Viewpoint

An Ethics Framework for Autonomous Weapon Systems

By Professor Peter Lee

Biography: Peter Lee, Professor of Applied Ethics, is the Director, Security and Risk Research and Innovation, at the University of Portsmouth. His research interests span the politics and ethics of war, the ethical and other human aspects of RPAS operations in military, policing and wider security contexts, and the ethics of autonomous weapon systems. In 2016 he was granted unprecedented research access to the two RAF Reaper squadrons for his book, *Reaper Force: Inside Britain's Drone Wars* (October 2018). He is currently an Expert Adviser to the UK All Party Parliamentary Group on Drones. From 2008 to 2017 he taught Ethics and Law of War at Royal Air Force College Cranwell for King's College London and the University of Portsmouth respectively. Peter holds a PhD in War Studies from King's College London which explores the emergence of Western war ethics. From 2001 to 2008 he served as a Royal Air Force chaplain.

Abstract: This essay draws on extensive personal engagement in numerous national and international events and discussions, to explore some of the key ethical challenges presented by the development and deployment of autonomous weapon systems (AWS). Some of these ethical reflections are process driven, while others are outcome focused. I sketch out a technological, political and operational landscape, bringing together several of the elements that must combine to provide AWS capability within an ethical framework. Practically, each of these elements is hugely complex; this paper can only hope to provide an overview of the challenges, rather than an in-depth analysis. Whilst distinct legal questions will also be raised by these elements, the priority here is to outline an AWS ethics framework for current and future discussion. Throughout, I assume that war ethics are *comparative* – making better or worse choices – rather than *ideal* – making simple choices between good and evil – in complex, seemingly impossible situations.

Introduction[1]

Since 2011 the main strand of my research and writing has addressed the ethical aspects of remote air warfare through the lens of the RAF Reaper Force. In parallel, I have participated as an ethicist in numerous national and international discussions and events concerned with rapid technological advances towards autonomous weapon systems (AWS). Visualizing the breadth and complexity of the many challenges of AWS is essential in order to address the ethical considerations at scientific, policy and operational levels. Practically, each of these elements is hugely complex so, in working towards an ethics framework for AWS, this paper can only provide an overview rather than an in-depth analysis of several constituent parts.

In 2014 the International Committee of the Red Cross convened an Expert Meeting on 'Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects' and has hosted regular discussions since.[2] Separately, in 2017 the UN convened a 'Group of Governmental Experts related to emerging technologies in the area of lethal autonomous weapons systems (LAWS)'.[3] However, any discussion of AWS is problematic, for several reasons, across multiple fields. These include: the difficulty of describing and operationalising autonomy; the different types and purposes of AI that are needed to enable autonomy, including predictability; human-machine system working and robustness; future communications which rely on nascent quantum entanglement capabilities; and human and machine bias. Whilst each of these also raises legal questions, my emphasis below is on ethical considerations. In addition, new ethical considerations are added to already-disputed ethics of war. Just War theory has, for centuries, challenged two opposing approaches to war: *realism* and the unconstrained use of force in the pursuit of power on the one hand, and on the other, *pacifism*, which rejects all use of military force and the killing involved as inherently wrong.[4] Like any weapon, AWS could also be used outside the domain of war, but this paper focuses on its use in international armed conflicts and non-international armed conflicts.

To provide the basis of an outline ethics framework for AWS, this paper conceptualises a future lethal autonomous aircraft system (LAAS) – popularly referred to as lethal autonomous drones – by taking the MQ-9 Reaper Remotely Piloted Aircraft System as a starting point and considering the implications of how a future variant might operate when elements of its functionality are delivered autonomously using AI. In the subsequent sections, the following key terminological, theoretical and other challenges will be outlined: the difficulty of describing and operationalising autonomy; the different types and purposes of AI that are needed to enable autonomy, including predictability; human-machine system working and robustness; future communications which rely on nascent quantum entanglement capabilities; and human and machine bias. The paper concludes by outlining an ethics framework for AWS under four headings: overarching ethics of war, human responsibility, the methods and means of war, and risk assessment and mitigation. Whilst the potential for the proliferation of AWS technology beyond state actors clearly exists, as do associated legal questions, the emphasis is on developing a proposed ethics framework which applies to states and not to individuals,

groups or other non-state actors. Further, the ethics framework is rooted in just war reasoning as an ethic of comparative, not absolute, justice.[5]

Autonomy – a contested concept

Definitions of autonomy have important operational, legal, ethical and political dimensions which go far beyond simple semantics. For example, the UK Government's official position on fully autonomous weapons was set out by the Parliamentary Under-Secretary of State for Foreign and Commonwealth Affairs in a 2013 Parliamentary debate on Lethal Autonomous Robotics:

> [T]he Government of the United Kingdom do not possess fully autonomous weapon systems and have no intention of developing them. Such systems are not yet in existence and are not likely to be for many years, if at all. Although a limited number of defensive systems can currently operate in automatic mode, there is always a person involved in setting the parameters of any such mode. As a matter of policy, Her Majesty's Government are clear that the operation of our weapons will always be under human control as an absolute guarantee of human oversight and authority and of accountability for weapons usage.[6]

The Under-Secretary went on to state that 'We cannot develop systems that would breach international humanitarian law, which is why we are not engaged in the development of such systems'.[7] Note, however, that the UK Government's 'intention' did not exclude several possibilities: development of 'fully autonomous weapon systems' at some point in the future; the development and deployment of systems that fall short of 'fully autonomous'[8]; or fully autonomous non-weapon systems. Despite these potential caveats – and the lack of a definition of what 'fully autonomous' means – there is also reference to weapons 'always be[ing] under human control'. Not in the sense of individual weapon engagements but 'as an absolute guarantee of human oversight and authority and of accountability for weapons usage.'

Neither conventional just war ethics, nor compliance with IHL, specifies a requirement for continuous human control of weapons.[9] However, the point is politically and culturally sensitive and the lack of direct human control of weapon release is a key component of the argument by the Campaign to Stop Killer Robots that AWS are 'abhorrent, immoral, and an affront to the concept of human dignity and principles of humanity'.[10] While there is not the scope here to develop this 'human dignity' argument, Amanda Sharkey provides a useful starting point for further enquiry: 'If it is accepted that there are many weapons, artifacts, and human behaviours that are held to be against human dignity, then this itself becomes a reason for not relying too heavily on human dignity in arguments against AWS, as distinct from other means and weapons of warfare.'[11] Extending my 'comparative justice' approach to the idea of 'comparative dignity', human dignity has been violated through enslavement, rape, torture and myriad abuses throughout the history of war, and I remain unconvinced that AWS provide a special case which trumps all other violations.

The Campaign to Stop Killer Robots also used similar language to that of the UK Government when, in 2013, it called for a 'pre-emptive and comprehensive ban on the development, production, and use of fully autonomous weapons.'[12] Fully autonomous weapons were those which had the capability to 'choose and fire on targets on their own'.[13] It is not clear if The Campaign to Stop Killer Robots was simply demanding what the UK Government appeared to be offering – human control of weapon systems – or if the campaign was driven in part by a broader pacifist ideology which opposes all military violence, with 'fully autonomous weapons' serving as a campaign focus. Every government, however, will adopt its own position and not necessarily be open to human control of AWS, so the Campaign to Stop Killer Robots provides a timely global challenge.

Degrees of autonomy

There are further semantic points of which to be aware in the burgeoning literature on autonomy, which extend to the context of AWS. For now, however, this paper suggests that a crucial terminological battle is – or should be – over the word 'fully' and what it means in relation to autonomy in weapon systems. I have written elsewhere about what can be referred to as AI and autonomy in an idealised philosophical sense and also in the sense of more 'limited machine autonomy'.[14] In the former – idealised autonomy – AI achieves or exceeds human equivalence (the 'singularity') in functioning, reasoning and decision-making, and is represented in science fiction films. But even these science fiction autonomous robots do not claim or allude to the next and ultimate level of artificial intelligence – Artificial General Intelligence (AGI). AGI occurs when computer-driven AI reaches a state of self-learning that exponentially increases its knowledge on a continuous, positive-feedback learning loop.[15]

Machine limited autonomy based on AI more closely reflects the practical experience of AI theorists and practitioners. In a military context, a UK Joint Doctrine Note captured that limited scope in 2011:

> An autonomous system is capable of understanding higher level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.[16]

Recent developments and testing of military technology in the UK indicates the future trajectory when it comes to AWS. In 2016, the Royal Navy hosted Exercise Unmanned Warrior which, in conjunction with 40 industry and other partners, enabled the testing of 'unmanned and autonomous vehicles… to remove human operators from the most tedious, unpleasant and dangerous activities, such as mine-laying and recovery and anti-submarine operations.'[17] These practical applications of emerging technology are consistent with the UK Government's stated aims set out above.

In 2018, the British Army hosted Exercise Autonomous Warrior with the specific aim of 'test[ing] and evaluat[ing] the effectiveness of robotic and autonomous systems (RAS) on the battlefield.'[18] Defence Minister Mark Lancaster reinforced the vision for the future place of autonomy in UK military doctrine and practice when he said, 'Autonomous Warrior sets an ambitious vision for Army operations in the 21st Century as we integrate drones, unmanned vehicles and personnel into a world-class force for decades to come.'[19] While the language of autonomy is used more prominently here than in Exercise Unmanned Warrior in 2016, there is still not explicit mention of autonomous weapon systems that could be deployed in a lethal attack role, let alone a 'fully autonomous weapon system' of the kind that UK policy (at least at the time of writing) and The Campaign to Stop Killer Robots both reject.

Yet, despite this flurry of activity – a mere glimpse of worldwide autonomous systems development – agreed definitions are no closer. A small sample of that literature highlights the extent of the challenges, definitions and practicality, of autonomous weapon systems. In a US military-focused context, Massie discusses the extent of autonomy, the role of AI, machine-human working and the importance of trust in systems.[20] In 2014, the Birmingham Policy Commission examined drone use in the UK and discussed many of the challenges of LAWS at great length. The Commission challenged the UK Government to 'take a leading role in discussions to build an international consensus around a set of norms to regulate, if not ban, LAWS'.[21] It helpfully defined AWS as 'ones that have the following properties: *automation*, *volition*, and *intention*.'[22] These properties were sourced from Marra and McNeil, who expand further:

> Autonomy also requires some decision-making agency, which is captured by *volition*, or "choice in action or thought," and *intent*, or deliberate "pursuit of goals." Truly autonomous machines may also actually be able to learn, meaning they can draw conclusions based on past experience and incorporate these lessons into future actions. This baseline distinction between automation and autonomy offers a useful starting point.[23]

These concepts of automation, volition and intention get even more interesting – and attributed with philosophical capabilities that are typically identified as human – when traced back even further to Clark's 1999 theorising on Cyborged Ecosystems.[24] The agency required for autonomy includes 'independence of comportment' and 'a sufficiently conscious mind'; the necessary automation includes 'the capacity to operate without outside intervention'; volition extends to 'defining its own goals and then formulating and executing strategies for attaining them'; before, finally, 'in order to be significantly autonomous an entity must be intentful, and actually exercise its volition.'[25]

Perhaps increasing the usefulness of these concepts is Clark's contention that 'autonomy should be measured on a continuous scale.'[26] At one extreme would we would find 'truly autonomous' systems or, to use more recent terminology, 'fully autonomous' systems referred to above by the UK Government, The Campaign to Stop Killer Robots, the Birmingham

Commission, as well as by Human Rights Watch.[27] Lower levels of autonomy would still be available, but would go beyond mechanised automation. Clarke's insights undermine an important statement in the Birmingham Commission report, where it says: 'Put simply, a weapons system is either autonomous or it is not – there is no spectrum of autonomy.'[28] The proposition that emerges from this brief discussion is that it is both practically and conceptually possible to have AWS that fall short of being 'fully autonomous' however that contested phrase is understood. AWS become practically possible because of the possibility of limited levels of autonomy which, in turn, would rely on varying applications and capabilities of AI as identified in the next section.

Different types and purposes of AI

Definitions of AI are proliferating but a useful description by Hopgood in 2003 is a good starting point: 'Artificial intelligence is the science of mimicking human mental faculties in a computer.'[29] While it may not always be helpful, comparisons of AI and human intelligence can provide a convenient shorthand to help non-experts grasp what AI is capable of.[30] With his emphasis on operationalizing AI, Hopgood avoids – even cautions against – hyperbolic claims about what AI is and what it can achieve. He provides understanding for the non-expert by offering a spectrum of intelligent behaviours (see Figure 1 below) and explaining how it is the middle of the spectrum that AI finds most difficult to replicate. It turns out that 'common sense' is vastly more complex than first appreciated by AI scientists several decades ago, and that the brain operations we put into understanding what we see are tremendously difficult to replicate artificially.

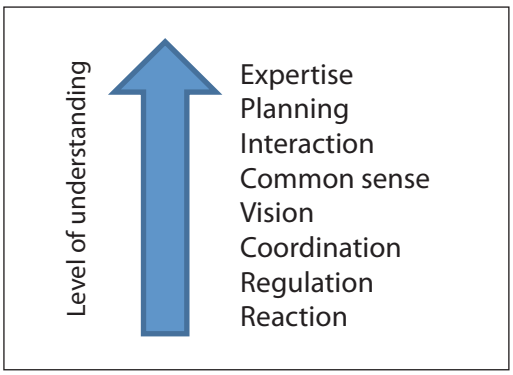

Figure 1[31]

He also sets out two broad AI categories, distinguishing between 'explicit modelling with words and symbols' and 'implicit modelling with numerical techniques,' then identifies a number of AI techniques: neural networks, genetic algorithms, simulated annealing, artificial immune systems and fuzzy logic.[32] In terms of applying these techniques in complex AI systems, Hopgood offers a 'blackboard system'.[33] Such a system seeks to replicate human teamworking by combining different AI techniques on an integrated system to address multiple elements of an overarching complex problem.

Such an approach offers a practical way ahead in the operationalising of AWS where different types of AI will most likely be needed for different elements of the system. For example, knowledge-based AI will be needed to interpret and apply complex rules of the kind that would enable a system to conform to proportionality constraints in the use of lethal force. In contrast, only a limited range of AI capability might be needed for the repeated, mundane and predictable tasks of automated take-off and landing of an air vehicle.[34] Furthermore, the AI requirements for human and object recognition will differ from those needed for flying safely in either civil airspace or the military-controlled battlespace.

One further AI technique that has gained a high public and scholarly profile is 'deep learning' in neural networks. Schmidhuber suggests one recurring use of deep learning is unsupervised learning which, in turn, 'can facilitate both SL [Supervised Learning] and RL [Reinforcement Learning]'.[35] Supervised learning is particularly effective at pattern recognition, while reinforcement learning occurs in the absence of a 'supervising teacher'.[36] Each of these is one of many techniques that can be refined, applied, self-refining, and so on. However, to be legal, ethical and operationally effective, in a lethal autonomous aircraft system each technique will have to be predictable and robust enough to underpin decision-making that may cost lives and inflict physical destruction.

Given the likely surveillance modes of any future autonomous aircraft system or drone, unsupervised learning would be useful for encoding vast amounts of video feed, and enabling more focused analysis on smaller search areas of interest. Any such analysis would need – in some kind of AI blackboard system or AI integrating control programme – to engage with other systems in any potential weapon use to conform to the legal requirements and policy constraints captured in rules of engagement for specific conflicts. Crucially, under current and foreseeable technological developments, in the UK's political environment at least, some degree of human involvement will likely be a necessary part of weapon use. Elsewhere, Morgan has helpfully set out a number of ways in which AI approaches might benefit or impact upon air power.[37] For now, however, discussion moves on to the *system* considerations in AWS.

Putting the 'system' into Autonomous Weapon System

One of the most immediate challenges militaries and governments face in attempting to operationalise AWS is to conceptualise what the system would comprise and how it would work. Part of the difficulty is that, if my experiences in numerous conferences and symposia are an indicator, many civilian experts in the multifarious domains that are needed if legal, ethical and operationally useful AWS are to be achieved, seem unaware of even basic operational requirements if, say, an air force wanted to deploy an armed autonomous aircraft. Marra and McNeil illustrate this disjuncture between theory and practice in their paper, 'Understanding the "Loop": Regulating the next generation of war machines', where they state:

> As drones develop greater autonomy, however, humans will increasingly be "out of the loop." Human operators will not be necessary to decide when a drone (or perhaps a swarm

> of microscopic drones) takes off, where it goes, how it acts, what it looks for, and with whom it shares what it finds.[38]

When Marra and McNeil claim here that humans will not be necessary to decide when a military drone takes off, where and how it operates, and what it does with the data it collects, there appears to be a disconnect from the practicalities of how advanced air forces operate and how they might incorporate autonomous aircraft/drones over time. Some basic practical assumptions need to be made for effective theorising and planning. So this paper assumes that any large future military armed drone, or LAAS, will operate – at least for a decades-long transition period – alongside manned aircraft.[39] Further, among many other requirements they will also need the following: hangars for storage and where repairs and servicing will be carried out; to share a dispersal, or airfield hard-standing, with manned aircraft; for armourers to still fit missiles and bombs under the wings; engineers to maintain the hardware, software, airframe, avionics, and sensors; and that their take-offs and landings will be integrated with the movements of manned aircraft and still be subject to an air traffic control system using detect-and-avoid technology. All of that just to get in the air and fly around without accidently colliding with manned aircraft – or other LAAS – in the sky, and before integration with a Combined Air Operations Command, Defence Intelligence, air battle managers (or future equivalents), Joint Terminal Attack Controllers, and more. The most realistic scenario envisaged here for the MQ-9's successors is for elements of its systems, or functions of the humans mentioned above, to be increasingly replaced with AI over time. A long time.

Taking the RAF's MQ-9 Reaper RPAS as an approximate template for now, the 'system' operates on two levels. First, at the higher level, the RPAS sits within a much larger system – the political-military complex itself. That system sits within a clear command hierarchy which is operationally and ethically accountable for the conduct of war. Second, the RPAS is an operational 'system', comprising airframe, avionics, multiple sensors, computing, communication, information, weapon, and human elements.

British WWII bombing command hierarchy	UK Reaper MQ-9 RPAS command hierarchy	Theoretical Autonomous Weapon System command hierarchy
1. Prime Minister 2. Defence Minister 3. Chairman of the Chiefs of Staff Committee 4. Chief of the Air Staff 5. AOC-in-C Bomber Command 6. Bomber squadron commanders 7. Bomber captain	1. Prime Minister 2. Defence Minister 3. Chief of Defence Staff 4. Chief of the Air Staff 5. AOC 1 Group 6. Reaper RPAS squadron commanders 7. Reaper captain (with continuous remote access to multiple support elements and resources)	1. Prime Minister 2. Defence Minister 3. Chief of Defence Staff 4. Chief of the Air Staff 5. AOC 1 Group 6. Lethal Autonomous Aircraft System squadron/military commanders 7. Lethal Autonomous Aircraft System – offensive and/or countermeasures (plus constellation of ethically implicated actors)

Table 2

Table 2 compares simplified versions of a World War II bomber command hierarchy, a current Reaper RPAS hierarchy, and a projected hierarchy for a LAAS. Each hierarchy is responsible, ethically and legally, for operational decision-making, with those near the top of the hierarchy bearing greater responsibility than those at the bottom.[40]

The autonomous weapon, say a lethal autonomous aircraft (or autonomous weaponised drone), will not be able to function on its own outside a 'system' that includes human elements. Politicians and commanders will still decide when and how they should be deployed; lawyers and others will still advise on ROE; coders will need to write programmes and update systems. Even if autonomous machine elements of the 'system' increasingly replace human elements over time, political, legal and ethical concerns – in responsible states – will always require a degree of human input as well as human accountability; accountability for deploying the system and accountability for the component parts of AWS.

Communications Assurance in quantum communications

(With contribution from Benjamin Davies, Theoretical Physicist, Loughborough University)[41]

Rapidly approaching from the technological horizon is quantum communications (QComms).[42] QComms work on the relationship between entangled photons. If two photons are entangled, measurement of the properties of one of them – say spin direction - has a direct effect on the properties of the other, entangled photon. This is true regardless of the distance between the photons. In a recent experiment, a video conference – via satellite – using QComms between Beijing and Vienna confirmed this technique.[43] While there is still much to be understood about the science and application of quantum technology, one major advantage it will offer is that such a system is substantially more secure from hacking, in the sense of an aggressor taking control of the system. Attempts to intercept the signal, however, would affect the quantum relationship between the photons and make QComms impossible, essentially a denial of service (DoS). In the event of such a potential signal interception, one benefit of QComms is that no information would be revealed to the eavesdropper.

In a contested environment, such a DoS could be sufficient to remove the capability of an autonomous weapon to communicate with human elements in the system. While independence of decision and action in such an environment would be part of the *raison d'être* for autonomous weapons, there are numerous reasons for wanting to retain a live comms link with the rest of the system: intelligence updates, revised ROE, system checks or the passing of other important information. Dual communication – classical and quantum – would provide the greatest possibility of maintaining communication between a LAAS and its command structure. They could be used for different functions or to provide redundancy. For example, a mission directive (or change of directive) could be sent using QComms, with standard encrypted comms providing further intelligence data. In this case, a DoS attack on the QComms would not provide an eavesdropper with the mission

information and any intercepted, classically encrypted information would be potentially useless, even if decrypted, due to that lack of mission information.

An AWS might not be immediately aware that its signal is being intercepted. A QComms system works on the basis of statistics and, potentially, there could be a delay while the AWS works out that the statistics are incorrect. Randomness sometimes matches expectations, which could result in false information being fed into the system even if it eventually shuts down. An attacker would not gain control in the same way as they would with hacking but there could be unforeseen effects. Consequently, a QComms channel may not be reliable enough to trust for immediate confirmation/refusal of a strike if there is a possibility of randomly-generated, false commands.

Bias avoidance in AWS

A significant challenge for operationalizing AI in AWS is the propensity for human bias to be programmed – inadvertently or otherwise – into its component algorithms. There are therefore two linked tasks: understanding the nature and potential for human bias; and coding in such a way as to avoid inputting that bias into an AWS. Such bias could have serious repercussions when distinguishing, for example, between combatants and noncombatants, or even in making judgements about the legitimacy of AWS in the first place.

Human bias

Perhaps the greatest bias that a person might have – if they are even aware of the human propensity for bias – is a sense that it does not affect them, that their own rationality can keep bias at bay. When it comes to AWS, bias operates on different levels: at a policy level there is the question of whether they should be allowed to exist, while at an operational level there are questions about how they can be used in ethical ways. Almost 50 years ago, Tversky and Kahneman described how bias in imagining the unknown can inform the extent to which an activity might be perceived as risky:

> The risk involved in an adventurous expedition, for example, is evaluated by imagining contingencies with which the expedition is not equipped to cope. If many such difficulties are vividly portrayed, the expedition can be made to appear exceedingly dangerous...Conversely, the risk involved in an undertaking may be grossly underestimated if some possible dangers are either difficult to conceive, or simply do not come to mind.[44]

Consider these words in the context of potentially building and operating an AWS. Discussion around AWS necessarily involves imagination because future systems that are being conceived and developed do not exist yet, even though the legal, ethical and operational challenges must be considered during the ongoing developmental process. Take two possible opposed views. On the one hand there is implacable opposition to AWS, where they are 'made to appear exceedingly dangerous',[45] drawing on science fiction tropes

and imaginings that are informed by films like *Terminator* and *I-Robot*. On the other hand, technical experts and experienced military figures might be less concerned about the potential of AWS, perhaps because 'some possible dangers are either difficult to conceive, or simply do not come to mind' as a result of familiarity with the use of lethal force in a military context and the multiple legal and practical constraints that they operate within.[46] It seems highly unlikely that either position, as they have been exaggerated here, is without bias.

Coding bias

When it comes to using AWS, bias has the potential to surface in different guises. One potential widespread risk is that the subjective bias of the coder is somehow encoded into the system through the particular lines of code that are used as the building blocks of the autonomous elements of the system. As long ago as 1996, Friedman and Nissenbaum highlighted three different categories of bias in computer systems: 'preexisting bias, technical bias, and emergent bias. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use.'[47] Each of these types of bias is a field of study on its own, so consider the implications for AWS if coding bias was to influence their operations.

Clark observes that 'complete independence in an entity requires a structure that is free of any implicit design objectives or behavioral biases that might influence the definition or pursuit of goals'.[48] One of the potential challenges of AWS is the adoption of sensor systems that have different degrees of recognition effectiveness across different age, gender and racial characteristics. Introna and Wood point out algorithms that display evidence of bias in facial recognition: males being more accurately recognised than females, and older people being easier to recognise than young people.[49] Introna and Wood were developing the work of Givens et al, whose study found that Asians, African Americans and 'other race members are [all] easier to recognise than whites'.[50] Even setting aside additional practical difficulties like face coverings and the wearing of spectacles, there are clear ethical consequences for the potential uses of AWS and the identification of targets if the reliability of facial recognition is greater in some parts of the world than in others. In an assessment of the Metropolitan Police Service's (MPS) trial of Live Facial Recognition (LFR) technology in London, Fussey and Murray highlighted failures and potential risks in the system which have implications for AWS.[51] Watchlist accuracy poses such a risk, for example where the system correctly identifies a person on the list but for a minor offence which would not normally be deemed serious enough to warrant a place on the watchlist in the first place.[52] This may have an impact on the liberty of a criminal in London but the consequences are potentially more lethal for anyone on a watchlist to be targeted by an AWS. Even more problematic is accuracy of the system in recognising faces: 'Across the six test deployments MPS officers engaged with 22 individuals as a direct result of a computer generated match judged to be credible by a human operator. Fourteen of these (63.64%) were verified incorrect matches, eight were verified correct matches (36.36%).'[53]

Raising further ethical questions about using AI to identify human targets in an AWS, an AI system used by American judges to predict if an offender is likely to reoffend in the future is alleged to be biased towards minorities: 'The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants'.[54] Transposing this degree of potential bias into an intelligence gathering context has obvious risks in terms of ensuring that individuals are not wrongly criminalised or targeted. Even if a misidentification was recognised and cancelled as part of the process of positively identifying targets, any such errors in an AI system would create and increase time inefficiency.

Predictability of the AI 'black box'

Given all of the preceding complexities, this final section examines another of the significant difficulties in creating and deploying an AWS which incorporates self-learning AI algorithms. Namely, that AI – especially where a deep learning or self-learning characteristic is included – is often seen as an unpredictable 'black box'. It is possible to see inputs and outputs but not fully understand – or be able to replicate – the AI decision-making process in between. Importantly, if data is entered into an AI programme or an interlinked family of AI programmes, the output, or decision, from that programme needs to be understood and predictable if the weapon system is to conform reliably to ROE and international law. A reliable degree of predictability would, however, at least be able to engender a degree of trust that the system would function in a militarily consistent way. One difficulty presented by AI is reverse-engineering any 'black box neural networks',[55] or auditing an outcome to see what decisions the AI programme(s) made at every decision point along the way from the moment of the inputs. Bathaee, describes two different reasons for the inability of humans – currently – to understand the AI black box:

> First, a lack of transparency may arise from the complexity of the algorithm's structure, such as with a deep neural network, which consists of thousands of artificial neurons working together in a diffuse way to solve a problem…Second, the lack of transparency may arise because the AI is using a machine-learning algorithm that relies on geometric relationships that humans cannot visualize, such as with support vector machines. This reason for AI being a black box is referred to as 'dimensionality.'[56]

This 'complexity' and 'dimensionality' may eventually be understood but these technical challenges do not excuse a government or military force from their ethical and legal responsibilities with regard to weapons and their use. One operational-level purpose of conducting weapon reviews in accordance with Article 36 of 1977 Additional Protocol 1 to the Geneva Conventions is to provide military commanders with the assurance that their use of specific weapons and weapon systems is lawful.[57] In addition, Doshi-Velez, Finale and Kortz et al have considered a number of ways to legally hold AI systems to account, from which some implications for AWS emerge.[58]

Considerations for Approaches for Holding AIs Accountable

Approach	*Well-suited Contexts*	*Poorly-suited Contexts*
Theoretical Guarantees	Situations in which both the problem and the solution can be fully formalized (gold, standard, for such cases)	Any situation that cannot be sufficiently formalized (most cases)
Statistical evidence	Problems in which outcomes can be completely formalized, and we take a strict liability view; problems where we can wait to see some negative outcomes happen so as to measure them	Situations where the objective cannot be fully formalized in advance
Explanation	Problems that are incompletely specified, where the objectives are not clear and inputs might be erroneous	Situations in which other forms of accountability are not possible

Table 3[59]

While they suggest a number of tools for 'increasing accountability in AI systems' – theoretical, statistical and explanation – the one that they put forward as most practical is 'explanation'. Given the complexity of AI approaches, even explanation has limitations, although they argue that '[b]y exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute'.[60] Just as military commanders, RPAS operators and others today need to provide explanations for decisions they make surrounding missile or bomb strikes, future weapon strikes by AWS would require the same accountability. The discussion then becomes about the degree of explanation necessary. Humans are currently not required to explain the neural processes they used (which probably cannot be fully explained anyway) to underpin decisions to approve or conduct a lethal strike. They do, however, need to be consistent and provide enough explanation of the context as they perceive it and how they applied legal considerations in deciding a particular course of action for accountability to be served. So the *degree and nature* of explanation required of an AWS needs to be considered as technology develops.[61]

Moving away from law, from a deontological (rule-based) ethics perspective, a commander needs to have confidence that an AWS is capable of conforming to multiple rules and processes, including IHL, ROE, air battle management systems and air traffic control. Within these various rules and codes, the system will also need to be able to discern whether a particular use of force is proportionate, discriminate, militarily necessary and not likely to inflict outcomes or harms that are *mala in se* – evil in itself. That is, to also consider a consequentialist (or outcomes-orientated) perspective as well. However, Western Just War ethics does not merely provide a rigid set of rules. As Elshtain argues, 'Just war thinkers do not propound immutable rules – they are not, to repeat, deontologists – so much as clarify the

circumstances that justify a state's going to war'.[62] Just War ethics also clarify the circumstance, means and methods of engaging in warfare, and can be extended to autonomous weapon systems. While military commanders and other combatants will necessarily be required to conform to the law and the question: 'Is this action legal?' I suggest that it will be a rare individual who, when using lethal force in war, does not also consider the associated ethical question, 'Is it right?'

But if ethical considerations go beyond the observation of rules, codes and obligations, then outcomes are the other key element. From a consequentialist (outcome-based) ethical perspective, the results of a particular action or strike are paramount. There is little point in deploying a system which, with the uncertainty of war and the capacity for things to go wrong even in the most favourable of circumstances, still leads to an unwanted, harmful, unethical outcome. The programmed or AI elements of the system can still have unintended negative ethical repercussions. Maner observes that 'the smallest possible perturbations – i.e., changes of a single bit – can have the most drastic consequences…[and] gives rise to a unique ethical difficulty, at least for those who espouse a consequentialist view of ethics'.[63] He goes on to describe how a single missing hyphen in a line of code caused the destruction of an Atlas Agena rocket and, with it, a Venus probe. Perhaps self-learning algorithms will learn to self-correct, or perhaps control programmes will protect against such errors or unintended outcomes, but the ethical, legal and operational consequences of coding errors or oversights cannot be avoided.

A more recent, less costly, but higher profile example of an AI self-learning algorithm learning the 'wrong' things is Microsoft's 'Tay' chatbot. Released in 2016, the purpose of publicly releasing Tay to engage with the public over the internet was partly about Tay learning to understand millennials and their language and culture, and partly so Microsoft could learn how such a self-learning algorithm performed.[64] What Tay learned was to be racist and genocidal, before being taken offline 24 hours later. Tay was put under a sustained attack – a form of denial-of-service attack – from trolls who subverted the original Microsoft intention. Tay's successor, Zo, has been released with a number of controls in place to prevent a recurrence of pro-Hitler sentiments and other offensive views. One criticism, however, is that Zo is now *too constrained* and 'politically correct'.[65] While Zo might not make racist statements like Tay, the self-learning element of the AI is limited by control functions. The ability to limit more complex self-learning AI in this way will be important for autonomous weapons, but the limitations may well be too constricting to achieve the effects that militaries desire.

Summary

The challenge of creating a predictable, militarily effective AWS which conforms to both ethical and legal requirements is vast. Functional decision-making in AI programmes – and therefore decisions in the machine element of AWS – is not underpinned by the consciousness, cultural embeddedness, motivations, conscience, self-reflexivity and other elements of human essence that inform ethical choices. That functional decision-making in AWS is shaped by the

humans who set the context for the coding and especially those who carry it out, as well as the programming process which has its own ethical consequences. It is the human element in the autonomous weapon *system* that provides the possibility of those systems making decisions with an ethical dimension. Despite the lack of clarity on definitions, the UK has at least indicated one possible direction in the debate and in the practical development of autonomous weapon systems: 'The UK believes that the level, nature and primacy of human control over specific functions is the key consideration in the LAWS debate rather than technology, which is likely to change rapidly'.[66] In light of the foregoing discussion of technical, human, operational, legal, ethical and political factors, this paper concludes by proposing the following outline ethics framework for AWS, which includes *jus ad bellum*, *jus in bello* and *jus post bellum* considerations.

An Ethics Framework for Autonomous Weapon Systems[67]

Section 1: Guiding ethical principles on the use of Autonomous Weapon Systems

1. There is an ethical requirement to ensure that international law is applied to all weapon systems, including the development, use and disposal of autonomous weapon systems (AWS).
2. The motives and methods by which AWS are designed and deployed should be governed by ethical principles as set out below.
    a. AWS should be deployed in support of the jus ad bellum principle that war is waged in order to obtain just peace and security, as set out in the UN Charter (1945).
    b. War ethics permits the minimum force necessary to achieve legitimate military goals, and requires that AWS be used within an ethics framework.
    c. AWS ethics demand the discriminate and proportionate use of force in pursuit of militarily necessary objectives.
    d. As weapons of war, AWS which are capable of the proportionate and discriminate use of lethal force in line with international law, and are designed and deployed to achieve both, are not inherently unethical.
3. Wars should be fought only when necessary. AWS should be used in support of the jus ad bellum principles of last resort and military necessity.
4. AWS should conform to the principle that suffering in war should be minimised where possible.
5. Ethics of war require that judgements be made on the relative goods and harms of deploying AWS and is a *comparative* justice.
6. AWS ethics recognises the principle of asymmetry, which has been the strategic and tactical aim of military commanders throughout history. Asymmetry is ethically neutral; it is in the means of deploying asymmetric advantage that ethical judgements are required.
7. Asymmetric disadvantage with respect to AWS does not confer the right to ignore the requirement for ethical conduct and decision-making in war.
8. AWS violence, like that of fully human controlled weapons, is not inherently or necessarily a moral wrong.
    a. Under certain circumstances, AWS violence – including killing – can be a moral

necessity if its use results in less harm than would occur without its use.
   b. The harms caused by AWS violence may be offset by the good accomplished on behalf of others.
9. These guiding principles recognise and value the rights and dignity of every person.
   a. In deciding whether or not to deploy AWS, states should consider the impact of their use, or non-use, on individual human dignity.
   b. AWS should not be deployed where their use is expected, on balance, to cause greater loss of human dignity than would occur if they were not used.

Section 2: Human responsibility for Autonomous Weapon Systems

1. Humans are required to retain ethical responsibility and accountability for AWS at every stage of the weapon life cycle, from design to decommissioning. Ethical responsibility or AWS extends beyond the operational chain of command to enabling parties which include, but are not limited to, designers, engineers, software programmers, AI developers, military and security intelligence personnel, scientists and weapons manufacturers.
2. Where AWS incorporate any self-learning AI or combination of AI techniques, the individual AI elements and collective AI network should be constrained by rules embedded in the AI knowledge systems. It should be possible to vary the ratio of human-to-machine functioning in AWS according to operational context and ROE.
3. Self-learning algorithms in AWS must not have the ability to change sides in any conflict. Responsibility for AWS must remain with the deploying state.
   a. States which export AWS must ensure that purchasing states have the capability and intent to assure system integrity and legal and ethical practice.
   b. Notwithstanding the principle that autonomous elements of AWS must not have the decision-making capacity to change sides, AWS should have the capacity and authority to intervene with lethal force to protect civilians from deliberate attack by 'friendly' or allied forces.
4. AWS should be capable of recognising *hors de combat* human targets and acting in accordance with relevant delegated legal and operational authority.
5. 'Intention' is central to the doctrine of double-effect, which plays a crucial role in the defence or criticism of civilian deaths and 'collateral damage' during military operations.
   a. Artificial Intelligence and robotic elements within AWS should not be attributed with the capacity for conscious intent.
   b. 'Intention' in the application of lethal force remains the domain of the human operators and chain of command which deploy AWS.
6. AWS should be used proportionately and in a discriminating manner against specified and identified targets.
7. AWS should not be used against geographical areas containing unidentified human targets.
8. Positive identification of human targets required to be confirmed in AWS 'kill chain', comprising suitably qualified elements.

9. A Final Authority Officer is required to approve each AWS mission. That authorizing officer must confirm that suitable safety checks have been conducted on each LAWS prior to each mission. Primary approval for each of these individual elements can be granted by the specialists concerned before overall final mission authority is granted.
    a. In multinational operations, a National Asset Authority is required in order to retain state responsibility for the deployment of AWS.
10. AWS should be predictable to the extent that the deploying state can have the reasonable certainty that its operational intent will be carried out. The AWS should not have the capacity to override or ignore its programmed function or embedded operational commands. A human is required to make such changes to operational parameters and functions.

Section 3: Autonomous Weapon Systems and the methods and means of war

1. State representatives are responsible for ensuring that the development, testing, acquisition and deployment of AWS conforms to international law at every stage.
2. AWS should not employ methods or means that are *mala in se*. These include:
    a. Munitions or actions that are intended to inflict unnecessary suffering on human beings or cause unnecessary harm to the natural environment.
    b. Deployment of munitions that are indiscriminate in nature.
    c. Deliberate abdication or concealment of human responsibility for AWS.
3. As well as conforming to international law, the methods and means of AWS should conform to the following ethical principles:
    a. Force must be used proportionately in relation to the importance of any military objective.
    b. Discrimination between legitimate targets and protected people and objects, in both weapon mode and surveillance mode.
    c. The basic humanity and dignity of people is not a legitimate target.
    d. AWS should be deployed against targets within the concept of military necessity.
    e. Where AWS can vary the ratio of human-to-machine input, that ratio will be guided by military necessity and Rules of Engagement.
4. AWS must, with reasonable certainty, be capable of positive identification of prohibited civilian targets such as schools, hospitals and places of religious worship.
5. AWS should be used only against identified and legitimate targets, either human or objects. They should not be targeted against geographical areas or civilian populations.
    a. In a theatre of operations, AWS must be able to distinguish civilian criminal actions from martial actions, with authorised military force being used only against the latter.
6. The default operational requirement for AWS should be 'zero civilian casualties' unless a State approves a higher permissible civilian casualty level consistent with legal advice, rules of engagement and approval from a suitable operational commander.
7. In defence-of-friendly-forces engagements which may incur civilian casualties, a human with appropriate delegated authority must be actively involved in setting collateral damage and civilian death parameters.

Section 4: Risk assessment and mitigation measures for AWS

1. Ethical use of AWS requires that states incorporate risk assessment and mitigation measures at every stage of a weapon life cycle. These include:
    a. Developing the concepts and doctrines that frame AWS requirements and aims.
    b. Design and manufacture of AWS should recognise and incorporate the ethical requirement to protect civilians in times of war.
    c. Reliability, assurance and security of both hardware and software underpins accountability and reliability of AWS. Where multiple AI techniques are used within AWS, each individual technique and the integrating control programme must each be reliable, assured and secure.
    d. Testing of new systems in such a way as to conform to IHL prior to being declared operational.
    e. AWS deployed in operational theatres should be used proportionately and discriminately.
    f. Maintenance and upgrading of hardware and software should be coordinated and controlled by humans. The AI within AWS should be limited through control functions, in the extent to which it can self-upgrade any aspect of its capabilities.
    g. Ethical decommissioning of AWS requires that noxious substances are disposed of in such a way as to minimise risk to humans. Hardware should be recycled appropriately in a way that will not harm the natural world. All lines of computer code must be destroyed and made irretrievable.
2. AWS should be able to assess, mitigate or avoid harm to the natural environment.
3. AWS are not to cause harm to the natural environment which are long-term, widespread or severe.
4. AWS should have active safety systems:
    a. A fail-safe manual override should be installed in all AWS.
    b. Airborne AWS should operate an automated fail-safe Return-to-Home system where possible.
    c. Maritime and land-based AWS should be capable of safe shut-down where malfunction could cause the untargeted release of munitions.
5. System security should be highly assured to prevent hostile acquisition or control of AWS by unauthorised persons, groups or states.
6. In pre-planned operations, the speed and volume of information provided to authorised personnel to approve a lethal strike against a human target should not exceed the capacity of the decision-maker to make a reasonable assessment and subsequent decision.
7. In self-defence reactive operations, the parameters for speed and lethality of AWS response will be set in accordance with national rules of engagement.

Notes

[1] For brevity, this paper refers to 'Autonomous Weapon System (AWS)', the term used at the 2014 ICRC Expert Meeting on Autonomous Weapon Systems in which I participated. The term

'Lethal Autonomous Weapon System (LAWS)' is preferred by the United Nations Group of Governmental Experts and elsewhere. I assume here that an autonomous weapon system is intended to be lethal.

[2] International Committee of the Red Cross, 1 November 2014, 'Report on Expert Meeting on Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects, Geneva, 26-28 March 2014', https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014, accessed 13 May 2019.

[3] United Nations, 23 October 2018, 'Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', located at https://www.unog.ch/80256EDD006B8954/(httpAssets)/20092911F6495FA7C125830E003F9A5B/$file/CCW_GGE.1_2018_3_final.pdf, accessed 10 April 2019.

[4] Elshtain, Jean B., 'The Third Annual Grotius Lecture: Just War and Humanitarian Intervention', *American University International Law Review*, Vol. 17, No. 1 (2001-2002), pp. 1-2.

[5] See Jean Bethke Elshtain, 'Just War and Humanitarian Intervention', *Ideas*, Vol. 8, No. 2 (2001) pp. 1-21.

[6] UK Parliamentary Debate, 17 June 2013, 'Lethal Autonomous Weapons', House of Commons Hansard, Vol. 564, Column 733, https://hansard.parliament.uk/Commons/2013-06-17/debates/13061744000002/LethalAutonomousRobotics, accessed 25 May 2019.

[7] ibid., Column 735.

[8] ibid.

[9] International Committee of the Red Cross, 2010, 'The Geneva Conventions of 1949 and their Additional Protocols', https://www.icrc.org/en/doc/war-and-law/treaties-customary-law/geneva-conventions/overview-geneva-conventions.htm, accessed 2 June 2019.

[10] Campaign to Stop Killer Robots, 31 August 2018, 'Majority call to negotiate a new treaty', https://www.stopkillerrobots.org/2018/08/sixthmeeting/, accessed 15 July 2019.

[11] Amanda Sharkey, 'Autonomous weapons systems, killer robots and human dignity', *Ethics and Information Technology*, Vol. 21 (2019) p. 85. For further reading see Dieter Birnbacher (2016) 'Are autonomous weapon systems a threat to human dignity?' In N. Bhuta, S. Beck, R. Geiss, H. Liu, & C. Kress (Eds.), *Autonomous weapons systems: Law, ethics, policy* (Cambridge: Cambridge University Press, 2016) pp. 105-121; Ozlem Ulgen, 'Human dignity in an age of autonomous weapons: Are we in danger of losing an 'elementary consideration of humanity?' *European Society of International Law conference paper series*, Vol. 8, No. 9 (2016) pp. 1-19.

[12] Campaign to Stop Killer Robots, 23 April 2013, 'Urgent Action Needed to Ban Fully Autonomous Weapons', Press Release, http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_LaunchStatement_23Apr2013_fnl.pdf

[13] ibid.

[14] Peter Lee, 'Armed Drones: Automation, autonomy, and ethical decision-making' in Ryan Kiggins, Ed., *The Political Economy of Robots: Prospects for Peace and Prosperity in the Automated 21st Century* (Basingstoke: Palgrave Macmillan, 2017) p. 302.

[15] For further reading see Goertzel, Ben and Pennachin, Cassio, *Artificial General Intelligence* (Berlin: Springer, 2007); Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies* (Oxford:

Oxford University Press, 2014). An excellent critique of hyperbolic claims about AI can be found at, Goertzel, Ben, 'Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's The Singularity is Near, and McDermott's Critique of Kurzweil', *Artificial Intelligence* Vol. 171 (2007) pp. 1161–1173.

[16] Ministry of Defence. 2011. *Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems* (Swindon: Development, Concepts and Doctrine Centre) pp. 2-3.

[17] Royal Navy, 2016, 'Unmanned Warrior', **https://www.royalnavy.mod.uk/unmannedwarrior**, accessed 30 May 2019.

[18] MOD, 20 June 2018, 'British Army set to redefine warfare with joint Autonomous Warrior', **https://www.gov.uk/government/news/british-army-set-to-redefine-warfare-with-joint-autonomous-warrior**, accessed 20 May 2019.

[19] ibid.

[20] Massie, Andrew, 'Autonomy and the Future', *Strategic Studies Quarterly*, Vol. 10, No. 2 (Summer 2016) pp. 134-147.

[21] Birmingham Policy Commission, 'The Security Impact of Drones: Challenges and Opportunities for the UK', *Birmingham Policy Commission Report*, October 2014, p. 7.

[22] ibid., p. 19.

[23] Marra William and McNeil, Sonia K., 'Understanding the Loop: Regulating the Next Generation of War Machines', *Harvard Journal of Law and Public Policy*, Vol.36 No. 3 (2013) p. 1150-1151.

[24] Clark, O. Grant, 'Characterization of Cyborged Ecosystems', PhD Thesis, McGill University, August 1999, p. 113-4.

[25] ibid.

[26] ibid., p. 113.

[27] Human Rights Watch and the International Human Rights Clinic, *Losing Humanity: The Case against Killer Robots*, Human Rights Watch, November 2012, p.42, **www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf**, accessed 10 June 2019.

[28] Birmingham Policy Commission Report, p. 66.

[29] Hopgood, Adrian, A., 'Artificial Intelligence: Hype or Reality?' *IEEE Computer*, Vol. 36, No. 5, p. 24.

[30] See Susan Fourtané, 25 August 2019, 'The Three Types of Artificial Intelligence: Understanding AI', *Interesting Engineering*, **https://interestingengineering.com/the-three-types-of-artificial-intelligence-understanding-ai**, accessed 18 September 2019.

[31] ibid., p. 25.

[32] ibid., p. 26.

[33] ibid., p. 26-7. Hopgood also provides here a brief history of the evolution of blackboard systems.

[34] Reem Nasr, 26 March 2015, 'Autopilot: What the system can and can't do', CNBC, **https://www.cnbc.com/2015/03/26/autopilot-what-the-system-can-and-cant-do.html**, accessed 5 June 2019.

[35] Schmidhuber, Jürgen, 'Deep Learning in Neural Networks: An Overview,' preprint, *Neural Networks*, Vol. 61 (2015) pp. 89.

[36] ibid., p. 86.

[37] Morgan, Phillip, *Putting AI into Air: What is Artificial Intelligence and what it might mean for the Air Environment*.
[38] Marra and McNeil, 'Understanding the Loop', p. 1141-2.
[39] I am referring to large aircraft-sized, weapon-bearing drones like the Reaper which require a runway and not to small, hand-held or shoulder-launched drones.
[40] For further reading on the idea of moral hierarchy in these contexts seen Peter Lee, 'Armed Drones: Automation, autonomy, and ethical decision-making' in Ryan Kiggins, Ed., *The Political Economy of Robots: Prospects for Peace and Prosperity in the Automated 21st Century* (Basingstoke: Palgrave Macmillan, 2017) pp. 291-315.
[41] To ensure the scientific accuracy of my references to the mysterious world of quantum communications I approached Benjamin Davies – PhD candidate in Theoretical Physics at Loughborough – to review my initial draft. With his permission, I have retained his corrections and contributions and thank him for the wording in this section. He is not responsible for the views expressed elsewhere throughout this paper.
[42] For an introduction to the topic of quantum physics see John Gribbin, *In Search Of Schrodinger's Cat* (London: Black Swan, 1985).
[43] Philip Ball, 'Focus: Intercontinental, Quantum-Encrypted Messaging and Video', *Physics*, Vol. 11, No. 7 (19 January 2018), https://physics.aps.org/articles/v11/7, accessed 1 August 2019. See also Juan Yin, Yuan Cao, Yu-Huai et al, 'Satellite-based entanglement distribution over 1200 kilometers', *Science*, Vol. 356, Issue 6343 (2017) pp. 1140-1144, DOI: 10.1126/science.aan3211, https://science.sciencemag.org/content/356/6343/1140, accessed 15 May 2019.
[44] Tversky, Amos and Kahneman, Daniel, 'Judgement Under Uncertainty: Heuristics and Biases', *Oregon Research Institute Research Bulletin*, Vol. 3, No. 1 (1973) p. 18.
[45] ibid.
[46] ibid.
[47] Friedman, Batya and Nissenbaum, Helen, 'Bias in Computer Systems', *ACM Transactions on Information Systems*, Vol. 14, No. 3 (July 1996) p. 332.
[48] Clark, 1999, p. 114.
[49] Introna, Lucas and Wood, David, 'Picturing algorithmic surveillance: the politics of facial recognition systems', *Surveillance & Society*, Vol. 2, No. 2/3 (2004) p. 190. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-200675.
[50] Givens, G., J.R. Beveridge, B.A. Draper and D. Bolme, (2003), 'A Statistical Assessment of Subject Factors in the PCA Recognition of Human Faces', Colorado State University, located at http://www.cs.colostate.edu/evalfacerec/papers/csusacv03.pdf, accessed 15 June 2019, p. 8.
[51] Pete Fussey and Darragh Murray, 'Independent Report on the Metropolitan Police Service's Trial of Live Facial Recognition Technology', *The Human Rights, Big Data and Technology Project*, Human Rights Centre, University of Essex (July 2019), https://48ba3m4eh2bf2sksp43rq8kk-wpengine.netdna-ssl.com/wp-content/uploads/2019/07/London-Met-Police-Trial-of-Facial-Recognition-Tech-Report.pdf, accessed 5 August 2019.
[52] ibid., p. 11.
[53] ibid., p. 75.
[54] Angwin, Julia; Larson, Jeff; Mattu, Surya; and Kirchner, Lauren, 'Machine Bias', 23 May 2016,

ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed 20 June 2019.

[55] Oh S.J., Schiele B., Fritz M. (2019) 'Towards Reverse-Engineering Black-Box Neural Networks', in Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, Vol. 11700, Springer, Cham, pp. 121-144.

[56] Bathaee, Yavar, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation', *Harvard Journal of Law & Technology*, Vol. 31, No. 2 (Spring 2018) p. 901.

[57] Ministry of Defence, 2016, *UK Weapon Reviews*, p. 4, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/507319/20160308-UK_weapon_reviews.pdf, accessed 12 October 2019.

[58] Doshi-Velez, Finale and Kortz, Mason et al, 'Accountability of AI Under the Law: The Role of Explanation', *arXiv*, 21 November 2017, Berkman Klein Center Working Group on Explanation and the Law, https://arxiv.org/abs/1711.01134, accessed 10 June 2019.

[59] ibid., p. 11.

[60] ibid., p. 2.

[61] ibid., p. 12.

[62] Elshtain, 'The Third Annual Grotius Lecture', p. 6.

[63] Maner, William, 'Unique Ethical Problems in Information Technology', in Bynum, Terrell W., and Rogerson, Simon (Eds) *Computer Ethics and Professional Responsibility* (Oxford: Blackwell, 2004) p. 54.

[64] Abby Ohlheiser, 25 March 2016, 'Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac', *The Washington Post*, https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/?noredirect=on&utm_term=.a3bc2cf8d89c, accessed 16 June 2019.

[65] Chloe Rose Stuart-Ulin, 31 July 2018, 'Microsoft's politically correct chatbot is even worse than its racist one', *Quartz*, https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/, accessed 16 June 2019.

[66] United Kingdom, 8 August 2018, 'Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles', Submission to the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, CCW/GGE.2/2018/WP.1, p. 2.

[67] This proposed ethics framework – like IHL – is intended to apply to states, not to individuals, groups or other non-state actors. Consideration of these elements can be added as part of a wider debate.