Autonomous Weapons in Humanitarian Law: Understanding the Technology, Its Compliance with the Principle of Proportionality and the Role of Utilitarianism

Elliot Winter*

DOI: 10.21827/5b51d56abd19e

Keywords
INTERNATIONAL HUMANITARIAN LAW; LAW OF ARMED CONFLICT; DISARMAMENT; LETHAL AUTONOMOUS WEAPONS; AUTONOMOUS VEHICLES; PROPORTIONALITY; UTILITARIANISM; BENTHAM

Abstract

Autonomous machines are moving rapidly from science fiction to science fact. The defining feature of this technology is that it can operate independently of human control. Consequently, society must consider how 'decisions' are to be made by autonomous machines. The matter is particularly acute in circumstances where harm is inevitable no matter what course of action is taken. This dilemma has been identified in the context of autonomous vehicles driving under the regulation of domestic law and, there, governments seem to be moving towards a utilitarian solution to inevitable harm. This leads one to question whether utilitarianism should be transposed into the context of autonomous weapons which might soon operate on the battlefield under the gaze of humanitarian law. The argument here is that it should because humanitarian law includes the core principle of 'proportionality', which is fundamentally a utilitarian concept – requiring that any gain derived from an attack outweighs the harm caused. However, while human soldiers are always able to come to a view on proportionality, albeit subjective, there is much doubt over how an autonomous weapon might determine what is proportionate. There is a very large gap between our embryonic understanding of utilitarianism in relation to autonomous vehicles manoeuvring around a city on one hand; and what would be required for armed robots patrolling a battlespace on the other. Bridging this gap is fraught with difficulty but perhaps the best starting point is to take Bentham's expression of utilitarian mechanics and build upon them. With conscious effort and, ideally, collaboration, states could use the process of applying his classic theory to this very modern problem to raise the standard of protection offered to those caught up in conflict.

Introduction

'Suppose there is a driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed.'[1]

The above extract is the classic iteration of the 'tram problem' and was posed by Foot to demonstrate the ethical conundrum that arises in situations where harm of some

* The author is a Teaching Fellow at Newcastle Law School (elliot.winter@newcastle.ac.uk). The article is dedicated to the memory of Dr James Upcher.

1 Foot, P, "The Problem of Abortion and the Doctrine of the Double Effect" *in* Foot, P, ed, *Virtues and Vices and Other Essays in Moral Philosophy* (University of California Press 1978).

sort is inevitable, and where a decision must be made to determine which harm is allowed to manifest. In recent years, the problem has come back into focus with the rise of 'autonomous' technology where, instead of a human being having to decide whether to pull the proverbial lever, it will be left to a machine to make the call. The matter is perhaps of greatest moment in the context of autonomous vehicles where manufacturers and governments alike are struggling to come up with definite solutions to this most vexed of problems. The earliest indication available is that states might move towards a utilitarian solution to the tram problem when it comes to autonomous vehicles.[2]

Of course, autonomous technology is not confined to vehicles. The principal question for present purposes is whether a utilitarian solution is right for 'autonomous weapons'. These are machines that can act independently in the battlespace and whose deployment inherently involves artificial intelligence assuming some degree of responsibility for critical assessments.[3] Such weapons used in an international armed conflict will be governed by humanitarian law, a core principle of which is 'proportionality'. In essence, this principle requires that the harm caused by an attack must not exceed the gain garnered from it. While the concept itself is clear, the practicalities of determining whether harm exceeds gain in any particular scenario are not.

This article will explain that the best starting point is to recognise that the principle of proportionality is analogous to the principle of utility – the former requiring more gain than harm; the latter more pleasure than pain. From there, it becomes clear that the various mechanisms developed by Bentham in the eighteenth century to enable application of utilitarianism can now be carried over and used to apply proportionality.[4] Of course, these mechanisms must be taken from their abstract form and given more concrete meaning based on the sorts of harm and gain that might be expected to arise in the context of armed conflict. Thereafter, the matter can be passed to policy makers, military officials and computer programmers to create algorithms that can implement the relevant mechanisms on the battlefield. This process presents an opportunity for states, acting alone or in concert, to hold this emerging technology to tougher standards than presently demanded by humanitarian law. Indeed, amid failure to achieve an outright ban on autonomous weapons, this is perhaps the best compromise available.

It should be noted that the ambit of this article is strictly limited. Proportionality will be considered only in its humanitarian law (*or jus in bello*) sense and as it would apply in the context of an international armed conflict. Proportionality in its other myriad contexts, such as *jus ad bellum*, *jus post bellum*, human rights and so on, will not be considered as, in those areas, it has evolved with nuanced differences in meaning.[5] Similarly, there are other rules of humanitarian law which have a bearing on the use of autonomous weapons. For example, the rule of 'distinction' is a fundamental rule which requires parties to a conflict to discern military objectives from civilians and civilian objects.[6] Clearly, distinguishing targets from non-targets is a prerequisite of any proportionality assessment; however, that is a separate issue for another article, as is the ethical nature (or otherwise) of an attack by a machine.[7] Finally, the article will not

---

2 As will be explained below, Germany is the first state to head in this direction.

3 The working definition proposed by the US for autonomous weapons is supplied below.

4 Bentham, J, *The Principles of Morals and Legislation* (Prometheus Books 1988).

5 For full discussion of proportionality, see Newton, M and May, L, *Proportionality in International Law* (Oxford University Press 2014).

6 *Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of International Armed Conflicts*, 8 June 1977, 1125 UNTS 3, Article 48 [*Additional Protocol I*].

7 See Leveringhaus, A, *Ethics and Autonomous Weapons* (Palgrave Pivot 2016).

attempt to specify what the final algorithms should look like, only to elucidate one principle underpinning them.

I. The Technology

A. Understanding autonomous weapons

Humanity's level of technological sophistication continues to grow at an exponential rate and much innovation can currently be found in the area of automation. Indeed, Bagrit predicted decades ago that we would witness an 'age of automation' where machines increasingly take over activities performed by humans.[8] The autonomy phenomenon can be encountered in factory production processes, vehicular transportation and even space exploration, but there are also important developments in military technology. It is important to grasp the meaning, novelty and significance of autonomy in military technology to understand why it has prompted the present line of enquiry.

The starting point is to define what is meant by 'autonomy', yet this effort can quickly deteriorate into a confusing metaphysical conundrum. Donne observed that 'no man is an island, entire of itself'[9] and the same holds true for autonomous weapons which are never completely autonomous – there will always be dependence on some external element such as other machines or soldiers in the field, intelligence operatives scouting locations, trainers or programmers at base and so on.[10] Furthermore, as Bradshaw *et al* put it, autonomy is not a 'unidimensional concept' (which, at its simplest, could be said to be comprised of self-direction and self-sufficiency) and instead has a broad range of potential meanings.[11] As a result of these considerations, states and academics have grown less enthusiastic about trying to define autonomy and there is therefore no accepted international definition of what constitutes an autonomous weapon. Nonetheless, in 2012, the US Department of Defense adopted a useful working definition providing that an autonomous weapon is a 'weapon system that, once activated, can select and engage targets without further intervention by a human operator.'[12] This definition was widely cited and certainly manages to capture the essence of what is meant by an autonomous weapon for the purposes of this article: namely, a machine that can be assembled with hardware, imbued with software and then released into the battlespace to perform its function independently. The point is that it is the absence of direct human involvement in operation that most clearly separates autonomous weapons from the more familiar technology found in 'drones' which, while 'unmanned', are still piloted by a human, albeit from a distant military base.[13] This distinction has led one of the leading actors in humanitarian law, the International Committee of the Red Cross, to comment that the deployment of autonomous weapons represents a 'paradigm shift' in the way hostilities are conducted.[14]

---

8 Bagrit, L, "The Age of Automation" 17(1) *British Journal for the Philosophy of Science* (1966) 80.

9 Alford, H, ed, *The Works of John Donne*, Volume III (John W Parker 1839) 574-575.

10 United States Department of Defense, *Task Force Report: The Role of Autonomy in DoD Systems*, 19 July 2012, 59, *at* <bit.ly/2pwXT9C> (accessed 20 March 2018) [*Task Force Report*].

11 Bradshaw, JM, Hoffman, RR, Johnson, M and Woods, DD, "The Seven Deadly Myths of 'Autonomous Systems'" 28(3) *IEEE Intelligent Systems* (2013) 54.

12 United States Department of Defense Directive, "Autonomy in Weapons Systems", Number 3000.09 of 21 November 2012, Glossary, Part II ("Autonomy in Weapons Systems").

13 For a full analysis of the problems posed by drones, see Casey-Maslen, S, "Pandora's box? Drone strikes under *jus ad bellum*, *jus in bello*, and international human rights law" 94 *International Review of the Red Cross* (2012) 597.

14 International Committee of the Red Cross, *Autonomous weapon systems - Q&A*, 2014, *at* <bit.ly/2ixib2p> (accessed 20 March 2018).

Of course, weapons with limited autonomy have already been employed widely by states for *defensive* purposes – even a landmine could be said to fulfil the basic requirements. On a more sophisticated level, there are sentry guns and missile interception technologies that repel incoming targets without the need for any additional human authorisation such as 'Phalanx' and 'Super aEgis-II'.[15] However, when it comes to the more *offensive*, advanced and mobile technologies that are the focus of this article, states have been much more cautious. For potential examples, one might consider 'Taranis', an aerial combat vehicle being developed by BAE Systems plc (a UK-based aerospace manufacturer), or 'Atlas', a humanoid-like machine being developed by Boston Dynamics (a US-based private robotics company).[16] In short, although we are yet to see the completion of any 'offensive' autonomous weapons, there seems little doubt from a technical perspective that they will soon be available.

B. The inevitable deployment of autonomous weapons

Of course, the mere availability of a particular technology does not necessarily mean that states must employ it. For example, 'blinding laser weapons' were developed in the late twentieth century but were pre-emptively banned by a protocol to the Conventional Weapons Convention.[17] Turning to autonomous weapons, the official line of a number of states at present is that 'critical decisions' (ie decisions to strike) will not be delegated to a machine and that there will always be a human 'in the loop' (to authorise a strike) or 'on the loop' (with the ability to abort it). Most recently, in its (somewhat overdue) 2017 Joint Doctrine Publication, the Ministry of Defence confirmed that 'current UK policy is that the operation of our weapons will always be under human control as an absolute guarantee of human oversight and authority and of accountability for weapon usage.'[18] The US, for its part, had earlier affirmed that 'autonomous … weapons systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.'[19]

However, scratch beneath the surface and these assertions become less convincing. The UK has opted to set a very high bar when defining what would actually constitute an 'autonomous system' in requiring that it would need to be 'capable of understanding higher-level intent and direction.'[20] Of course, as discussed below, there is no suggestion of machines such as Taranis or Atlas being able to genuinely 'understand' what is going on around them – that level of artificial intelligence remains confined to science fiction. Therefore, the UK has deftly created a lacuna within which to develop weapons that it does not consider to be 'autonomous'. Similarly, the US language of 'appropriate levels of human judgment' is deliberately ambiguous and has been heavily criticised on the basis that, in some cases, the 'appropriate' level of human judgment may

---

[15] Raytheon, *Last Line of Defense for Air, Land and Sea*, *at* <raytheon.com/capabilities/products/phalanx> (accessed 21 April 2018); Dodaam, *Combat Robot (Lethal), at* <dodaam.com/eng/sub2/menu2_1_4.php> (accessed 21 April 2018).

[16] BAE Systems, *Taranis, at* <baesystems.com/en/product/taranis> (accessed 21 April 2018); Boston Dynamics, *Atlas, at* <bostondynamics.com/atlas> (accessed 21 April 2018).

[17] *Protocol IV to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects* (1980) 1342 UNTS 137, Article 1.

[18] United Kingdom Ministry of Defence (Development, Concepts and Doctrine Centre), *Joint Doctrine Publication 0-30.2, Unmanned Aircraft Systems,* August 2017, 42, *at* <bit.ly/2pvFQkn> (accessed 20 March 2018) [*Joint Doctrine Publication*].

[19] "Autonomy in Weapons Systems", *supra* nt 12, 2.

[20] *Joint Doctrine Publication*, *supra* nt 18, 13.

be none at all.[21] Indeed, recently, a US Department of Defense report has recommended that the US must *accelerate* its exploitation of autonomy on the basis that, *inter alia*, autonomous technology will 'increase the quality and speed of decisions in time-critical operations.'[22] It is difficult to see how there can be space for *any* human judgment in such 'time-critical' cases – rather the implication seems to be that the quality of the determination will be higher *without* human meddling. Keeping the door ajar for autonomous weapons is not unique to the West. The Russian Federation made it clear in a recent position paper that, until there are working examples of autonomous weapons, any regulation is premature and that to stifle the development would be to preclude a whole range of associated technologies that are emerging thanks to automation and that are legitimate and desirable.[23]

The reluctance of states to act decisively on autonomous weapons has bled into proceedings at the UN. It had been decided unanimously in December 2016 at the UN's Review Conference of the Convention on Certain Conventional Weapons to establish a Group of Governmental Experts (GGE) to discuss autonomous weapons. Although the GGE was formed, and talks are indeed being held, very little progress has been made. After being postponed from April 2017, the first meeting in November failed to deliver much in the way of tangible progress and, in the words of one commentator, deteriorated into 'a chaotic and ultimately inconsequential discussion of AI generally.'[24] Indeed, the Campaign to Stop Killer Robots, a leading NGO, went so far as to call the whole of 2017 a 'lost year for diplomacy' and has also criticised the decision to hold further meetings across ten days in April and August 2018 as 'unambitious' and observed that it was 'unlikely to result in significant steps forward'.[25] All of this diplomatic hesitation is encapsulated by the fact, at present, only twenty two states have signalled their support for a ban and none of these are global military powers.[26] This low number should also be read against a backdrop of ninety countries around the world operating unmanned aircraft – all of which could be augmented through the incorporation of autonomy.[27] Furthermore, there has not been any progress towards the looser sort of 'standard of operation', advocated by the likes of Kastan, that might act as a voluntary military manual for the use of autonomous weapons.[28]

---

[21] Sauer, F, ''Stopping 'Killer Robots': Why Now Is the Time to Ban Autonomous Weapons Systems'', *Arms Control Association*, October 2016, *at* <http://bit.ly/2Goa2rZ> (accessed 20 March 2018).

[22] United States Department of Defense, *Report of the Defense Science Board Summer Study on Autonomy*, Washington DC, 1 June 2016, 1, *at* <bit.ly/2Goa2rZ> (accessed 20 March 2018) [*Report of the Defense Science Board*].

[23] Russian Federation, *Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the Convention*, CCW/GGE.1/2017/WP.8, 10 November 2017, *at* <bit.ly/2ufPjSx> (accessed 20 March 2018).

[24] Tucker, P, "Russia to the United Nations: Don't Try to Stop Us from Building Killer Robots" *Defense One*, 21 November 2017, *at* <bit.ly/2B7J8yc> (accessed 20 March 2018).

[25] Campaign to Stop Killer Robots, *2017: A Lost Year for Diplomacy*, 22 December 2017, *at* <bit.ly/2ptumi3> (accessed 20 March 2018).

[26] Algeria, Argentina, Bolivia, Brazil, Chile, Costa Rica, Cuba, Ecuador, Egypt, Ghana, Guatemala, Holy See, Iraq, Mexico, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela and Zimbabwe.

[27] Sayler, K, "A World of Proliferated Drones: A Technology Primer" *Center for a New American Security*, 10 June 2015, *at* <bit.ly/2G1kR3Q> (accessed 20 March 2018).

[28] Kastan, B, "Autonomous Weapons Systems: A Coming Legal Singularity?" 1 *Journal of Law, Technology and Policy* (2013) 45, 62.

In conclusion, it seems clear that Anderson and Waxman were correct in their prediction that the deployment of autonomous weapons is inevitable.[29] There are many justifications for this conclusion ranging from the existing financial investment in the technology to the tactical benefits it offers. Ultimately, though, the reality is that the freedom to develop autonomous weapons is already viewed by the major military powers as a strategic imperative. As Vladimir Putin said, 'artificial intelligence is the future, not only for Russia, but for all humankind … Whoever becomes the leader in this sphere will become the ruler of the world.'[30]

C. The limits of autonomous cognition

Assuming what has been said above is correct and that the deployment of autonomous weapons is indeed inevitable, we must deal with the full spectrum of challenges that it presents. For the purposes of this article, the focus is on the principle of proportionality in humanitarian law and so this means grasping how autonomous weapons might arrive at proportionate determinations on the battlefield. This is problematic because, hitherto, the application of proportionality has revolved around the decisions of *human* combatants and so replacing these with determinations made by machines would appear to remove a key pillar upon which the principle is based. However, while it is indeed true that autonomous weapons pose some serious challenges, if one is to understand the proper extent of these challenges it is important to be realistic about what autonomous weapons *will* be able to do and what they *will not* be able to do.

There are, of course, many science fiction books and films which depict autonomous machines, usually on the rampage, with human-levels of understanding of their environment. To some extent, this sort of background material has influenced academic consideration of the topic. For example, Wallach and Allen argue that society is on a quest to build a machine that can tell right from wrong with the effect that existing theories of ethics and agency are not adequate and, therefore, that we must begin constructing new conceptual frameworks to provide autonomous machines with even rudimentary ethical sensitivity.[31] However, although there are research projects aimed at producing artificial intelligence which would be equivalent to human intelligence and capable of full moral agency, the attainment of this goal is estimated to be a long way off. To illustrate this, Müller surveyed hundreds of artificial intelligence experts at a series of conferences and asked, 'by what year would you see a (10%; 50%; 90%) probability for … high level machine intelligence to exist?'. The median response for 10% probability was 2022, the median response for 50% probability was 2040 and median response for 90% probability was 2075.[32] In short, according to artificial intelligence experts, advanced robot cognition is at least twenty years away.

In the meantime, what we might realistically expect to see is a more superficial form of artificial intelligence that merely *appears* to make decisions. On this basis, it has been argued that autonomy is not 'a widget or discrete component' but rather a 'capability of the larger system enabled by the integration of human and machine

---

[29] Anderson, K, and Waxman, MC, "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can" *Stanford University The Hoover Institution (Jean Perkins Task Force on National Security and Law Essay Series)*, 10 April 2013, *at* <bit.ly/2pBFnO9> (accessed 20 March 2018).

[30] Putin, V, Speech to Yasoslavl University, 1 September 2017, *at* <bit.ly/2Go1LUS> (accessed 20 March 2018).

[31] Wallach, W and Allen, C, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press 2008).

[32] Müller, VC and Bostrom, N, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion" *in* Müller, VC, ed, *Fundamental Issues of Artificial Intelligence* (Springer 2016).

abilities.'[33] The significance of this conclusion for present purposes cannot be overstated. It means that, even if machines physically replace humans on the battlefield, human judgment will remain essential to matters such as determinations on proportionality. In short, human judgments and values will be *implemented* by machines. This has led Bradshaw *et al* to argue that the idea of an 'autonomous *system*' is a myth and what we are really dealing with are, at most, machines with limited autonomous *capabilities*.[34] This focuses the debate about autonomous weapons and brings it back from science fiction to reality. Machines would operate on the battlefield simply by carrying out calculations (albeit very complex ones), without having to understand what the overall concept of proportionality is actually about.

D. Autonomous cognition in autonomous vehicles

Given that, as a matter of technological limitation, autonomous cognition will for the next few decades only extend as far as the implementation of human judgments and values, society must begin to crystallise these into clear, objective, standards with which machines will be able to work. By way of comparative analysis, it is enlightening to explore how this issue is being handled in the context of autonomous vehicles.[35] Interest in this technology has been intense and one prediction has driverless vehicles accounting for 40% of car manufacturers' profits by 2035.[36] Of course, such vehicles also face the conundrum of making determinations about harm with, in the example of an inevitable collision, either the occupant or a pedestrian being injured depending on what action the car takes. It should be noted that the recent tragedy in Arizona involving the death of a pedestrian in a crash with an autonomous Uber vehicle does not appear to have been the result of any 'determination' by the vehicle but, rather, a simple failure of its sensors to detect her. Of course, the investigation is only in its nascent stages.[37]

Until very recently, there has been a legislative vacuum in this area and so it has been left to manufacturers to come up with their own views on what should happen in situations where harm must inevitably fall on someone. The first manufacturer to candidly set out its position on this issue was Mercedes-Benz. Speaking at the 2016 Paris Motor Show, the company's manager of driver assistance systems, Christoph von Hugo, said, 'If you know you can save at least one person, at least save that one. Save the one in the car. If all you know for sure is that one death can be prevented, then that's your first priority.'[38] In other words, if there is a risk of death to both the driver and a third party outside the car, a Mercedes will simply prioritise the driver. The bluntness of this 'prioritisation' approach is quite shocking but, on reflection, it should not be all that surprising. Mercedes is in the business of selling cars and pleasing the people who buy them. Those people would, the company presumes, prefer that they and their families are prioritised over other road users no matter what. Indeed, there is empirical evidence to

---

33 *Task Force Report*, *supra* nt 10, 23.

34 "Seven Deadly Myths", *supra* nt 11, 54.

35 For examples of autonomous car development see: Atiyeh, C, "Google's Latest Search is for Automaker Partners" *Car and Driver*, 15 January 2016, *at* <bit.ly/2pFnfD7> (accessed 20 March 2018).

36 Boston Consulting Group, "By 2035, New Mobility Tech Will Drive 40% of Auto Industry Profits", Press Release, 11 January 2018, *at* <on.bcg.com/2I5bInu> (accessed 20 March 2018).

37 National Transportation Safety Board, "NTSB Update: Uber Crash Investigation", News Release, 20 March 2018, *at* <bit.ly/2I3olPQ> (accessed 20 March 2018).

38 Taylor, M, "Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians" *Car and Driver*, 07 October 2016, *at* <bit.ly/2G9MCU0> (accessed 20 March 2018).

suggest that manufacturers would indeed lose customers if they did not take this approach.[39]

Of course, leaving regulation to corporations is not necessarily the best tack from the point of view of society generally or pedestrians in particular. Governments have been slow to set clear guidelines on the matter with even the US, at the forefront of development, merely requiring that algorithms for resolving these situations should be 'developed transparently using input from Federal and State regulators, drivers, passengers and vulnerable road users.'[40] This *laissez-faire* view may change soon as a result of the events in Arizona. However, a more advanced position has already been reached by Germany which formed a Commission of Experts to investigate the challenges of autonomous vehicles. In 2016, the Commission reported back with twenty rules that car manufacturers must consider when developing automated driving systems.[41] The German rules begin with the sensible position that collisions should be avoided.[42] However, where a collision is inevitable, the protection of human life takes the highest priority, including over damage to animals and property.[43] When it comes to assessing potential physical injury to multiple people, general programming aimed at reducing the *number* of personal injuries is permitted however it is made clear that manufacturers are barred from attaching any weight to personal characteristics such as age, gender, physical or mental constitution.[44] It is important to note one very significant limitation to the German rules – they do not deal with the toughest problems. It is concluded that 'life-versus-life' decisions are so abstract that general *ex-ante* rules cannot be imposed upon them.[45] As a consequence, in such cases the intention is to immediately return control to the driver who is then faced with making the decision – although it is stipulated that such abrupt transfers of control should occur as seldom as possible.[46] The same difficulty has been identified beyond Germany, with Bonnefon *et al* noting that defining the relevant algorithms presents a 'formidable challenge.'[47]

The question for present purposes is whether any of the lessons learned from the debate on autonomous vehicles can be carried over to autonomous weapons. The answer is mixed. The prioritisation model offered by manufacturers can be safely discounted for a number of reasons: autonomous weapons have no passengers to prioritise; the approach appears to have been rejected by the first state actor to set out a definitive position and, as will be explained shortly, the model runs contrary to humanitarian law. The German state's approach, as it stands, is also incapable of adequately regulating autonomous weapons. By stopping short of articulating rules that deal with life-versus-life situations and instead requiring return of control to the driver, the German rules exclude precisely the sort of problems that must be resolved for autonomous weapons. Such weapons are, by their very nature, going to be involved in situations where harm, desired or otherwise, is caused to humans. Furthermore, the option of returning control to a human operator will not always be available if, for example, the timing is too tight or

---

39 Bonnefon, JF, Shariff, A and Rahwan, I, "The Social Dilemma of Autonomous Vehicles" 352 *Science* (2016) 1573, 1574.

40 United States, Department of Transportation, *Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety*, September 2016, 26-27, *at* <bit.ly/2dgxqJV> (accessed 20 March 2018).

41 Bundesministerium für Verkehr und digitale Infrastruktur, *Bericht der Ethik-Kommission Automatisiertes und vernetztes Fahren*, 20 June 2016, *at* <bit.ly/2IBhZZ8> (accessed 20 March 2018).

42 *Bericht der Ethik-Kommission Automatisiertes und vernetztes FahrenId*, Rules 2 and 5.

43 *Bericht der Ethik-Kommission Automatisiertes und vernetztes FahrenId*, Rules 2 and 7.

44 *Bericht der Ethik-Kommission Automatisiertes und vernetztes FahrenId*, Rule 9.

45 *Bericht der Ethik-Kommission Automatisiertes und vernetztes FahrenId*, Rule 8.

46 *Bericht der Ethik-Kommission Automatisiertes und vernetztes FahrenId*, Rule 17.

47 "The Social Dilemma of Autonomous Vehicles", *supra* nt 39, 1573.

if communication signals are being jammed. Having said that, the work done by Germany is promising. In permitting vehicle manufacturers to take account of the number of personal injuries, we can see the beginnings of a (qualified) utilitarian solution to the problem of competing harm which is compatible with, and arguably mandated by, humanitarian law. As will be explained, the various metrics proposed first by Bentham to explain utilitarianism could be used to provide the clearer picture of proportionality that is so urgently required. However, before turning to that endeavour, the substantive rule of proportionality in humanitarian law requires some exposition.

II. Proportionality

A. The Basics of proportionality

Humanitarian law (or, variously, the law of war or the law of armed conflict) is typically viewed as having two branches. The first branch, 'Hague law', is the elder of the two and seeks to regulate the means (weapons) and methods (tactics) of warfare. The 1868 St. Petersburg Declaration is an early modern example of a Hague rule that we would recognize today as being grounded in the principle of proportionality.[48] The Declaration prohibited the use of projectiles weighing less than 400g that exploded upon contact with soft surfaces such as human flesh. The twenty state parties noted that in war, 'it is sufficient to disable the greatest possible number of men [and] this object would be exceeded by the employment of arms which uselessly aggravate the sufferings of disabled men or render their death inevitable'. The rule itself was later incorporated into the 1907 Hague Convention.[49] Today, by virtue of Additional Protocol I, Hague law takes a much broader view of proportionality and prohibits any means and methods of warfare that cause 'superfluous injury or unnecessary suffering.'[50] The second branch, 'Geneva law', is more recent and seeks to protect victims of conflict such as wounded combatants,[51] civilians[52] and others. In this context, Additional Protocol I prohibits attacks that would cause collateral damage 'excessive in relation to the concrete and direct military advantage anticipated.'[53] Similar iterations of the proportionality principle can be found in a number of other humanitarian instruments such as the Convention on Certain Conventional Weapons[54] and the San Remo Manual (on conflicts at sea).[55] Consequently, in its heavily influential study, the International Committee of the Red Cross confirmed proportionality to be a customary rule of humanitarian law.[56]

---

48 *Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight* (adopted and entered into force 11 December 1868) 138 CTS 297, *at* <bit.ly/2Gdwz7I> (accessed 20 March 2018).

49 *Convention (IV) Respecting the Laws and Customs of War on Land* (18 October 1907) 205 CTS 277, *Regulations Concerning the Laws and Customs of War on Land*, Article 23(e), *at* <bit.ly/2pFR3zI> (accessed 20 March 2018).

50 *Additional Protocol I*, Article 35(2).

51 *Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field* (12 August 1949) 75 UNTS 31, *at* <bit.ly/2pGIKUg>, (accessed 20 March 2018).

52 *Convention (IV) Relative to the Protection of Civilian Persons in Time of War* (12 August 1949) 78 UNTS 287.

53 *Additional Protocol I*, Articles 51(5)(b) and 57(2)(a)(iii).

54 *Protocol II to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects*, 10 April 1981, 1342 UNTS 137, Article 3(3)(c).

55 Doswald-Beck, L, ed, *San Remo Manual on International Law Applicable to Armed Conflicts at Sea* (Cambridge University Press 1995), Article 46(d).

56 Henckaerts, JM and Doswald-Beck, L, *Customary International Humanitarian Law, Volume I: Rules* (Cambridge University Press 2005), 46-48.

The crystallisation of proportionality into a customary rule is certainly to be welcomed. It acts as a restriction on states, encouraging them to adopt more moderate methods and means and, in certain cases, may save the lives of civilians who would otherwise have been caught up as 'collateral damage'. Furthermore, any violation of the principle will constitute a war crime under the Rome Statute which provides the sanction of individual criminal responsibility where the principle has been violated (albeit with the modification to 'clearly excessive' discussed below).[57] However, it must also be remembered that proportionality can act as an *enabler*, providing states with legitimate sanction for particular attacks. For these reasons, and for both attacker and target, a clear understanding of what the principle means is vital. It should be noted that, for present purposes, proportionality does not only have a bearing on autonomous weapons if they are deployed at some point in the *future*. Rather, the principle *already* applies to this nascent technology as states are obliged, in the study, development, acquisition or adoption of new weapons to determine whether their employment would be prohibited by international law.[58]

B. The current approach to proportionality

Vital though it is to have a clear picture of what proportionality means, understanding that the principle is not straightforward.[59] As a starting point, it can be said that Additional Protocol I attempts to ensure that proportionality is assessed in the context of each individual 'attack'.[60] This is an attempt to confine the assessment of proportionality and to ensure that it relates only to the immediate objective in question: for example, the destruction of an enemy fuel depot. This narrow objective can be contrasted with the overall strategic goal of a campaign: for example, to bring about regime change.[61] If the latter was to be used as the yardstick, then a very high number of casualties would seem to be 'proportionate' even for minor military gains. How states have approached this delineation issue is discussed further below. Once the ambit of the relevant attack has been determined, the task is then to perform the proportionality assessment proper. It will be recalled from above that proportionality prohibits attacks that would cause collateral damage which would be 'excessive'[62] or which would cause 'unnecessary' or 'superfluous' suffering.[63] However, these terms possess inherent ambiguity as it is not immediately clear what might be considered excessive, unnecessary or superfluous from one case to the next.

The presence of ambiguity at the most fundamental level of the proportionality principle begs the question of how it has been capable of any meaningful application on the battlefield. The answer is simple - *human beings*. Human judgment has been the agent to imbue a flat concept with substance and to create a fully-formed and workable rule for any particular situation. The individuals exercising this human judgment could be military officers deciding on whether to launch an attack, or judges deciding if an attack was disproportionate. In any event, humans can be thought of as the yeast to proportionality's bread and relatively recent jurisprudence confirms this. The

---

[57] *Rome Statute of the International Criminal Court* (1998) 2187 UNTS 90, Article 8(2)(b)(iv), *at* <bit.ly/2pHbk83> (accessed 20 March 2018) [*Rome Statute*].

[58] *Additional Protocol I*, Article 36.

[59] See generally Newton and May, *Proportionality in International Law* (Oxford University Press 2014), chapters 5-12.

[60] *Additional Protocol I*, Articles 51(5) and 57(2).

[61] For a rare, overt, expression of desire for regime change, see: United States Congress, *Iraq Liberation Act of 1998,* PL 105-338, 112 Statute 3178, 31 October 1998, *at* <bit.ly/2pIztuS> (accessed 20 March 2018).

[62] *Additional Protocol I*, Articles 51(5)(b) and 57(2)(a)(iii).

[63] *Additional Protocol I*, Article 35(2).

International Criminal Tribunal for the former Yugoslavia found in *Galic* that, in assessing proportionality, 'it is necessary to examine whether a reasonably well-informed *person* in the circumstances of the actual perpetrator, making reasonable use of the information available to *him* or *her*, could have expected excessive civilian casualties to result from the attack [emphasis added]'.[64] In fact, beyond providing a mere mechanism for resolving the ambiguities inherent in proportionality, it has been argued that human judgments have, over time, helped to *clarify* the principle. As Newton and May put it, repeated application of proportionality by countless individuals over a long period of time has consolidated proportionality and there is, as a consequence, 'a core of *jus in bello* proportionality that has remained fixed for generations.'[65] This, in turn, means that proportionality is no longer a vague notion but, rather, a clear 'fixed standard' setting limits on what commanders and soldiers can do and removing unwanted discretion that might be exploited or abused.[66]

If correct, this assertion is positive in the sense that it means humans operating in this field have a commonly understood meaning of which actions are 'proportionate' and which are 'disproportionate'. There is, of course, a significant problem with the current approach. This tacit understanding shared by people is of little utility in the context of autonomous weapons which sees part of the implementation of proportionality delegated to machines. The question therefore remains as to how proportionality will work in this new context.

C. The need for a new approach to proportionality

It has already been explained above that, for some decades at least, there is no question of truly thinking machines operating in any context.[67] Consequently, what we will certainly not see are instances of machines making 'judgments' about proportionality. Instead, modern artificial intelligence is limited to taking clearly defined rules (whereby any necessary judgments have already been made by humans) and then applying those rules to factual scenarios.[68] The distinction between *making* judgments and *applying* judgments may seem to be a fine one, but it is important. However, in the context of autonomous vehicles, the reality of trying to define algorithms that capture all of those judgments (and might thereby allow machines to replicate human behaviour) has proven to be a 'formidable challenge.'[69] This is what lay behind Germany's retreat when drawing up guidelines on the matter and prompted the requirement that, in life-versus-life cases, the controls have to be passed back to a human.[70] Again though, this option is not practicable in the case of autonomous weapons and so an alternative is needed.

One reaction might be to return to the position proposed by vehicle manufacturers to the effect that some sort of blanket prioritisation model should be adopted –this time favouring one of the actors on the battlefield rather than favouring the occupant of a vehicle.[71] However, it was indicated above that this would not be appropriate and, now that an exposition of proportionality has been supplied, the reason for this should be self-

---

64 International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Stanislav Galic*, Judgment (Trial Chamber), Case No. IT-989-29-T, 05 December 2003, paragraph 58, *at* <bit.ly/1kYGPkL> (accessed 20 March 2018).

65 *Proportionality in International Law*, *supra* nt 5, 4.

66 *Id*, 3.

67 "Future Progress in Artificial Intelligence", *supra* nt 32.

68 *Task Force Report*, *supra* nt 10.

69 "The Social Dilemma of Autonomous Vehicles", *supra* nt 39, 1573.

70 *Bericht der Ethik-Kommission Automatisiertes und vernetztes Fahren,* Rule 8.

71 "Self-Driving Mercedes", *supra* nt 38.

evident. Proportionality in humanitarian law involves consideration of, on one hand, any military advantage that an attack would deliver and, on the other, any suffering the attack would cause. There must then be a check to ensure that the harm does not exceed the gain. A blanket presumption in favour of one side – either the target *or* the attacker – would not be compatible with this approach. Furthermore, there has been a suggestion by academics such as Fagnant and Kockelman that the introduction of autonomous vehicles will one day see the reduction of harm on the roads.[72] Bennefon *et al* have even suggested that switching to 'self-driving' vehicles will eliminate 90% of accidents.[73] In essence, the suggestion is that harm on the roads will one day be a moot point. This belief, whether right or wrong, is simply not applicable in relation to autonomous weapons. Attacks are by their very nature dangerous and intended cause harm – eliminating harm to enemy combatants or civilians is simply not an option. Suggestions of incorporating black boxes into autonomous technology to record the circumstances in which harm occurs do not change the fact that harm will always occur.[74]

The conclusion that can be drawn from all of this is that there is a need for a new, bespoke, approach to proportionality in relation to autonomous weapons. Experiences with autonomous vehicles have generally been either inapposite or under-developed. That said, as was noted above, Germany's position permits manufacturers to programme their vehicles to assess the number of casualties that may occur in any potential crash.[75] This opens the door to a quasi-utilitarian solution to the problem that seems to offer a plausible way forward. This is because proportionality is based upon the twin pillars of 'gain' and 'harm' that are analogous to the pillars of 'pleasure' and 'pain' upon which utilitarianism was founded by Bentham. It is to his work that we now turn.

III. Utilitarianism

A. The theory of utilitarianism

The classic utilitarian model of decision making is generally recognised as having first been expressed by Bentham as the basis for a new penal code – although similar notions can be traced as far back as Plato.[76] Utilitarianism is satisfied by any action 'when the tendency it has to augment the happiness of the community is greater than any it has to diminish it.'[77] From the point of view of the individual, Bentham summarised utility as being 'that principle which approves or disapproves of every action … according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question.'[78] The parallel between utilitarianism and proportionality is eminently apparent. Just as utilitarianism is satisfied when a given action will furnish an individual with more happiness than unhappiness, so too is proportionality satisfied when an attack will furnish the attacker with more gain than the attendant harm it will cause. The fact that these two principles share the same foundations offers an alluring prospect - that the painstaking work poured into utilitarianism by Bentham at the end of the eighteenth century in a bid to reform English criminal law might be used at the dawn

---

[72] Fagnant, DJ and Kockelman, K, "Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations" 77 *Transportation Research Part A: Policy and Practice* (2015) 167.

[73] "The Social Dilemma of Autonomous Vehicles", *supra* nt 39, 1573.

[74] Endsley, MR, "Building Resilient Systems: Incorporating Strong Human-system Integration" 45(1) *Defense Acquisition, Technology and Logistics Magazine* January/February (2016) 6, *at* <bit.ly/ACd2EK> (accessed 20 March 2018).

[75] *Bericht der Ethik-Kommission Automatisiertes und vernetztes Fahren*, Rule 9.

[76] Barrow, R, *Plato, Utilitarianism and Education* (Routledge 2009).

[77] *Principles of Morals and Legislation*, *supra* nt 4, 3.

[78] *Id*, 2.

of the twenty-first century to resolve a vexed philosophical and technological question. There would be some poetry in that.

Before proceeding, it is important to acknowledge that utilitarianism is not without its criticisms. Ayer believed that he had identified a fundamental problem with the principle – that it is not always contradictory to say that some pleasant things are bad, or that some painful things are good.[79] Consequently, he argued, one cannot equate the fact that an action brings pleasure with it being 'right' or 'desirable' or *vice versa* – for example, a doctor might advise a patient of a terminal illness causing the latter emotional pain; but few would regard it as the wrong thing to do.[80] Ayer thus concluded that 'the validity of ethical judgements is not determined by the felicific tendencies of actions, any more than by the nature of people's feelings; but that it must be regarded as 'absolute' or 'intrinsic', and not empirically calculable.'[81] Kant, for his part, favoured moral philosophy grounded in deontology, or the logic of duty, and so created a series of principles which would stand in contradistinction to utilitarianism such as 'act only according to that maxim whereby you can at the same time will that it should become a universal law.'[82] In addition to logical or philosophical criticisms, utilitarianism comes with some heavy baggage owing to its occasional adoption by politicians, medical professionals and others to justify ruthless actions – the eugenics programme of the National Socialist Party being the most heinous example.[83] As a consequence, utilitarianism has seen a backlash that is generally based around the notion that it can be used to compromise the interests of the few for the benefit of the many.[84]

Many of the criticisms of utilitarianism are well-founded, some less so. For example, some policies that purport to follow a utilitarian model are in fact based on a perversion of the principle twisted by those in power to achieve nefarious ends.[85] In this sense, it is unfair to brand utilitarianism itself as inherently immoral. Indeed, Bonnefon *et al* found in the context of autonomous vehicles (through a detailed survey) that members of the public 'overwhelmingly expressed a moral preference for utilitarian AVs programmed to minimize the number of casualties'.[86] Of course, that result should be read against the finding that people would not actually want to *buy* utilitarian vehicles and prefer their vehicle to prioritise their own life in any scenario – although that probably says more about the survival instinct of humans than the morality of utilitarianism.[87] Nonetheless, the overall picture seems to be that utilitarianism may be better received than one might have thought and that people are generally cognisant of the difficult questions new technology can bring. In truth though, the point is not to determine which moral philosophy has the stronger case or to defend utilitarianism – that task has been undertaken in myriad contexts.[88] Instead, the key point is that utilitarianism here needs no defence. Proportionality in humanitarian law, rightly or wrongly, simply *is* a utilitarian principle – this is a *fait accompli*. The most fruitful endeavour is therefore to

---

79 Rogers, B, ed, *Ayer: Language, Truth and Logic* (Penguin 2001).

80 *Id*, 107.

81 *Ibid*.

82 Gregor, M, ed, *Kant: The Metaphysics of Morals* (Cambridge University Press 1996).

83 LaChat, MR, "Utilitarian Reasoning in Nazi Medical Policy: Some Preliminary Investigations" 42(1) *The Linacre Quarterly* (1975) 14.

84 Brooks, S, "Dignity and Cost-Effectiveness: A Rejection of the Utilitarian Approach to Death" 10 *Journal of Medical Ethics* (1984) 148.

85 Alexander, L, "Medical Science under Dictatorship" 241(2) *New England Journal of Medicine* (1949), 39.

86 "The Social Dilemma of Autonomous Vehicles", *supra* nt 39, 1574.

87 *Ibid*.

88 Bagaric, M, "In Defence of a Utilitarian Theory of Punishment: Punishing the Innocent and the Compatibility of Utilitarianism and Rights" 24 *Australian Journal of Legal Philosophy* (1999) 95.

try to understand how it might operate in the context of autonomous weapons. Germany side-stepped this vexing issue in relation to vehicles, essentially handing the choice back to the driver in difficult cases, and so it remains a lacuna that needs to be addressed. Fortunately, Bentham explored this issue, albeit in a very general way, and his work offers a starting point for ensuring that autonomous weapons comply with proportionality in humanitarian law.

B. The hedonic calculus – gain and harm

For Bentham, the application of utilitarianism was achieved through the 'hedonic' or 'felicific' calculus.[89] This is essentially a simple process of weighing the amount of pleasure and pain a given course of action will cause. To complete the calculation, one is required to total pleasure on one hand and pain on the other – ensuring that all those individuals affected, and all of the resultant effects, are captured. If the balance is on the side of pleasure then the act is 'good' (ie right). Conversely, if the balance is on the side of pain then the act is 'evil' (ie not right). This can be transposed across to proportionality: if the balance is on the side of military gain then the action is proportionate (and permitted); if the balance is on the side of harm then the action is disproportionate (and not permitted). Of course, we must first know what pleasures and pains we wish to weigh against each other. Bentham discussed these matters and made effort to enumerate the various pleasures and pains that one might experience, with some of these 'perceptions' capable of manifesting as either. The pleasures are those of sense, wealth, skill, amity, good name, power, piety, benevolence, malevolence, memory, imagination, expectation, association and relief.[90] The pains are those of privation, senses, awkwardness, enmity, ill name, piety, benevolence, malevolence, memory, imagination, expectation and association.[91] Some thought must be given to how these 'pleasures' and 'pains' might manifest in a conflict involving 'gains' and 'harms'.

Turning first to 'gain' or, more properly, 'military advantage'. The starting point for understanding what is included here is Additional Protocol I which makes it clear that the parties to a conflict must distinguish between civilians and combatants and between civilian objects and military objectives and only direct operations only against military objectives.[92] In other words, there is no military advantage to be had from killing or injuring civilians or damaging their property, so nothing of that nature would qualify as 'gain' for the purposes of the hedonic calculus. Civilians are defined negatively as any individuals who are not members of the armed forces, militias, volunteer corps or organised resistance movements and who do not spontaneously take up arms to resist invading forces.[93] Persons falling within those categories are generally combatants and so are legitimate targets.[94] On the other hand, military objectives are articles which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.[95] Special protection is afforded to cultural objects, places of worship, objects indispensable to the survival of the civilian population,

---

89 *Principles of Morals and Legislation*, *supra* nt 4, 34.
90 *Id*, 34-37.
91 *Id*, 37-41.
92 *Additional Protocol I*, Article 48.
93 *Id,* Article 50(1).
94 Zimmermann, B *et al*, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (International Committee of the Red Cross 1987), 620.
95 *Additional Protocol I*, Article 52(2).

the natural environment and sites containing dangerous forces.[96]

Of course, it is necessary to move beyond these broad notions of military advantage and refine them into narrower rules if the goal is to lay down clear standards with which autonomous weapons might comply. However, this is difficult because, as Dinstein put it, 'the spectrum of military advantage is necessarily wide' and there is ambiguity over whether certain results or effects should be considered as rendering legitimate military advantage. [97] For example, there is some doubt over whether political or economic advantage garnered from an attack should be considered. At first blush, the answer would appear to be 'no' as such benefits are not strictly 'military' in nature.[98] However, it has been argued by Fleck that the purpose of any military action 'must always be to influence the political will of the adversary.'[99] Indeed, this argument was cited by the Eritrea-Ethiopia Claims Commission when it found that an Ethiopian attack on a power station was proportionate partly because it was intended to exert political pressure on Eritrea to agree to a cease fire.[100] Even if the Commission was wrong, there remains ambiguity over which particular articles are 'military'. Some articles are military by their very nature such as tanks, command centres and munitions centres.[101] However, ostensibly neutral articles may be capable of military use. For example, a vacant site which could be used as barracks for enemy combatants still qualifies on the basis of its potential military use in the future.[102]

Turning next to 'suffering' and 'harm', again the starting point for understanding what is included is Additional Protocol I which provides that death, injury and property damage qualify.[103] Again though, there are ambiguities as there is no exhaustive definition for what sorts of 'injury' fall to be measured under humanitarian law. For example, conflict not only causes injury by kinetic weapons but can also expose victims to toxic air produced by burning buildings and to psychological harm as a result of witnessing the aftermath of attacks. On the latter point, Additional Protocol I includes reference to 'health' as being something distinct from injury so this suggests that there is scope for consideration of impact on mental health.[104] Furthermore, the Protocol prohibits 'all acts or threats of violence the primary purpose of which is to spread terror' and this might be seen as alluding to psychological harm.[105] However, it is important to bear in mind that this instrument was drafted at a time when the level of attention paid to mental health generally was much lower than it is today and so there are limits on just how much one can read into these provisions without simply indulging in speculation. Indeed, there is good reason for states being reluctant to include psychological harm in the hedonic calculus as 'from a military medical point of view the most obvious defect of the concept of "suffering" is that it cannot be … related to wounding' and so is inherently

---

96 *Additional Protocol I*, Articles 53-56.

97 Dinstein, Y, *The Conduct of Hostilities under the Law of International Armed Conflict* (3rd ed, Cambridge University Press 2016), 106.

98 *Id*, 107.

99 Fleck, D, ed, *The Handbook of Humanitarian Law in Armed Conflict* (Oxford University Press 1995), 157.

100 Eritrea-Ethiopia Claims Commission, *Partial Award: Western Front, Aerial Bombardment and Related Claims*, 19 December 2005, Volume XXVI, 291-349, 335, *at* <bit.ly/2uDOLGc> (accessed 20 March 2018).

101 *The Conduct of Hostilities*, *supra* nt 97, 104.

102 *Id*, 107.

103 *Additional Protocol I*, Article 51(5)(b).

104 *Id*, Article 85(3).

105 *Id,* Article 51(2).

harder to forecast.[106]

The consequence of the above is that, although there is a relatively definite core when it comes to deciding what might constitute gain and harm in humanitarian law, lingering ambiguities remain. It will be a matter for each individual state to resolve these difficulties when defining algorithms for any autonomous weapons – making countless decisions on what qualifies for inclusion in the hedonic calculus and what does not. However, utilitarianism offers a clear starting point – that *all* gain and *all* harm should be considered otherwise the result of the hedonic calculus and in turn, the application of proportionality, may be flawed. In short, the default position must be inclusion. Admittedly though, even if states do adopt this position, further difficulties remain to be overcome.

C. The hedonic calculus – context

While the task of identifying each of the various gains and harms to be included in the hedonic calculus presents a very complex task for states in its own right, it is not the end of the process. Before a utilitarian answer can be supplied, those gains and harms must be weighed against each other. Weighting depends on two key points: context and quantification.[107] Context will be considered first and, in essence, it must be recognised that proportionality assessments cannot be made in a vacuum. As Bradshaw *et al* put it, 'autonomy is relative to the context of activity. Functions cannot be automated effectively in isolation from an understanding of the task, the goals, and the context.'[108]

Perhaps the most significant contextual problem with assessing proportionality arises when attempting to ascertain the context within which the anticipated level of military gain is to be measured. Again, proportionality is violated by '*an attack* which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof [emphasis added].'[109] This indicates that, for example, the gain from the destruction of an 'enemy' tank should be the 'friendly' lives and property which might otherwise have been destroyed by that specific tank. However, a number of states have made declarations in respect of Additional Protocol I, or inserted text into their military manuals, to the effect that 'anticipated military advantage' is to be interpreted more broadly than this. For example, the UK has stated that 'the military advantage anticipated from an attack is intended to refer to the advantage anticipated from the attack considered *as a whole* and not only from isolated or particular parts of the attack [emphasis added]'.[110] Similarly, Canada has indicated that an advantage exists if the attack 'will make a relevant contribution to the success of the *overall operation* [emphasis added]'.[111] Many other states have promulgated this view including Australia, Belgium, France, Germany, Italy, the Netherlands, New Zealand, Nigeria, Spain and the US.[112] Under this approach, it is not the gain from the destruction of the single tank that

---

[106] Scott, R, "Unnecessary Suffering? A Medical View", *in* Meyer, MA, ed, *Armed Conflict and the New Law* (British Institute of International and Comparative Law 1989), 277.

[107] Barber, RJ, "The Proportionality Equation: Balancing Military Objectives with Civilian Lives in the Armed Conflict in Afghanistan" 15(3) *Journal of Conflict and Security Law* (2010), 467.

[108] "Seven Deadly Myths", *supra* nt 11, 56.

[109] *Additional Protocol I*, Article 51(5)(b).

[110] United Kingdom, *Reservations and Declarations made upon Ratification of the 1977 Additional Protocol I*, 28 January 1998, paragraph 1, *at* <bit.ly/2rj4Z1M> (accessed 20 March 2018). United Kingdom, *Declaration of the United Kingdom of Great Britain and Northern Ireland in respect of Additional Protocol I*, 2 July 2002, *at* <bit.ly/2E8eOVx> (accessed 20 March 2018).

[111] Canada, Office of the Judge Advocate General, *The Law of Armed Conflict at the Operational and Tactical Levels*, 13 August 2001, 415.1, *at* <bit.ly/2H1VrR6> (accessed 20 March 2018).

[112] *Customary International Humanitarian Law*, *supra* nt 56, 46-49.

is considered but, rather, the gain from the whole operation. The operation might have involved the destruction of an entire military base and so that is the gain to be considered. It should be noted that general, as opposed to multilateral, support for this position can be found in the Rome Statute which refers to the 'direct overall military advantage anticipated'.[113]

It seems clear that a truly utilitarian approach to proportionality favours the broader reading of military gain – anything less would warp the hedonic calculus by underestimating the potential advantage. Of course, this is a controversial position as it is likely to justify greater levels of harm and, in turn, might be viewed as contrary to the object and purpose of humanitarian law to reduce suffering when interpreted under the Vienna Convention.[114] Furthermore, in relation to the parallel with the Rome Statute, it must be conceded that this is technically only relevant to the determination of individual criminal responsibility and one might naturally expect a higher threshold than for 'normal' international wrongs which occasion only state responsibility. Indeed, during the drafting stages of the Rome Statute, the International Committee of the Red Cross made clear that the inclusion of the word 'overall' for the purposes of international criminal law must not be interpreted as modifying the standard under humanitarian law.[115]

However, it is submitted that adopting the broader reading of military gain will not, in fact, lower the level of protection offered by humanitarian law *provided* that the overall *harm* caused by an attack is also considered. This would involve taking a much broader view of the damaging effects an operation might have including, for example, the long-term physical and mental health implications for civilians who might become caught up in the conflict. This approach would help to rebalance the equation and ensure that the result produced by the hedonic calculus remains accurate. The problem at present is that states are incredibly reluctant to take a broader view of harm; but their position is simply perverse. As Barber put it, there is a 'logical inconsistency' in taking a broad view of military advantage but a narrow view of suffering.[116] Similarly, as McCormack and Mtaharu stated, 'to the extent that mid to longer term civilian damage resulting from an attack is expected, such damage should be taken into account in the application of the proportionality equation just as the campaign-wide military advantage is.'[117] In short, taking a fully utilitarian approach to the matter of context should favour neither the attacker nor those in the firing line, it should merely paint a clearer picture of the consequences of an attack and, in turn, permit better determinations to be made whether by humans or autonomous weapons.

D. The hedonic calculus – quantification

Once the relevant gains and harms expected to result from a strike have been identified and once they have been adequately contextualised, the final task for an autonomous weapon would be to quantify these factors and make the final proportionality assessment. However, like apples and oranges, gain and harm would appear to be incommensurable.

---

[113] *Rome Statute*, Article 8(2)(b)(iv).

[114] *Vienna Convention on the Law of Treaties*, 23 May 1969, 1155 UNTS 331, Articles 31-32.

[115] International Committee of the Red Cross, "Paper Submitted to the Working Group on Elements of Crimes of the Preparatory Commission for the International Criminal Court", 13 February 1997, paragraph 190.

[116] "The Proportionality Equation", *supra* nt 107.

[117] McCormack, T and Mtharu, P, "Expected Civilian Damage and the Proportionality Equation" *Third Review Conference of the States Parties to the Convention on Conventional Weapons*, CCW/CONF.III/WP.9, 9, *at* <bit.ly/2pRAaSG> (accessed 20 March 2018).

Therefore, as Walzer put it, 'proportionality turns out to be a hard criterion to apply, for there is no ready way to establish an independent or stable view of the values against which the destruction of war is to be measured.'[118] A number of authors have taken this apparent incommensurability to mean that proportionality assessments will remain vague undertakings. For example, Schmitt stated that, while there will be situations at each pole of the proportionality spectrum on which there will be broad consensus, 'the complexity emerges when one moves … along the proportionality continuum toward the centre'.[119] This would be a most undesirable position as it is precisely the sort of stumbling block that might undermine the ability of autonomous weapons to comply with humanitarian law. However, there is a solution as, according to Newton and May:

> The key to making proportionality manageable is to have weighing that can be done between things that are similar, not dissimilar whenever feasible. It is much easier if the value of the military objective can be couched in terms of lives to be protected or saved so that the costs of such an operation, also often drawn in stark terms of the risk of loss of non-combatant lives, can be assessed more straightforwardly.[120]

In other words, the solution to the incommensurability problem is, first, to measure harm as usual with reference to lives *lost* and injuries *inflicted* and, second, to express gain in terms of lives *saved* and injuries *prevented*. This is an incredibly useful mechanism, but it only takes one so far. It leaves open the finer detail of how to perform the proportionality balance. Certainly, the gains and harms are now much more amenable to comparison, but how, for example, is the loss of one life to be compared to the prevention of a dozen serious injuries? Again, we can turn to Bentham for guidance as he recognised that metrics were crucial to utilitarianism and created lists of 'circumstances' (now more commonly referred to as 'dimensions of value') to be used when measuring the pleasure and pain resulting from an action. The dimensions of value are as follows: 1. intensity; 2. duration; 3. certainty or uncertainty; 4. propinquity or remoteness; 5. fecundity (the chance of a sensation generating a later sensation of the same kind); 6. purity (the chance of a sensation not generating a later sensation of the opposite kind); and 7. extent (the number of people affected).[121]

Bentham's dimensions offer a very useful starting point to those tasked with developing autonomous weapons that are capable of complying with proportionality in humanitarian law. Each pleasure or pain is expressed by magnitude in terms of 'hedons' and 'dolors' (or 'positives' and 'negatives'). These are, in a sense, the raw figures that form the basis of the hedonic calculus. Admittedly, quantification in this way may seem crude, but there are two important points to bear in mind. Firstly, as explained above, it is not the autonomous weapons themselves that will determine the values attached to, for example, intensity or duration of harm. Machines are incapable of that level of understanding and will remain so for decades; instead, human beings will be tasked with setting these values.[122] Secondly, we should recall that precise values are already placed on highly sensitive matters such as human life. In the context of statistics, there is the 'cost of life' concept which is used to represent the cost of preventing death in different

---

[118] Walzer, M, *Just and Unjust Wars* (Basic 1977).
[119] Schmitt, M, "The Principle of Discrimination in 21st Century Warfare" 2 *Yale Human Right Development Law Journal* (1990) 143, 170.
[120] *Proportionality in International Law*, *supra* nt 5, 285.
[121] *Principles of Morals and Legislation*, *supra* nt 4, 30.
[122] "Future Progress in Artificial Intelligence", *supra* nt 32.

circumstances. For example, in 2016 the US Department of Transportation put the 'value of a statistical life' at USD 9.6 Million.[123] The US Environmental Protection Agency also uses the 'Value of a Statistical Life', although it is at pains to explain why this is not the same as placing a value on individual lives and, indeed, is seeking to move away from the concept.[124]

The point to be taken from this is that operating the hedonic calculus in the context of autonomous weapons *is* possible. More than that, it presents significant opportunities as quantification removes the subjectivity associated with an individual's judgment and replaces it with something more objective. Currently, proportionality assessments are made by human beings whose assessments might be coloured by the fact that they are operating in the fog of war and at personal risk. However, quantification in the context of autonomous weapons would be completed *before* the machine is deployed, with the input of policy makers, lawyers, ethicists, military officers and others and can therefore be achieved in a more considered manner. If the results of this process can be brought into the light in the same way that rules of engagement are publicised, it might even be possible to encourage the voluntary placement by states of heavier weight on the side of harm and therefore reduce collateral damage. Going a step further, in working to protect their citizens and improve their humanitarian credentials, states could reach agreement on quantification either multilaterally in peacetime or bilaterally at the outbreak of war. Moreover, this sort of international agreement may be essential in securing *domestic* support for autonomous weapons with, for example, the US identifying 'trust' of autonomous technology as a central priority for development in this area.[125] Military operations are increasingly under the media spotlight and it is easier than ever before for populations to find out what actions their respective states are undertaking in their names. Disproportionate attacks by autonomous systems would inevitably be uncovered and criticised by the press; lower the perceived trustworthiness of such machines in the estimation of the public (as well as military personnel using them) and could potentially compel states to remove them wholly or partially from the field. [126] In this sense, a transparent, utilitarian approach to autonomous weapons might simultaneously benefit both states and those caught up in armed conflict.

Conclusion

Autonomous weapons are those which can be built, programmed and then deployed to the battlefield to serve their function without further human involvement. The hitherto lack of political will to impose a pre-emptive ban, coupled with the recent deterioration in relations between Russia and the West, seems to confirm the inevitability of their adoption. This raises *inter alia* the difficult problem of how to ensure that their actions will comply with the principle of proportionality in humanitarian law. Proportionality requires that attacks do not result in superfluous or unnecessary suffering and, so far, the principle has only been applied by human beings who are able to give the abstract concept practical meaning in any given scenario. While autonomous weapons are certainly impressive pieces of technology, no software presently exists that might endow them with human levels of intelligence. Instead, these machines are limited to making

---

[123] US Department of Transportation, *Revised Departmental Guidance 2016: Treatment of the Value of Preventing Fatalities and Injuries in Preparing Economic Analyses*, 8 August 2016, *at* <bit.ly/2HiGOaI> (accessed 20 March 2018).

[124] US Environmental Protection Agency, *Mortality Risk Valuation*, *Online Questions and Answers*, *at* <bit.ly/2GrXaOk> (accessed 20 March 2018).

[125] United States Department of Defense, *Report of the Defense Science Board*, *supra* nt 22, 14-21.

[126] "Seven Deadly Myths", *supra* nt 11, 56.

fairly rudimentary calculations, albeit with vast amounts of information and at blistering speed. The question is therefore whether machines with this limited cognitive capability might be capable of applying proportionality.

The answer is a tentative 'yes'. The starting point is that proportionality and utilitarianism are expressions of the same concept: they seek to promote pleasure over pain and gain over harm respectively. As a result, proportionality can learn much from utilitarianism. In particular, utilitarianism shows us that assessments can be made using the 'hedonic calculus' but that, in order for the result to be accurate, there must be a willingness to include all of the relevant gains and harms rather than cherry picking some and ignoring others. Thus, while the current approach of certain states in assessing the 'overall' military advantage is appropriate, it must be matched with an assessment of the overall harm caused too. Utilitarianism also requires that those developing autonomous weapons must be able to assign values to each element on a battlefield if the calculus is to be completed. This can be achieved in terms of lives saved and lives lost, and there are past examples of states having ascribed specific values to human life, so the task is possible.

Aspects of utilitarianism might seem cold, however that is simply the law as it currently stands – proportionality is a utilitarian concept. Having said that, in this context it does in fact present an opportunity to raise the protection afforded to those embroiled in conflict. Rather than having proportionality assessments conducted by individual humans whose judgment will inevitably be clouded by the fog of war, the parameters are set in advance of the conflict by teams of individuals working together calmly as a collective. Hopefully, the necessary debates and conversations would result in a trend toward greater emphasis on humanitarianism. Furthermore, autonomous weapons would merely implement those parameters without the same desire for self-preservation or survival that can skew the application of proportionality when undertaken by humans. This more precise, mathematical, approach to proportionality could be exploited to its fullest if states work together to place greater weight on harm and thereby raise the bar that must be met before an autonomous weapon would engage in an attack. These sorts of agreements might represent a more achievable goal than the seemingly doomed discussions of a ban.

Funnily enough, Bentham himself did not think it would ever be possible to apply utilitarianism with mathematical precision to every judgment. Of course, he was writing well before an 'information age' in which calculations can be performed at high speed by machines fitted-out with microprocessors. Technology has finally caught up with Bentham; and society should make the most of it.

*