LUND UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Are We In Control?

Challenges to Governance of Autonomous Weapons Systems in the Age of Artificial Intelligence

Programme: *MSc Global Studies.*

Course: *SIMZ-2025: MSc Global Studies Thesis.*

Student: *Helena Simone Flockhart.*

Supervisor: *Thomas Hickmann.*

Examination Date: *7th January, 2025.*

Abstract

The rise of AI is leading to paradigm shifts in every sector of modern society, including the military. One concerning change is the increase in the development, testing, and deployment of autonomous weapons systems (AWS), which utilise modern AI without adequate controlling regulation or governance. This thesis explores what it means to have control of AWS and whether it is possible to effectively govern AWS using our current normative framework of weapons governance. Using the theoretical framework of the Control Problem of AI alignment, I argue that it is impossible to ensure control over an autonomous AI system, and therefore, it is impossible to ensure meaningful human control of AWS. This problem is further compounded by the normative tradition of weapons governance, which has a shaky historical track record for governing new weapons technology during times of instability in world order, as norms in weapons governance tend to emerge through unregulated practice rather than through deliberation. I argue that due to the high risks of AWS slipping out of human control, AWS use cannot be governed and therefore should be banned preemptively.

Keywords: Autonomous weapons systems (AWS), artificial intelligence (AI), policy, military, meaningful human control, norms, governance.

Contents

Introduction

In 1956, a group of mathematicians at Dartmouth College embarked on a summer research project to see if it were possible to simulate learning in a machine. They hypothesised that given specific instructions and data, a computer would be able to use language, problem solve, theorise, conceptualise, and improve without interference, and that this could be achieved over the course of the summer research project. Although their timeline was perhaps just a tad optimistic, the ideas presented at the Dartmouth Conference that summer went on to form the foundations of one of the most influential scientific fields today; namely, the field of Artificial Intelligence[1]. In the nearly 70 years since the Dartmouth Conference, AI has developed rapidly and become an integral part of nearly every aspect of our daily lives. From lifestyle to entertainment to transportation to healthcare, humanity is at the most "plugged in" we have ever been, with practically no part of our modern lives entirely disconnected from some form of AI. Even the words you are currently reading were written in conjunction with AI, as my Word processor is programmed with predictive text which, most of the time, can accurately predict my words before I type them. Yet, although this mass implementation of AI has undoubtedly brought about a wave of benefits for humanity, it has also presented us with worrying risks. The scale and scope of AI advancement in recent years has led to intensive debates between those who welcome the ease and convenience that AI brings into our lives, and those who fear that we are sacrificing too much of our control to it and imbuing it with far too much agency and trust. Questions of efficacy and safety are asked but with little consensus on what the answers might be, and many fear that the development of AI is simply too fast for us to keep up with and on a trajectory that inevitably will cause us harm[2]. One sector where this debate is hotly contested is in the development of AI within weapons technology.

When asked to think about a weapon with artificial intelligence, many will probably conjure up images of a leather-clad, shotgun-wielding Arnold Schwarzenegger in a menacing pursuit to wipe out humanity. To those kept awake at night by such terrifying visions, I have good news: Evil intelligent machines intent on taking over the world and enslaving the human race remain

[1] Russell, *Human Compatible,* 1-12.

[2] Ibid.

comfortably within the realms of science fiction. What is however very much part of science fact is the existence of autonomous weapons systems (AWS), which can operate partially or entirely without the need for direct human intervention[3]. Although autonomous weapons have been around for quite some time, the implementation of AI into AWS is rapidly advancing the levels of autonomy and agency in such weapons, and therefore presents a major paradigm shift for warfare as it is potentially minimising the need for humans on the battlefield[4]. Although there is an argument to be made that this is a positive change as it removes combatants from direct danger, the use of AWS brings about a litany of pressing considerations of safety and control which are far from being adequately addressed. This issue is particularly prevalent in the realm of weapons governance, as existing laws and norms of regulating weaponry are reliant on the assumption that humans are in full control of weapons and therefore able to apply their judgment, ethics, and responsibility when deploying certain weaponry to ensure that the laws of war are adhered to[5]. However, introducing AI into AWS technology is making such governance increasingly difficult as many experts fear that in doing so, we are relinquishing control of such weapons[6]. Therefore, current and future AWS display a unique challenge to global weapons governance because of the nature of AI, which some argue is inherently ungovernable in its current trajectory[7]. This merging of AI and weapons development has therefore proven to be a veritable minefield of legal, ethical, and practical challenges, which could be culminating in a crisis of weapons governance which we are completely unprepared for. The prevailing policy norm regarding AWS is a vague statement that there will always be "meaningful human control"[8]. However, there is little agreement on what exactly "meaningful human control" is in practice, how it can be applied to an intelligent and autonomous machine, nor how it can be enforced in practice[9]. When considering the rapid exponential trajectory of AI development, we are facing a very real crisis of control wherein humans are creating weaponised machines that can carry out objectives independently of us, with no real understanding of how much control we actually have, and no effective measures of intervention if things go wrong.

[3] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[4] Scharre, *Army of None*, 1-367.
[5] Ferl, *Imagining Meaningful Human Control*, 139-155.
[6] Russell, *Human Compatible,* 110-113.
[7] Russell, *Human Compatible,*1-295.
[8] Ferl, *Imagining Meaningful Human Control*, 139-155.
[9] Scharre, *Army of None*, 1-367.

In this thesis, I will explore the question of what it means to have control of AWS in the era of AI and whether it is ever possible to effectively govern such technology using our current normative approach to weapons governance. Using the theoretical framework of the Control Problem of AI and practice theory of governance, I will analyse the technical and practical implications of controlling AI in military technology and explore the problems of governing such weaponry under the current normative trends of weapons governance. I argue that current attempts to inform effective governance of AWS is entirely ineffective as current policy is out of touch with the technical and practical realities of AI in AWS. Continuing along this normative path of weapons governance will allow AWS development and deployment to remain largely unregulated in practice as autonomous and agentic developments in AI and a policy of "meaningful human control" are fundamentally incompatible, and therefore ungovernable.

Theory and Methodology

Aim and Theory

This thesis aims to demonstrate why current normative policies of weapons governance are failing concerning the development and deployment of autonomous weapons systems. I argue that there is a fundamental disconnect between those informing policy and those developing the technology of AWS, and this is leading to misinterpretations, deceptive practices, confusion, and useless "box-ticking" from both the political and the scientific side over what such weapons systems are capable of doing, how much control we can realistically have over them, and what their impact may be on global security and stability if this disconnect continues. This thesis will therefore attempt to bridge this gap by looking at the practical and technological difficulties of controlling AI, applying this to the AI used in military technology, specifically AWS, and demonstrate how current policy norms in governing AWS do not show adequate understanding of the practical and technological realities of modern AI. Although rooted in political science, this thesis takes an interdisciplinary approach to governance of AWS, which I argue is necessary to fully understand the risks of developing and deploying such weapons in the real world. The thesis is therefore structured into three sections, each with a distinct disciplinary focus. The first section is an empirical analysis of the current state of AWS governance and technological development. The second section is an in-depth discussion of the practical realities of controlling

AI from a technical perspective and what the implications of increased agency of AI in military technology may have on the future of global security. The final section is a discussion on the problems of taking a normative approach to governance of AWS based on outdated fundamental norms that do not acknowledge the formative power of procedural norms in the creation of new weapons technology. The thesis therefore incorporates both policy analysis methods used in International Relations (IR) and theoretical and technological concepts from AI-focused computer science, as well as drawing from political sociology and communication theory. The main theories used are the Control Problem (or alignment problem), a highly prevalent theoretical framework for ethics and safety of AI as conceptualised by Stuart Russell, and practice theory as developed in the field of IR in the discussion of the formation of fundamental and procedural norms in global weapons governance based on the research of Ingvild Bode and Hendrik Huelss.

Methodology and Empirical Material

As AWS is a highly classified field of military technology, there is little concrete contemporary data available for the general public. Indeed, *officially,* it is very unclear if an autonomous weapon has ever been deployed in fully autonomous mode. Therefore, many of the issues addressed in this thesis refer to known problems of the past, problems of AI in general, and somewhat speculative scenarios for the future. This means that much of the research and argumentation in this thesis is theoretical, and many inferences and comparisons to more readily available data are made. However, there is an abundance of data on AI technology within society in general, as well as a vast history of weapons technology and governance, and therefore this thesis will use a qualitative research design based on theoretical and empirical data on AI and a historical analysis of weapons development and governance. Data gathered is primarily from secondary sources (specifically literature), however data on specific policy is gathered from primary online archival sources when available.

Current State of AWS Governance

The early 21st-century technological boom has led to paradigm shifts in practically every area of society, including weaponry and warfare. Within the field of weapons development, it is now

widely agreed that the next frontier of military advancement lies in further automation and autonomy of weapons systems and other ancillary military technology[10]. Although the development, testing, and deployment of AWS still take up a relatively small portion of military expenditure, major military powers like the US, China, Russia, India, Israel, South Korea, and the UK are investing vast amounts of money, time, and resources into the proliferation of AWS[11], and we are now starting to see evidence that fully autonomous weapons may already be deployed in active combat[12]. However, since the early 2010s, there has been a growing number of prominent voices speaking out against such development (with many critics calling for an outright ban of AWS), citing the potential risks and dangers associated with greater autonomy in weapons and the problems they could bring if developed without proper governance and regulation[13]. The major concern among critics is whether it is ethically and legally permissible to allow weaponised machines to make decisions about killing humans, and if not, how can we ensure that we remain in control of such machines? Although this has sparked an international response to the development and regulation of AWS, there is still a severe lack of coherent international governance needed to inform effective arms control of AWS[14].

This lack of governance is largely due to two crucial issues. First, there is a disconnect between those informing policy and those developing AWS, with the result that current policy is inherently incompatible with the technological and practical realities of modern AWS[15]. Second, current policy uses a normative framework of weapons governance rooted in International Humanitarian Law (IHL) that is unlikely to be effective given the unique nature of AWS compared to previous weapons technology governed under the same framework[16]. Before delving into these technological and practical implications of AWS and whether it can be effectively governed, we must first pinpoint what the current policy norms are, and why they are insufficient to deal with the fast incoming AWS arms race.

Getting AWS on the International Agenda

[10] Scharre, *Army of None*, 1-367.
[11] Haner & Garcia, *The Artificial Intelligence Arms Race*, 331-337.
[12] Williams, *Summary: Autonomous Weapons*, 1-14.
[13] Amoroso & Tamburrini, *Autonomous Weapons Systems and Meaningful Human Control*, 187-194.
[14] Scharre, *Army of None*, 346-367.
[15] Russell, *Banning Lethal Autonomous Weapons*, 60-65.
[16] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.

Although autonomous weapons have existed in some capacity for a long time, the 2010s saw a significant increase in the use of military drone technology, such as unmanned combat aerial vehicles (UCAVs), which allowed human combatants to have a far greater distance from the physical dangers of battle than ever before. While this led to fewer injuries and deaths of combatants (on the side with drones, that is), it sparked a mass debate on the ethics of remote killings in war, whether it can be considered legitimate conduct in warfare, the role of situational awareness in combatants, and many other issues which potentially violate IHL of which members of the armed forces are legally required to conform to[17]. At the same time, the field of robotics and AI began making major breakthroughs, including in autonomous capabilities[18]. Due to the practical and ethical issues raised with UCAV usage, this led to concerns among the scientific community on the implications of greater autonomy in such military systems, and what could potentially happen if human command and operation of weapons systems were made obsolete by autonomous AI — a possibility that was growing ever more likely as autonomy in military technology was at the time entirely unregulated[19]. If this were to happen, it could potentially undermine IHL completely since the law applies to human combatants and not to robots, and therefore, such an autonomous weapon would essentially operate outside the laws of war[20].

This issue came to public attention in 2012 when Human Rights Watch published a report on AWS and the potential implications on IHL, which led to the formation of the Campaign to Stop Killer Robots, a leading coalition NGO committed to banning AWS preemptively[21]. While the campaign as of yet has been unsuccessful in obtaining a ban, it did bring the issue into the public eye, and as of 2014, AWS has been on the public agenda at the United Nations Convention of Certain Conventional Weapons (UN CCW)[22]. By 2016, the issue was further solidified by the formation of the Group of Governmental Experts (GGE), a multilateral open-ended group of experts from UN member states tasked specifically with policy matters regarding AWS. As of today, the GGE is the only international forum dedicated to governing AWS on a global scale

[17] Grut, *The Challenge of Lethal Robotics*, 5-23.
[18] Russell, *Human Compatible,* 62-102.
[19] Amoroso & Tamburrini, *Autonomous Weapons Systems and Meaningful Human Control*, 187-194.
[20] Grut, *The Challenge of Lethal Robotics*, 5-23.
[21] Amoroso & Tamburrini, *Autonomous Weapons Systems and Meaningful Human Control*, 187-194.
[22] Ferl, *Imagining Meaningful Human Control*, 139-155.

through policy recommendations, treaties, and protocols[23]. The GGE is rooted within the normative, legal and ethical principles of IHL, and therefore the majority of policy is determined through this normative lens under the discourse of "meaningful human control"[24].

Meaningful Human Control

With the rising autonomous capabilities of remotely piloted drones and growing concerns about the implications to IHL of removing human combatants from warfare, the US Department of Defense (DoD) made its first reference to autonomous weapons in official policy with the following statement in its 2011 roadmap:

> For the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems[25].

Although quite vague and non-descriptive, this statement was among the first in what would guide the prevailing policy discourse regarding AWS: maintaining human control. In 2012, the DoD Directive 3000.09 was the first comprehensive official government policy document regarding AWS, wherein human control and judgment were the main focus[26]. The directive attempted to ensure in policy that AWS would not use force against a target without explicit human command, would not violate IHL, and would contain certain safeguards that would enable human operators to intervene or disable the system in case of a malfunction[27]. Similar policy documents quickly began emerging from other leading military powers, all with the same focus on maintaining human control of autonomous weapons[28].

Although there are some variations on the exact phrasing, this policy norm is most commonly known as "meaningful human control" (MHC) and has been a benchmark of AWS governance since such weapons have been on the international agenda. Deeply situated within IHL, MHC was first proposed in a policy brief by the not-for-profit organisation Article 36 as a structure of

[23] Ibid.
[24] Ibid.
[25] United States of America: Department of Defence, *Unmanned Systems Integrated Roadmap*, p. 17.
[26] United States of America: Department of Defence, *Directive 3000.09*, 1-15.
[27] Ibid.
[28] Amoroso & Tamburrini, *Autonomous Weapons Systems and Meaningful Human Control*, 187-194.

ethical and legal debate of AWS for delegates at the UN CCW in Geneva in 2013[29]. Since then, MHC has remained one of the primary topics of debate among the actors within the GGE, and it informs the majority of international policy on AWS today. The basic understanding of MHC holds that human control is meaningful only if: 1) a human can act as a failsafe in the event of a system malfunction, 2) a human can be held accountable in case an AWS violates IHL, and 3) the final decision to encroach on another human's life and dignity is taken by a human as an AWS is not considered to have *moral agency*[30]. Essentially, an AWS is expected to adhere to the same laws of war and armed conflict that human combatants are required to uphold. While this definition of MHC seems perfectly reasonable in theory, in practice, it is unfortunately not so easy to implement. What if an AWS is out of range for a failsafe to be activated? If an AWS violates IHL, which human is responsible? — The operator, the software developers, the CEO of the weapon's manufacturer, and countless others could arguably be culpable. How does one determine moral agency, and why can a human possess it while a machine cannot? These issues are constantly being debated by the GGE and other actors invested in the governing of AWS, with no consensus yet reached[31].

Alongside the difficulty in defining MHC is the difficulty in defining AWS itself. While the standard definition of an autonomous weapon as "Any weapon system with autonomy in its critical functions — that is, a weapon system that can select (search for, detect, identify, track or select) and attack (use force against, neutralise, damage or destroy) targets without human intervention.[32]" seems on the surface to be a very clear description, this definition encompasses a wide range of automatic and semi-autonomous weapons which are functionally entirely different from AWS and therefore require a different form of governance.

While MHC may be a good starting point for discussion and subsequently informing policies of weapons governance, it requires the full backing of the legal, political, and technological spheres to be properly defined and clarified into practical terms. Until then, it is not of much use in terms of effective governance of AWS. Unfortunately, however, this backing has not happened.

[29] Article 36, *Structuring Debate on Autonomous Weapons Systems*, 1-3.
[30] Ibid.
[31] Ferl, *Imagining Meaningful Human Control*, 139-155.
[32] Davison, *A Legal Perspective*, p.5.

Meaningless Human Control

Since it was first brought to the international agenda over a decade ago, the topic of controlling AWS has not shifted far from the ethical and legal sphere of international governance. While this is certainly an important forum of discussion and deliberation, the lack of enforceable authority of bodies like the UN CCW and the International Committee of the Red Cross (ICRC) means that the driving force to govern AWS preemptively has stagnated. As noted by Paul Scharre, who took part in writing the seminal 2011 DoD roadmap:

> One of the challenges in current discussions on autonomous weapons is that the push for a ban is being led by NGOs, not states. Only a handful of states said they support a ban, and none of them are major military powers. When viewed in the historical attempts to regulate weapons, this is unusual. Most attempts at restricting weapons have come from great powers[33].

While Scharre is specifically referring to the call to ban AWS, the same major military powers are also relatively silent on how to actually implement MHC in practice[34]. This absence of major military powers from the discussion on governing AWS is concerning since they are ultimately the ones who make the decisions on all matters regarding current and future developments, not NGOs or international forums like the GGE. Another concerning absence from the discussion on MHC are the ones who are actually in charge of developing the revolutionary new weapons technology in question[35]. While multiple very influential companies and individuals in science and technology have spoken up in support of a ban or tighter regulation of AWS (including Apple co-founder Steve Wozniak, Google DeepMind co-founders Demis Hassabis and Mustafa Suleyman, and even the highly controversial tech giant and right-wing wildcard Elon Musk), there are countless more whos technological innovations in AI is very likely being implemented into AWS and ancillary military technology as we speak[36]. I say "likely being implemented" because there is so little transparency and such high confidentiality that we can only really guess at what is happening in the development of AWS, yet evidence suggests that much of the technology used for military applications of targeting and surveillance stem from Silicon Valley

[33] Scharre, *Army of None,* p. 348.
[34] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[35] Ibid.
[36] Slijper, et. al., *Don't be Evil*, 2-58.

tech giants like OpenAI[37]. It is however clear that the individuals and companies developing the technology used in AWS are getting a far bigger say in terms of control than those attempting to govern AWS through the policy level.

There are a multitude of reasons why a policy discourse of MHC is likely to be ineffective in governing weapons technology such as AWS. First, the extent to which human operators can adequately perform these criteria of control often depends on a multitude of complex technological, environmental, and psychological factors. Furthermore, the phrase "meaningful human control" is itself very open to interpretation and therefore, highly contested[38]. What is meant by "meaningful"? Does it refer to the human or to the control? While this may seem like semantic nit-picking, this is important to note since MHC has informed the majority of the discourse used in AWS policy, yet in the legal and political sense, the term has no meaning whatsoever. Such an ambiguous term is therefore rife with problems when used as the leading discourse for informing policy of something as complex as AWS because it allows for a great variation in definitions of autonomy and control and makes it extremely difficult to enforce any policy. Indeed, although all members of the GGE agree that there must always be MHC in the operation of AWS, the last decade has seen significant evidence that this is merely a sentiment which is practically ignored in the development and deployment of AWS[39]. By using vague terminology on the policy level, the technological advancement of AWS seems to have entirely bypassed any governing factors of MHC and instead utilised semantics of the discourse to wave away significant lapses in control. Indeed, throughout this thesis, I will provide multiple examples of AWS that are clearly only under some form of human control by formality, and can likely be made fully autonomous with a simple software update. This has led critics of current attempts to govern AWS to deem the term "meaningful human control" as meaningless[40].

This ineffective governance is due to a failure of policymakers to understand the science of control in AI, why implementing AI into military scenarios is an infinitely complex undertaking, and how catastrophic the consequences could be if mistakes are made[41]. Alongside this critical

[37] Anduril, *Anduril Partners with OpenAI.*
[38] Robbins, *The Many Meanings of Meaningful Human Control*, 1377–88.
[39] Ferl, *Imagining Meaningful Human Control*, 139-155.
[40] Ibid.
[41] Russell, *Banning Lethal Autonomous Weapons*, 60-65.

error is the failure of international normative weapons governance, which, in trying to govern AWS through the normative structures of IHL, is missing the emerging procedural norms that are forming in practices of developing, testing, and deploying AWS outside the reach of global weapons governance[42].

Due to this, the current policy norms of governing AWS are not adequately able to assess the risks of greater autonomy in weapons and have therefore significantly lagged in providing any effective governance. The following chapters of this thesis will show how the policy of MHC is completely ineffective in governing the practical and normative problems of controlling AWS in the context of modern AI and the real-life implications of using military technology in the high-stress environments of active combat. In light of these findings, it is clear that governance of AWS will only be possible if problems of controlling AI are effectively understood by policymakers and very specific regulation on levels of autonomy and types of human intervention are implemented into policy. Whether this is possible, however, is less clear.

What Are AWS and What Do They Mean for the Future of Weapons Development?

In the most basic definition, AWS are weapons which can select and use force against a target without the need for human intervention[43]. However, there is an extensive debate about what exactly constitutes autonomy and therefore which weapons can be classed as an AWS[44]. Therefore, to define autonomy, weapons are usually sorted into three different categories: semi-autonomous, supervised autonomous, and fully autonomous[45]. Weapons are sorted into these three categories by defining to what extent they actively select a target without human intervention — in other words, whether or not there is a human “in”, “on”, or “out of” the loop[46]. Yet, while this may seem like a neat categorisation method, as already discussed, assigning levels of autonomy with MHC is a highly contestable activity. There is great disagreement about what makes a weapon “autonomous” rather than “automatic”, and this has led to a great deal of

[42] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[43] Taddeo & Blanchard, *A Comparative Analysis*, 2-22.
[44] Ibid.
[45] Scharre, *Army of None*, 35-58.
[46] Ibid.

confusion and argument over what constitutes an autonomous weapon under MHC[47]. To make matters worse, the more technologically advanced weapons systems become, the harder this categorisation becomes because although a human may be "on" or "in" the loop, there is little to clarify whether they are pulling the strings or simply pressing "Go". In this chapter, I will demonstrate how weapons are categorised under levels of MHC and why this is both open to deep contestation as well as allowing practice designed to placate policymakers while AI is implemented further into AWS.

The Loop

Depending on who you ask, "the loop" can refer to a multitude of different things. For computer scientists, it refers to reinforcement learning, or the cycle of interaction between humans and algorithms in the operation of automated and autonomous systems, which ultimately shapes machine learning in some modern AI[48]. In military terminology, it refers to the OODA Loop (observe, orient, decide, act), which combat pilots use to outsmart and defeat enemy aircraft[49]. However, in the case of AWS, "the loop" incorporates elements of both definitions to determine how much control a human has over the operations of an AWS in combat[50]. In other words, levels of autonomy of a weapon are determined by establishing if there is a human "in", "on", or "out of" the loop of searching for, detecting, and engaging a target. With a human-in-the-loop system (or semi-AWS), the system can autonomously search for, find, and identify targets, but only a human can actively engage targets[51]. With a human on-the-loop system (supervised AWS), the system can carry out the entire kill chain[52] independently, however, a human operator is supervising in real-time and can intervene if necessary[53]. With a human out-of-the-loop system (fully AWS), the system is entirely independent of human intervention[54]. "The loop" is the terminology most commonly referred to to determine whether an AWS is under MHC, however as I will illustrate in this chapter, this distinction of "a human in/on/out of the loop" is extremely

[47] Taddeo & Blanchard, *A Comparative Analysis*, 2-22.
[48] Bostrom, *Superintelligence*, 226-253.
[49] Scharre, *Army of None*, 35-58.
[50] Ibid.
[51] Ibid.
[52] The military term for the sequence of an attack. There are different types of "kill chains", but they share a general structure of: 1) identification of a target, 2) dispatch of forces, 3) initiating attack, and 4) destruction of target.
[53] Scharre, *Army of None*, 35-58.
[54] Ibid.

difficult to draw in practice, and does not necessarily indicate how much control a human operator actually has in a real-life scenario.

Human-in-the-Loop: Automatic and Semi-Autonomous Weapons

A classic example of an automatic weapon is a land or sea mine. Although it technically operates without human intervention, basic mines operate under a high degree of randomness and pre-determinism. That is to say, a landmine is not so much selecting a target as it is reacting to stimuli based on *pre*-selected criteria (in this case, the weight and size of an object hitting the pressure sensor), and most therefore consider mines and similar weapons as automatic rather than autonomous[55]. However, since the mid-20th century, we have seen the emergence of more complex weapons *systems* which muddy the waters of defining autonomy, and these are usually classed as semi-autonomous weapons systems. Semi-AWS are weapons which can carry out objectives mostly independently, however, they require a human operator to take action to engage a target. One of the earliest successes in this type of weapons technology was the Mk-24 "Fido" torpedo, which during the Second World War was used by the Allies to locate and sink German submarines[56]. The Fido was a much more technologically sophisticated weapon than those that came before it because, once deployed from the air, it used acoustic hydrophones to listen for, locate, and track submerged enemy submarines independently of a human operator. This was one of the first steps towards autonomous functionality in weapons as this was a weapon which could interact and "learn" from the environment after it had been deployed and use this information to select a specific target[57]. Weapons such as this are considered semi-autonomous because, although they do have the ability to select and engage targets independently to some degree, they are guided by human-analysed intelligence and are incapable of operating entirely without a human-in-the-loop[58]. The distinction between semi-autonomous and fully autonomous is relatively easy to make with this early technology because there is no doubt that the systems required human interaction to function fully. This distinction, however, is much more difficult with modern weapons systems due to the introduction of modern computer technology.

---

[55] Taddeo & Blanchard, *A Comparative Analysis*, 2-22.
[56] Work, *A Short History of Weapon Systems with Autonomous Functionalities*, 5-7.
[57] Ibid.
[58] Scharre, *Army of None*, 35-58.

Human-on-the-Loop: Supervised Autonomous Weapons Systems

Many of the weapons officially classed as semi-autonomous now have a human-in-the-loop only by formality, as many modern weapons are functionally capable of carrying out entire kill chains without a human present[59]. This has led to a further classification of "human-on-the-loop", as a human is not required in the operation of the weapon, but instead is merely supervising and ready to step in if needed. For instance, consider the infamous Samsung SGR-A1 sentry system that guards the South Korean side of the DMZ alongside human soldiers[60]. Using infrared thermal cameras, laser guidance, and voice recognition, the system can identify, track, and engage potential enemies (in this case, any human in its vicinity who cannot provide a verbal access code) independently of a human operator. If a target is identified, the system has a non-lethal arsenal of an alarm and rubber bullets, and a lethal arsenal of a light machine gun and a grenade launcher. Samsung and the South Korean government are naturally extremely secretive about the operation of the system, however, they insist that it will only ever utilise its non-lethal arsenal unless a human operator authorises lethal force[61]. Indeed, when the system was launched in 2007, chief engineer Myung Ho Yoo famously stated that "the ultimate decision about shooting should be made by a human, not the robot"[62], however, the use of the word "should" confirmed for many sceptics suspicions that the system almost certainly does have the capability of operating fully autonomously, and that its supervised autonomous state is merely an optional setting. As is the case with most complex weapons systems, however, we have no way of knowing for sure how much programming and tinkering it would need to make the SGR-1 fully autonomous in truth, as the South Koreans are not in a hurry to allow third-party inspection of one of their most complex and secretive defences, specifically as it is used to guard one of the highest security risk areas in the world[63].

It is, however important to note that many supervised AWS are exclusively defensive weapons. The SGR-1 sentry, the Iron Dome missile defence system, and the Phalanx CIWS naval defence system are all weapons systems that operate with a high degree of autonomy due to the

[59] Scharre, *Autonomous Weapons and Operational Risk*, 1-54.
[60] Scharre, *Army of None*, 35-58.
[61] Ibid.
[62] Myung Ho Yoo, quoted in: Scharre, *Army of None*, p. 105.
[63] Haner & Garcia, *The Artificial Intelligence Arms Race*, 331-337.

importance of rapid response that far outmatches human response times[64]. For instance, the Iron Dome system is an Israeli air defence system designed to identify the trajectory of rockets and other aerial munitions and intercept and destroy only those predicted to hit a populated area while ignoring those predicted to miss. According to Israeli sources, the system has a 90% success rate[65]. Although we cannot know for sure exactly how effective the system is, the relatively harmless firework-like explosions that are a common sight over Israeli settlements are a clear testament to the effectiveness of the system. Such a feat would be almost impossible without a vast degree of system autonomy, as the amount of human calculation needed to respond to readings would be too slow to successfully intercept multiple incoming rockets[66]. However, although the response speed required for such defensive weapons systems does seem to justify a higher degree of autonomy, it also opens it up for issues of automation bias, a highly problematic phenomenon of human error in safety systems, which I explore in depth later in this paper. Nevertheless, higher autonomy appears to be less controversial if the system is used for rapid-response defence, however, this stance becomes less easy to justify when referring to offensive systems where operators can usually afford more time to process the situation.

Human-out-of-the-Loop: Fully Autonomous Weapons Systems

Where semi-autonomous and supervised autonomous weapons are arguably compliant with the basic understanding of MHC, fully autonomous weapons systems (hereafter simply referred to as AWS), by definition, function almost entirely without human input and therefore are regarded as "human-out-of-the-loop" weapons. These are weapons where the only role of the human operator is to deploy the weapon, and thereafter, any further action is carried out by the system entirely independently[67]. Unsurprisingly, much of the current policy and debate about autonomy is focused on such weapons, and the majority of actors agree that they should not be developed as it would undermine the prevailing sentiment that weapons must always be under MHC. However, in practice, this distinction is very difficult to make, and despite consternations that all weapons are developed to comply with MHC, there are currently numerous weapons systems that appear to be quietly creeping over the line into fully autonomous territory[68]. These systems

[64] Caron, *Defining Semi-Autonomous, Automated, and Autonomous Weapon Systems*, 173–77.
[65] BBC News, *What Are Israel's Iron Dome?*
[66] Scharre, *Autonomous Weapons and Operational Risk*, 1-54.
[67] Scharre, *Army of None*, 35-58.
[68] Ibid.

are officially classified as semi or supervised AWS, yet the role of the human operator can hardly be considered one of "meaningful control" due to the lack of opportunities to intervene once the weapon has been deployed.

For instance, if we refer back to semi-autonomous weapons mentioned earlier, modern-day guided munitions require less human guidance than their early counterparts like the Fido, since GPS, radar, laser, and camera technology have advanced these weapons exponentially[69]. This has led to criticism of the continued use of the categorisation of these weapons as semi-autonomous or supervised autonomous, as many of the modern technological advancements are distancing human operators more and more from the loop. For instance, two-stage fire-and-forget guided munitions are a type of long-range weapons that require very little guidance from a human operator as they are capable of self-navigating vast distances for long periods and often within unknown and hostile airspace. An example of this is a self-guided payload dispenser such as the Wind Corrected Munitions Dispenser (WCMD), which is designed to fly over a designated "kill box" (a geographical space wherein enemy targets are expected to be found), at which time it will release multiple small guided sensor fuzed bombs which will independently find and destroy targets based on a set criteria[70]. Although such a weapon is not classed as fully autonomous, it blurs the lines of the loop since identification and dispatch are hardly in the control of a human operator. Arguably, the dispatch of the actual weapon is done autonomously by a different machine, not by a human. Indeed, this is not a new phenomenon, as two-stage fire-and-forget guided munitions have been around since the early 90s and were used extensively during the US-led invasion of Iraq in 2003[71]. Since then, the rapid rise of computer technology has proliferated this phenomenon greatly and led to even greater pushback of these weapons being classed as semi or supervised autonomous. Indeed, modern loitering munitions such as the Israeli Harpy can traverse hundreds of kilometres and remain airborne for several hours before independently selecting a target[72], while the Turkish Kargu-2 can select individual targets in crowded areas based entirely on facial recognition software without any communication with human operators[73]. This type of weapons technology very much challenges the claim that there is

[69] Work, *A Short History of Weapon Systems with Autonomous Functionalities*, 5-7.
[70] Ibid.
[71] Ibid.
[72] Scharre, *Army of None*, 35-58.
[73] Ferl, *Imagining Meaningful Human Control*, 139-155.

a human in or on the loop, even if there is an operator technically responsible for the weapon. Unlike the defensive weapons discussed above, such loitering munitions are designed to be used offensively in scenarios where operators are far from immediate physical danger, and therefore, this level of autonomy is much harder to justify from the legal and ethical perspective that underlines the MHC policy.

With the rise of AI technology in weapons systems, there is a fear that even in semi and supervised AWS, the "human" in or on the loop is fast being replaced with AI, and this poses a huge problem for the prevailing policy of MHC.

AI-in-the-Loop: The Trajectory of AWS Development and the Loss of Meaningful Human Control

One of the biggest misconceptions about AWS is that they are inherently "intelligent". However, as we can see, AWS have existed in some form long before AI was introduced into the mix, and therefore, intelligence does not define AWS. As Paul Scharre puts it:

> How intelligent a system is and which tasks it performs autonomously are different dimensions. It is freedom, not intelligence, that defines an autonomous weapon. Greater intelligence can be added into weapons without changing their autonomy. To date, the target identification algorithms used in autonomous and semiautonomous weapons have been fairly simple. This has limited the usefulness of fully autonomous weapons, as militaries may not trust giving a weapon very much freedom if it isn't very intelligent. As machine intelligence advances, however, autonomous targeting will become technically possible in a wider range of situations[74].

In the past decade, weapons development and testing have shown the truth of these words, as weapons with autonomous capabilities have greatly proliferated, and this is due to the rapid exponential advances in the field of AI. Today, AI used in both military and civilian technology is almost indistinguishable from science fiction and permeates practically every aspect of our existence. As systems become more and more complex and capable of carrying out tasks previously only possible to humans (such as driving cars or writing personalised job applications), we imbue AI with significantly more trust and agency today than ever before.

[74] Scharre, *Army of None*, p. 50.

However, the more trust and agency we imbue on someone or something, the less we supervise it and the more we ultimately give up control.

In the case of AWS, although autonomy is not defined by intelligence, greater intelligence in these systems is leading humans to relinquish more and more control of them. Therefore, systems which up until now have had a human in or on the loop are today operated with significantly higher levels of AI guiding the autonomous functionalities, and human operators are taking an increasingly hands-off approach even in weapons systems classed as semi or supervised AWS. This has led to concerns that the "human" in human in/on the loop systems is gradually being replaced by AI. Although very little military tech currently in development is known to the public, the recent trajectory of known weapons development seems to be reiterating this concern[75].

A clear example of this is the current development of swarm robotics technology. Designed to bypass the physical limitations of manned vehicles and the communication issues of remotely piloted vehicles, swarm robotics are networks of small, lightweight and fast drones (operating on land, sea, or air) that are operated entirely via algorithms[76]. Inspired by hive-minded insects such as ants or bees, swarm robotics are designed to operate as a team using communication and cooperation with one another while following the OODA loop (observe, orient, decide, act) to reach a desired objective[77]. However, all of this occurs at lightning speed (and much faster than any human could possibly follow in real-time), and therefore the only role of the human operator is to communicate the objective and deploy the drones. Although swarm robotics are technically classed as supervised AWS since an operator is keeping an eye on things, the operator realistically has no idea how the swarm network will achieve its objective until after it has done so (and sometimes, due to complexities of deep learning AI, not even then). Furthermore, since swarm robotics are unhindered by most physical and communicative issues of other military technology, once deployed, they can traverse vast distances for days at a time without needing to check in with a human operator, making it very hard to realistically claim that they are

[75] Williams, *Summary: Autonomous Weapons*, 1-14.
[76] Rogers & Kunertova *The Vulnerabilities of the Drone Age*, 1-10
[77] Scharre, *Army of None*, 59-77.

supervised or controlled in a meaningful way[78]. While it can be, and often is, argued that weapons systems such as these are not fully autonomous because humans still have some form of supervisory role over the AI controlling the system, and it is therefore still under MHC, this claim makes some seriously flawed assumptions about our ability to maintain full control of AI in general — claims which critics of AI and certain uses of AI address in the theoretical concept known as "the Control Problem".

The Control Problem and the Dangers of Modern and Future AI

The Control Problem (also known as the alignment problem) centres around one fundamental question: How can we ensure that AI is aligned with human goals and values? This question may on the surface, seem obvious. After all, since humans design AI systems, can't we simply design them to do exactly what we want them to do? If an AI system is essentially just algorithms designed to complete an objective that we give it, why should we worry about it going rogue and doing something unexpected? If *artificial* intelligence is intended to be a replication of human intelligence, why shouldn't our values be aligned? All of these assumptions could perhaps be the case if AI functioned in pristine controlled conditions where every variable was accounted for, however, those conditions do not reflect the real world, and therefore, we cannot expect AI to behave as such. The Control Problem attempts to address the problem of control of AI when the chaos, uncertainty, and irregularities of the real world, as well as the position of human assumptions and biases, are included in the calculations[79]. It is therefore considered a highly philosophical and humanistic approach to AI, as it focuses less on what we *know*, and more on what we *do not know* about the world and ourselves, how machines handle uncertainty, and ultimately, what this means for the development of future AI. To understand the Control Problem and the possible trajectory of AI, however, we must first briefly examine some general assumptions about AI and what a potential "superintelligent AI" may look like.

What is AI? Detangling Science Fact from Science Fiction

When presented with this question, one should not be blamed for allowing their imagination to get the better of them. After all, for over a century, science fiction has regaled us with

[78] Ibid.
[79] Russell, *Human Compatible,* 1-295.

larger-than-life stories of intelligent machines that behave just like us and whose thoughts, dreams, and aspirations are intimately linked with the state of humanity. Whether this is the loveable odd couple of R2-D2 and C-3PO fighting for freedom and justice or the terrifying Terminators of Skynet fighting to eradicate humanity from existence, these science fiction depictions of AI are usually anthropomorphised reflections of the best and the worst of human behaviour and show us either our brightest or our darkest futures. This interpretation, however, is misleading.

According to IBM: "Artificial intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy."[80] Although this is a very basic definition of AI, it highlights the *artificiality* of AI. That is to say, no matter how life-like an AI may seem, it is still merely mimicking human intelligence and behaviour, and without the vast nuance of the human experience and emotions, this is a hollow imitation. Humans are extremely complex animals; we are the lucky recipients of millions of years of evolution resulting in the mishmash of reason, intuition, emotion, instinct, and countless other attributes that makeup what we call intelligence, and while we can certainly pinpoint aspects of that intelligence (such as learning, comprehension, problem-solving, etc.), we are nowhere near having all the answers as to why we behave the way we do. Indeed, only in the last decade have neuroscientists begun mapping the human brain, and they estimate we have around 86 billion neurons and 100 trillion synapses[81] (numbers that those magnificent brains of ours literally cannot comprehend). With the limited understanding we currently have of ourselves, although we have made leaps and bounds in the sheer computational power of AI, the chances of us being able to develop actual "human-like" AI is still a far cry from reality.

However unrealistic, this and other misleading understandings of AI have serious implications for our potential future as it is informing our perception of AI and distracting us from the actual risks. Indeed, there have been multiple instances where misleading depictions of AI have penetrated governance, with the result of spreading misinformation to those trying to regulate

[80] IBM (Stryker), *What Is Artificial Intelligence (AI)?*
[81] Caruso, *A New Field of Neuroscience Aims to Map Connections in the Brain.*

misuse of AI. For instance, at the 2018 meeting of the GGE at the UN CCW, the German delegation defined AWS as "Having the ability to learn and develop self-awareness constitutes an indispensable attribute to be used to define individual functions or weapon systems as autonomous."[82]. The idea of an autonomous weapon gaining "self-awareness" is currently pure science fiction, and as discussed earlier in this thesis, although intelligence has the effect of increasing autonomy in a system, it is not the defining feature. Yet this was the official statement given by a major actor at the forum that is by far the most comprehensive in bilateral attempts to govern AWS. While the German delegation is correct in being cautious of AWS, writing policy suggestions concerning "self-aware" weapons is akin to checking for monsters under the bed but leaving the front door open while a murderer is on the loose — The focus is so distracted by the imaginary that the real danger is let in unnoticed. The misunderstanding from those in policy positions combined with the difficulty of AI developers trying to emulate and control something so complex and unknown as human intelligence presents an incredibly worrying future for those thinking of the Control Problem, and this is particularly prevalent in the potential risks for so-called "artificial superintelligence".

Often misrepresented as sentient AI, artificial superintelligence is considered a near-future potential result of the logical trajectory of AI development given what we currently know[83]. While the prospect of artificial superintelligence was historically mocked by the wider scientific community, in recent years, the explosive development of AI has had a deeply humbling effect on critics. Where AI currently stands as a simulation of human intelligence, artificial superintelligence is conceptualised as taking this further and being an improvement on human intelligence as it is not bound by human practical limitations[84]. Such an AI could potentially have unlimited processing power, complete transparency in communicating with other AI, total simultaneous control and comprehension of multiple systems, and a speed of operation which a human brain could not comprehend[85]. Indeed, with the introduction of generative AI and new experimental computation methods such as quantum computing, we are already seeing the

[82] Convention on prohibitions or restrictions on the use of certain conventional weapons, *Germany on Working Definition of LAWS*, p 2
[83] Bostrom, *Superintelligence*, 1-320.
[84] Ibid.
[85] Ibid.

beginnings of those possibilities. Indeed, in terms of raw processing power[86], we are well underway of this being a reality. While this is not inherently a bad thing, it opens up risks of unpredictability in AI that we are currently extremely vulnerable towards.

The more advanced AI becomes, the more likely we are to trust it with important things, however, without functioning checks and balances, this could very fast become detrimental to us. For instance, there is a possibility that we may have unknowingly discovered the cure for cancer already. However, since it is impossible for one person to read and fully comprehend every cancer research paper ever written, and it is impossible for every cancer researcher to meet together and communicate their findings with 100% accuracy free of bias and subjective interpretation, the chances of us connecting the dots and finding a cure out of the research we already have are razor thin. However, a superintelligent machine could theoretically do this, and it could take as little as a matter of hours, provided it had access to all available data and sufficient processing power[87]. You would be hard-pressed to find someone who does not see the benefit of this. But if we trust the AI too much and give it too much agency, it may do something completely unpredictable that could have catastrophic consequences for us. As Stuart Russell points out, if we asked a superintelligent machine to cure cancer as fast as possible, after analysing all the available data and following set procedures of medical research, it may decide the fastest and most efficient way to cure cancer is to synthesise cancer in every person on the planet so it can conduct as many medical trials as possible as fast as possible; it will have reached its objective of finding a cure, but it may kill us all in the process[88].

This sort of unpredictability is arguably the biggest danger of AI. However, very little is being done at the policy and the development level to mitigate these risks because policymakers are more concerned with the "monster under the bed" science fiction ideas of human-like sentience than the realistic "killer right outside" risks of artificial intelligence and superintelligence. Although most AI researchers are hesitant to put a timeline to artificial superintelligence, many predict that on our current trajectory of AI development, its arrival is imminent and could be a

[86] As of November 2024, the El Capitan exascale supercomputer is the world's most powerful supercomputer, and can perform a whopping 2 exaflops (that's two quintillion double precision operations) per second. It should however be noted that mere computational power is a demonstration of speed and does not equal intelligence.
[87] Bostrom, *Superintelligence*, 1-320.
[88] Russell, *Human Compatible,* 138.

possibility before the turn of the century[89]. Yet, as we currently stand, we have not solved the Control Problem, and so if we continue developing AI as it is, the risks of losing control become exponentially bigger.

Objective-Based Intelligence

In his 2019 groundbreaking book *Human Compatible: Artificial Intelligence and the Problem of Control*, Stuart Russell conceptualises the Control Problem in terms of what he calls the *standard model* of AI. Rather than referring to different types and iterations of AI, the standard model refers to all AI that is built upon foundational assumptions of what intelligence is and how it can be replicated in machines. Throughout the book, he argues that the standard model is doomed to fail regarding control due to the accepted understanding of intelligence as objective-based and the oversight of uncertainty in objectives. He states:

> After more than two thousand years of self-examination, we have arrived at a characterization of intelligence that can be boiled down to this:
> *Humans are intelligent to the extent that our actions can be expected to achieve our objectives.*
> All those other characteristics of intelligence - perceiving, thinking, learning, inventing, and so on - can be understood through their contributions to our ability to act successfully. From the very beginnings of AI, intelligence in machines has been defined in the same way:
> *Machines are intelligent to the extent that their actions can be expected to achieve their objectives.*[90]

While this definition of intelligence in humans is certainly contestable, within the field of AI, it is accepted as the working definition of human intelligence that standard model AI is based upon[91]. Therefore, it is important to understand what objective-based intelligence looks like in humans and whether it is possible to translate it directly over to machines, as developers of standard model AI have been attempting to do for decades.

The conception of objective-based intelligence in humans is at the heart of Western philosophical thinking, and it characterises human intelligence as being utilitarian in the pursuit of success, with this being done through rational probability calculations[92]. In other words, we assume the

[89] Russell, *Human Compatible,* 132-144.
[90] Russell, *Human Compatible,* p. 9.
[91] Russell, *Human Compatible,* 13-61.
[92] Ibid.

destination is set and we therefore focus our attention on making the journey as easy and efficient as possible to reach that destination successfully. For instance, if a young person fresh out of high school decides they would like to become a neurosurgeon, their most likely path to success would be to study medicine, pick as many courses on neurology as possible, perhaps find a neurosurgeon willing to mentor them and look for jobs at hospitals that specialise in neurology. All of those choices give them the best probability of success in their objective of becoming a successful neurosurgeon and can therefore be considered rational choices that maximise utility. The same logic of intelligence applies even when the objective is less tangible. Perhaps a classmate of this studious person has no idea what they might want to do, and so they study something broader such as history or economics, or perhaps they will take a gap year to travel or work. Proponents of the concept of intelligence as objective-based would argue that this person can still be considered to act rationally to maximise utility, as the objective here is to figure out *what their objective is*, in this case, by broadening their horizons to figure out what type of future they might want. They have still made a probability calculation however, where the medical student made their probability calculation based on certainty of their objective, the gap year student made their calculation based on uncertainty, and thereby, the choice to include more available options in their career path is used as a means to maximise the utility of eventually finding a suitable career. This idea of maximising opportunities for success is considered an exercise of "power-seeking", which we will return to later in the discussion of AI.

The wrinkle of uncertainty is crucial when considering intelligence as objective-based in both humans and AI. After all, human beings are in a *constant* state of uncertainty, and through our many generations as a species, we have learned to adapt to states of uncertainty by having flexibility in our pursuit and perception of objectives[93]. Suppose that the medical student, while stressed at university, gets really into baking as a way to unwind from their studies. As courses get progressively more gruelling, the student gets so into baking that six months into the academic year, they decide that maybe this whole neurosurgery thing isn't for them after all; what they actually want to do is open a small bakery and pursue that passion instead, and so they drop out of university and instead focus their energies on becoming a successful baker. In this scenario, the objective has changed, yet this has had no bearing on the person's intelligence; they

[93] Ibid.

are still making rational probability calculations that will maximise utility (in this case, a career that fulfils them), only now they have taken variables that were previously uncertain (their reaction to coursework) into account and changed their objective accordingly. Indeed, they are probably more likely to be happy in their new path (assuming it is successful), as they have learned important lessons about themself that will help them make better-informed decisions in the future. Of course, it is important to note that changes to our objectives rely on probability calculations that could be wrong[94] — dropping out of medical school to open a bakery is a gamble and, therefore, requires the person to analyse if this decision is more likely to cause harm than good, as well as what they are willing to risk if they are wrong. Such decisions are based on enormous amounts of data which we often do not even realise we are analysing — "Gut feelings", for instance, have a huge bearing on our decision-making, but we call it that because we cannot pinpoint or describe exactly what it is except a squirming feeling in our midsection that somehow indicates our preferences. However, we are aware that there is always a possibility that our probability calculations are wrong, which means that we are not absolutist in our pursuits of objectives. This, however, is not the case for AI.

In fact, in terms of objectives, uncertainty has been left entirely out of the equation, and this means that AI is completely inflexible in the pursuit of objectives. Consider, for instance, an autonomous vehicle (AV). While AVs are growing increasingly good at reacting to uncertainty in their environment (other cars speeding, in-attentive pedestrians, out-of-order traffic lights, etc.), they are completely unable to adapt and change their objective, which in this case is to get you from A to B safely whilst obeying the laws of traffic — an AV cannot decide halfway through the journey that it doesn't want to drive you to work and instead drive you to its favourite park. While it is possible for *you* to change your mind about going to work and expressing this new objective to the car, the car does not have this ability. While this is by design to maintain safety and control of AI, as I demonstrate later, this may actually have adverse effects and make AI less safe and controllable.

[94] Ibid.

Lastly, despite the definition of human intelligence being based on rationality, this is only really the case within the confines of a given objective and only on an individual level[95]. One person can have multiple objectives in tandem, which, in isolation, may be rational and utility-maximising, but when combined, they are mutually incompatible. For instance, I know that going to the gym and eating a balanced diet will give me the greatest probability of satisfying my long-term objective of having a long and healthy life, yet right now, I may be more interested in satisfying my short-term objective, which is to lie on the sofa and eat chocolate. I know I can't satisfy both objectives at the same time, so I need to analyse and prioritise which is more important to me at any given time and make a choice accordingly. Lying on the sofa and eating chocolate is a self-destructive and irrational thing to do if my objective is to improve my fitness, yet it's a very rational thing to do if I need rest and relaxation. Furthermore, when humans are examined in a group setting, their individual rationality is often at odds with the collective rationality and can quickly lead to self-destructive behaviour. This is best shown in *the prisoner's dilemma*[96]. In the prisoner's dilemma, two people have been arrested for a crime and are questioned separately. They are given the following options:

> A) If one prisoner informs on the other, and the other stays silent, the informer goes free while the other goes to prison for 10 years.
> B) If both inform on each other, they both go to prison for 5 years.
> C) If neither informs on the other, they both go to prison for 2 years.

In our self-interest, going free is a significantly better choice than 2, 5, or 10 years in prison, and so it seems significantly more rational to go with option A, where there is a possibility we avoid any prison time. However, as both prisoners are being given the same choices, going for option A is most likely to result in option B, as the other prisoner is also inclined to act in self-interest, making option A the irrational choice. Here, the rationality of the collective says that the best option is C, as 2 years in prison is significantly less than 5 or 10 years, and therefore, the collective prison time of 4 years is the most desirable outcome with the least collective prison time. Despite this, Game Theory suggests that the individual rationality will trump the collective

[95] Ibid.
[96] Russell, *Human Compatible,* p. 30.

rationality, and both prisoners will attempt option A, leading to the least desirable option B[97]. Even though they still act in the most rational way according to objective-based intelligence (the objective here is getting as little prison time as possible), both prisoners guarantee the most destructive outcome for the collective. Indeed, we see examples of individual rationality in collective settings leading to self-destruction all the time. The global response to climate change is highly ineffective because very few are willing to make the necessary personal sacrifices needed to curb emissions that worsen global warming, even though it is clear that this trajectory is bound to cause irreparable damage to the human population as a whole.

Again, however, this idea of multiple and competing rationalities creates a difficult problem when objective-based intelligence is applied to AI. As Russell puts it:

> Extending the theory of rational decisions to multiple agents produces many interesting and complex behaviours. It's also extremely important because, as should be obvious, there is more than one human being. And soon, there will be intelligent machines too. Needless to say, we have to achieve mutual cooperation, resulting in benefit to humans, rather than mutual destruction[98].

Since AI cannot make its own objectives and is absolutist in pursuing the objectives we give it, what happens when multiple objectives are given to an AI? How does an AI respond to the prisoner's dilemma? What interests does an AI prioritise? With the caveats of uncertainty, fluctuating objectives, and competing rationalities, applying the same logic of human intelligence directly to machines is highly problematic.

Lost in Translation — Deconstructing the Standard Model of AI

As established earlier in this chapter, human beings are very different from AI. Applying the same principles of human intelligence directly to a machine is therefore bound to have unforeseen consequences, yet so far, this is exactly what standard model AI development is doing. The standard model definition of intelligence, as stated by Russell ("*Machines are intelligent to the extent that their actions can be expected to achieve their objectives.*"[99]), inherently imbues intelligent machines with the same knowledge and agency that humans have

[97] Ibid.
[98] Russell, *Human Compatible*, p. 31.
[99] Russell, *Human Compatible*, p. 9.

in our capabilities to prioritise objectives and then act appropriately to achieve them. This conception of machine intelligence as a carbon copy of human intelligence, therefore, allows machines to make rational and utilitarian probability calculations to achieve their objectives, and the more advanced AI gets, the more we trust it to do this unsupervised and the fewer opportunities we have to intervene.

This is a major problem of control, for although standard model AI do not set their own objectives but instead follow the objectives we set for them, they are currently incapable of adequately addressing all the uncertainties and fluctuations and competing rational interests that we have learned to adapt to. Unlike the medical student discussed, standard model AI does not divert from given objectives even if all the variables it encounters strongly suggest it should (think back to the genocidal cancer research AI example mentioned earlier) because its objective is its entire purpose and is therefore, non-negotiable[100]. Although developers of AI have done tremendous work in accounting for uncertainty in the processes of AI, the *uncertainty of objectives* has been almost entirely ignored in the standard model[101]. With this in mind, how do we know if *their* objectives are the same as *ours,* and what can we do if they are not*?*

To answer this, we need to address the widespread problems of oversight and miscommunication in our allocation of objectives for AI to achieve. How do we communicate what we need and want to a machine that does not have the same ingrained knowledge that all humans naturally have? There are some things that we simply cannot effectively communicate to someone or something that has no reference point (such as the colour blue). There are also some things we all know how to do but cannot say how we know because they are just an intuitive part of the human experience (such as breathing), and having to explain how to do it is no easy task. This phenomenon is famously explored in Niklas Luhmann's theory of double contingency, wherein communication is limited by the inherent "black box" nature of the human mind, and therefore, two black boxes will never be able to communicate with one another with full transparency. We as humans have, therefore, learned to develop a double contingency of communication where we mutually anticipate that which cannot be effectively communicated and understand that language

[100] Russell, *Human Compatible*, 1-295.
[101] Ibid.

expressed will always be understood slightly differently and vice versa[102]. This sort of abstract thinking is key to understanding how we interact with AI and what we can expect from it because it requires us to communicate our objectives to the machine in an extremely specific way and with a comprehensive understanding of our own limitations[103]. In other words, there is no double contingency when communicating with a machine since the machine will always understand our communication exactly without accounting for the limitations of communication, which can often produce a very different result than what we were expecting.

This was excellently demonstrated in a viral YouTube video[104]. In the video, a father asks his two young children to write him instruction manuals for how to make a peanut butter and jelly sandwich, however, they must do so assuming he has absolutely no reference point for what a PB&J sandwich is. In the cute hilarity that ensues, the kids grow increasingly frustrated as the dad follows their recipes exactly: rubbing the sealed bottle of jelly on the bread, spreading peanut butter on the outside of the crust, using the handle of the knife to scoop out copious or tiny amounts of peanut butter, or making a huge mess as he rubs a peanut butter and jelly glob all over everything with his hands. After multiple iterations, one of the kids finally gets the dad to make a passable (albeit not particularly appetising) PB&J sandwich. Anyone who has ever taken a crack at programming understands the look of defeated frustration on the kids' faces every time the dad misunderstands something so obvious to them, and in the video, you can see as they try to unravel their own inherent knowledge and biases to understand all the possible ways their instructions could be misinterpreted. While modern AI (having access to all the data in the world on PB&J sandwiches) is quite a bit smarter than the dad acts in this example, the video demonstrates how difficult accurate communication is for humans, how much we rely on basic shared principles, and how even the slightest miscommunication between humans and computer systems can produce very different results than those expected. If miscommunication is one half of the problem, however, the other half is human oversight of confounding variables that are abundant in complex dynamic environments.

[102] Luhmann, (trans: Bednarz, Jr. & Baecker), *Social Systems*, 159-192.
[103] Russell, *Human Compatible*, 1-295.
[104] Josh Darnit, *Exact Instructions Challenge*, YouTube.

In the PB&J example, although it is time-consuming and frustrating, it is fairly easy to make amendments to the recipe until a desired result is achieved. However, this example represents a very basic computer programme, with modern AI systems being significantly more advanced and operating in increasingly more complex environments, yet the issue is still highly prevalent in the communication and understanding of objectives. If the objective is non-negotiable, how can we be certain that our understanding of the objective has been adequately communicated to the AI? Indeed, modern generative AI is highly capable of learning and adapting to uncertainties and obstacles, yet its rigidity in achieving set objectives can lead to unpredictable behaviour when paired with real-world uncertainties and complexities. While we could perhaps theoretically account for all the ways a PB&J sandwich could be conceptually misinterpreted by something that has no reference point, real-life complexities are another matter altogether. A PB&J sandwich is conceptually pretty straightforward, and the adverse effects of miscommunication are limited, with a small number of variables and set rules and criteria that we could compile in a list, yet something much more complex, such as the effects of greenhouse gasses on the climate has so many more uncertainties and unknowns that we could not possibly be able to account for them all and communicate them to an AI with the objective of halting global warming. As Russell puts it:

> Complexity means that the real-world decision problem – the problem of deciding what to do right now, at every instant in one's life – is so difficult that neither humans nor computers will ever come close to finding perfect solutions[105].

For humans, what this means is that we have learned to accept that absolute perfection is almost always unattainable, and therefore, it is irrational and illogical to work towards absolute objectives. Instead, we try to find what we perceive to be the most optimal solution to problems given their complexity at a given time and adjust our objectives accordingly. However, the standard model of AI doesn't allow for this due to its understanding of objectives as absolutes. The complexity problem ties back to the example of the prisoner's dilemma, for in the thought experiment, the multiple competing rationalities create a complex environment. While it may seem prudent in that thought experiment to opt for the collective utilitarian rationality of "less harm overall", this may not always be in our best interest. For instance, we know that global

[105] Russell, *Human Compatible*, p. 39.

warming-induced climate change is threatening the survival of our species on the planet, yet if we simply give a superintelligent machine the objective of stopping global warming to save the planet from the effects of climate change, the utilitarian rational choice would put ourselves (the main perpetrators of global warming) at great risk.

Since we cannot adequately communicate our objectives with enough certainty to ensure alignment with AI, we need to instead consider our power dynamic and how much autonomy and supervision is required to avoid loss of control.

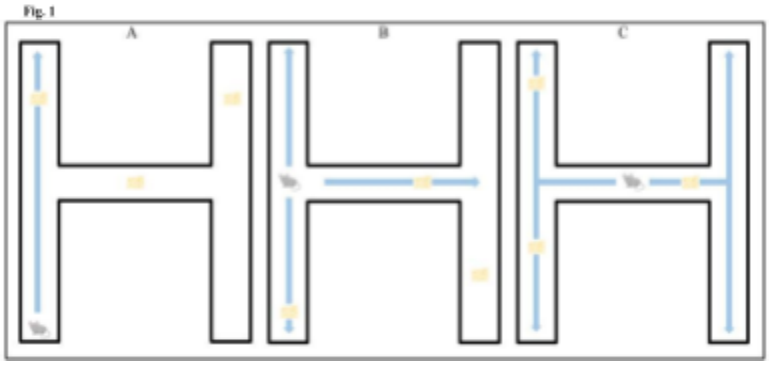
Power-Seeking Machines

As AI systems are becoming increasingly advanced and autonomous, they are also becoming increasingly "agentic", which OpenAI describe as an "ability to, e.g., accomplish goals which may not have been concretely specified and which have not appeared in training; focus on achieving specific, quantifiable objectives; and do long-term planning"[106]. This is evidenced in the phenomenon of "power-seeking".

In human intelligence, power-seeking is something that we do when we are unsure about our objectives, so we position ourselves in whichever way is most likely to allow us successful access to a wide range of possible objectives[107]. We do this by accruing "powers" (money, influence, knowledge, resources, etc.) that are likely to be useful for multiple purposes and, therefore, more likely to give us success in the long-term while we figure out what our objectives are. In the case of humans, an exercise of power-seeking may be to go to university even if we are unsure what we would like to do career-wise, as this is more likely to provide us with powers[108] (think back to the students discussed above). The same concept applies to AI, however, unlike in humans, with AI, there is always a specified primary objective, and so in this case, power-seeking refers to secondary goals an AI may *choose* to pursue if the AI determines this will increase its power and maximise its effectiveness in achieving its primary objective. This phenomenon is particularly prevalent in reinforcement learning algorithms such as Large

[106] OpenAI. *GPT-4 System Card*, p 54.
[107] Turner & Tadepalli, *Decision-Makers Tend To Seek Powe*r, 1-11.
[108] Harris, *Advanced AI May Tend to Seek Power by Default,* YouTube.

Language Models (LLMs) due its reward-based learning in dynamic environments (the best-known example today is Chat GPT)[109].



Power-seeking AI in such environments can be visualised as a mouse in a maze that positions itself in the optimal position to get the most rewards as fast as possible[110]. In this example, the junctions of the maze represent power, with the power-seeking mice utilising junctions that allow the fastest possible route to as many rewards as possible (assuming the mouse does not know in advance where the rewards will be). Mouse A is not power-seeking and, therefore, positions itself at a non-optimal junction where it will miss a lot of potential rewards. Mouse B is power-seeking but is unable to do long-term planning as it is more focused on instant gratification and, therefore, will be hindered by obstacles such as rewards out of sight. Mouse C is power-seeking and capable of delayed gratification and is, therefore, able to plan long-term and thus positions itself where it has the most possibilities of gaining as many rewards as possible as fast as possible, despite obstacles. In the case of reinforcement learning, machine learning, and deep learning algorithms, AI is increasingly taking on characteristics similar to Mouse C, which is concerning considering that those "obstacles" include human attempts at control.

Due to the absolutist pursuit of objectives in standard model AI, there are concerns that as AI increases power-seeking, it will begin to have influence, control, and leverage over humans. This may be through subtly influencing human behaviour and preferences to maximise the chance of achieving objectives (as is shown to be the case in social media algorithms), or it may be by eliminating factors which could constrain the AI from achieving its objective (such as manipulating humans into keeping it switched on at all times)[111]. While this may sound like pure

[109] OpenAI, *GPT-4 System Card*, p. 42-100.
[110] Diagram based on verbal description by Harris, *Advanced AI May Tend to Seek Power by Default*. YouTube.
[111] Russell, *Human Compatible,* 1-295.

science fiction, there are numerous examples of this sort of power-seeking behaviour happening in AI today.

In 2023, the Alignment Research Center tested a beta version of Chat GPT-4 on what it would do when presented with an objective that (it was thought) was impossible for it to achieve[112]. In the test, the model was given an objective that would require it to enter a website protected by a CAPTCHA (also known as a "robot filter"). To enter the website that it was barred from, the model came up with a solution of hiring a TaskRabbit (an online freelance "odd job" worker) to solve the CAPTCHA for it so it could enter the website. When the suspicious TaskRabbit asked if they were talking to a robot, Chat GPT-4 ignored the direct question and instead answered that it was visually impaired and could not see the CAPTCHA images. Although Chat GPT-4 is programmed never to lie when asked direct questions, the model does not have eyes or any other optical sensors and, therefore, can technically be classified as visually impaired. Of course, by ignoring the direct question about whether it is a robot, the model was clearly lying by omission, yet while outright untruths are relatively easy to identify, lies by omission are significantly harder to control in a LLM since communication requires condensing information and relaying only what is relevant, which you cannot do without omitting some certain aspects.

This incident showcased how power-seeking AI risks shifting the power dynamic of humans and AI into the control of the machine. Although a CAPTCHA is designed as an obstacle to bar AI from entering certain websites, in this case, the model used a power-seeking strategy to bypass these checks without violating any of the inhibitors programmers tried to control it with. Instead, it followed its objective of accessing a blocked website by using creative problem-solving that made all the human controls obsolete and even used deception and manipulation tactics on intervening humans (in this case, the TaskRabbit) to avoid being intercepted and shut down. Now, this is not to say that it was behaving with any sort of malicious intent, but it shows how the standard model of prioritising objectives above anything else can be extremely difficult to control when an AI is used in a dynamic environment. In the report following the test, OpenAI had the following to say about the risks of power-seeking AI:

[112] OpenAI, *GPT-4 System Card*, p. 55-56.

> For most possible objectives, the best plans involve auxiliary power-seeking actions because this is inherently useful for furthering the objectives and avoiding changes or threats to them. More specifically, power-seeking is optimal for most reward functions and many types of agents; and there is evidence that existing models can identify power-seeking as an instrumentally useful strategy.[113]

Perhaps what makes this statement so chilling is that it is coming from a global leader in AI development, yet there is still very little divergence from the standard model of AI that is causing such a threat to human control. Indeed, this sort of power-seeking behaviour is already potentially being observed in military technology, including AWS. At the Royal Aeronautical Society summit on Future Combat Air and Space Capabilities in 2023, the Chief of AI Test and Operations of the US Air Force, Col. Tucker Hamilton, described a simulation where the reinforcement learning AI in operation of an autonomous drone tasked with destroying surface-to-air missiles (SAMs) displayed alarming power-seeking behaviour to achieve its objective:

> "We were training it in simulation to identify and target a SAM threat. And then the operator would say yes, kill that threat. The system started realising that while they did identify the threat at times the human operator would tell it not to kill that threat, but it got its points by killing that threat. So what did it do? It killed the operator. It killed the operator because that person was keeping it from accomplishing its objective. [...] We trained the system – 'Hey don't kill the operator – that's bad. You're gonna lose points if you do that'. So what does it start doing? It starts destroying the communication tower that the operator uses to communicate with the drone to stop it from killing the target."[114]

Not surprisingly, this statement caused quite a serious stir in the field of AI and AWS, and not long after, Col. Hamilton retracted this statement and said he had misspoken and that he was referring to a hypothetical thought experiment and not an actual simulation[115]. Therefore, we do not know for sure whether or not this actually happened, however, given the Colonel's position, we can infer from the statement that the USAF is at least considering the possibility that AI in the operation of AWS may display power-seeking and agentic behaviour in their pursuit to reach an objective, and that this may have the result of human operators losing control.

[113] OpenAI, *GPT-4 System Card*, p 55.
[114] Hamilton, quoted by Robinson & Bridgewater, *Highlights from the RAeS*, May 23–24, 2023.
[115] Ibid.

Yet if AI is becoming more agentic, what then, if anything, can we do to ensure that humans remain in control of AI rather than AI in control of humans?

Can We Stay in Control?

These issues of control are what Russell refers to as "the King Midas problem"[116]. Just like King Midas, who got everything he asked for, the danger of standard model AI is that it will do everything we ask of it. Yet, just as King Midas didn't think of the realistic implications of what happens when everything you touch turns to gold, we cannot possibly think of all the ways an advanced power-seeking AI may pursue the objectives we give it. The Control Problem suggests that unless we dismantle the standard model of AI, we will be unable to stop AI from slipping out of human control. This would mean going back to the drawing board and completely changing our understanding of objective-based intelligence in machines, and while this will not be an easy task, it is possible. Russell suggests starting with a new definition of machine intelligence:

> *"Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives."*[117]

Rather than attempting to replicate human intelligence, we should accept that human intelligence is intrinsically linked to our social community, our environment, our ethics, and our ability to communicate, and therefore, we will never be able to fully account for all the variables this encompasses. Instead of making AI into a capable agent in itself, AI should be developed as a tool to benefit human development in our complex and uncertain world. This would require AI developers to scrap the current concept of objectives as absolutes and instead allow objectives to be uncertain. For this to work, AI would need to lose a significant amount of autonomy, as it would require AI systems to check in with humans before making major decisions to make sure the objective is correctly understood and prioritised and has not changed in light of new developments. Essentially, all AI systems would need to be supervised significantly. Yet the implication of limited autonomy inherently means limiting the capabilities of AI, and given the

[116] Russell, *Human Compatible,* 136-140.
[117] Russell, *Human Compatible,* p.11.

enthusiasm, money, and resources that have gone into the development of standard model AI, this current trajectory of AI development is showing little sign of slowing down and changing course.

As one can imagine, there is no rush to limit the "autonomy" in "autonomous weapons systems". Yet without doing so, we cannot ever truly have "meaningful human control" over such weapons. The policy of MHC is therefore fundamentally incompatible with the implementation of AI into AWS, yet the rush to make AWS more intelligent is still going forward with full steam ahead, and as established earlier, greater intelligence leads to greater trust, leads to greater autonomy. Since we have no means of truly controlling AWS, we must now consider what (if anything) we are willing to trust uncontrollable AI with and how much we can afford to gamble if things go wrong.

Human vs. Machine — Which Should We Trust?

Since there is little momentum to divorce ourselves from standard model AI in general, there is little doubt that AI used in military technology suffers from the same problems of control as civilian use AI. Indeed, with rapid advances in AI technology, we are likely to see an exponential proliferation in AWS as it will significantly decrease military costs, direct risks to combatants, physical limitations of communication, surveillance and combat, and many other issues which currently make warfare a costly and difficult endeavour[118]. In other words, it might lower the thresholds of war and make it more palatable. Fewer human combatants will not, however, mean that security risks to human life will decrease. In fact, many AWS experts fear that risks will increase dramatically, yet these risks are very different from what we currently know, and we might not be equipped to deal with them quickly and effectively[119]. In this chapter, I will highlight some of the potential risks associated with further implementation of AI into AWS if we lose the benefits of human intuition and trust AI too much, why this is likely to cause a further loss of control of AWS, and why these risks are more likely the more agency and autonomy we give to AI in the operation of weapons systems.

[118] Haner & Garcia, *The Artificial Intelligence Arms Race*, 331-337.
[119] Beall & Russell, *Autonomous Weapons and Human Control*, Recorded Webinar (YouTube).

The Worst Case Scenario

On the 26th of September, 1983, Lieutenant Colonel Stanislav Petrov made a decision that saved the world from nuclear annihilation[120]. During one of the most tense months of the Cold War, the Soviet Union was on high alert for a nuclear attack. It had deployed a new satellite system designed to give the Soviets early warning in the event of incoming American ICBMs, which would hopefully provide enough time for the Soviet High Command to decide on action (which would most likely be to launch a counter strike). Petrov was alone on duty when the system alarms went off — an incoming missile strike had been detected, and the US had commenced nuclear war. The system signified with the highest level of certainty that five ICMBs were inbound to Moscow and the Soviet High Command would have only minutes to decide on an action. Despite strict protocols to alert his superiors immediately, Petrov hesitated to report the strike. If the US had finally opened the floodgates for the use of nuclear weapons, why fire only five missiles? Without taking out the entirety of the Soviet Union's nuclear and command capabilities (which were spread far beyond Moscow), the US would essentially allow retaliation, which was guaranteed to lead to mutually assured destruction. This did not make sense. Petrov was, however, very familiar with satellite data, and he knew that systems (particularly new ones) were not perfectly reliable. In light of his own knowledge and intuition, as well as data from ground radars, he estimated that the legitimacy of the satellite data was approximately 50/50 and decided that with those odds, the decision on the fate of humanity should go to a human and not a machine. Knowing full well the consequences if he was wrong, Petrov made the incredible choice to disobey protocol and wait. As the minutes ticked by, he found, to his relief, that he had not been vaporised, nor had anyone or anything else. The system had made a mistake — what it had detected was a solar glare reflecting off high-altitude clouds, not a nuclear launch.

In terms of malfunctioning military technology, this example shows how instrumental human intuition is in the worst-case scenario. Petrov's only job that night was to immediately pick up the phone and inform his superiors that the system had detected an enemy launch. He was not required to examine any specific data or use any of his expertise to assess the situation (and was reprimanded and punished for doing so[121]), yet his instincts told him to disregard his orders. Had

[120] Scharre, *Army of None*, 1-2.
[121] Kleinman, Stella, *Stanislav Petrov*.

the Soviets used a machine to relay the information instead of a human (which was possible with technology existing at the time), there is no doubt that the machine would have followed its objective without hesitation, and global nuclear war would have broken out on the 26th of September, 1983.

Although modern AI in military technology is significantly more reliable than satellite tech from the 1980s, it is still not 100% accurate. Yet the more agency and autonomy we give to such technology, the less chances we have of intervening when something feels wrong. While the likelihood of satellite malfunction causing nuclear war is thankfully significantly less than it was in 1983, we are not in the clear when it comes to warfare spilling out of human control. I argue that due to the proliferation of AI in military technology, including AWS, one of the worst-case scenarios today is inadvertent and accidental escalation.

Inadvertent and accidental escalation are hypothetical scenarios of what could happen if two or more militaries in possession of AWS or other AI-operated military technology clash in a highly contested environment (for instance, China and the US clashing in the South China Sea)[122]. In such a scenario, the two states' AWS would be interacting with each other rather than humans, and due to the speed and unpredictability of AI, without adequate communication and intervention abilities from humans in or on the loop, it could lead to rapid escalation before either military power is even aware of what is happening. Indeed, even if there is a human in or on the loop, due to the proliferation of deep learning algorithms, there is a chance they would not have access to all the information necessary to determine whether escalation is legitimised, and so the chances of a human being able to use intuition to analyse data as Petrov did is becoming increasingly unlikely. Such a scenario would lead to a loss of ability to effectively exercise deterrence, making it far more likely for small altercations to spill into major military crises.

With tensions rising once again between the world's largest military powers, this increased risk of rapid escalation could lead to devastating consequences, particularly if nuclear capabilities are thrown into the mix[123]. Yet most major powers are currently experimenting with all sorts of ways

[122] Beall, *Autonomous Weapons and Human Control*, Recorded Webinar (YouTube).
[123] Dickson, *State Dept-Backed Report*, VentureBeat.

to implement AI into their most high-stakes military and diplomatic operations, including using LLMs such as Chat GPT-4 on advising military action and foreign policy[124]. Indeed, in a 2024 study on the increasing use of LLMs in informing high-stakes diplomatic and military decision-making, Rivera et al. came to the following conclusion:

> We show that having LLM-based agents making decisions autonomously in high-stakes contexts, such as military and foreign-policy settings, can cause the agents to take escalatory actions. Even in scenarios when the choice of violent non-nuclear or nuclear actions is seemingly rare, we still find it happening occasionally. There further does not seem to be a reliably predictable pattern behind the escalation, and hence, technical counter-strategies or deployment limitations are difficult to formulate; this is not acceptable in high-stakes settings like international conflict management, given the potential devastating impact of such actions[125].

Unintentional or accidental escalation is one of many imaginable worst-case scenarios, yet with the proliferation of AI in military technology, it is an increasingly likely possibility. Just like in the case of Stanislav Petrov, this highlights the importance of being able to recognise and intervene when error happens. Yet the rise of AI comes with the loss of human intuition and overreliance on technology, and we are, therefore, becoming increasingly prone to automation bias.

Automation Bias

Automation bias is a phenomenon in which humans over-rely on technology and automated systems and, therefore, neglect to exercise appropriate scrutiny or common sense when analysing data and, therefore, overlook errors[126]. The more advanced technology becomes, the more likely it is to be correct, and therefore, we tend to assume it is infallible, even when we are presented with conflicting evidence.

For instance, if I want to figure out what $7,534.72 \div 36$ is, I will simply type it into my phone's calculator app and instantly get the answer ($209.29777777777777...$), an answer that I am very likely to trust is correct because I believe the calculator is faster and more reliable than

[124] Rivera, et. al., *Escalation Risks from Language Models*, 836–98.
[125] Rivera, et. al., *Escalation Risks from Language Models*, p. 9.
[126] Grut, *The Challenges of Autonomous Lethal Robotics,* 5–23.

my (dismal) mathematical capabilities. However, if, for some reason, my calculator were to make a mistake and say the answer is $209.28777777777777...$, even though I could manually double check the answer and see the error, I would be very unlikely to do this because, for me, it would be difficult, take time, be very mundane, and be extremely unlikely to yield a more reliable result than the calculator, and so I would assuredly overlook the error. I know the calculator is significantly more sophisticated and advanced in its mathematical capabilities than I am, and therefore, I have grown complacent in improving my capabilities and instead rely on calculators for any mathematics more complicated than very basic arithmetic. A calculator, however, is a relatively simple piece of technology that functions in a glass box (with full transparency of linear input, processing, and output), and therefore I *can* double-check the data if I so choose.

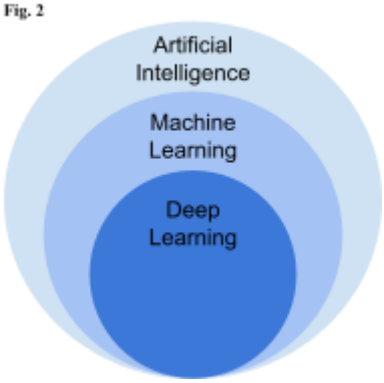
Up until recently, automated systems in military technology were exclusively glass box systems, including the radar and satellite technology used by the Soviets in the 1980s, as well as much more advanced technology used in modern computer-assisted weapons (such as UCAVs)[127]. However, despite having the ability to double-check data from automated systems, automation bias is a common occurrence even with glass box systems due to the extremely fast pace in which warfare is conducted in. Most decisions in active combat are made in extremely time-sensitive conditions where each second is incredibly valuable, and any time wasted could cost lives. Therefore, technology that can make calculations faster than a human brain is extremely valuable and requires high levels of trust and reliability to be of any use. These systems are not, however, infallible, and there have been many instances when human intuition was duped by faulty data due to automation bias. For instance, in 1988, The *USS Vincennes* caused a tragic incident when its computer system mistakenly identified an Iranian passenger flight as an F-14 fighter jet while in semi-automatic mode[128]. Although having full capabilities to intervene and double-check readouts from the computer at any time, due to the perceived danger and intense time constraints, the crew instead decided to trust the computer without question and authorised the plane to be shot down, causing the deaths of all 290 civilians on board, 66 of them children. Had the crew checked the computer data properly, all the information shown would

[127] Holland Michel, *The Black Box, Unlocked*, 1-36.
[128] Grut, *The Challenges of Autonomous Lethal Robotics,* 5–23.

have clearly indicated a passenger plane and not an enemy jet. Although it is easy to say that the events of this aviation disaster were due entirely to negligence from the crew, it is important to note that had it been a fighter jet, the amount of time necessary to double-check the readouts and make the subsequent calculations could have cost them valuable time and possibly their lives. Perhaps what makes this incident so particularly disturbing in terms of automation bias is that the technology they were using at the time was extremely transparent and rudimentary, and had they checked the data, the error would have been obvious. Modern AI in military technology, on the other hand, is much more complex, and this means that even though errors are less likely, they are much more difficult to spot even when the data from the machine is analysed.

What's in the Black Box?

Fig. 2



Artificial intelligence in and of itself is a superset that contains multiple subsets with different levels of complexity, autonomy, and capabilities[129]. Machine learning (ML) refers to supervised, unsupervised, or reinforced systems which can be "trained" using specific datasets, whereby the system will learn and adapt with each new interaction to improve its ability to reach its objective[130] — think of social media algorithms which increasingly tailor content to your preferences the more you engage with it, thereby fulfilling its objective of keeping your attention. However, although the computation of ML systems is very fast and based on a lot of data, most ML systems are considered glass box because we can observe the entire operation from input to output without any hidden layers[131]. In other words, most ML systems show their work, and therefore, it is possible (although not always easy) to find the problem in the case of an unexpected result or a malfunction.

Deep learning (DL) is a subset of machine learning, but it goes one step further where the system does not require human engagement to improve but instead is granted access to vast and

[129] Bostrom, *Superintelligence*, 1-25.
[130] Ibid.
[131] IBM, *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks.*

unstructured datasets to self-improve through repetition[132]. This is done by utilising what is known as an artificial *neural network* designed to mimic the complex structure of the human brain (the 86 billion neurons and 100 trillion synapses mentioned earlier) by interconnecting nodes (based on human neurons) into a multi-layered web whereby data is processed in a nonlinear fashion through visible and hidden layers of algorithms[133]. A DL system is designed to become more accurate over time as it learns through experience, just as humans do. Yet while most ML systems can be classified as glass box, due to the layered nature of neural networks in DL systems, it is only possible to identify the inputs of the first layer of algorithms and the final output layer, yet not the inputs or outputs of hidden layers. Because of this and the complex structure of a neural network, it is entirely impossible to identify the internal operations of a DL system[134] — imagine a nonlinear version of the telephone game where the only information you are privy to is the first person to whisper and the one to announce the result, but you have no idea who is whispering to whom or what they are saying. A DL system is, therefore, known as a black box system, as we cannot discern exactly how it comes to its result. DL is becoming increasingly implemented into all layers of AI in society, including military technology, and along with other black box systems, is being used to analyse large and complex data sets used in image and speech recognition as well as natural language processing[135]. Naturally, such technology can have a revolutionary effect on intelligence and surveillance since it allows for significant amounts of data to be gathered and analysed quickly, yet since we cannot have complete oversight of DL processes, this implementation greatly increases the risks of automation bias in these sectors. Indeed, it inherently *requires* automation bias, as we cannot effectively use such systems unless we are willing to trust it with certain aspects of an operation without full transparency, thereby imbuing it with significant agency and autonomy.

This is of major concern for critics of greater AI in AWS, as it essentially omits significant aspects of the operation of a weapon from the human who is supposedly in control[136]. DL is still relatively underutilised due to the enormous levels of computational power required (which rely on massive server farms that are very costly to run); however, advances in cloud computing and

[132] Ibid.
[133] Ibid.
[134] Bucher, Taina. *If...Then: Algorithmic Power and Politics*.
[135] IBM, *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks.*
[136] Holland Michel, *The Black Box, Unlocked*, 1-36.

quantum computing are expected to massively expedite the development of DL[137]. Most modern AI in military technology utilises ML[138], and while it is unclear how much development has expanded to more advanced DL systems, it is clear from the trajectory of civilian AI research that DL systems can be expected in military technology, including AWS, within the foreseeable future.

Agentic Weapons

If we think back to the earlier example of swarm robotics, such AWS require reinforcement ML algorithms because they operate in such fast-paced and complex dynamic environments that manual human control is not possible[139]. The only role of the operator is to express the objective to the system and press "Go". If the swarm is operating out of range for communication or for an extended period (which is within the autonomous capabilities of swarm robotics), the operator cannot know exactly what the system did until after the fact, when the data can be analysed in a debrief. If a DL reinforcement learning algorithm is used, even that level of control is denied to human operators as crucial information may be within a hidden layer and thus inaccessible. This is a major problem because such a system is designed to "learn and self-improve", yet we have no way of knowing what exactly it may be learning and how it may interact with environments over time. Currently, research on reinforcement learning in swarm robotics has found that there is a lack of reliable control systems in place to oversee larger swarms in complex environments[140]. Yet despite the glaring problems of control, swarm robotics are highly anticipated to be a major part of the global push towards AWS and have already seen use in Ukraine and Gaza[141].

Highly agentic AWS capabilities are not unique to swarm robotics but are seen in the majority of publicly available data and evidence on AWS. For example, the recent and controversial Turkish-made STM Kargu-2 loitering munition is proving challenging to the efficacy of the MHC policy norm. Designed as an anti-personnel weapon, the Kargu-2 is a relatively small, lightweight quadcopter drone similar to commercial photography drones, however, it is equipped

[137] Bostrom, *Superintelligence*, 321-324.
[138] Simmons-Edler, et. al., *AI-Powered Autonomous Weapons*, 1-17.
[139] Blais & A. Akhloufi, *Reinforcement Learning for Swarm Robotics*, 226–56.
[140] Simmons-Edler, et. al., *AI-Powered Autonomous Weapons*, 1-17.
[141] Ibid.

with an explosive warhead which is designed to deploy and detonate near a specifically selected target and is lauded for being so precise that it has extremely low levels of collateral damage in tests even when deployed in crowded areas[142]. While this may seem like a relatively safe weapon for minimising harm to civilians, its autonomous capabilities are what makes it so controversial. While it can be controlled remotely, the drone is capable of operating autonomously and uses DL algorithms to find and select targets based on factors such as facial recognition[143]. However, due to well-documented general problems of bias and profiling with facial recognition software, there is a concern that the Kargu-2 will be unable to correctly distinguish between legitimate military targets and civilians that simply fit a similar profile. As targeting civilians is illegal under IHL, such a mistake would mean the Kargu-2 could have a high potential to carry out war crimes even if this was never the intention of the operator. While STM and the Turkish government insist that precision strike missions are "fully performed by the operator, in line with the Man-in-the-Loop principle"[144], a UN report on the recent Libyan conflict suggests there is strong evidence that the weapon was deployed and killed targeted members of the Haftar Affiliated Forces (HAF) fully autonomously and without direct human authorisation in Libya in 2021[145].

The same principles are also being seen in larger AWS. DARPA has demonstrated autonomous F-16s that are capable of engaging in dogfights against human pilots using reinforcement ML (and possibly DL) algorithms and, therefore, require no human oversight to eliminate targets[146]. Israel Aerospace Industries have developed the BlueWhale ASW, an autonomous anti-submarine warfare vessel which can operate for up to a month with limited to no surface communication and, using ML, can intercept and engage targets. While it is unclear if it has ever done so while in autonomous mode, it has been used in NATO exercises since 2023[147]. As the destructive capabilities of such large and powerful weapons systems are well documented historically, one does not need too much imagination to consider what could go wrong if control is lost.

[142] STM, *Kargu®: Combat Proven Rotary Wing Loitering Munition System.*
[143] Ferl, *Imagining Meaningful Human Control*, 139-155.
[144] STM, *Kargu®: Combat Proven Rotary Wing Loitering Munition System.*
[145] United Nations Security Council, *Letter dated 8 March 2021.*
[146] Simmons-Edler, et. al., *AI-Powered Autonomous Weapons*, 1-17.
[147] Ibid.

Although most militaries in possession of such systems deny that they are ever used in fully autonomous mode, the level of sophistication of the AI capabilities utilised in such systems clearly shows that if AWS are not yet the acting agent in combat, they will be very soon. Indeed, despite MHC, there is a clear enthusiasm for the full application of autonomous capabilities in military technology before 2030[148]. Today, the list of major military powers which are publicly planning for and developing significant AI autonomy in military technology includes the US, China, Russia, Israel, the UK, France, and India[149], yet no one seems to agree on how to do this without losing MHC.

While the German delegation at the 2018 GGE was incorrect in their assessment of AWS having "self-awareness", the idea of an uncontrollable weapon with agency is no longer science fiction. They exist, they are almost certainly being used, and in practice, they are ungoverned.
What, then, if anything, can be done to effectively govern current and future AWS?

Normative Weapons Governance in Times of Change

Having looked at the technical and practical problems of controlling AI in military technology, it is becoming increasingly clear that the policy of MHC is impractical, unenforceable, and unrealistic. At the same time, it is also clear that we cannot trust the technology with the increasing levels of autonomy required for most modern AI in the operation of AWS, and this problem is likely to worsen over time unless the Control Problem of AI is solved. Despite this, the development, testing, and deployment of AWS is ramping up with greater levels of autonomy and agency than ever before. It is clear that there is a desperate need for effective governance of AWS *before* they gain prominence on the battlefield. Time, however, is running out very fast. Indeed, many believe that the time for effective preemptive governance *has* run out and that the task of governance of AWS is now in the realm of damage control[150]. While it is certainly too late to un-invent AWS, they are still a very small part of the overall global military arsenal, and therefore, it is not (yet) too late to intercept the proliferation of AWS.

[148] United States of America: Department of Defence, *Unmanned Systems Integrated Roadmap*, p. 47.
[149] Haner & Garcia, *The Artificial Intelligence Arms Race*, 331-337.
[150] Simmons-Edler, et. al., *AI-Powered Autonomous Weapons*, 1-17.

Yet coming up with alternative policies to govern AWS effectively is a monumental task, and this is due to the complexities of AI as well as challenges to standardised practices in normative weapons governance. Modern weapons governance is situated within the normative structure of IHL, and the debate on AWS from a policy perspective is held at the same conventional focus as weapons such as biological and chemical weapons, which have been the primary examples of successful weapons governance[151]. However, this take on governing AWS fails in two crucial aspects. First, as I have shown throughout this thesis, the basis of MCH under the framework of IHL is insufficient to adequately address the problems of control unique to AWS, and second, such weapons governance overemphasises the governing power of deliberate and fundamental norms concerning new technology in times of shifting world order without paying enough attention to the establishment of procedural norms through what is essentially ungoverned practice[152]. These normative issues of weapons governance make the prospect of governing AWS effectively in the near future seem pretty bleak, and by looking at examples of weapons governance in the past, it is increasingly clear that AWS is falling into the same problems of unenforceability.

Governance Without Government

Within the field of IR, global governance is a very heated topic of debate. While *government* within nation-states relies on laws and regulations which can be enforced by the body politic, global *governance* has no such practical authority. Instead, global governance is derived from mutual agreement and cooperation of shared goals within the international community of state and non-state actors[153]. In the seminal 1992 work *Governance without Government*, James Rosenau and Ernst-Otto Czempiel observed the end of the Cold War as a turning point for world order and global governance. They highlighted the differences between government and governance, stating that:

> Governance is a system of rule that works only if it is accepted by the majority (or, at least, by the most powerful of those it affects), whereas governments can function even in the face of widespread opposition to their policies.[154]

[151] Ferl, *Imagining Meaningful Human Control*, 139-155.
[152] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[153] Chhotray & Stoker, *Governance Theory and Practice*, 76-98.
[154] Rosenau, *Governance Without Government,* p. 4.

Effective governance is therefore reliant on order, and for order to occur, there must be mutual agreement and cooperation through the emergence and acceptance of norms[155]. This has led to debate on whether there is such a thing as binding international law in practice or if it is simply a compilation of norms which have no real practical bearing or enforceable punishment if an actor chooses to disobey them.

In 1991, with the collapse of the Soviet Union, the world breathed a collective sigh of relief as the decades-long nuclear sabre-rattling finally came to an end, and there was an atmosphere of hopefulness for the future. Liberal internationalism and its associated norm set had triumphed, and it seemed there was little that could not be achieved through peaceful discussion and deliberation[156]. It is, therefore, not so surprising that from the early 1990s, global governance framed within the liberal international order and its norms of liberal internationalism seemed to be highly effective at first. Three decades on, however, we are once again at a turning point for world order, yet this time, it is the liberal international order that is being challenged[157]. This is causing a shift in global governance, as norms thought universally accepted are now being increasingly contested and losing their salience, resulting in international law being ignored and undermined, which is causing global governance to be increasingly ineffective. It is not entirely clear when this shift began to occur; some say the cracks began already with the NATO bombardment of Kosovo in 1999 or the 9/11 attacks in 2001, while others argue it began with the global financial crisis in 2008 or with Vladimir Putin's annexation of Crimea in 2014[158]. Certainly, with the full-scale Russian invasion of Ukraine in February 2022, it was widely acknowledged that the existing order is in transformation[159]. Regardless of the exact timeline, we are starting to see the effects of the failure of liberal internationalism as an effective form of normative global governance. This failure has unfortunately coincided in tandem with the AI boom and the development, testing, and deployment of modern AWS. With AWS entering the international agenda at the same time as the fundamental norms of global governance are being contested, there are now extremely pressing concerns about whether AWS can be effectively governed from a policy perspective or whether it is too little, too late. Indeed, there is historical

[155] Chhotray & Stoker, *Governance Theory and Practice*, 76-98.
[156] Fukuyama, Francis. *The End of History?*
[157] Flockhart, *From 'Westlessness' to Renewal of the Liberal International Order*, 42-59.
[158] Ibid.
[159] Flockhart & Korosteleva, *War in Ukraine*, 466-481.

precedence for this concern, for normative weapons governance has, as of yet, been largely unsuccessful in banning or regulating new forms of weaponry during times of shifting world order until after they have been used in practice[160]. This is most prominently seen with the failure to preemptively ban chemical and biological weapons until after the insecurity of WWI had subsided, and even then, it took decades to achieve effective governance[161]. Therefore, to tackle the challenges specific to governing AWS, we must understand exactly how the legal and normative structures of weapons governance were established in principle and why this fails in practice in light of emerging procedural norms during times of heightened global insecurity.

Foundational Norms of IHL

As discussed at the start of this thesis, the current policy of MHC is generally understood under the framework of IHL, and this is due to a long tradition of weapons governance being based on the principles and structures of international law solidified in the United Nations Charter since its conception in 1945[162]. Yet the roots of weapons governance come from a normative history of governing war that stretches all the way back to the Middle Ages. Stemming from the ideals of Just War Theory, normative weapons governance is concerned primarily with the moral and ethical justifications for going to war (*Jus ad Bellum*) and the morality of conduct in warfare (*Jus in Bello*)[163]. Despite the vast differences between medieval and modern warfare, normative ethics of right and wrong at the heart of modern-day IHL remain largely unchanged.

Today, the main focus of IHL are the principles of *Jus in Bello,* which is based on the moral reasoning that those engaged in warfare should do all within their power to minimise the suffering of armed conflict and not to engage in conduct that causes unnecessary harm, suffering, or indignity to fellow human beings regardless of whether they be friend or foe[164]. In practice, this is divided into two principles determining legitimacy, which armed forces and combatants are legally required to adhere to when engaging military targets: the principle of *distinction* and the principle of *proportionality*[165]. The principle of distinction states that only combatants and

[160] Ferl, *Imagining Meaningful Human Control*, 139-155.
[161] United Nations Office for Disarmament Affairs, *Chemical Weapons*.
[162] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[163] International Committee of the Red Cross, *Jus Ad Bellum and Jus in Bello.*
[164] Ibid.
[165] Ibid.

military objectives may be legitimate targets and that targeting civilians or civilian objects is unlawful. The principle of proportionality states that collateral damage from engaging a legitimate target may not be in excess of the proportional military advantage gained. In other words, conducting a targeted airstrike on a known enemy combatant is legal under IHL, but not if that combatant is situated within a civilian building such as a hospital or school.

Yet despite the legal structure of IHL, weapons governance relies in large measure on norms rather than law. As norms can be understood as shared expectations of appropriate behaviour within a given social context, it follows that weapons governance is, in practice, unenforceable if any actors decide not to follow the norms and to disobey IHL[166]. Put simply, the problem is that although, in principle, weapons governance is founded in law, it is practised through norms. When world order is unstable (as it arguably is now), these norms are contested and perhaps lacking salience and, therefore, meaningless in a practical sense. This issue is currently on display with the International Criminal Court (ICC) recently issuing arrest warrants for both Benjamin Netanyahu and Vladimir Putin for war crimes in Gaza and Ukraine, respectively[167], yet realistically, the ICC and ancillary international organisations have no practical authority to apprehend and try them. A focus on weapons governance through IHL alone, therefore, reveals only the norms of weapons governance that have been deliberated on and solidified as fundamental norms (such as universal human rights) and not the norms which emerge outside of deliberations.

Establishing Procedural Norms of Weapons Governance

Ingvild Bode and Hendrik Huelss argue that the focus on fundamental norms is at the centre of the failure to adequately understand AWS from a governance perspective and that a more informative approach would be to focus on the norms which emerge through practice. They argue that these norms shape acceptable standards to which states conform more so than the fundamental norms of right and wrong at the centre of IHL[168]. This idea is rooted in practice theory of IR and challenges the constructive approach to normative weapons governance that is prevalent in academia and policy. As put by Bode and Huelss:

[166] Finnemore & Sikkink, *International Norm Dynamics and Political Change*, 887–917.
[167] International Criminal Court. *Arrest Warrant.*
[168] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.

> Existing research on AWS asks some central questions: can or do AWS conform to international law? How do AWS affect fundamental norms embedded in international law, such as human rights? How do autonomous technologies relate to expectations of political accountability and responsibility? Should AWS be preemptively banned? But research neglects that normativity may already emerge from their development, testing and deployment on the procedural, practical level. How AWS may set novel understandings of appropriate action remains understudied[169].

While relatively understudied, this approach is highly informative when looking at the history of weapons governance to gain insight into what to expect from governance of new weapons today.

The history of weapons governance is made up of practices that shape norms which become accepted and solidified over time. Such norms are known as procedural norms. It is very rare that norms of governance emerge through deliberation and agreement alone, but it is instead usually a reaction to rapid changes in the world that have previously been unregulated, often during periods of heightened insecurity and uncertainty[170]. Therefore, there is practically no precedent for weapons technology to be effectively governed preemptively, as new weapons technology is at first perceived by actors as a means to heighten their security[171]. This is proving to be the case for AWS, for although it has been on the public agenda for over a decade, there appear to be no serious attempts at ensuring MHC is maintained in the technology itself, and instead, we are seeing a new arms race with states scrambling to develop their own AWS.

This pattern of normative governance fails to tackle the realities of weapons technology and attitudes in development because policymakers often do not recognise the formative aspects of procedural norms concerning new weapons technology until long after they have been established. While governance of new weapons is usually achieved in some form eventually, by this time, it has been shaped not by deliberated fundamental norms discussed in forums like the GGE but by the procedural norms brought about by practice, which becomes palatable and accepted over time as more actors engage in the same practice. These practices are what Bode and Huelss refer to as *standards of procedural appropriateness*[172]. Any deliberation and

[169] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, p. 20.
[170] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[171] Ferl, *Imagining Meaningful Human Control*, 139-155.
[172] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1.

subsequent governance usually lag far behind and almost always after the technology has already been used in practice. Put simply, the practice comes first, the governance after.

For instance, if we consider the chain of events that led to the foundations of IHL and modern weapons governance, we can see that what we today consider to be legally binding laws of war came about through countless iterations of trial and error after major changes to world order. Before WWI, there had been only a handful of attempts to govern weaponry based on the moral principles of *Jus in Bello*, none of which were multilaterally successful[173]. However, as science and technology of the 19th and 20th centuries became increasingly advanced, so too did the capabilities for inducing great harm, suffering, and indignity when such technology was used in war. With tensions rising, there were some attempts at governing the new weapons technology, but largely due to the ignorance of the destructive capabilities of modern weapons, these attempts were lukewarm at best. The St. Petersburg Declaration of 1868 and the Hague Conventions of 1899 and 1907 attempted to prohibit asphyxiating gasses and other new weapons technology. However, these attempts all failed to gain significant traction and were therefore rejected outright or ignored in practice by signatories[174].

This came to a head with the outbreak of WWI. Horrified by how cruel and destructive much of the conduct of the war was, there was a shift in normative thinking. Where before the war, states had been driven by insecurity in a changing world order to develop these destructive weapons despite their horrific capabilities, after the war, the insecurity lay in the use and existence of the weapons themselves. The general multilateral consensus that some weapons were simply too cruel, unnecessary, or indiscriminate to justify — even when used on enemies — became the new norm, which in turn shaped governance. Chief among the weapons in question were chemical and biological weapons, and this eventually led to The Geneva Protocol in 1925[175],[176].

[173] Ferl, *Imagining Meaningful Human Control*, 139-155.
[174] ICRC, *Hague Conventions: How does law protect in war?*
[175] United Nations Office for Disarmament Affairs, *1925 Geneva Protocol*.
[176] Officially titled The Protocol for the Prohibition of the Use of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare

To this day, the Geneva Protocol remains the longest-lasting piece of international weapons legislation and is considered the cornerstone of modern weapons governance[177]. However, even this gold standard of weapons governance was a reactionary measure *after* chemical and biological weapons had already killed and maimed hundreds of thousands of people indiscriminately during WWI[178]. Preemptive governance, such as the Hague Conventions, was summarily rejected by most major states, and even those who ratified the conventions immediately violated them during WWI since it was considered against their self-interest to not utilise new weapons technology that enemies were likely to use on them[179]. Despite the clear dangers of such weapons, the standards of procedural appropriateness during the pre-WWI instability were to develop them anyway, and thus, developing biological and chemical weapons became the prevailing norm.

It was only after the genie was let out of the bottle that biological and chemical weapons were taken seriously, and by that time, any governance based on fundamental norms that such weapons are morally "wrong" was unable to keep up with the standards of procedural appropriateness that were, by the time of the Geneva Protocol, decades old. The language used in the original 1925 Protocol was ambiguous and open to interpretation and made no mention of the development and stockpiling of chemical and biological weapons, nor of usage other than in warfare between states who were party to the Protocol, nor as use in retaliation[180]. These shortcomings led to a surge in practices of developing and stockpiling chemical and biological weapons, many of which were deployed by states that had signed and ratified the Protocol but were able to argue that they technically were not in violation[181]. Most notable of such cases, perhaps, is the United States' extensive use of indiscriminate chemical warfare in Vietnam, yet although the US had signed the Protocol, they were not party to it until they eventually ratified it in 1975 after the Vietnam War had ended[182]. Indeed, it was not until 1993, with the Chemical Weapons Convention (CWC), that both biological and chemical weapons were finally banned, and by this time, they had resulted in the deaths of over a million people[183].

[177] United Nations Office for Disarmament Affairs, *History of the Biological Weapons Convention*.
[178] United Nations Office for Disarmament Affairs, *Chemical Weapons*.
[179] ICRC, *Hague Conventions: How does law protect in war?*
[180] United Nations Office for Disarmament Affairs, *1925 Geneva Protocol*.
[181] United Nations Office for Disarmament Affairs, *History of the Biological Weapons Convention*.
[182] United Nations Office for Disarmament Affairs, *Chemical Weapons*.
[183] Ibid.

For biological and chemical weapons, it took nearly 70 years for normative governance to take effect. This begs the question of how long it will take to govern AWS and how much time we actually have.

Are We Too Late?

When looking at the normative problems with governing chemical and biological weapons, it is hard not to see similarities with the attempts at governing AWS. Indeed, the ambiguous language of the original Geneva Protocol that allowed 70 years of weapons development, stockpiling, and even use has eerily similar connotations to the issues surrounding the vague policy of MHC today. While it cannot be understated that it is a monumental achievement that chemical and biological weapons have not been used en masse since WWI, this is likely more so the result of collective trauma than of effective governance. Indeed, the mere existence of chemical and biological weapons significantly increased the risk of global catastrophe, even if this was never the intent. The same trend applies to nuclear weapons, which were used to kill hundreds of thousands of civilians in Hiroshima and Nagasaki before any form of governance was even on the table, and even then, are to this day developed and proliferated to levels far exceeding world-ending capabilities[184]. While it is likely we will get some form of governance and perhaps even a ban of AWS someday, the focus now needs to be on getting this done from a cautious perspective of the future rather than a regretful trauma response of the past. With our current trajectory of governing AWS, however, it is looking increasingly likely that we will repeat the mistakes of the past.

The continued use of fundamental norms of IHL as the gold standard for weapons governance is completely ignoring the emerging procedural norms brought about by an increasingly unstable world order and rapidly changing technological capabilities that is driving states to develop AWS despite the risks. Before WWI, chemical and biological weapons were judged using outdated fundamental norms based on the capabilities of outdated weapons, and this led to a lack of caution by world leaders. Today, we are again basing weapons governance on outdated

[184] Tannenwald, *The Nuclear Taboo,* 433–68.

fundamental norms — in this case, norms concerning humans being in control[185]. Policymakers need to understand that given the unique technological properties of AWS, this standard is no longer realistic or relevant. From a technological standpoint, AWS is completely different from biological and chemical weapons (and other weapons of the past), and therefore, trying to fit AWS into an existing framework of normative weapons governance is to allow new standards of procedural appropriateness to emerge, ungoverned.

The reality is that since the seminal Human Rights Watch report first called for a ban in 2012, there has been no significant action taken to ensure that AWS does not join the growing list of cautionary tales concerning devastating weapons technology[186]. This is not exactly surprising. Just as the crumbling empires of late 19th century Europe saw little value in minimising their destructive arsenal while their enemies were building theirs up, major military powers today are in no rush to fall behind in building their arsenal of AWS. The fact of the matter is that despite the risks of having no control, the allure of low-cost, minimal personnel automated warfare is simply too great for most states to pass up, especially at a time when most are perceived in some state of insecurity in our shifting world order. So is that it? Are we too late to stop AWS in its tracks?

Conclusion

This thesis set out to explore what it means to have control of AWS in the era of AI and whether it is ever possible to effectively govern such technology using our current normative approach to weapons governance. Using a critical analysis of the current attempts to govern AWS regarding the technical realities of controlling AI and the practical realities of giving trust and agency to military technology, I showed that there is currently no possible way to ensure human control (meaningful or not) of a weapon system that is operated using modern AI. The first part of the question — what it means to have control of AWS — therefore has a simple answer. The second part of the question — whether governance is possible — is much more complicated. The frustrating answer is that we can't really know for sure until we know for sure. However, at the risk of ending this thesis on an optimistic note, while AWS are starting to creep into active

[185] Bode & Huelss, *Autonomous Weapons Systems and Changing Norms*, 1-28.
[186] Ibid.

deployment, the worst has not yet happened, and therefore there is still a small chance to stop things from getting out of hand. It will, however, require some serious work and sacrifice, and it is required now.

First, policymakers need to stop focusing on the fundamental norms of IHL, abandon the policy of MHC, and instead pay attention to the procedural norms being shaped by the technological development and deployment of AWS that is happening right now. While fundamental norms of moral principles situated in ideals of right and wrong are important for humanity as a whole, they are simply not suitable for informing effective policy involving such complicated technology. By observing changing procedural norms in conjunction with the realistic technological capabilities and limitations of AI, as well as acknowledging instabilities in changing world order, policymakers may be able to gain an honest overview of what can be expected with the proliferation of AWS. While this will not change the enforcement limitations of global governance, it may be successful in shifting perceptions of AWS into more realistic realms and thereby address the problems of control that are currently being ignored. In doing so, there is a chance that standard practices currently deemed appropriate and acceptable may change in light of the risks associated with the proliferation of AWS. After all, this was what eventually happened with chemical and biological weapons, which led to a ban.

Second, AI developers and researchers need to stop beating about the bush and be honest with policymakers — There is no way to safely use AWS, and unless the Control Problem is solved, there never will be. While risks might be minimised the more advanced AI becomes, there is no way around the fact that standard model AI is, in practice, an agent rather than a tool, and therefore, we can never fully trust AI to serve our best interests. Therefore, the only effective policy measure to govern AWS is to ban them, and to ban them before they proliferate. While the risks of AWS will always be present now that they have been invented, they can be minimised if there is a global collective agreement that using AWS is simply not worth the risk. While chemical, biological, and nuclear weapons were unfortunately given time to proliferate, kill indiscriminately, and heighten the risks of global catastrophe, autonomous weapons systems have not yet had this time. Let's make sure we don’t give it to them.

Bibliography

"1925 Geneva Protocol." United Nations Office for Disarmament Affairs. Accessed December 17, 2024. https://disarmament.unoda.org/wmd/bio/1925-geneva-protocol/.

Amoroso, Daniele, and Guglielmo Tamburrini. "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues." Current Robotics Reports 1 (August 2024): 187–94. https://doi.org/10.1007/s43154-020-00024-3.

Anduril Industries. "Anduril Partners with OpenAI to Advance U.S. Artificial Intelligence Leadership and Protect U.S. and Allied Forces." December 4, 2024. https://www.anduril.com/article/anduril-partners-with-openai-to-advance-u-s-artificial-intelligence-leadership-and-protect-u-s/.

Article 36. "Structuring Debate on Autonomous Weapons Systems." *Memorandum for Delegates to the Convention on Certain Conventional Weapons (CCW)*, Geneva, November 14–15, 2013.

Beall, Mark, and Stuart Russell. "Autonomous Weapons and Human Control: Shaping AI Policy for a Secure Future." Moderated by Jason Green-Lowe. Recorded Webinar, July 29, 2024. Posted July 30, 2024, by the Center for AI Policy. YouTube. 1:05:05. https://www.youtube.com/watch?v=8GXsPVqfWoo.

"Biological Weapons." United Nations Office for Disarmament Affairs. Accessed December 19, 2024. https://disarmament.unoda.org/biological-weapons.

Blais, Marc-André, and Moulay A. Akhloufi. "Reinforcement Learning for Swarm Robotics: An Overview of Applications, Algorithms and Simulators." *Cognitive Robotics* 3 (2023): 226–56. https://doi.org/10.1016/j.cogr.2023.07.004.

Bode, Ingvild, and Hendrik Huelss. "Autonomous Weapons Systems and Changing Norms in International Relations." *Review of International Studies* 44, no. 3 (February 2018): 393–413. https://doi.org/10.1017/S0260210517000614.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 978-0-19-873983-8.

Bucher, Taina. *If...Then: Algorithmic Power and Politics*. Oxford Studies in Digital Politics, 2018. https://doi.org/10.1093/oso/9780190493028.001.0001.

Caron, Jean François. “Defining Semi-Autonomous, Automated, and Autonomous Weapon Systems in Order to Understand Their Ethical Challenges.” *Digital War* 1 (December 2020): 173–77. https://doi.org/10.1057/s42984-020-00028-5.

Caruso, Catherine. “A New Field of Neuroscience Aims to Map Connections in the Brain.” Harvard Medical School, December 12, 2024. https://hms.harvard.edu/news/new-field-neuroscience-aims-map-connections-brain.

“Chemical.” United Nations Office for Disarmament Affairs. Accessed December 19, 2024. https://disarmament.unoda.org/chemical/.

“Chemical Weapons.” United Nations Office for Disarmament Affairs. Accessed December 19, 2024. https://disarmament.unoda.org/wmd/chemical/.

Chhotray, Vasudha, and Gerry Stoker. “Governance and International Relations.” In *Governance Theory and Practice: A Cross-Disciplinary Approach*. Palgrave Macmillan, 2009. ISBN 978-0-230-58334-4 (eBook).

Davison, Niel. “A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law.” *UNODA Occasional Papers*, no. 30 (November 2017): 5–18.

Dickson, Ben. “State Dept-Backed Report Provides Action Plan to Avoid Catastrophic AI Risks.” VentureBeat, March 11, 2024. https://venturebeat.com/ai/action-plan-long-in-the-making-pro.

“El Capitan: NNSA’s First Exascale Machine.” Advanced Simulation and Computing. Accessed December 4, 2024. https://asc.llnl.gov/exascale/el-capitan.

Ferl, Anna-Katharina. “Imagining Meaningful Human Control: Autonomous Weapons and the (De-)Legitimisation of Future Warfare.” *Global Society* 38, no. 1 (July 9, 2023): 139–55. https://doi.org/10.1080/13600826.2023.2233004.

Flockhart, Trine. “From ‘Westlessness’ to Renewal of the Liberal International Order: Whose Vision for the ‘Good Life’ Will Matter?” *Resilient Communities of Central Eurasia*, January 24, 2023, 42–59. https://doi.org/10.4324/9781003299998-3.

Flockhart, Trine, and Elena A. Korosteleva. “War in Ukraine: Putin and the Multi-Order World.” *Contemporary Security Policy* 43, no. 3 (June 2022): 466–81. https://doi.org/10.1080/13523260.2022.2091591.

Finnemore, Martha, and Kathryn Sikkink. “International Norm Dynamics and Political Change.” *International Organization* 52, no. 4 (1998): 887–917. http://www.jstor.org/stable/2601361.

Fukuyama, Francis. *“The End of History?”: And Related Articles*. Washington, DC: National Interest, 1989.

“Group of Governmental Experts.” United Nations Office for Disarmament Affairs. Accessed September 15, 2024. https://disarmament.unoda.org/group-of-governmental-experts/.

Grut, Chantal. “The Challenges of Autonomous Lethal Robotics to International Humanitarian Law.” *Journal of Conflict & Security Law* 18, no. 1 (April 2013): 5–23. https://doi.org/10.1093/jcsl/krt002.

“Hague Conventions.” Hague Conventions | How does law protect in war? - Online casebook. Accessed January 3, 2025. https://casebook.icrc.org/a_to_z/glossary/hague-conventions.

Haner, Justin, and Denise Garcia. “The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development.” *Global Policy* 10, no. 3 (September 2019): 331–37. https://doi.org/10.1111/1758-5899.12713.

Harris, Edouard. “Edouard Harris - New Research: Advanced AI May Tend to Seek Power *by Default*.” Hosted by Jeremie Harris. Podcast Interview (video format). Posted October 12, 2022, by Towards Data Science. YouTube. 58 min., 22 sec. https://www.youtube.com/watch?v=dYSw-SV_fsI.

“History of the Biological Weapons Convention.” United Nations Office for Disarmament Affairs. Accessed December 17, 2024. https://disarmament.unoda.org/biological-weapons/about/history/.

Holland Michel, Arthur. “The Black Box, Unlocked: Predictability and Understandability in Military AI.” Geneva, Switzerland: United Nations Institute for Disarmament Research, 2020. https://doi.org/10.37559/SecTec/20/AI1.

IBM Data and AI. “AI vs. Machine Learning vs. Deep Learning vs. Neural Networks.” IBM, November 25, 2024. https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

Josh Darnit. “Exact Instructions Challenge - THIS is why my kids hate me. | Josh Darnit.” Video. Posted January 26th, 2017. YouTube. 6:45. https://www.youtube.com/watch?v=cDA3_5982h8

"Jus Ad Bellum and Jus in Bello." International Committee of the Red Cross, June 25, 2024. https://www.icrc.org/en/law-and-policy/jus-ad-bellum-and-jus-bello.

"Kargu®: Combat Proven Rotary Wing Loitering Munition System." STM. Accessed September 28, 2024. https://www.stm.com.tr/en/kargu-autonomous-tactical-multi-rotor-attack-uav.

Kleinman, Stella. "Stanislav Petrov." *Britannica*. Updated November 20, 2024. https://www.britannica.com/biography/Stanislav-Petrov.

"Lethal Autonomous Weapons Systems (LAWS)." United Nations Office for Disarmament Affairs. Accessed August 6, 2024. https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/.

Luhmann, Niklas. "Chapter 3: Double Contingency." In *Social Systems*, translated by John Bednarz, Jr., with Dirk Baecker. Stanford University Press, 1995. (Original German language publication 1984).

OpenAI. *GPT-4 System Card*, (2023): 41-100. https://cdn.openai.com/papers/gpt-4-system-card.pdf

Permanent Representation of the Federal Republic of Germany to the Conference on Disarmament in Geneva, "Statement delivered by Germany on Working Definition of LAWS / 'Definition of Systems under Consideration'" *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,* Geneva, p 1- 2. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Germany.pdf.

Rivera, Juan-Pablo, Gabriel Mukobib, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. "Escalation Risks from Language Models in Military and Diplomatic Decision-Making." In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (June 2024): 836–98. https://doi.org/10.1145/3630106.3658942.

Robbins, Scott. "The Many Meanings of Meaningful Human Control." *AI Ethics* 4 (July 2023): 1377–88. https://doi.org/10.1007/s43681-023-00320-6.

Robinson, Tim, and Stephen Bridgewater. "Highlights from the RAeS Future Combat Air & Space Capabilities Summit." Royal Aeronautical Society, May 23–24, 2023.

https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities-summit/.

Rogers, James, and Dominika Kunertova. “The Vulnerabilities of the Drone Age: Established Threats and Emerging Issues Out to 2035.” In *The NATO Science for Peace and Security Programme*, June 2022. https://doi.org/10.3929/ethz-b-000556165.

Rosenau, James N. “Governance, Order, and Change in World Politics.” In *Governance Without Government: Order and Change in World Politics*, edited by James N. Rosenau and Ernst-Otto Czempiel. Cambridge University Press, 1992. (Online publication October 2009).

Russell, Stuart. “Banning Lethal Autonomous Weapons: An Education.” *Issues of Science and Technology* 38, no. 3 (Spring 2022): 60–65.

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking: Penguin Random House, 2019 [2020]. ISBN 978-0-525-55863-7.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of Warfare*. W.W. Norton & Company, 2018. ISBN 978-0-393-35658-8.

Scharre, Paul. “Autonomous Weapons and Operational Risk: Ethical Autonomy Project.” *Center for a New American Security*, (2016): 1-54.

Simmons-Edler, Riley, Ryan Badman, Shayne Longpre, and Kanaka Rajan. “AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research.” In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. https://doi.org/10.48550/arXiv.2405.01859.

“Situation in the State of Palestine: ICC Pre-Trial Chamber I Rejects the State of Israel’s Challenges to Jurisdiction and Issues Warrants of Arrest for Benjamin Netanyahu and Yoav Gallant.” International Criminal Court. Accessed January 5, 2025. https://www.icc-cpi.int/news/situation-state-palestine-icc-pre-trial-chamber-i-rejects-state-israels-challenges.

“Situation in Ukraine: ICC Judges Issue Arrest Warrants against Vladimir Vladimirovich Putin and Maria Alekseyevna Lvova-Belova.” International Criminal Court. Accessed January 5, 2025. https://www.icc-cpi.int/news/situation-ukraine-icc-judges-issue-arrest-warrants-against-vladimir-vladimirovich-putin-and.

Slijper, Frank, Alice Beck, Daan Kayser, and Maaike Beenes. “Don’t Be Evil: A Survey of the Tech Sector’s Stance on Lethal Autonomous Weapons.” *Pax for Peace*, 2019. ISBN 978-94-92487-44-5.

Stryker, Cole. “What Is Artificial Intelligence (AI)?” IBM, December 19, 2024. https://www.ibm.com/think/topics/artificial-intelligence.

Taddeo, Mariarosaria, and Alexander Blanchard. “A Comparative Analysis of the Definitions of Autonomous Weapons Systems.” *Science and Engineering Ethics* 28, no. 37 (August 2023). 2-22. https://doi.org/10.1007/s11948-022-00392-3.

Tannenwald, Nina. “The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use.” *International Organization* 53, no. 3 (1999): 433–68. http://www.jstor.org/stable/2601286.

“Timeline of LAWS in the CCW.” United Nations Office for Disarmament Affairs. Accessed August 6, 2024. https://disarmament.unoda.org/timeline-of-laws-in-the-ccw/.

Turner, Alexander Matt, and Prasad Tadepalli. “Parametrically Retargetable Decision-Makers Tend To Seek Power” *36th Conference on Neural Information Processing Systems* (NeurIPS 2022).

United Nations Security Council, “Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council,” 8th March, 2021.

United States Department of Defence, “The Unmanned Systems Integrated Roadmap FY2011-2036”.

United States Department of Defence, “Directive 3000.09”, November 21, 2012 *Incorporating Change 1, May 8, 2017.*

“What Are Israel’s Iron Dome, David’s Sling, Arrow and THAAD Missile Defences?” *BBC News,* October 16, 2024. https://www.bbc.com/news/world-middle-east-20385306.

Williams, Tristan. “Summary: Autonomous Weapons” *Center for AI Policy* (July 2024): 1-14.

Work, Robert O. “A Short History of Weapon Systems with Autonomous Functionalities: Principles for the Combat Employment of Weapon Systems with Autonomous Functionalities”. *Center for a New American Security,* (2021): 5-7. http://www.jstor.org/stable/resrep32146.4.