VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

IN DEFENSE OF (VIRTUOUS) AUTONOMOUS WEAPONS

Don Howard

IN DEFENSE OF (VIRTUOUS) AUTONOMOUS WEAPONS

*Don Howard**

INTRODUCTION

In 2012, Human Rights Watch (HRW) issued a call for a global ban on autonomous weapons.[1] A new NGO, the Campaign to Stop Killer Robots (CSKR) was formed in October 2012 to promote such a ban. In 2015, the Future of Life Institute (FLI) issued a new call for a ban, though now restricted to offensive autonomous weapons.[2] The FLI proposal garnered the support of tens of thousands of signatories, including such prominent figures as Elon Musk and Stephen Hawking, and generated considerable attention in the international press and on social media. Meanwhile, the CSKR helped to organize "informal meetings of experts" starting in 2014 in Geneva under the auspices of the UN's 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 (CCW) for the purpose of exploring the possibility of adding an autonomous weapons ban to existing bans on land mines and blinding lasers, among other banned or restricted weapons.[3] In 2017 these sessions were elevated to the level of annual and still ongoing meetings of a formally constituted Group of Governmental Experts (GGE).[4] Against the background of these developments on the international legal front, an extensive literature on the ethics and policy of autonomous weapons has emerged and media attention to the debate has intensified. At least in the public arena, momentum seems to be building for some kind of ban.

Is a ban the right way to go? I think not. There are obvious questions of law, policy, and ethics that must be weighed regarding autonomous weapons. But, in my opinion, imposing a total ban, even if

*Professor, Department of Philosophy, University of Notre Dame.

[1] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

[2] *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, FUTURE OF LIFE INST. (July 28, 2015), http://futureoflife.org/open-letter-autonomous-weapons/.

[3] Campaign (2015). "Step up the CCW Mandate." Campaign to Stop Killer Robots. http://www.stopkillerrobots.org/2015/06/mandateccw/.

[4] *Convention on Certain Conventional Weapons – Group of Governmental Experts on Lethal Autonomous Weapons*, UNITED NATIONS, https://meetings.unoda.org/meeting/ccw-gge-2017.

only a ban on offensive autonomous weapons, risks our depriving ourselves of tools that can continue the progress already made with the advent of "smart" weapons in reducing the suffering that will always be part of war, especially by way of still further reductions in harm to non-combatants. Moreover, as I will argue, we can construct effective means for norming the use of autonomous weapons short of a total ban by building upon the foundation of existing requirements stipulated in Article 36 of Protocol I to the Geneva Conventions that all new weapons technologies be reviewed for compliance with the International Law of Armed Conflict (ILOAC) and International Humanitarian Law (IHL).

I begin with a critical review of several of the most commonly encountered arguments in favor of a ban. That is followed by a discussion of the moral opportunities afforded by enhanced autonomy. I conclude with a concrete policy proposal based upon the principle of Article 36 review.

I. Arguments for a Ban on Autonomous Weapons

Many arguments have been adduced for some kind of ban on autonomous weapons. They are too numerous and diverse all to be reviewed here. I have, therefore, chosen to focus on six of the most compelling arguments, as judged by their prominence in the literature and their seeming effectiveness in moving public opinion.

A. Morality, Emotions, and Robots

The original HRW call for an autonomous weapons ban placed surprisingly heavy emphasis on an argument that invites skepticism if not outright scorn. The argument is this: Morality requires an emotional capacity. Robots cannot feel emotions. Therefore, robot weapons are inherently immoral.[5]

One understands the idea behind this argument. In many situations, the ability to feel emotions makes possible an empathic relation to those affected by our actions, which includes an appreciation of their needs and fears. One feels oneself into the place of the other. And my Roomba cannot do that. Moreover, it is an empirical fact of

[5] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots. In fairness, this is my distillation of an extended argument that includes acknowledgment of, if not an adequate response to, some of the critical points that I make. But it is an accurate representation of the main thrust of the report's argument.

considerable importance that emotional responses can be powerful enablers of moral action and powerful brakes on immoral action. One might well argue that merely knowing the good does not suffice for doing the good, that knowledge is ineffective without the will to act. The HRW report touches upon all of these points. But there are still at least three serious problems with this argument.

First, there is a long tradition in moral philosophy, from Plato to Kant and beyond, that holds that emotion is an impediment, not an aid to morality, because emotion clouds reason. That that can be so is obvious from long experience. Sometimes emotions of misplaced sympathy lead one to act more kindly toward some than reason would dictate, as when one male faculty member declines to report a case of possible sexual abuse or gender discrimination by another male colleague out of sympathy for that friend, whose career might suffer. Emotions do not always connect us in proper measure to everyone whose interests are involved. Second, and far more importantly, not all emotions move us to sympathy or kindness. Some move us to do truly horrible things, as when fear motivates racist violence. To this point I will return a bit later.

The third problem with this argument is that there can be no first principles proof for the claim that robots cannot sense or express emotions, unless one simply defines emotions as something distinctly human. But that is an evasion, not an argument. No, this is an empirical question, the answer to which depends on progress in research and development. In the ten years since the original HRW call, some developers have claimed considerable progress in designing robots that are said to be able to read human emotions and respond in emotionally appropriate ways. The most widely publicized early example was the robot, Pepper, that was announced in 2014 and brought to market in 2015 by Aldebaran.[6] And while she prefers the language of "sociability" to that of "emotion," the development of such robots has long been the focus of Cynthia Breazeal's highly innovative Personal Robotics group in MIT's Media Lab.[7] It goes without saying that none of these robots yet evince anything like a full, human-like, emotional capacity. But a lot of progress has been made, and that is just the point. Only time will tell to what extent robot emotions will be realized.

[6] *SoftBank Mobile and Aldebaran Unveil "Pepper" – the World's First Personal Robot that Reads Emotions*, SOFTBANK (June 5, 2014), https://www.softbank.jp/en/corp/group/sbm/news/press/2014/20140605_01/.

[7] *Cynthia Breazeal*, MIT MEDIA LAB PEOPLE, https://www.media.mit.edu/people/cynthiab/overview/ (last visited . . .).

Serious conceptual confusions also plague discussions of the potential emotional capacities of robots. That current robotic technology cannot produce in robots the kind of emotional capacity that we recognize in ourselves is, as noted, not worth disputing, if only because human emotion requires the biology of an endocrine system. But is that the kind of competence needed in autonomous weapons? Do we really need weapons that cry? No. If an emotional capacity is needed, it might only be the ability to read human emotion and to respond in emotionally appropriate ways. One can well imagine that a sentry-bot might do its job more reliably were it able to sense fear, nervousness, or anger, even if it does not, itself, experience such. It is important to keep the difference in mind, because designing robots that read emotion and respond in emotionally appropriate ways is, from an engineering point of view, a much more tractable problem than designing robots that genuinely feel sadness or remorse. So the argument about emotional capacity proves little or nothing about the wisdom of developing and fielding autonomous weapons. That might be why one hears it less frequently today.

B. Discrimination and Proportionality

The other major argument in the 2012 HRW call for an autonomous weapons ban was that robots are inherently incapable of respecting the International Law of Armed Conflict and International Humanitarian Law because they lack the ability to distinguish combatants from non-combatants and the ability to make judgments about proportionality.[8] There is no disputing the fact that no current weapons system has the ability to make all of the subtle distinctions that human combatants must and often do make between, say, a nervous suicide bomber walking up to a checkpoint in Tikrit and a pregnant woman on her way home from shopping made anxious by all of the foreign force on display everyday in what was once her happy home town. But that obvious fact does not settle the question.

First, as with the question of robots and emotion, what capabilities we might engineer into weapons systems in the future is an empirical question, not one of principle. Will a robot ever be able to make the distinction just discussed between the suicide bomber and the pregnant shopper? Only time will tell, but it has to be noted that one of the areas

[8] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

of most rapid progress in artificial intelligence is pattern recognition. If a deep learning system can teach itself the difference between cats and dogs, why is it not conceivable that such a system can also learn to distinguish a gathering of Taliban leaders from a wedding party?

Second, the question is not whether perfection is possible in a robot's making such discriminations. The question is whether an autonomous weapons system can reach a reasonable threshold of success. After all, however high our expectations, we have to acknowledge that humans make far too many mistakes, sometimes costing the lives of the innocent, sometimes costing the lives of our own personnel. If an autonomous weapons system can consistently outperform human soldiers in distinguishing combatants from non-combatants, then there would be a moral gain. I would set the threshold higher still. But wherever that threshold lies, whether it can be met is an empirical question to be answered only by further research and development.

Third, the argument as stated seems to assume that discrimination is a context-independent competence. But this is not true. In fact, the kind of discrimination that is needed is highly context dependent. Consider, for example, the British Brimstone air-launched ground-attack missile system.[9] First deployed in 2005, it was originally designed as a fire-and-forget missile for use mainly against tanks and other mobile, armored vehicles. The original design assumed operation within a highly-circumscribed fire zone, one in which there was a reasonably low probability of encountering non-combatants. On-board sensor systems and programming, including active radar homing, handled target identification, acquisition, tracking, and firing, all based upon a set of situation specific targeting data uploaded before launch by a weapons system officer (WSO). Most importantly, the Brimstone system was designed to be capable of distinguishing between, say, a tank and a passenger vehicle, with the decision to fire based entirely on that distinction. If a suitable target was not found, the missile would self-destruct. The crucial fact is that, in this original configuration, Brimstone is an autonomous offensive weapons system capable of making context-specific discriminations between permissible and impermissible targets.

But the rules of engagement in Afghanistan required a person-in-the-loop, precluding the use of Brimstone in its original form. This led to

[9] *Brimstone*, MISSILE THREAT (July 30, 2021), https://missilethreat.csis.org/missile/brimstone/; *Brimstone Advanced Anti-Armour Missile*, ARMY TECH. (July 16, 2021), https://www.army-technology.com/projects/brimstone/.

the development in 2008 of a new, dual-mode model, with an added laser-targeting system that could be used by the pilot of the launch aircraft (so far only British Tornado and Typhoon aircraft) to guide the munitions to the target, the choice of mode being in the hands of the pilot.

Brimstone has now been modified for use also as a ground-based, antitank weapon, with the capacity of being mounted on unmanned ground vehicles. A variant model, Sea Spear, has been developed for use against swarms of small boats, in either a ship-launched or helicopter-launched version. Dual-mode Brimstone systems have been sold to Saudi Arabia, and there has been discussion of supplying the Sea Spear system to both Estonia and Ukraine.

There are many questions that one might ask about Brimstone. Should mode selection be in the hands of the pilot of the launch aircraft? What should be the constraints on the targeting data uploaded by the WSO? In what kinds of conflict arenas is such a system appropriate? But the main point, again, is that Brimstone is an example of an autonomous offensive weapons system about which it is claimed that, within an appropriately circumscribed context, it is capable of making the kind of discrimination required by ILOAC and IHL.

Whether the claimed discrimination capability is as robust as has been asserted and whether still more stringent constraints are appropriate are, of course, relevant questions. But I want to defer those questions to when I take up the proposal of an Article 36 based certification system for autonomous weapons. For now, let us just use the Brimstone example to illustrate the point that discrimination is a context-dependent issue and that, in some contexts of deployment, we might already have hardware and software capable of making the necessary discrimination.

C. Human Dignity

Of all of the arguments against autonomous weapons that are known to me, the most moving is perhaps that which asserts that the decision to kill must be left to a human being because, only thus, do we respect the essential human dignity of the human target and of all of those humans otherwise implicated in the use of violence in war. The idea is that a combatant makes him- or herself less than human by delegating a kill decision to an artificial system that cannot understand the victim's suffering and that one also, thereby, denies the human dignity of the victim. This argument takes center stage in the 2018 HRW report updating the call for a ban on autonomous weapons and it has been widely discussed in the literature.[10]

That the argument from human dignity has the power to persuade is obvious. But is it a cogent argument? One curious feature of many invocations of the argument from dignity is the frequency with which its proponents openly acknowledge the difficulty of clearly articulating the core concept of human dignity. For example, Amanda Sharkey, one of the leaders of CSKR, devotes four dense pages of her 2019 paper, "Autonomous Weapons Systems, Killer Robots, and Human Dignity," to a surprisingly detailed cataloguing of the contradictions, ambiguities, and other muddles to be found in the literature, concluding that "it should be apparent that not only have some specific questions been raised about the impact of AWS on human dignity, but also that there is a lack of a clear consensus about what dignity is."[11] Equally noteworthy, however, is the fact that, having acknowledged the inherent lack of clarity of the concept of human dignity, the proponents of the dignity argument still commend its usefulness from a rhetorical point of view. Sharkey is straightforward about this. Having asked whether the dignity argument would help the campaign against killer robots, she responds:

> "There could be some campaigning advantages. Saying that something is against human dignity evokes a strong visceral response. Even though dignity is difficult to define clearly, people have an intuitive understanding of its meaning, and of the importance of maintaining and preserving it. Reference to human dignity can highlight a repugnance to the idea of machines having the power of life or death decisions over humans."[12]

Elvira Rosert and Frank Sauer make a similar rhetorical point in their 2018 paper, "Prohibiting Autonomous Weapons: Put Human

[10] *Heed the Call: A Moral and Legal Imperative to Ban Killer Robots*, HUMAN RIGHTS WATCH (Aug. 21, 2018), https://www.hrw.org/report/2018/08/21/heed-call/moral-and-legal-imperative-ban-killer-robots. *See e.g.*, Michael Horowitz*, The Ethics and Morality of Robotic Warfare: Assessing the Debate Over Autonomous Weapons*, 145 J. OF THE AM. ACAD. OF ARTS & SCI., no. 4, 2016, at 25–36 (2016); Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 75–87 (2018); Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOBAL POLICY, no. 3, 2019, at 370–75.

[11] Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 82 (2018).

[12] Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 83 (2018).

Dignity First," writing: "From a strategic communication point of view, adjusting the message toward the infringement on human dignity would have the general benefit of dampening the overall level of contention."[13]

Those dedicated to a cause are not to be faulted for thinking carefully about the rhetorical impact of their arguments. But we must remember that the ultimate aim of the campaign for a ban on autonomous weapons is the crafting of new international law or other ways of norming the use of such weapons, and premises that work by eliciting a visceral response might not serve well as a basis for that latter enterprise, one in which clarity is most definitely a virtue. Some champions of the dignity argument, such as the authors of the 2018 Human Rights Watch call for a ban,[14] will respond by claiming that the appeal to human dignity already serves well as a basis for International Humanitarian Law (IHL) and the International Law of Armed Conflict (ILOAC) in the form of the Martens Clause, which was incorporated in the 1899 Hague Convention and added to the Geneva Conventions in Additional Protocol 1 of 1977:

> "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience."[15]

Appeals to the Martens clause have played an important role in the arguments leading to the adoption of several additions to ILOAC, such as the ban on blinding lasers. But it is well to remember what led to the adoption of the Martens clause in the first place. It was added to the Hague Convention precisely to paper over issues about which the delegates could not reach consensus by reasoning from other, clear, legal principles, and there has since been a long history of debate and disagreement over how to interpret the clause, precisely because of the mentioned unclarity in such notions as essential human dignity.[16]

[13] Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOBAL POLICY, no. 3, 2019, at 370–75.

[14] *Heed the Call, supra* note 10, 8-43.

[15] Geneva Conventions in Additional Protocol 1 of 1977

[16] Rupert Ticehurst, *The Martens Clause and the Laws of Armed Conflict* 317 INT'L REV. OF THE RED CROSS, April 1997, at 125–34.

Still, such arguments work well by way of stirring the emotions. When I think about the argument from human dignity, my mind goes immediately to a remarkable moment in the climactic battle sequence of the movie, "Saving Private Ryan," where, at the end of a harrowing, hand-to-hand struggle, a tough German soldier rolls atop the exhausted Private Stanley Mellish, taking a bayonet from Mellish's hand. Mellish begs the German, "Listen to me. Listen to me. Stop. Stop." But the German slowly pushes the bayonet into Mellish's heart, holding him almost tenderly and gently whispering to him, "Shhh. Shhh," like a father to a frightened son, until Mellish breathes his last. The German understood Mellish's suffering and fear, and one wants to think that Mellish might have taken comfort at the end from the warmth of the German's embrace. I think that Steven Spielberg was trying to make a complicated point about morality in war with that scene. We are supposed to despise the German soldier, but, ironically, his act of killing becomes an act of love. There can be no more essentially human moment in war than such an intimate, face-to-face act of violence.

I am so moved by such a scene that even just describing it leaves me emotionally and psychologically drained. I have to take a deep breath. I have to recenter and relax. Only then can I stop and think clearly.

What do I think? When emotion subsides and my head clears, I am horrified by the suggestion that, because Mellish was killed by a human who sought to comfort him in his dying moment, there was, therefore, in that act, respect for human dignity of a kind that would be missing were Mellish killed by a robot. On the contrary, one can argue that killing in any form, even in war or self-defense, entails the denial of human dignity, if there is such. But the problem is that killing in war and killing in self-defense are sometimes necessary, however fundamentally inhumane that killing might be. Kill we must, but let's not make killing out to be anything other than what it really is, namely, a horrible, if unavoidable, denial of both our own and the victim's humanity. This is why even people fighting on the "good" side in a perfectly just, defensive war experience killing in war as morally corrosive. I think that any attempt to make it appear that humans killing humans in war is more humane than robots killing humans in war is to lose sight of our humanity in a most profound way.

What, then, of the argument against autonomous weapons from the premise of essential human dignity? I think that, killing in war being the denial of human dignity, the morally responsible thing to do is to minimize it, to do no more killing, to inflict no more harm than is absolutely necessary for the achievement of proper ends. That principle has long been fundamental in International Humanitarian Law and the International Law of Armed Conflict going all the way back to the 1868 St. Petersburg Declaration, which banned weapons and practices that cause unnecessary suffering, and it is now codified as Rule 70 of Customary IHL.[17] I think that I respect the dignity of my enemy and of all of those who suffer in war by doing everything that I can to minimize violence and the harm that I do to others. If autonomous weapons further that end, then so be it. Were I in Mellish's situation, knowing that I am going to die, what I would want most would be for it to be a quick and painless death. Soothing words from my killer would only add to the insult.

D. Increasing the Temptation to Engage in Conflict

If autonomous weapons promise both to minimize a nation's own casualties and to minimize harm to non-combatants, will there not be an added incentive to initiate conflict, say by intervening in conflict situations where, previously, the threat to one's own troops or worries about collateral casualties would have made the risk not worth the gain or the intervention politically unacceptable? Could we imagine that a high-minded effort to minimize death and suffering might, in this way, ironically, increase death and suffering by increasing the number of fights in which we engage?

The worry is not new to autonomous weapons. The same concern has often been expressed about the "smart" weapons that featured so prominently already in the First Gulf War. Thus, more than one critic of US military policy has argued that we would not have intervened in the Libyan conflict had it required troops on the ground and that Obama judged it politically feasible to intervene because "smart" munitions gave us an ability to assist the anti-Gaddafi forces without seriously risking the lives of US troops.[18] While that intervention toppled the Gaddafi regime, the long-term consequences, including bloody civil war and Libya's becoming a terrorist haven, proved to be catastrophic.

That the availability of autonomous weapons might increase the temptation to engage in conflict cannot be denied. But, as with so many of the other arguments against autonomous weapons, the first response

[17] *Rule 70. Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering*, IHL DATABASE, https://ihl-databases.icrc.org/customary-ihl/eng/docindex/v1_rul_rule70 (last visited . ..).

[18] *See, e.g.* Lawrence Kaplan, *More Questions than Answers: Obama, Libya, and the Dubious Ethics of Modern Air Wars*, THE NEW REPUBLIC (Mar. 22, 2011), https://newrepublic.com/article/85555/obama-libya-air-war-qaddafi-ethics.

is that, whether in fact such an effect occurs is an empirical question, as is the question of the magnitude of the effect. However, in this case, it is also a political and a moral question. It is not just whether such actions do occur, but whether they should. There are other examples of military intervention made politically and militarily easier by technology that have a very different moral valence than the Libyan conflict. The NATO intervention in Kosovo in 1998 is one such.[19] Opinion differs strongly about the net benefit of NATO intervention, but I am on the side of the argument that sees NATO's role in Kosovo as an exemplary model for the future. Mistakes were made and innocent civilians suffered. But an ethnic war of possibly catastrophic proportions was prevented. European and US public opinion would not have tolerated a massive NATO ground involvement in Kosovo. The good that was achieved was made possible by our ability to apply force with minimal risk to our own personnel and to non-combatants. Did we kill civilians who otherwise would not have died? We did. But how many Kosovar and Serbian lives did we save in the process? That is the proper question. And my reading of the evidence suggests that we probably saved many tens of thousands of lives.[20]

So the question is not whether the even greater reduction in suffering promised by autonomous weapons would lead to more military interventions. The question is, rather, what kinds and numbers of interventions would such a capability facilitate. If such a capability could have made it politically and militarily feasible to stop the slaughters in Rwanda, Cambodia, and Biafra - to name only the most horrific wars of the last several decades - then that would have been a moral gain.

E. An Autonomous Weapons Arms Race

The 2015 Future of Life Institute call for an offensive autonomous weapons ban foregrounded an argument mentioned but not as much emphasized in the 2012 HRW call for a total ban. This is the argument that, absent a ban, we will see a global autonomous weapons arms race that will make the nuclear weapons arms race pale by comparison.[21]

That there would be an autonomous weapons arms race is likely. After all, it is declared US policy to seek and maintain technological

[19] BENJAMIN LAMBETH, NATO'S AIR WAR FOR KOSOVO: A STRATEGIC AND OPERATIONAL ASSESSMENT (2001).

[20] Agon Maliqi, *Remembering the U.S. Intervention That Worked*, WASH. POST. (June 8, 2019), https://www.washingtonpost.com/opinions/2019/06/08/remembering-us-intervention-that-worked/.

[21] *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, FUTURE OF LIFE INST (July 28, 2015), http://futureoflife.org/open-letter-autonomous-weapons/.

dominance over all of our potential adversaries, and adversaries such as China and Russia have made clear their determination to narrow the gap if not to surpass US capabilities in at least some modes of conflict, space-based weapons being an especially noteworthy example.[22] We know that Russia has been developing various types of autonomous weapons. Other actors are also getting into the game. For example, in October 2015 reports on a recent military exercise, Iran announced that it was testing what it called "kamikaze robots," whatever that means.[23] In June 2021, it was reported that the Libyan government used Turkish-made, autonomous, weaponized drones in an attack on rebels.[24] Moreover, history has shown that adversaries capable of competing with innovative US military technologies have done so. Soviet era competition with the US in ballistic missile and space technology is probably the most famous example, because, the US did not always lead in that competition, certainly not in its earliest years, with Sputnik having been the first earth satellite and Yuri Gagarin the first human in space (see Wolfe 2013), and some argue that the US is now trailing behind Russia and China in the development of hypersonic weapons.[25] But the history of competition in weapons technology goes back far beyond the Cold War to the earliest days of technologized warfare. One thinks of competition in submarine and tank technology in World War II, or the tragic competition in poison gas weapons in World War I.

Competition in weapons development has, thus, been the norm for a long time. Why, then, would one think that there would be something importantly different about an autonomous weapons arms race? Cost might be one factor, some robotic systems being cheap by comparison with both conventional arms and human combatants. So there might be more players in a robot weapons arms race. But the cheap

[22] *See* GIAN GENTILE ET AL., A HISTORY OF THE THIRD OFFSET, 2014–2018 (2021); Abraham Mahshie, *Russia and China Could Team Up to Challenge US Space Superiority, Experts Say*, AIR FORCE MAG. (June 29, 2021), https://www.airforcemag.com/russia-china-team-up-challenge-us-space-superiority/.

[23] *Straight Truth, 'Kamikaze' robots debut in Iran Army Drill*, TEHRAN TIMES (Oct. 21, 2015), https://www.tehrantimes.com/news/250250/Kamikaze-robots-debut-in-Iran-Army-drill).

[24] Joe Hernandez, *A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says*, NPR (June 1, 2021), https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d.

[25] McLeary, Paul and Alexander Ward (2021). "U.S. 'Not as Advanced' as China and Russia on Hypersonic Tech, Space Force General Warns." *Politico*. November 20, 2021. https://www.politico.com/news/2021/11/20/hypersonic-technology-us-behind-china-russia-523130.

weapons are likely to be the less worrisome ones, there being a rough correlation between cost and destructive potential. But the price of entry for large-scale autonomy – with, say, integrated, autonomous command and control combined with autonomous weapons platforms for multiple levels and modes of combat across a large field of combat – will keep out all but a few actors, the US, Russia, and China being the main candidates.

Competition at that level could be worrisome. And one would expect to see greater levels of autonomy and increased system integration. But then the question is whether such development is likely to take those actors to a level where serious fears about a loss of human control is conceivable. Here, again, history is helpful. For, during the Cold War, both the US and the Soviet Union took steps in the direction of automating their nuclear attack response capabilities, the idea being that, if human operators do not survive, then the computers can launch the retaliatory strikes. Hollywood had fun with this theme, in movies like "War Games." But the attendant risks of such automated response capabilities were well understood, which is why we never went too far down that road and why we engineered multiple layers of checks and controls. We learned important lessons about the vulnerabilities of engineered systems to unanticipated failure modes. To be sure, we came close to nuclear Armageddon on too many occasions, but those were mostly human failures, and we came close to serious nuclear accidents on many more, and from those near-misses we learned still more about how to engineer against failure (see Schlosser 2013).

One final feature of the analogy between the nuclear arms race and an autonomous weapons arms race puzzles me greatly. The destructive capability of nuclear weapons is such that even a medium-scale, regional nuclear exchange could have globally catastrophic consequences. But autonomous weapons are, from one point of view, the next phase in a history of steadily dialing back destructive power thanks to our technology's making possible the ever-more-accurate delivery of force on a target. This trend line is no accident. It is deliberate policy at least in the US military. If competition in autonomous weapons development were to accelerate this trend, would that not be a moral gain rather than a loss (U.S. Mission Geneva 2019)?

F. Autonomous Weapons and an Artificial Intelligence Apocalypse

I can construct only one scenario through which an autonomous weapons arms race could leave us in a worse place than the nuclear weapons arms race did, and that scenario is the one envisioned in the newest, and, I think, most curious, argument for an autonomous weapons ban. This argument asks us to imagine a time when, the singularity having arrived, the artificial intelligence is smarter than us and decides to use enhanced autonomous weapons capabilities either to eliminate humankind altogether or wreak comparable mayhem in service of a goal that we mere humans cannot comprehend. This is the Skynet apocalypse, famous from the "Terminator" movie series.

When first I saw this argument, I could not believe that serious people would promote it, because I tell all my students and all of my audiences never to look to Hollywood science fiction for guidance, for the obvious reason that Hollywood purveys, well, fiction, not fact, and fiction that preys on our deepest irrational fears, not reasonable extrapolations from current technology. Imagine my even greater surprise, therefore, when, in 2015, AI specialist, Toby Walsh, the main engine behind the new Future of Life Institute call for an autonomous weapons ban wrote, in an op-ed at CNN: "Once this genie is out of the bottle, there will be an arms race to improve on the initially rather crude robots. And the end point of such an arms race is precisely the sort of terrifying technology you see in 'Terminator. Hollywood got that part right."[26] Seriously? Are we really debating such an important issue on the basis of Hollywood nightmare films?

But let us be serious about the question and ask whether the "Terminator" apocalypse is a realistic scenario of such a kind that it should guide our thinking about weapons development policy. Is a "Terminator" apocalypse possible? Of course it is, from a purely logical point of view. There is nothing inherently contradictory in the concept of such a future. But if it is possible, and if it would mean the end of all human life, then must we not do everything possible to prevent it, starting with an immediate ban on all autonomous weapons development? However unlikely the possibility, the consequences would be so dire that all other possible futures are irrelevant. That seems like a reasonable argument. No?

The very reasonableness of the argument, or its seeming reasonableness, is the problem. If, in any policy debate, one assigns an infinite negative utility to a given possible outcome, such as the death of all humankind, then, no matter how tiny the probability, the product of negative infinity times that tiny probability totally overwhelms every

[26] Toby Walsh, *The Rise of the Killer Robots - And Why We Need to Stop Them*, CNN (October 26, 2015), http://www.cnn.com/2015/10/26/opinions/killer-robots-walsh/index.html.

other term in the expected utility calculation, rendering such a calculation useless for policy purposes. In other words, the invocation of an apocalypse means – and this is sometimes the goal – the end of rational deliberation.[27]

Another way to think about this is to realize that there are many conceivable apocalypse scenarios. Global climate change might render the planet uninhabitable for all higher life forms within a few hundred years if we pass a climate tipping point in the very near future. That is another, possible future. Must we, therefore, immediately subordinate all other human purposes to effecting not just an immediate end to CO2 equivalent emissions but also the active removal of CO2 equivalents from the atmosphere?

Of course it is also possible that a new "terminator" pathogen might evolve tomorrow, one vastly more virulent and lethal than the Spanish flu of 1918 or Ebola or COVID-19, one that could eliminate all human life. Therefore, instead of redirecting all of our resources to combating climate change, we should stop all travel, all meetings of two or more strangers, all animal farming, all raising of pets, all activities that might facilitate viral transmission among individuals and species. And we should redirect all of our research efforts to studying viral evolution and to the development of new vaccines and disease treatments. But wait a minute. We cannot do such research, because the research, itself, might accidently create such a "terminator" pathogen that might be accidentally released into the wild. It could happen.

Perhaps our demise might be caused not by our actions but by our inaction. It is possible that a heretofore undiscovered space rock of a size capable of causing an extinction-level event might be found next year to be hurtling toward a collision with Earth that could cause a catastrophe on the scale of that which produced the cretaceous extinction. Again, Hollywood loves this scenario. But it is possible, as witness the sudden appearance in 2015 of a previously unknown asteroid, 2015 TB145, large enough, at 400m, to cause continent-scale devastation, that passed nearer to Earth on Halloween than any other object of that size since

[27] Don Howard, *On the Moral and Intellectual Bankruptcy of Risk Analysis: Garbage In, Garbage Out,* SCIENCE MATTERS BLOG. (Sept. 26, 2014), http://donhoward-blog.nd.edu/ 2014/09/26/on-the-moral-and-intellectual-bankruptcy-of-risk-analysis-garbage-in-garbage-out/#.VjeO-H6rT4Y; Casadevall, Arturo, Michael Imperiale, Don Howard, *The Apocalypse as a Rhetorical Device in the Influenza Virus Gain-of-Function Debate,* MBIO: AN OPEN ACCCESS JOURNAL PUBLISHED BY THE AMERICAN SOCIETY FOR MICROBIOLOGY 5 (5) e01875-14 (Oct. 14, 2014), http://mbio.asm.org/content/5/5/e02062-14.full.

1999.[28] If our doing nothing would seal the fate of humankind, should we not redirect all of our resources to the most rapid possible development of a technology for deflecting such space objects from a collision course with Earth?

I trust that the point is clear. Invocations of apocalypse make rational policy decisions impossible. It is well to be mindful of all such possible catastrophes. But balanced good judgment would place more emphasis on the extremely low probabilities than on the infinite, negative utilities of such events. Prudence might well dictate our taking steps to minimize those probabilities still further or to mitigate harm in the case of events beyond our control. But neither prudence nor reason should lead us to act merely on the basis of possibility.

What, then, should we say about Terminator-AI apocalypse scenarios? What we should say is that, contrary to Toby Walsh's confident assertion that such an apocalypse is the "end point" of an autonomous weapons arms race, such an extrapolation from current technology is not supported by any evidence or compelling argumentation.

History has shown that forecasting technology development is a nearly impossible task. This point is forcefully made in a 1983 paper by Charles Townes, co-inventor of the transistor, in which he reflected on our poor record of technology forecasting in the twentieth century. He points to a 1937 report of a committee of experts assembled at the behest of President Roosevelt to assess technology trends as they might affect national policy and planning. Among the revolutionary technologies of the near future totally missed by the committee were: nuclear energy, radar, antibiotics, jet aircraft, rocketry, space exploration, computers, microelectronics, and genetic engineering. And Townes notes that the scientific and technical bases for nearly all of these developments were already in place in 1937.[29]

But what about the development of AI in particular? In fact, opinion is strongly divided over the pace and nature of advances in AI. There have been notable achievements in recent years, thanks, especially, to machine learning algorithms and neural nets. Some predict that the AI singularity – the point at which AI is supposed to surpasses human

[28] Todd Leopold, John Newsome, Jareen Imam, *Halloween Asteroid Resembling Skull Narrowly Misses Earth*, CNN (Oct. 21, 2015), https://www.cnn.com/2015/10/21/us/asteroid-earth-nasa-halloween-feat/index.html.

[29]Charles H. Townes, *Science, Technology, and Invention: Their Progress and Interactions*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES (1983) at 80: 7679–7683.

intelligence – might arrive within a couple of decades. But others push the date back to the end of the century.[30] Still others - and count me in this group - point out that the very question is probably not well posed, since human intelligence is not some one thing that will be achieved in artificial systems at a magical moment when the light of consciousness suddenly turns on in Watson. My expectation is that more and more human abilities will be better approximated, some even eclipsed, in the coming years, but that this will happen in a context-dependent and task specific way, not in the form of general AI. We should closely monitor all of these developments for their potential impact on human well being, and we need to be able to respond nimbly and quickly should serious threats emerge. But no one can pretend now to know that a Terminator-AI apocalypse is inevitable or even likely.

G. *Differences Between Offensive and Defensive Weapons Systems.*

An interesting twist in the 2015 Future of Life Institute call for an autonomous weapons ban is that it proposes to ban only offensive autonomous weapons. One might guess that one reason for this modification is that defensive autonomous weapons have been proving their effectiveness and basic safety for a number of years. The US Navy's fully autonomous, Phalanx, ship-borne, anti-missile defense system, which fires 20 mm projectiles at a rate of between 3,000 and 4,500 rounds per minute from six, revolving barrels, was first developed in late 1970s.[31] Israel first used its autonomous, Iron Dome anti-missile defense system in 2011.[32] And South Korea introduced the Samsung SGR-A1 border patrol robot, which has both a fully autonomous and a person-in-the-loop mode, in September of 2014.[33]

[30] In November 2015, Microsoft's head of research, Eric Horvitz, opened MIT's annual EmTech conference by noting that "the mastery of AI has been much harder than expected." (http://www.techrepublic.com/article/mastery-of-ai-has-been-harder-than-expected-and-future-is-uncertain-says-microsofts-ai-chief/)

[31] John Pike, MK 15 Phalanx Close-In Weapons System (CIWS), FAS Miltary Analysis Network (January 9, 2003), https://man.fas.org/dod-101/sys/ship/weaps/mk-15.htm.

[32] Missile Defense Project, "Iron Dome (Israel)," *Missile Threat*, Center for Strategic and International Studies (April 14, 2016), https://missilethreat.csis.org/defsys/iron-dome/.

[33] David Crane, *Samsung SGR-A1 Armed/Weaponized Robot Sentry (or 'Sentry Robot') Remote Weapons Station (RWS). Finally Ready for Prime Time?,* DEFENSE REVIEW (September 17, 2014), https://defensereview.com/samsung-sgr-a1-armedweaponized-robot-sentry-or-sentry-robot-remote-weapons-station-rws-finally-ready-for-prime-time/.

Many people seem to share the intuition that the rules of defensive warfare might differ from those for offensive action, I suppose on the grounds that killing in self-defense differs from initiating killing in morally relevant ways. That may be so in cases of individual self-defense, as when I am allowed to use deadly force to protect myself and others from an imminent threat of death, when the pre-emptive taking of life would not be permissible. But it is not at all obvious that there is a morally-relevant difference when it comes to the employment of autonomous weapons in war. Start with the fact that the tradition of Just War Theory and the body of international law founded upon it assumes that going to war is morally justified only to right a wrong, meaning that, in a sense, the only permissible war is one of defense against an aggressor. Of course, bad actors initiate conflict for bad reasons all the time (however much they might convince themselves that they are righting wrongs), and ILOAC and IHL seek to norm all such conflict. Consider next the fact that, in war of any kind, there is no perfect or even very clean distinction between offensive and defensive action. If someone shoots at me and I shoot back, that is clearly a defensive act, no? But what if I provoked the first shot by some tactic like reconnaissance in force, aimed at eliciting enemy fire? On the other hand, my initiating combat to secure an objective seems the epitome of an offensive action. But what if the ultimate goal were to secure, say, a high point for better defense against possible future assaults? Examples such as these are the daily bread of courses on ILOAC for young ROTC cadets and students of military law. They all go to prove the point that what makes an act offensive or defensive is highly context dependent and depends also on the larger aims and intentions of the actors.

But what about the weapons themselves? Surely there is no imaginable offensive use for Iron Dome or Phalanx. They were designed as defensive systems and have only been deployed for purposes of defense. Or have they? Iron Dome is an especially interesting example. It has been used so far mainly only in defense against Hamas missile attacks originating from within Gaza. While its effectiveness has been disputed, it has made many impressive kills and has surely prevented damage if not also saved lives. What could be a more morally just use of high technology? In fact, Iron Dome is only the first layer of Israel's evolving, multi-layer, anti-missile, defense system, that also includes the Arrow 2, Arrow 3, Arrow 4, and David's Sling systems that are designed to defend against not only Hamas's crude, short-range missiles but also against tactical and intermediate-range ballistic missiles of the kind that Iran has developed.[34] Some observers think that the real goal of the overall program is to provide a comprehensive defensive shield against Iranian ballistic missiles so as to insulate Israel against retaliation if, for example, Israel chose to launch a pre-emptive strike against Iranian nuclear weapons facilities.[35] If so, then what appears a defensive capability becomes an offensive one by making possible offensive actions that would otherwise lead to unacceptable risk to one's own nation. The logic here is much like that in the debate about Star Wars in the 1980s. Who could not welcome a perfect defense against nuclear armed ICBMs? The Soviets, for one. They regarded Star Wars as a highly destabilizing technology because they feared that it would embolden the US to launch a pre-emptive strike secure in the faith that a Soviet retaliatory strike would fail.

The Phalanx system challenges the offensive-defensive distinction in the same way as Iron Dome, for a defense against anti-ship missiles facilitates offensive action in, say, the Straits of Hormuz, that otherwise might be too risky. But Phalanx also challenges the distinction in a more straightforward way. During the Iraq War it was already adapted for use by ground forces in such settings as perimeter defense against mortars and other small, fast munitions.[36] Mount it on a mobile platform, alter a few lines of code, and it would become a fearsome offensive weapon, obliterating bodies, buildings, and even heavy armor that might be in its path.

So there is no clear-cut distinction between offensive and defensive autonomous weapons sufficient to support the restriction of the Future of Life Institute's proposed ban to offensive weapons alone. The offensive-defensive distinction is functional and contextual, not structural, a matter not so much of technology as of human intention. The contextual nature of the offensive-defensive distinction reminds us of the point made earlier about the contextual nature of discrimination, and both points will be relevant when, shortly, we turn to the question of an Article 36-based alternative to a wholesale ban.

[34] Gili Cohen, *Why Does Israel Need Three Different Missile Defense Systems*?, HAARETZ (April 2, 2015), https://www.haaretz.com/.premium-why-does-israel-need-3-anti-missile-systems-1.5346632.

[35] John Hannah, *Israel Needs Weapons to Stop Iran's Bomb*, FOREIGN POLICY (October 15, 2021), https://foreignpolicy.com/2021/10/15/israel-idf-iran-nuclear-arms-weapons/.

[36] *20 mm Phalanx Close-in Weapon System (CIWS)*, NAVWEAPS (last updated, Jan. 6, 2022), http://www.navweaps.com/Weapons/WNUS_Phalanx.php.

II. Moral Advantages of Autonomy

All of the main arguments in favor of an autonomous weapons ban have been found wanting. Let us turn to the other side of the argument and remind ourselves about the claimed moral gains from the introduction of autonomous weapons. By far the most compelling case of this kind is that made by Ronald Arkin in his 2009 book, *Governing Lethal Behavior in Autonomous Robots*.[37] Do not dawdle over the particular architecture that Arkin suggests in that book, some of which is already dated, though his idea of the "ethical governor" is still worthy of attention.[38] Appreciate, instead, his main point, which is that humans are notoriously unreliable systems, that human combatants commit war crimes with frightening frequency, and that what we must ask of autonomous weapons systems is not moral perfection, but simply performance above the level of the average human soldier.

There is not space here to review in detail the study by the United States Army Medical Command's Office of the Surgeon General from the Iraq War upon which Arkin mainly bases his assessment of human combatant performance.[39] Suffice it to say that the numbers of admitted war crimes by US troops, the numbers of unreported but observed war crimes, and the self-reported ignorance about what even constitutes a war crime are staggering. With such empirical evidence as background, Arkin's claim to be able to build a "more moral" robot combatant seems far more plausible than one might initially have thought. Why?

Start with the obvious reasons. Autonomous weapons systems suffer from none of the human failings that so often produce immoral behavior in war. They feel no fear, hunger, fatigue, or anger over the death of a friend. Move on to the slightly less obvious reasons. Thus, a robot, not fearing for its own well-being, can easily err on the side of caution, choosing not to fire in moments of doubt (think of the suicide bomber/pregnant shopper scenario above), where a human might rightly have to err on the side of self-defense. Then consider still more important design constraints, such as those embodied in Arkin's "Ethical Adaptor," into which are programmed all relevant parts of ILOAC, IHL, and the

[37] Ronald Arkin, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS, Boca Raton, FL: Chapman Hall/CRC (2009).

[38] *Id.* at 127-133.

[39] Office of the Surgeon General, *Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07. Final Report,* DEPARTMENT OF COMMERCE, NATIONAL TECHNICAL REPORTS LIBRARY (November 7, 2006), https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB2010103335.xhtml#.

rules of engagement specific to given conflict arena or a specific action.[40] The Ethical Adapter blocks the "fire" option unless all of those prescriptions are satisfied. Arkin's robots could not fire (absent an override from a human operator) at all, unless the most stringent requirements are met. In the face of uncertainty about target identification, discrimination, applicability of rules of engagement, and so forth, the robot combatant defaults to the "no fire" option. Of course, other militaries could design the robots differently, say, by making "fire," rather than "no fire," the default. But hold that thought until, again, we turn to the discussion of an Article 36 regulatory regime.

Arkin illustrates the functioning of the Ethical Adaptor with several scenarios, one of which – a Taliban gathering in a cemetery for a funeral[41] – bears an eerie similarity to the horrific US attack on a Doctors without Borders (Médecins Sans Frontières - MSF) hospital in Kunduz, Afghanistan in October of 2015.[42] The rules of engagement as uploaded to the Ethical Adaptor would typically include specific coordinates for areas within which no fire would be permitted, including hospitals, schools, important cultural monuments, and other protected spaces. Likewise, no fire could be directed at any structure, vehicle, or individual displaying the red cross or the red crescent. This assumes, of course, sensor and AI capabilities adequate for spotting and correctly identifying such insignia, but, especially with structures and vehicles, where the symbol is commonly painted in large, high-contrast format on the roof, that is not a difficult problem. A fully autonomous drone designed as per Arkin's model that was tasked with the same action that led to the bombing of the MSF hospital in Kunduz simply would not have fired at the hospital. A human might have overridden that decision, but the robot would not have fired on its own. Moreover, the kind of robot weapon that Arkin has designed would even remind the human operator that a war crime might be committed if the action proceeds.

Another kind of moral gain from autonomous weapons was once pointed out to me by an undergraduate student – an engineering major – in my "Robot Ethics" class. He recalled the oft-expressed worry about the dehumanization of combat with standoff weapons, such as remotely

[40] Arkins, *supra* note 36 at 138-143.

[41] *Id.* at 157-161.

[42] Alissa J. Ruben, *Airstrike Hits Doctors Without Borders Hospital in Afghanistan*, NEW YORK TIMES (October 3, 2015), https://www.nytimes.com/2015/10/04/world/asia/afghanistan-bombing-hospital-doctors-without-borders-kunduz.html.

piloted drones. The concern is that the computer-game-like character of operator interfaces and controls, and the insulation of the operator from the direct risk of combat, might dull the moral sensitivity of the operator. But my student argued with deliberate and insightful irony, that the solution to the problem of dehumanization might be to take the human out of the loop, because it is the human operator who is, thus, dehumanized. For the record, I would dispute the dehumanization argument in the first place, because the typical drone operator often watches the target for many minutes, if not hours, and gets to know the humans on the receiving end of the munitions – including the wives, husbands, and children – far better than does, say, an artillery officer, a bombardier in a high-altitude bomber, or even the infantryman who gets, at best, a fleeting and indistinct glimpse of an enemy combatant across a wide, hazy, busy field of combat. That drone operators get to know their targets so well is part of the explanation for the extremely high reported rates of PTSD and other forms of combat stress among them.[43] Still, my student's point was a good one. If dehumanization is the problem, then take the dehumanized human operator out of the loop. This is really just a special case of Arkin's point about how stress and other contextual circumstances increase the likelihood of mistakes or deliberate bad acts by humans in combat and that, since robots are unaffected by such factors, they will not make those mistakes.

One of the most common criticisms of Arkin's model is the same voiced in the original HRW call for a ban, namely, that sensor systems and AI are not capable of distinguishing combatants from non-combatants, so that, even if the principle of discrimination is programmed into a robot weapon, it still cannot satisfy the requirements of international law. But we dealt with that point above, the two main responses having been: (1) what is or is not technically feasible is an empirical question to be decided by further research, not on a priori grounds, and (2) discrimination is usually a highly context-dependent challenge, and in some contexts, such as finding and identifying a Red Cross or Red Crescent symbol, the problem is easily solved.

The other major criticism of Arkin's model is that, since it assumes a conventional, structured, top-down, decision tree approach to programming ethics and law into autonomous weapons, it cannot deal

[43] Chappelle, Goodman, Reardon, Thompson, *An analysis of post-traumatic stress symptoms in United States Air Force drone operators*. J ANXIETY DISORD. 2014 Jun;28(5):480-7. doi: 10.1016/j.janxdis.2014.05.003. Epub 2014 May 17. PMID: 24907535.

with the often bewildering complexity of real battlefield situations. The basis of the objection is a simple and old worry about any rule-based or algorithmic approach to ethical decision making, such as deontology or consequentialism. It is that one cannot write a rule or build a decision tree to cover every contingency and that the consequentialist's calculation of benefit and risk is often impossible to carry out when not all consequences can be foreseen. The objection is a good one, at least by way of pointing out the limited range of applicability of Arkin-type autonomous weapons systems.

But Arkin's model for ethical autonomous weapons design is only a beginning. This last objection – that one cannot write a rule to cover every contingency – is the main reason why some of us are hard at work on developing a very different approach to ethics programming for artificial systems, one inspired by the virtue ethics tradition and implemented via neural nets and machine learning algorithms. The idea – already explored in concept by Wendell Wallach and Colin Allen in their 2010 book, *Moral Machines* (Wallach and Allen 2010) – is to supplement Arkin's top-down approach, involving rules and perhaps a consequentialist algorithm, with a bottom-up approach in which we design autonomous systems as moral learners, growing in them a nuanced and plastic moral capacity in the form of habits of moral response, in much the same way that we mature our children as moral agents.[44] There is considerable debate about this approach via moral learning. Arkin, himself, objects that neural nets and learning algorithms "black box" the developed competence in such a way as to make impossible both the robot's reconstructing for us either a decision tree or a moral justification of its choices, which he regards as a minimum necessary condition on moral machines, and the operator's reliably predicting the robot's behavior.[45] We respond that human moral agents are also somewhat unpredictable and that what they produce, when pressed for a justification of their actions, are after-the-fact rationalizations of moral choices. Why should we demand more of moral robots? How to produce after-the-fact rationalizations is an interesting technical question, one currently being vigorously and successfully investigated under such headings as "rule extraction," "interpretable AI," and "explainable AI."[46]

[44] Ioan Mutean & Don Howard, *Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency,* PHILOSOPHY AND COMPUTING (Thomas Powers, ed. Cham, Switzerland: Springer, 2017) at 121-159.

[45] Arkins, *supra* note 36 at 67, 108.

[46] Wojciech Samek, et al., eds. (2019). *Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning. Cham, Switzerland: Springer.*

Others object that there is no consensus on what morality to program into our robots, whether through learning or rule sets. We respond that moral diversity among robots should be prized in the same way that we prize human moral diversity. We learn from one another because of our moral differences. But, at the same time, in the constrained space of autonomous weapons, there is consensus in the form of the international support for extant international law and the just war moral theory upon which it is based. Saudi Arabian health care robots might rightly evince different habits with respect to touching and viewing unveiled bodies from those evinced by North American or European health care robots. But Saudi Arabia has ratified the main principles of the Geneva Conventions just as has the United States.

Earlier, we touched directly or indirectly upon other potential moral gains from autonomous weapons, such as facilitating military intervention to prevent genocide or other human rights abuses, minimizing risk of death or injury to our own troops, and sparing drone operators and other personnel both psychological damage and moral corrosion from direct participation in combat. One can imagine still more, such as employing weaponized autonomous escort vehicles to protect aid convoys in conflict zones. The conclusion is that there are, in fact, noteworthy potential moral gains from the development and deployment of both offensive and defensive autonomous weapons. Of course this must be done in such a way as to insure compliance with all existing international law and in a manner that minimizes the likelihood of the technology's being put to the wrong uses by bad actors. Short of a ban on autonomous weapons, how do we do that?

III. AN ARTICLE 36 REGULATORY REGIME

The goal is regulating the development and deployment of autonomous weapons in a way that ensures compliance with international law and minimizes the chance of misuse. Moreover, we need to do this in a politically feasible way, using regulatory structures that will be accepted by the international community. This last point is important, because one common criticism of the proposed ban on autonomous weapons is, precisely, that it stands little chance of ever being incorporated in international law.

Even in the talks under the aegis of the UN's Convention on Certain Conventional Weapons (CCW) that have being going on since 2014 in Geneva, it is mainly only nations with little or no prospect of becoming significant participants in the development and use of autonomous weapons that have shown support for moving forward with consideration of a ban. The major players, including the United States, have repeatedly indicated that they will not support a ban. In December of 2021, the United States representative in Geneva, Josh Dorosin, said it again, while adding that a non-binding, international code of conduct might be appropriate.[47] That sufficiently strong support for a ban was unlikely ever to emerge from the Geneva talks was already clearly sensed six years ago by the most energetic proponents of the ban. Thus, in a 2016 press release, the Stop Killer Robots campaign subtly shifted the discourse, hinting at a tactical retreat, by urging a focus on "meaningful human control" (whatever that might mean), though talk of a ban still dominates the headlines.[48] If the goal is regulating the development and use of autonomous weapons in a politically feasible way, then seven years of talks have been wasted by the continued insistence on a ban.

What could the international community having been discussing instead? The discussion should have focused on what might be done within the compass of extant international law. There is already in place since 1977 Article 36 of Protocol I to the Geneva Conventions, which stipulates:

> "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."[49]

[47] John Bowden, *Biden Administration Won't Back Ban on 'Killer Robots' Used in War*, THE INDEPENDENT. (December 8, 2021), https://www.independent.co.uk/news/world/americas/us-politics/biden-killer-war-robots-ban-b1972343.html.

[48] Clare Conboy, *Focus on Meaningful Human Control of Weapons Systems – Third United Nations Meeting on Killer Robots Opens in Geneva*, STOP KILLER ROBOTS (April 11, 2016), https://www.stopkillerrobots.org/news/press-release-focus-on-meaningful-human-control-of-weapons-systems-third-united-nations-meeting-on-killer-robots-opens-in-geneva/.

[49] *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I),* INTERNATIONAL COMMITTEE OF THE RED CROSS (Jun. 8, 1977), https://ihl-

174 states have ratified Protocol I, including Article 36, and three states, Pakistan, Iran, and the United States, are signatories but have not formally ratified the Protocol.[50] But the United States has promised to abide by nearly all provisions, including Article 36, and has established rules and procedures in all three branches of the military for insuring legal review of new weapons systems.[51] The countries having ratified Protocol I include every other major nation, among them China, the Russian Federation, and all NATO member states. I would argue that, since Article 36 is already a widely accepted part of international law, it is the best foundation upon which to construct a regulatory regime for autonomous weapons.

Concerns have been expressed about the effectiveness of Article 36 in general, chief among them being that the prescribed legal reviews are sometimes perfunctory and that it is too easy to evade an Article 36 review by declaring that a weapon is not new but just a minor modification of an existing and already authorized weapon. Those are serious worries, as evidenced by the recent controversy over whether the US's redesign of the B61 nuclear warhead with a tail assembly that makes possible limited, real-time steering of the warhead, the configuration designated now as B61-12, constituted a new weapon, as critics allege, or merely a modification, as the US asserts.[52] Another worry is that only a small number of states have certified that they are regularly carrying out Article 36 reviews. Equally serious are concerns that have been expressed about the effectiveness of Article 36 specifically with respect to autonomous weapons, as in a briefing report for delegates to the 2016 meeting of experts, which argued that what is at issue with autonomous weapons is not so much the conformity of individual weapons systems with international law, but the wholesale transformation of the nature of warfare wrought by the "unprecedented shift in human control over the

databases.icrc.org/applic/ihl/ihl.nsf/Treaty.xsp?action=openDocument&documentId=D9E6B6264D7723C3C12563CD002D6CE4/.

[50] *Id.*

[51] ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, INTERNATIONAL REVIEW OF THE RED CROSS (2006) at 88, 931-956; *see also* U.S. Army, *Legal Review of Weapons and Weapon Systems." Army Regulation 27–53*, DEPARTMENT OF THE ARMY (Washington, DC: Headquarters), (Sept. 23, 2019), https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN8435_AR27-53_Final_Web.pdf.

[52] Adam Mount, *The Case against New Nuclear Weapons*, CENTER FOR AMERICAN PROGRESS (May 4, 2017), https://www.americanprogress.org/article/case-new-nuclear-weapons/.

use of force" that autonomous weapons represent. The magnitude of that change was said to require not individual state review but the engagement of the entire international community.[53] All such concerns would have to be addressed explicitly in the construction of an autonomous weapons regulatory regime based on Article 36.

How would a new Article 36 regulatory regime be constructed? Most important would be the development of a set of clear specifications of what would constitute compliance with relevant international law. This could be the charge to a Group of Governmental Experts under the auspices of the CCW.

First in importance among such guidelines would be a detailed articulation of what capabilities an autonomous weapon must possess for handling the problem of discrimination, bearing in mind the point made repeatedly above that this is not an all-or-nothing capability, but, rather, one specific to the functions and potential uses of an individual weapons system. Thus, as discussed above, for use within its intended missions, the Brimstone missile need only the capability to distinguish different categories of vehicles within its designated field of fire. An autonomous check-point sentry, by contrast, would have to be capable of much more sophisticated discriminations. Similarly detailed specifications would have to be developed for determinations of proportionality, recognition of a human combatant's having been rendered hors de combat, recognition of a target's displaying insignia, such as the Red Cross or Red Crescent, that identify a structure, vehicle, or individual as protected medical personnel, and so forth.

Just as important as developing the specifications would be the development of protocols for testing to insure compliance. Optimal, but politically unachievable, for obvious reasons, would be the open sharing of all relevant design specifications. It is highly unlikely that states and manufacturers are going to let the world community look under the hood at such things as new sensor technologies and accompanying software. The alternative is demonstrations of performance capability in realistic testing scenarios. We already have considerable relevant experience and expertise in safety and effectiveness testing for a wide range of engineered systems, especially pertinent being the testing protocols for

[53] CCW, *Article 36 Reviews and Addressing Lethal Autonomous Weapons Systems*, Briefing Paper for Delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, at 11-15 (April 2016), http://www.article36.org/wp-content/uploads/2016/04/LAWS-and-A36.pdf.

certifying control systems in commercial aircraft and industrial systems. One might think that weapons developers would be just as shy about showing off the weapon at work in realistic scenarios, lest adversaries and competitors infer confidential capabilities and technologies. But, in fact, most weapons developers are proud to show off videos of their new systems' doing impressive things and to display and demonstrate their products at international weapons expositions. What would be required would not be the sharing of secrets but simply demonstrations of reliability in complying with the detailed guidelines just discussed.

As with the existing Article 36 requirements, certification of compliance will surely have to be left to individual states. But it is not unreasonable to begin an international conversation about a more public system for declaring that the required certifications have been carried out, even if that consists in little more than asking signatories and states parties to file such certifications with the UN, ICRC, or another designated international entity.

The good news is that, within just the last few years, serious discussion of precisely such concrete elaborations of Article 36 protocols for autonomous weapons has begun to appear in the scholarly, policy, and legal literatures.[54] Equally encouraging is the willingness of some governments to underwrite such work. Thus, the German Auswärtiges Amt (Foreign Office) subsidized a 2015 expert seminar under the auspices of the Stockholm International Peace Research Institute (SIPRI) that had representation from France, Germany, Sweden, Switzerland, the United Kingdom and the United States (Boulanin 2015).[55]

What have been the fruits of such work? Many good ideas have emerged. Especially thoughtful are the main recommendations contained in a 2017 report that was also sponsored by SIPRI covering Article 36 elaborations for cyber weapons, autonomous weapons, and soldier enhancement. Their approach was to focus on advice to reviewing

[54] Ryan Poitras, *Article 36 Weapons Review & Autonomous Weapons Systems: Supporting an International Review Standard*, AMERICAN UNIVERSITY INTERNATIONAL LAW REVIEW 34, at 465-495; *see* Cochrane, Jared M. (2020). "Conducting Article 36 Legal Reviews for Lethal Autonomous Weapons." *Journal of Science Policy & Governance* 16;1 (April 2020). https://www.sciencepolicyjournal.org/uploads/5/4/3/4/5434385/cochrane_jspg_v16.pdf.

[55] Vincent Boulanin, *Implementing Article 36 Weapon Reviews in the Light of Increasing Autonomy in Weapon Systems*, STOCKHOLM INT'L PEACE RESEARCH INST. (Nov. 2015), https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf.

authorities in individual member states, and they emphasize two broad categories of advice: (1) Building on best practices already being employed by states that have well-developed review procedures. (2) Strengthening transparency and cooperation among states. Under the first heading, they advise, for example:

> 1. Start the review process as early as possible and incorporate it into the procurement process at key decision points.
> 2. Provide military lawyers involved in the review process with additional technical training. Engineers and systems developers should also be informed about the requirements of international law so that they can factor these into the design of the weapons and means of warfare.[56]

About increased transparency and cooperation they say that it would become a "virtuous circle," and they observe that:

> 1. It would allow states that conduct reviews to publicly demonstrate their commitment to legal compliance.
> 2. It would be of assistance to states that are seeking to set up and improve their weapon review mechanisms and thereby create the conditions for more widespread and robust compliance.
> 3. It could facilitate the identification of elements of best practice and interpretative points of guidance for the implementation of legal reviews, which would strengthen international confidence in such mechanisms.

They add:

> Cooperation is also an effective way to address some of the outstanding conceptual and technical issues raised by emerging technologies. Dialogues, expert meetings and conferences can allow generic issues to be debated and addressed in a manner that does not threaten the national security of any state.[57]

[56] Vincent Boulanin & Maaike Verbruggen, *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*, STOCKHOLM INT'L PEACE RESEARCH INST., viii (Stockholm, Sweden) (2017).

[57] *Id.*

When it comes specifically to Article 36 reviews involving autonomous weapons, they identify as the foremost challenge verifying "the predictability of autonomous weapon systems' compliance with international law".[58]

I am not at all naive about how strict compliance with Article 36 requirements would be. But existing Article 36 requirements have already created a culture of expectations about compliance and a space within which states can and have been challenged, sometimes successfully, to offer proof of compliance, as with the widely expressed concerns about truly indiscriminate weapons, such as land mines and cluster munitions. We begin to norm such a space simply by putting the relevant norms in front of the world community and initiating a public conversation about compliance. This is what we should be talking about in Geneva if we are serious about building some measure of international control over autonomous weapons.

CONCLUSION

War is hell. It will always be an inherently immoral form of human activity. The goal of international law is to minimize the otherwise inevitable death and suffering that war entails. Advances in technology can contribute toward that goal by making weapons more accurate, less lethal, and more selective. The advent of autonomous weapons promises still further moral gains by removing the single most common cause of war crimes, the too often morally incapacitated human combatant. We cannot let unrealistic fears about a Terminator-AI apocalypse prevent our taking advantage of the opportunities for moral progress that properly designed and deployed autonomous weapons afford. We must, of course, ensure that such systems are being used for good, rather than malign purposes, as we must with any technology, and especially technologies of war. Indeed, with autonomous weapons we need to be more vigilant, still. But minimizing death and suffering in war is the ultimate goal. If autonomous weapons can contribute to progress toward that goal, then we must find a way to license their use in full compliance with what law and morality demand.

[58] *Id.* at xi.

ACKNOWLEDGMENTS

This paper was much improved thanks to helpful feedback from discussants at conferences and colloquia at the University of Zurich in November of 2015, Purdue University in December of 2015, and the University of Minnesota in April of 2016 where an early version was presented. Special thanks also go to my University of Notre Dame Ph.D. student, Kevin Schieman, for very constructive, critical feedback.