„Civitas. Studia z filozofii polityki" 2018, nr 23
ISSN 1428-2631
DOI 10.35757/CIV.2018.23.10

Michał Klincewicz

Autonomous weapon systems, asymmetrical warfare, and myths[1]

Introduction

Intelligence embodied in inanimate matter is not a new idea. Many creation myths, including those in the Judeo-Christian tradition, involve divine power that grants intelligence to clay and mud. It is also not a particularly new idea that people may, through magic or science, embody intelligence in inanimate matter themselves. This idea is often presented as part of a warning about the dangers of

Michał Klincewicz (ORCID 0000-0003-2354-197X) – Currently Assistant Professor in Tilburg University in the Department of Cognitive Science and Artificial Intelligence and Assistant Professor in Jagiellonian University in the Department of Cognitive Science, Institute of Philosophy. Past post-doctoral researcher in the Berlin School of Mind and Brain, he obtained a PhD in philosophy in 2013 from the City University of New York.

[1] I am grateful to the anonymous reviewers of this journal for helpful and challenging comments. I would also like to thank Marek Pokropski; this paper is at least in part a response to the challenges he aired during a seminar meeting in the Polish Academy of Science in 2015. An earlier version of arguments in this paper were presented as a part of Tomasz Żuradzki's online Filozofia w Praktyce (Philosophy in Practice) project in 2016. An earlier version of this paper was also presented at the Philosophy of Risk seminar organized by Sven Nyholm as a part of The Dutch Research School of Philosophy (OZSW) in the Technical University of Eindhoven. I am grateful to all participants of that meeting for their useful feedback. Subsequent work on the paper was financed by the Polish National Science Centre (NCN) SONATA 9 Grant, PSP: K/PBD/000139 under decision UMO-2015/17/D/HS1/01705.

technology: there are the golem of Jewish folklore, Mary Shelly's *Frankenstein*, Fritz Lang's *Metropolis*, the *Terminator* movies, and numerous other science-fiction novels, movies, comic books, and so on, with a similar theme. What is relatively new in this narrative is we now have the means to embody what may pass for intelligence in otherwise inanimate matter with the help of artificial intelligence (AI) research.

Military applications of this research in sophisticated lethal autonomous weapon systems (AWS) have been particularly visible in recent popular media, scholarship, and even policy debates. Autonomous devices that can kill people are not a completely new technology. Traps, mines, self-directing projectiles, and missiles that can select and engage on their own have all been used in warfare. However, there are scholars that would disagree with this application of "autonomous" and urge that such weapons are merely semi-autonomous[2]. AWS coupled with AI may present a special sort of autonomy that comes with their ability to reason and learn in a way that is qualitatively similar to the way that humans reason and learn. Still, it remains unclear what a claim about autonomy in AWS amounts to beyond their ability to select, target, and fire a weapon. "Autonomy" is either a legal or philosophical term of art, meant to capture something about complex human behaviour in the context of some domain-specific theory. How to operationalise human autonomy is controversial, which means there are no settled criteria to reference to verify whether a computer system or even in a non-human animal is autonomous.

For the purpose of this article, AWS are defined as computer systems that select, target, and engage targets without human control, and can eventually become replacements for humans in the battlefield. The main assumption of this definition is that satisfying both criteria can at the present time be done only with the help of sophisticated AI, including, but not limited to symbolic

[2] H.M. Roff and D. Danks: *Trust but Verify: The difficulty of trusting autonomous weapons systems*, „Journal of Military Ethics" 2018, t. 17, nr 1, s. 3.

systems and machine learning. Among these systems we can count autonomous drones, autonomous tanks, autonomous submarines, and other autonomous mobile systems that can function without any guidance from human operators but can behave as if they had human operators.

What is different about this category of AWS compared to other autonomous weapons, such as missiles, mines, and so on, is the potential precision and lethality with which they can autonomously (in a weak sense) select and engage without human supervision. With a few exceptions that will be discussed later, no other technologies have been able to achieve that kind of autonomy coupled with that level of precision and lethality. As the cost of their development and deployment decreases, such systems become natural candidates for replacing human soldiers. This presents a qualitative difference from other autonomous weapons, such as mines and missiles, which cannot play that role. Given this, the remainder of this paper does not consider (weakly) autonomous weapons such as autonomous stationary gun turrets, missile defence systems, and any type of lingering munitions, such as mines, autonomous missiles, and traps to be its subject. These weapons are not autonomous in the relevant sense. Furthermore, the issue of the ontological category for AWS, as autonomous in the sense of being able to reason or learn, is not addressed here at all.

Many dangerous weapons, such as anti-infantry mines, flechette bombs, or biological weapons, have been banned or restricted by international treaties. Special sessions of the United Nations on AWS resulted in debate, but no outright rejections, limits, or bans. It is not difficult to find reasons why AWS have not been banned. AWS offer many political, tactical, and strategic benefits at little apparent cost. First, targeting in AWS is likely to be more accurate, more effective, and their performance on the battlefield would likely outmatch that of their human adversaries. Second, AWS are not people. This means that AWS are not influenced by stress, fatigue, or pain; they will not intentionally kill civilians; they will not disobey

orders; and they will not come home from war with post-traumatic stress disorder, physical injury, or die in battle. Consequently, one may argue, AWS will likely limit or eliminate many political and social problems associated with war, all while delivering an unprecedented level of military force.

The aim of this paper is to make the case that a future world with AWS is more likely to be closer to the grim scenarios from popular science fiction than to the rosy picture of wars fought by virtuous machine soldiers. The key step in the argument to that conclusion are predictions about AWS being taken over with electronic and/or programming means, in other words, being hacked. Such predictions are typically thought to channel fears that drove all the myths about intelligence embodied in matter. One of these is the idea that the technology can get out of control and ultimately lead to horrific consequences, as is the case in Mary Shelley's classic *Frankenstein*. Given this, predictions about killer robots are sometimes dismissed as science-fiction fear-mongering. On the other hands, unlike *Frankenstein* AWS are not a myth but matters of fact, so the fears that they give rise to should be taken seriously. Myths about technology running amok can also be helpful in generating possible scenarios where AWS run amok by being hacked.

The next section of the paper defends the idea that an assessment of the likely consequences of AWS being hacked in the near future can be made by making well-motivated analogies to other weapons systems. The paper proceeds by considering several such analogies and ultimately offers an argument that nuclear weapons and their effect on the development of modern asymmetrical warfare are the best analogy to the introduction of AWS. The final section focuses on this analogy and offers speculations about the likely consequences of AWS being hacked. These speculations tacitly draw on myths and tropes about technology and AI from popular fiction, such as *Frankenstein*, to project a convincing model of the risks and benefits of AWS deployment.

Hacked AWS

Should we welcome ever more sophisticated autonomous weapon systems (AWS) into the armed forces? Ronald Arkin, who has been pioneering ethical military robotics for decades, is among those that have said "yes"[3]. There are good reasons supporting this answer. As already mentioned, AWS are not influenced by stress, fatigue, or pain; they will not intentionally kill civilians; they will not disobey orders; and they will not come home from war with post--traumatic stress disorder. Some scholars even argue that they are the best way forward in eliminating human involvement in war[4]. If all this is right, deploying AWS is not only morally permissible, but perhaps also morally imperative. Their introduction will lessen the morally abhorrent consequences of war. This is perhaps also why it is difficult to convince policy makers to ban them.

While AWS could potentially replace human soldiers in the field and do a better job at soldiering, they also introduce risks that undermine any moral imperative or moral permissibility to deploy them. But the idea of virtuous AWS comes with a trade-off. In order to be sophisticated enough to be better than soldiers, AWS will have to be so complex that they will be hack-able[5]. Software complexity inevitably leads to bugs, that is, errors in the logic of a program that typically manifest themselves only in very specific circumstances. Bugs, on the other hand, are inherent vulnerabilities that a hacker or a hacker team look to exploit. Unfortunately, there is no perfect certainty that bugs and with them affordances for hacking can be eliminated in complex software. This creates a situation in which we can be certain that the extremely complex software that is likely to operate AWS will simultaneously make them eminently hack-able.

[3] R.C. Arkin: *The Case for Ethical Autonomy in Unmanned Systems*, „Journal of Military Ethics" 2010, t. 9, nr 4, s. 334.

[4] S. Umbrello, P. Torres, and A.F. De Bellis: *The future of war: could lethal autonomous weapons make conflict more ethical?* „AI & Society" 2019, s. 2.

[5] M. Klincewicz: *Autonomous Weapon Systems, The Frame Problem, and Computer Security*, „Journal of Military Ethics" 2015, t. 14, nr 2, s. 169.

What is not clear, however, is that hacked AWS present a serious problem.

The first way to undermine the claim that hacked AWS are indeed problematic is to point out that it assumes that consequences are the relevant measure of what makes something permissible or imperative. A nonconsequentialist moral theorist would likely disagree with this approach. This is not in itself a fatal objection, since consequentialism can be defended, but it may weaken worries about hacked AWS, since it shows them to be hostage to a particular normative theory being true.

This objection can be overcome in two ways. The first is to deny that the notion of the consequences responsible for worries about hacked AWS is the same as the one at work in consequentialist normative theories. For these theories, the kind of consequences that matter are amounts of pleasure or harm, satisfaction or dissatisfaction, or simply utility. If we interpret "consequences" as whatever consequences that may be, but need not be, morally relevant, then we remain neutral about normative theory. Non--consequentialist theories espouse this wider notion of consequence, too.

The second and related way to resist this objection is to point out that the reason why it is presumably permissible or perhaps even imperative to deploy AWS in battle is an elimination of certain kinds of consequences of battle and war involving human soldiers. AWS will not act outside the chain of command, will not get post--traumatic stress, etc., so, the argument goes, their deployment will eliminate sources of harm/dissatisfaction/loss of utility that are consequences of deploying human soldiers. If what matters to the permissibility of AWS deployment is some calculus of consequences, then what matters in undermining the permissibility of AWS deployment should also be such a calculus. It has to be shown that consequences of AWS deployment are in some relevant way more problematic than those associated with human soldiers. Hacking is on its own not necessarily problematic and, furthermore, human

soldiers can also be hacked, metaphorically speaking, by religions, ideologies, and brainwashing.

It is not at all clear how we could argue with any degree of confidence that hacked AWS will lead to consequences that are more problematic than those that may be caused by human soldiers. One reason to believe that we cannot assess the risks posed by AWS deployment comes from their technological novelty. Speculations about risk typically involve predictions based on past performance. Past performance licences probabilistic inferences that can use more or less sophisticated mathematical methods to give a measure of risk. This option is not available for AWS deployment, because there is not enough data about their performance to form reliable probabilistic inferences. AWS are too novel technologically. So, what we are left with is uncertainty, not risk.

This puts pressure on the claim that we can ever come to a reasonably justified conclusion regarding the use of AWS and their potential risks. We may be in an epistemic situation not unlike people in the 19th century at the precipice of electricity in the home[6]. At that time, doomsday scenarios based on accidents were common. As we know, electricity in the home led to significant improvements instead. To avoid repeating such an error with AWS, we may want to refrain from drawing far-reaching conclusions about their positive or negative impact on society.

On the other hand, there are good reasons to believe that today we are in a significantly better epistemic situation with respect to AWS than people in the 19th century with respect to electricity in the home. We are not in a complete state of uncertainty when it comes to AWS. Just as we can look to the impact of the landline telephone to speculate about the impact of mobile phones[7], we can find an appropriate basis for speculation about the impact of AWS by

[6] G. Gooday: *Domesticating Electricity. Technology, Uncertainty and Gender, 1880–1914*. Routledge, London 2015.

[7] A. Lasen: *History Repeating? A Comparison of the Launch and Uses of Fixed and Mobile Phones*, w: L. Hamill, A. Lasen (red.): *Mobile World. Past, Present and Future*, Springer, London 2005, s. 30.

looking at the impact of similar weapons. The social and moral risks associated with the introduction of new kinds of weapons can be successfully assessed and compared with analogies to sufficiently similar devices introduced in the past.

There are different forms of analogical reasoning, but its basic structure is well understood. Analogical reasoning is a type of inference that depends on a shared set of elements that belong to both the source domain (S) and the target domain (T). For example, take an object *F*, with property *p* and another object *G*. If G is sufficiently like F, in that (S) and (T) overlap, then this can be the basis of the following line of reasoning:

1) F is p (S)
2) G is like F (T)
3) So, F is q (where q is a property of G)

Analogical arguments can be used as probabilistic sources of justification, if the elements in common between S and T are relevant. For example, if we know that seas are salty then we can conclude by analogy that oceans are similarly salty. The key to the inference is the amount of relevantly similar characteristics of seas and oceans, so in domains (S) and (T).

Strong analogical arguments allow probabilistic conclusions about novel cases based on knowledge we already have. In the case of making predictions about the future consequences of deploying new weapon technologies, (T) may include a selection of relevantly similar weapons introduced in the past that we know about. Based on that, we may obtain strong analogies to the weapon under consideration – in our case AWS – and then make some probabilistic inferences. This can yield an analysis of risk beyond uncertainties.

Assessment using analogies

One place to start looking for the relevant properties for (T) may be in technologies that radically changed the nature of war. The first

such example is the crossbow[8]. Predictions regarding the crossbow, which rendered most armour at the time useless, included an end to all wars and we know how that turned out, so not just any historically important weapon would do.

A better analogy then may be the machine gun. Its use during World War I, which occurred at a time when offensives were thought to give one a tactical advantage, resulted in massive casualties and limited the use of cavalry and concentrated infantry formations[9]. The consequence of its mass introduction was the invention of a new way of fighting a land war.

The possibility of massive casualties in pitched battles is unlikely to play out with the introduction of AWS. The dramatic consequences of introducing machine guns were at least in part the result of the anachronistic military tactics of the time, which actually projected high levels of casualties, just not *that* high. The tactics used in the 19th century or at the beginning of World War I are out of place on the contemporary battlefield. Soldiers do not assault fortified trenches in large numbers, running into machine gun fire. Given this context, it is highly unlikely that any major human force will engage in all-out battles with AWS. The relevantly similar historically introduced weapon for (T) cannot be the machine gun.

The fact that AWS technology can deliver violence at a distance makes it similar to crossbows or machine guns. But there are several other properties that make AWS qualitatively different from them. Most importantly, AWS can deliver a great deal of lethal force without any direct threat to their user. The only historically introduced autonomous weapon that is similarly lethal at a distance is the nuclear fission device coupled with a delivery system, such as a ballistic missile. Analysis of the risks of AWS deployment could therefore be based on an analogy with them.

[8] B. Brodie, F.M. Brodie: *From Crossbow to H-bomb*, Indiana University Press, Bloomington 1973, p. 37.

[9] S. Van Evera: *The Cult of the Offensive and the Origins of the First World War*, „International Security" 1984, t. 9, nr 1, s. 59.

The introduction of nuclear weapons with delivery systems to military arsenals eventually led to the doctrine of Mutually Assured Destruction (MAD). MAD has two principles: 1) ensure that a response to a nuclear attack is overwhelming and 2) ensure that a response to a nuclear attack focuses on population centres as well as military installations or only on population centres. These principles are either explicitly or implicitly a part of the nuclear military doctrines of the United States, Russia, and France. MAD was a radical departure from typical war strategy since it projects a confrontation with potentially catastrophic consequences for all of life on earth, not just belligerents in a conflict. The main motivation for this apocalyptic approach is deterrence of a nuclear first strike[10]. This deterrence is potent because the negative consequences of a nuclear war far outweigh the benefits anyone may hope to gain by starting it.

Whatever we may think of that logic, MAD contributed to half a century of relative peace where no major nuclear power went to war with another. This suggests that a similar prediction is warranted about the introduction of AWS technology. We can express this prediction in an analogical argument with the following form:

4. AWS are very lethal at a distance (S)

5. Nuclear weapons are very lethal at a distance and their introduction resulted in the MAD strategic doctrine (T)

6. So, the introduction of AWS will likely result in a MAD-like strategic doctrine

If armies of the world possess large amounts of AWS in the future, then indeed a possible long-term outcome may be more peace and stability, supported by something like the doctrine of MAD.

However, this analogy to MAD still fails to hit the mark. Deployment of AWS technology is not likely to be so catastrophic as to destroy life on earth. So, even if the development of AWS results in a MAD-like strategic doctrine, this would be for different

[10] C.W. Morris: *A Contractarian Defense of Nuclear Deterrence*, „Ethics" 1985, t. 95, nr 3, s. 484.

reasons than those that drove the Cold War-era powers to a stand--off. In addition, 4–6 falls short of supporting the claim that AWS will be more problematic than human soldiers. In fact, what it demonstrates is that the list of possible positive outcomes of AWS deployment could be supplemented with a prediction of more peace and stability.

Even if we accept the analogy contained in 4–6, there are other historically well-known consequences that followed from MAD, which lead us directly to a much better analysis. To get to it, we need to focus on how war was waged by nuclear powers during the Cold War and since then. First, typical post-MAD military conflicts did not involve major nuclear powers directly, but indirectly, *via* proxy-wars. Proxy-wars involved third-parties that had tacit or explicit financial, political, and military support from the major powers, which could thereby compete against each other by proxy, without the direct danger of nuclear engagement[11]. Secondly, if nuclear powers engaged in direct combat at all, they usually did so with opponents of disproportionately less military capability and strength. In short, major military powers of the MAD era engaged mostly in asymmetrical wars, such as those in Vietnam and in Afghanistan[12].

The most relevant feature of asymmetrical warfare is the way in which it is typically carried out by the asymmetric adversary:

> [The asymmetric adversary] will often conduct strikes at the lower, tactical level in the hope that they can produce enormous impact at the much higher, strategic level: bombs at one point designed to engineer a change of government policy at another, a hacker attack on one computer designed to shut down a whole economy, the downing of one aircraft to stop a whole bombing campaign, the disabling of one warship to stop a whole armada,

[11] E. Melander *et al.*: *Are 'New Wars' More Atrocious? Battle Severity, Civilians Killed and Forced Migration Before and After the End of the Cold War*, „European Journal of International Relations" 2009, t. 15, nr 3, s. 508, 531.

[12] M.N. Schmitt: *Asymmetrical Warfare and International Humanitarian Law*, w: W. Heintschel von Heinegg, V. Epping (red.): *International Humanitarian Law Facing New Challenges*, Springer, Berlin 2007, s. 2.

> the killing of a few troops to engineer a general 'pull-out', and the dragging out of wars so that interest in outcomes and the will to win are lost[13].

Small-scale tactical strikes that have big-scale strategic consequences are a part of every conflict. What is special about them in the context of asymmetrical warfare is that they are the main and sometimes only strategy for the asymmetrical adversary to take. Asymmetrical tactics have become the preferred tactics of non-state actors but have also been adopted as an option for states, an example being the recent Russian involvement in Ukraine.

Perhaps the best-known successful use of asymmetrical tactics are the terrorist attacks in the United States on September 11th 2001[14]. In that attack, a group of highly motivated terrorists hijacked four commercial airplanes and crashed three of them into strategically important sites: the World Trade Center in New York City and the Pentagon building in Washington, DC. The remaining plane crashed in a field in Pennsylvania during an attempt by passengers to regain control of the plane. Subsequently, the United States engaged in long and costly wars that, some argue, achieved many of the strategic aims that the perpetrators of those attacks were aiming to achieve[15].

The weapon of choice for the modern asymmetric adversary is an inexpensive solution that leverages the inherent weaknesses of complex systems. Hijacking a plane, kidnapping a journalist, or, as Thornton suggests, hacking a strategically important computer system. This last possibility of hacking is particularly important in the context of AWS given the inherent complexity of their software. This complexity along with the need to maintain communication with AWS during a mission makes them eminently hack-able. AWS hacking can be characterised as a small-scale tactical strike with the potential for big-scale strategic consequences.

---

[13] R. Thornton: *Asymmetric Warfare. Threat and Response in the 21st Century*, Polity, Cambridge UK 2007, s. 22.

[14] I am grateful to Sven Nyholm for suggesting this example to me.

[15] B.H. Fishman: *The Master Plan. ISIS, al-Qaeda, and the Jihadi Strategy for Final Victory*, Yale University Press, New Haven 2016, s. 64.

To sum up, modern conflicts create conditions in which tactics adopted from asymmetrical warfare will inevitably exploit the sort of vulnerabilities that AWS happen to possess. These conditions are an indirect result of the introduction massive nuclear arsenals and the MAD doctrines. Support for the claim that hacked AWS will introduce affordances for consequences that are more problematic than whatever human soldiers do comes from the very nature of asymmetrical tactics, which aim at maximal strategic impact. Any AWS deployed will present an ideal opportunity for a tactical hacking strike with high impact strategic consequences.

The remaining issue is whether such consequences would indeed have significant strategic import and, with it, morally abhorrent consequences. To demonstrate this some speculation needs to happen about what an asymmetrical adversary or an adversary that relies on tactics adopted from asymmetrical warfare may want to do with a hacked AWS to achieve strategic success. The myths and stories of intelligent machines running amok are a particularly fruitful source of material for such speculation.

Consequences of hacked AWS

a) Hacked AWS unleashed on a civilian population. Imagine a scenario in which AWS are hacked while not yet deployed, perhaps in a barracks or at a base, and then made to fire indiscriminately on a defenceless civilian population outside the conflict zone. The terror caused by the situation would count as a tactical victory, but also as a strategic success. An indiscriminate attack of killer robots on an unsuspecting civilian population would be treated in the media as a harbinger of science-fiction terminators and would receive massive news coverage. Its likely consequence would be international condemnation and public outrage and likely success at the strategic level.

b) Hacked AWS unleashed on a military installation. Imagine a scenario in which AWS are hacked during deployment and then made to fire on friendly units or destroy their base of operation, including infrastructure. This would greatly undermine the morale of the human force and damage whatever trust human soldiers may have had in AWS technology, with all the possible negative consequences that that would have on future tactical-level engagements. The consequence is likely success at the strategic level.

c) Hacked AWS unleashed in false-flag attacks. Imagine a scenario in which AWS are hacked and made to attack a civilian population within a conflict zone, perhaps during a time in which a campaign to win "hearts and minds" is under way. The backlash among the civilian population and media coverage of such an attack would likely paint it as a war crime. In defence, the owner of an AWS would have to argue convincingly that they are, in fact, not responsible for the tragedy and that it was due to a hacker. Whatever happens in that regard, this would likely end up being a high impact event at the strategic level.

d) Hacked AWS used as a weapon of terror. Imagine a scenario in which AWS are hacked and not used to attack immediately, but the possibility of having them attack is communicated through mass media, maybe with concrete demands of tactical significance. A situation in which AWS could be unleashed on civilians would not only bring the kind of media attention that terrorists aim for, but would likely result in general distrust of AWS. Again, the consequence is likely success at the strategic level.

None of the risks presented in these speculations are present in the same scale, if at all, with human soldiers or with weapons under total human control. Even "hacked" human soldiers do not present the same level of potential lethality and media coverage as hacked AWS. AWS add a level of potential lethality and media visibility that may not be possible with any other asymmetrical strike. Given this, asymmetric adversaries as well as state actors that deploy

asymmetric tactics will do their best to exploit AWS for maximum media coverage and carnage. This very affordance is the main reason why we should expect deployment and eventual hacking of AWS to lead to morally abhorrent consequences that far surpass those that may be caused by human soldiers. Maximally morally abhorrent consequences are what the people that eventually hack AWS will aim for.

e) Another likely consequence of hacked AWS is the uncontrolled acquisition and proliferation of AWS technology to parties that would use them to their own ends. This would mean that regimes that oppress and non-state actors that do not care about human rights would have a new way to deliver unprecedented levels of violence at the press of a button. AWS will not refuse to fire on civilians or protestors. Given this likely outcome of AWS proliferation, the dystopian scenario of a human rebellion against terminator robots would likely take the form of a struggle between a beleaguered population and a human tyrant that controls AWS. In this context the hackability of AWS may be a welcome consequence, as it is likely to be the only way to even the odds against an oppressive regime. In any other case, however, the hackability of AWS presents affordances for disaster, as outlined in a)–d).

Given the high cost of research and development in AI and robotics, there is some hope that AWS could be made difficult to acquire. Regardless, scenario e) will inevitably come to be a reality sooner or later as the technologies that are at the heart of lethal AWS will become cheaper and more accessible. The only instruments that could prevent this are international treaties like those that govern the use and sale of weapons of mass destruction and land mines, so the very instruments that have so far failed to deliver bans and controls on AWS technology. In addition to preventing the spread of lethal AWS technologies to states that may use them against their populations, such treaties would have a preventative function in helping democratic and free states avoid scenarios a)–d).

In sum, AWS are not influenced by stress, fatigue or pain, which are all bad consequences for human soldiers. AWS will also not intentionally kill civilians, which is certainly a way to avoid a morally abhorrent consequence. AWS will also not come home from war with post-traumatic stress disorder and they will not disobey orders. We can also imagine them becoming more ethical on the battlefield than humans could ever be. On the other hand, AWS can be the opposite of all these good things by being hacked. The grim scenarios a)–d) listed above are a potent counterweight to whatever positives AWS bring to the battlefield. In conclusion, the risks presented by AWS are far more morally abhorrent than those presented by human soldiers.

We can predict that AWS will be able to select targets and engage them with greater accuracy using whatever weapons are at the disposal of human soldiers. But they don't have to be discriminate, if that is not the aim of their deployment. AWS can and will be deployed to carry out atrocities. This likely abhorrent consequence is exacerbated by the tactical advantages that AWS bring to the field. AWS can function in conditions that human soldiers would find difficult or impossible, such as in space, underwater, in environments with high levels of radiation or those affected by chemical or biological agents. AWS will also be able to use technology that is currently unusable by human soldiers without concern for their own survival. Most importantly, AWS will be able to do all this on their own, with human guidance only being optional, and potentially with a mere press of the button.

If the analogy to asymmetrical tactics is strong, the vulnerability of AWS to hacking will inevitably lead to unprecedented levels of abhorrent consequences. We can assess these potential moral risks and confidently conclude that they far outweigh those presented by human soldiers. In consequence, it is not morally permissible to deploy them. Human soldiers, whatever their foibles, are morally better at war than weaponised artificial intelligence embodied in robots. The main reason for that is human soldiers cannot, at the

present time, be hacked except metaphorically and that human soldiers can refuse to follow an illegal order.

Abstract

Predictions about autonomous weapon systems (AWS) are typically thought to channel fears that drove all the myths about intelligence embodied in matter. One of these is the idea that the technology can get out of control and ultimately lead to horrific consequences, as is the case in Mary Shelley's classic Frankenstein. Given this, predictions about AWS are sometimes dismissed as science-fiction fear-mongering. This paper considers several analogies between AWS and other weapon systems and ultimately offers an argument that nuclear weapons and their effect on the development of modern asymmetrical warfare are the best analogy to the introduction of AWS. The final section focuses on this analogy and offers speculations about the likely consequences of AWS being hacked. These speculations tacitly draw on myths and tropes about technology and AI from popular fiction, such as Frankenstein, to project a convincing model of the risks and benefits of AWS deployment.

Keywords: ethics of artificial intelligence, autonomous weapon systems, war, asymmetrical warfare, hacking.