonomous Weapon Systems, their Ethics, and their Regulation under international law.

Arkin, R. (2009). Governing lethal behavior in autonomous robots. *Chapman and Hall/CRC*.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, *94*(886), 687-709.

Barnes, J. E., & Chin, J. (2018, Mar 2). The New Arms Race in AI*. The Wall Street Journal*. Retrieved from: https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261

Bartlett, M. (2019). Solving the AI Accountability Gap. *Towards Data Science.* Retrieved from: https://towardsdatascience.com/solving-the-ai-accountability-gap-dd35698249fe

Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.*, *31*, 889.

BBC Media Action. (2012). Conceptualising Accountability: An Approach to Measurement. *BBC Media Action.*

Berberich, N., & Diepold, K. (2018). The Virtuous Machine-Old Ethics for New Technology? *Cornell University,* 1-25.

Berman, M. N. (2012). Introduction: Punishment and Culpability. *Ohio State Journal of Criminal Law, 9:44, 441-448.*

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135- 148.

Brink, D. O. (2014). Two Kinds of Culpability. *San Diego Legal Studies Paper*, (14-178).

Brożek, B., & Jakubiec, M. (2017). On the legal responsibility of autonomous machines. *Artificial Intelligence and Law*, *25*(3), 293-304.

Brunnée, J. (2005). International legal accountability through the lens of the law of state responsibility. *Netherlands Yearbook of International Law*, *36*(1), 21-56.

Buchanan, A., & Keohane, R. O. (2015). Toward a Drone Accountability Regime. *Ethics & International Affairs,* 29(1), 15-37.

Cass, K. (2014). Autonomous Weapons and Accountability: Seeking Solutions in the Law of War. *Loy. LAL Rev.*, *48*, 1017.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks.

Association for the Advancement of Artificial Intelligence, 36-40.

Crawford, N. (2013). *Accountability for Killing: Moral Responsibility for Collateral Damage in America's Post-9/11 Wars*. Oxford University Press.

Crootof, Rebecca. "War torts: Accountability for autonomous weapons." *U. Pa. L. Rev.* 164 (2015): 1347.

Cummings, M. (2017). Artificial intelligence and the future of warfare. *Chatham House for the Royal Institute of International Affairs*.

Davey, T. (2016). Who is Responsible for Autonomous Weapons? *Future of Life Institute.* Retrieved from: https://futureoflife.org/2016/11/21/peter-asaro-autonomous-weapons/

Diamond, L., & Morlino, L. (2004). The quality of democracy: An overview. *Journal of democracy*, *15*(4), 20-31.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Eshleman, A. (2001). Moral responsibility. *Stanford Encyclopaedia of Philosophy (Winter 2016 Edition).*

Retrieved from: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=moral-responsibility

Ferejohn, J. (1999). Accountability and authority: toward a theory of political accountability. *Democracy, accountability, and representation*, *131*.

Finer, H. (1941). Administrative responsibility in democratic government. *Classics of administrative ethics*, 5-26.

Foster, D. (2003). The Message in the Anthrax. *Vanity Fair*, (518), 180ff. Retrieved from: http://anthraxinvestigation.com/AnthraxFoster.pdf

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *Perspectives on thinking, judging, and decision making*, 21-35.

Freedman, R. (2018). Unaccountable: a new approach to peacekeepers and sexual abuse. *European Journal of International Law*, *29*(3), 961-985.

Future of Life Institute. (2019). Lethal Autonomous Weapons Pledge. *Future of Life Institute*. Retrieved from: https://futureoflife.org/lethal-autonomous-weapons-pledge/?cn-reloaded=1

Garcia, M. (2017, Feb 13). How to Keep Your AI from Turning into a Racist Monster. *Wired*. Retrieved from: https://www.wired.com/2017/02/keep-ai-turning-racist-monster/

Goldsmith, J. (2003). The self-defeating international criminal court. *U. Chi. L. Rev.*, *70*, 89.

Grant, R. W., & Keohane, R. O. (2005). Accountability and abuses of power in world politics. *American political science review*, *99*(1), 29-43.

Grut, C. (2013). The challenge of autonomous lethal robotics to International Humanitarian Law. *Journal of conflict and security law*, *18*(1), 5-23.

Guterres, A. (2018, Nov 6). 2018 Web Summit - António Guterres (UN Secretary-General) on Technological advances. *United Nations.* Retrieved from: https://www.youtube.com/watch?v=0B0UdN5DpD0 [8.50]

Helton, J. C. (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*, *57*(1-4), 3-76.

Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: an African perspective. *South African journal on human rights*, *33*(1), 46-71.

Himma, K. E. (2009). 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), 19-29.

Houvast, F., Timmerman, R., & Zwaan, Y. (2018). Exploring the Legal Rights and Obligations of Robots: A Legal Book Review of I, Robot by Isaac Asimov. *Utrecht University.*

Jeffries, A. (2014, Jan 28). Should a Robot Decide when to Kill? *The Verge.* Retrieved from: https://www.theverge.com/2014/1/28/5339246/war-machines-ethics-of-robots-on-the-battlefield

Joyner, C. C. (1996). Arresting impunity: The case for universal jurisdiction in bringing war criminals to accountability. *Law & Contemp. Probs.*, *59*, 153.

Joyner, C. C. (1997). Redressing Impunity for Human Rights Violations: The Universtal Declaration and the Search for Accountability. *Denv. J. Int'l L. & Pol'y*, *26*, 591.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... & Severson, R. L. (2012, March). Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 33-40

Kastan, B. (2013). Autonomous weapons systems: a coming legal singularity. *U. Ill. JL Tech. & Pol'y*, 45.

Kovic, M. (2018). The strategic paradox of autonomous weapons. *Zurich Institute of Public Affairs Research,* 1-16.

Lee, S. (2004). Double effect, double intention, and asymmetric warfare. Journal of Military Ethics, 3(3), 233-251.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. California Polytechnic State Univ San Luis Obispo.

Lindberg, S. I. (2009). Accountability: the core concept and its subtypes. *Africa Power and Politics Programme Working Paper*, *1*.

Lockheed Martin. (2019). Aegis: The Shield of the Fleet. *Lockheed Martin.* Retrieved from: https://www.lockheedmartin.com/en-us/products/aegis-combat-system.html

Lumbreras, C. (2014, May 2). Is Worker's Death by Lightning Strike Compensable? *Risk & Insurance*. Retrieved from: https://riskandinsurance.com/workers-death-lightning-strike-compensable/

Macintyre, B. (2008). *Agent Zigzag: A True Story of Nazi Espionage, Love, and Betrayal*. Broadway Books.

Margulies, P. (2016). Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts. *Research Handbook on Remote Warfare, Edward Elgar Press, Jens David Ohlin ed*.

Médecins Sans Frontières. (2015, Dec 12). Updated Death toll – 42 people killed in the US airstrikes on the Kunduz hospital. *Médecins Sans Frontières.* Retrieved from: https://www.msf.org/kunduz- updated-death-toll-%E2%80%93-42-people-killed-us-airstrikes-kunduz-hospital

Metz, C. (Mar 3, 2016). How Google's AI viewed the move no human could

understand. *Wired.* Retrieved from: https://www.wired.com/2016/03/googles-ai-viewed-move-no- human-understand/

Moore, M. S. (1996). Prima facie moral culpability. *BUL Rev.*, *76*, 319.

NavWeaps. (2018). 20mm Phalanx Close-in Weapon System (CIWS). *Naval Weapons, Naval Technology and Naval Reunions.* Retrieved from: http://www.navweaps.com/Weapons/WNUS_Phalanx.php

Noah, T. (2019, May 9). "Death at a Funeral" - Trevor Noah - (Crazy Normal) - Longer Re-Release. *Crazy Normal.* Retrieved from: https://www.youtube.com/watch?v=B50sVK_VT4A [1.45]

O'Donnell, G. (2004). The quality of democracy: Why the rule of law matters. *Journal of democracy*, *15*(4), 32-46.

Pollitt, C. C., & Hupe, P. P. (2011). The role of magic concepts. *Public Management Review: an international journal of research and theory*, *13*(5), 641-658.

Przeworski, A., Stokes, S. C., & Manin, B. (Eds.). (1999). *Democracy, accountability, and representation* (Vol. 2). Cambridge University Press.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). *Autonomous machines, moral judgment, and acting for the right reasons.* Ethical Theory and Moral Practice, *18*(4), 851-872.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, *18*(4), 851-872.

Red Cross. (2014). Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects. *International Committee of the Red Cross.*

Roff, H. M. (2019). Artificial Intelligence: Power to the People. *Ethics & International Affairs*, *33*(2), 127-140.

Sarch, A. (2017). Who cares what you think? Criminal culpability and the irrelevance of unmanifested mental states. *Law and Philosophy*, *36*(6), 707-750.

Scharre, P, & Horowitz, M. (2015). An Introduction to Autonomy in Weapon Systems. *Centre for a New American Security.*

Scharre, P. (2016). Autonomous weapons and operational risk. *Center for a New American Security*, 1-54.

Schedler, A. (1999). Conceptualizing accountability. *The self-restraining state: Power and accountability in new democracies*, *14*.

Sharkey, N. E. (2012). The Evitability of Autonomous Robot Warfare. *International Review of the Red Cross*, *94*(886), 787-799.

Shear, D., & Sengupta, S. (2015, Oct 7). Obama Issues Rare Apology Over Bombing of Doctors Without Borders Hospital in Afghanistan. *The New York Times*. Retrieved from: https://www.nytimes.com/2015/10/08/world/asia/obama-apologizes-for-bombing-of-afghanistan- hospital.html

Shrank, I. (1978). Determinism and the Law of Consent-A Reformulation of Individual Accountability for Choices Made without Free Will. *Suffolk UL Rev.*, *12*, 796.

Singer, P. (2008). *Robots at war.* Wilson Quarterly, 30-48.

Sloboda, J. (2003, May 27). 100 Names of Civilians Killed – and only 2% of a Vital Task Completed. *Iraq Body Count.* Retrieved from: https://www.iraqbodycount.org/analysis/beyond/100-names/

Smith, D. (2011, Jan 4). South African Politician says Number of Lightning Deaths is Rising. *The Guardian.* Retrieved from: https://www.theguardian.com/world/2011/jan/04/south-african-lightning- deaths

South, T. (2018). Here is how the army is fixing its M4 Misfire Problem. *Army Times*. Retrieved from: https://www.armytimes.com/news/your-army/2018/09/24/heres-how-the-army-is-fixing-its-m4-misfire- problem/

Stone, B. (1995). Administrative accountability in the 'Westminster'democracies: Towards a new conceptual framework. *Governance*, *8*(4), 505-526.

Stop Killer Robots. (2019). The Threat of Fully Autonomous Weapons. *Stop Killer Robots.* Retrieved from: https://www.stopkillerrobots.org/learn/

Susskind, J. (2018). *Future politics: Living together in a world transformed by tech*. Oxford University Press.

Teffer, P. (2018, Oct 24). Europe debates AI - but AI is already here. *EU Observer*. Retrieved from: https://euobserver.com/science/143137

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Thynne, K. (2008). The international criminal court: A failure of international justice for victims. *Alta. L. Rev.*, *46*, 957.

United Nations. (2019). War Crimes. United Nations Office on Genocide Prevention and the Responsibility to Protect. Retrieved from: https://www.un.org/en/genocideprevention/war-crimes.shtml

US Legal. (2019). Accountability Law and Legal Definition. *US Legal.* Retrieved from: https://definitions.uslegal.com/a/accountability/

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, *20*(1), 27-40.

Van Genderen, R. V. D. H. (2018). Do We Need New Legal Personhood in the Age of Robots and AI?. In *Robotics, AI and the Future of Law* (pp. 15-55). Springer, Singapore.

Van Inwagen, P. (2000). Free will remains a mystery: The eighth Philosophical Perspectives lecture. *Philosophical perspectives*, *14*, 1-19.

Velez-Green, A. (2015, Mar 1). The Foreign Policy Essay: The South Korean Sentry—A "Killer Robot" to Prevent War. *Lawfare.* Retrieved from: https://www.lawfareblog.com/foreign-policy-essay-south- korean-sentry%E2%80%94-killer-robot-prevent-war

Weir, R. (2015, Aug 12). Accountability on Drones Continues to Fall Short. *Open Society Foundations*. Retrieved from: https://www.opensocietyfoundations.org/voices/accountability-drones-continues-fall- short

West, N., & Tsarev, O. (Eds.). (2009). *Triplex: Secrets from the Cambridge Spies*. Yale University Press.

White, J., Lane, C., *&* Tate, J. (2006, Aug 28). Homicide Charges Rare in Iraq War Few Troops Tried for Killing Civilians. *The Washington Post.* Retrieved from: https://www.washingtonpost.com/archive/politics/2006/08/28/homicide-charges-rare-in-iraq-war-span- classbankheadfew-troops-tried-for-killing-civiliansspan/04a88326-c54c-4e2b-9ce7-501723f834a3/?noredirect=on

Whiteman, H. (2018, Feb 12). Alcohol 'More Damaging to Brain Health than Marijuana'. *Medical News Today.* Retrieved from: https://www.medicalnewstoday.com/articles/320895.php

Yudkowsky, E. (2016). The AI Alignment Problem: Why it is Hard, and Where to Start. *Symbolic Systems Distinguished Speaker*.