Defence Engagement Program – Final Report

Targeting Trust

A Workshop on the Nature and Role of Trust in Autonomous Weapon Systems (AWS) Policy Development

University of Ottawa
Ian Kerr, Jason Millar, Jacquelyn Burkell, & Katie Szilagyi

20 April 2018

Table of Contents

Executive Summary....2

PART I: Workshop Design....4

- Background:....4
- Workshop Format:....5

PART II: Summary of Discourse....6

- Introductory Presentations by Organizing Experts....6
- Lightning Round – Experts' "Big Questions" Presentations and Discussions....12
- Morning Breakout Sessions – Discussion Summaries....15
- Afternoon Breakout Sessions – Discussion Summaries....18
- Discussion of Policy Implications....21

PART III: Policy Considerations for Further Exploration....26

Appendix I....29

Appendix II....32

Appendix III....34

Appendix IV....35

Executive Summary

Semi-autonomous weapon systems, with varying degrees of human oversight, are already transforming battlespace. Although we have not yet faced the realities of autonomous weapon systems ("**AWS**")—systems that can target and kill without the need for direct human oversight, as the capabilities of AWS improve and the potential for autonomous decision-making increases, policymakers must articulate the level of human control that will be required in order to develop and deploy trustworthy AWS. Standards that require meaningful human control ("**MHC**") or "appropriate human involvement" have typically been considered preconditions for the permissible use of AWS in battlespace. Many believe that there are legal and moral requirements for such standards, and that they are necessary to instil sufficient trustworthiness in AWS.

Only a handful of scholars, activists, and stakeholders have linked the study of AWS policy directly to conceptions of trust and trustworthiness. Canada has an opportunity to provide leadership on the world stage by addressing the important challenges and questions presented by AWS. Recognizing this, *Targeting Trust* brought together an invited group of AWS experts from academia, government, and civil society. The goal was to investigate how perceptions of trust in AWS can impact Canada's policy choices, and vice versa.

Four key policy takeaways can be distilled from *Targeting Trust*:

1. **The importance of clearly defined definitions and terminology**. Collectively understood terminology is a necessary precondition of the AWS and MHC debate. Such clarity will greatly aid the debate by establishing among stakeholders common understandings of relevant, and in some cases novel, concepts such as trust, trustworthiness, MHC and "appropriate human involvement".
2. ***X trusts Y to do Z* – The Tripartite Nature of Trust.** In the case where AWS operates without MHC, one or more parties in the trust relationship is not human. This could impact how humans approach policies regarding target and kill decisions. Policymakers must consider the various types of trust relationships that are created by the presence of non-human operators, and how this might impact the proposed requirement of meaningful human control.
3. **Trustworthiness versus Trust in MHC.** Trustworthiness and trust are intertwined yet distinct concepts with significant implications for AWS development and deployment. While trustworthiness in the form of *reliability* is vital in order for human operators to appropriately rely on AWS, trusting these technologies can still

be dangerous. Humans acting as nothing more than rubber stamps will negate the meaningfulness of control.

4. **Transparency as a means of enhancing trust**. In general, transparency is positively correlated with and potentially causally related to trust. In the context of AWS, one aspect of transparency is the actual scrutability of machine decision-making. Ensuring scrutability of AWS decision-making processes will encourage trust in human operators and the broader citizenry. Such transparency will also encourage trust at the institutional level by demystifying machine decision-making, while simultaneously enhancing accountability frameworks.

PART I: Workshop Design

Background:

The arms race of the 21st century is embodied in the exponential increase of computing power. Artificial intelligence ("**AI**") is finding its way into battlespace through autonomous weapon systems (**"AWS"**)—"weapon systems that, once activated, can select and engage targets without further intervention by a human operator".[1] This has generated ongoing debate among governments, academics, and civil society, since AWS threaten to upend existing legal regimes and moral norms.

Parties to the debate over AWS continue to discuss the legal implications of the deployment of these weapon systems with respect to accountability. Lacking a "soul to be damned and a body to be kicked",[2] AWS cannot themselves answer for breaches of international humanitarian law ("**IHL**"). AI's lack of accountability invites us to ask questions on its behalf, *e.g.*, who then, in the chain of decision-makers, should be held accountable? And, would such accountability be congruent with national and international criminal law regimes?

The issues become even more complex when AWS are used to project lethal force, since this deployment highlights the moral and legal issues related to these technologies. Some have argued that the very notion of machines making lethal decisions—pulling the trigger—without direct human involvement or oversight is morally impermissible. This has led some parties to call for an outright ban on AWS that lack MHC, arguing that the delegation of life and death decision-making is an irreversible erosion of the basic principles of humanity. Others argue that the promise of better-than-human performance with concomitant decrease in non-combatant casualties provides sufficient justification for the use of AWS.

Meaningful human control ("**MHC**") has been suggested as a middle ground between the different positions on AWS. The idea is that the right level of human control over key aspects of the AWS decision-making process could address both legal and moral

---

[1] This is the definition used by the United States Department of Defence. See, *e.g.*, Ariel Conn, "The Problem of Defining Autonomous Weapons", Nov. 30. 2016, online: https://futureoflife.org/2016/11/30/problem-defining-autonomous-weapons/.

[2] This phrase is often used to bemoan the lack of accountability for robots. See, e.g., Peter M. Asaro, "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics" 169-186 in Patrick Lin, Keith Abney and George A. Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics*,(Cambridge: MIT Press, 2012).

concerns. Such hybrid systems are not, strictly speaking, "AWS": they are not fully autonomous since they involve humans either "in the loop" or "on the loop". Although the proposal for MHC has moved the debate along, not all parties agree on terminology or what exactly MHC should look like. For example, the US, Germany and other countries including Canada prefer the concept of "appropriate human involvement".[3] It is becoming increasingly clear that stakeholders within the debate are coalescing around the belief that an explicit definition of, and requirement for, meaningful human oversight/control is necessary for AWS. However, one has yet to emerge.

A level of trust between human and machine is necessary for either delegation of responsibility to the autonomous system or effective collaboration between humans and machines. To date, however, few scholars have addressed the concept of trust in research on AWS.[4] Nonetheless, as Roff and Danks state, "a failure to understand how humans can or cannot trust autonomous weapons systems has direct consequences for military strategy, the conduct of hostilities, and even long term economic impacts from investment and procurement choices."[5] Many problems remain to be solved if MHC is to effectively address the operational, ethical, and legal imperatives associated with AWS.

Workshop Format:

Targeting Trust was designed as a one-day interdisciplinary workshop. It began with "Primer and Provocation" presentations by the Principal Investigators of the project, Drs. Ian Kerr and Jason Millar, and co-organizer Dr. Jacquelyn Burkell. These presentations were intended to frame the discussion surrounding AWS, trust, and trustworthiness, as well as to provoke thoughts and reactions from participants, to be kept in mind for the balance of the day.

The presentations from the organizers were immediately followed by presentations from the four invited experts, each of whom presented two "Big Questions" to the

---

[3] Jean-Baptiste Jeangène Vilmer, "Autonomous Weapon Diplomacy: The Geneva Debates" 2016, 3 Politique Étrangère 119, online: https://www.cairn-int.info/article.php?ID_ARTICLE=E_PE_163_0119.

[4] Heather M Roff & David Danks, "The Necessity and Limits of Trust in Autonomous Weapons Systems" (2017, forthcoming) *J Military Ethics*; Thomas W Simpson, "Robots, Trust and War" (2011) 24:3 *Philosophy and Technology* 325; Brent J Steele & Eric A Heinze, "From Smart to Autonomous Weapons: Confounding Territoriality and Moral Agency" in Caron E Gentry & Amy E Eckert, eds, *The Future of Just War: New Critical Essays* (Athens; University of Georgia Press, 2014) 98 at 104.

[5] *Ibid*.

participants. The experts then together discussed all of the "Big Questions" that had been presented and answered additional questions from the participants in the room. Participants were then asked to vote on which of the "Big Questions" they wished to explore further in morning breakout sessions. From the eight proposed questions, the top three choices were identified as the topics for the three breakout sessions. Participants selected the breakout session that best aligned with their interests. Each breakout session was moderated by the expert who had proposed the "Big Question".

Lunch was followed by an afternoon breakout session. The workshop organizers curated the topics of discussion for the afternoon breakout session in advance. Three conversation topics were advanced: (1) the relationships between trust, trustworthiness, and MHC in the context of AWS; (2) the relationships between trust and various stakeholders in the maintenance of an operational system; and (3) the conditions that justify trusting a technology or delegating a task to that technology. Participants were once again free to participate in the breakout session of their choice.

Targeting Trust concluded with a discussion of policy implications guided by the workshop organizers. This discussion synthesized the many positions and propositions advanced during the day's interactions and discussions. The conversation in this plenary discussion focused on the role trust and related concepts can play in the debate surrounding AWS and MHC.

PART II: Summary of Discourse

Introductory Presentations by Organizing Experts

Dr. Jason Millar: Though it is rare to reach broad consensus on definitions, it is worthwhile to examine the role that definitions play in framing the AWS debate. Concepts of trust and trustworthiness play an important role in this definitional framework. A more nuanced understanding of these various concepts can shape the ethical landscape of the AWS debate and impact subsequent policy decisions. As long as "MHC" is used as a framing concept, it will be necessary to understand how it relates to human-machine interaction.

Trust, as an attitude held by humans, can be both subjectively and objectively held. Subjective trust arises based on a person's feelings. These feelings may not relate to any precise indicia of reliability in something or someone. Conversely, objective trust is held by a person based on specific knowledge about a person or thing, which allows them to assess that person's or thing's reliability. Trustworthiness is a measure of the extent to

which we are justified in trusting a person or thing. Conceptually, it is intertwined with both subjective and objective trust.

A person's trust may be violated conventionally or morally. A conventional violation of trust describes a situation in which you trust something, say a rope, and that rope unexpectedly breaks or fails, indicating that it was not as reliable as you previously thought and potentially not as trustworthy. Conversely, a moral violation of trust relates to situations in which you feel morally wronged by having had your trust violated, such as when a friend tells a secret of yours that they had promised not to tell.

Policymakers must consider not only the type of trust, but also its quantity. 'Overtrust' refers to situations in which too much trust is placed in a person or thing that is not in fact so trustworthy. Under-trust occurs when something is legitimately trustworthy, but a person nonetheless places insufficient trust in it.

A requirement for MHC may be linked to the perceived gravity of the activity being delegated to a system. The international community's concern about the delegation of decisions to use lethal force has led to the current debate about MHC. Many consider MHC to be a baseline requirement for the legal and moral use of AWS. If it is, a relatively common understanding of the meaning of MHC could allow parties to distinguish between moral and immoral, or legal and illegal, uses of an AWS by regulating the justifiable levels and points of automation in a system.

Some important questions to consider: (1) what is the relationship between these various concepts; and (2) is MHC a proxy for trust?

Dr. Ian Kerr: The relationship between human and machine has become increasingly complex. In many instances, humans and machines have a symbiotic relationship. However, there are also cases in which we choose whether a human should complete a task, or whether this task may be delegated to a machine. Finally, there are situations wherein humans and machines have an adversarial relationship, in which it is not clear which party is better trusted to accomplish a task.

Relinquishing human decision-making requires determining thresholds for when delegating certain tasks becomes permissible for humans. This decision involves both moral and technical thresholds. If society determines that it is morally permissible to delegate a task, then a consensus must also be developed around the standards a machine must reach to satisfy our concerns. Must a machine be as good as a human—or better? Answering this question reveals that we could have different expectations—and

define different trust levels—depending on what kinds of task or decision-making is being delegating.

The "Tripartite Nature of Trust", expressed as *X trusts Y to do Z*, conveys the interconnectedness between two entities (at least one of which is human) and the thing that is to be achieved. Y need not be human in this relationship (recall X trusting that the rope will not break); this could impact the trust relationship. A human may be more likely to trust another human to achieve a certain task based on subjective (affective) trust. Similarly, there are instances in which humans may be more likely to trust a machine with a given task. Humans tend to determine the appropriateness of trusting a machine based on reliability, rather than subjective "feelings" of trust. Currently, there is no human-like connectedness between humans and machines. However, as machines become increasingly anthropomorphized, the distinctions made by humans as to who or what they are willing to trust, and why, increasingly blur.

Humans have long sought to create predictable and reliable machines. Machine learning, and the "emergent" behaviour it sometimes displays, impact our notions of reliability as AI becomes better able to exceed its original programming. This is not to say that machines are the only party within the tripartite relationship that can be unpredictable. Humans develop reliability and trustworthiness through social forces that do not act upon machines. Humanity has even created social and physical systems to compensate for the unreliability of people.

"Trustworthy Computing" and "Trusted Computing" offer two competing visions of systems that address human unpredictability. Trustworthy Computing, which relies on security, privacy, and reliability, was introduced in the early 2000s as a means of encouraging people to access and use the internet. These features provided a necessary sense of security and reliability. Trusted Computing, on the other hand, was a response to a perceived lack of human reliability. By layering permission and rights, the systems themselves could become trusted, since the actions of human were not.

Trustworthy Computing was designed in the early days of the Internet to create the conditions for a corollary understanding of 'meaningful human control' by establishing the reliability necessary to convince people to engage in online activity. Trusted Computing, on the other hand, was an attempt to circumvent meaningful human control through an architecture and design of the Internet that would preemptively preclude certain forms of human action. Given these competing conceptions of trust in a technological system, one might therefore ask: under what conditions is MHC a feature and under what circumstances is it a bug in the system?

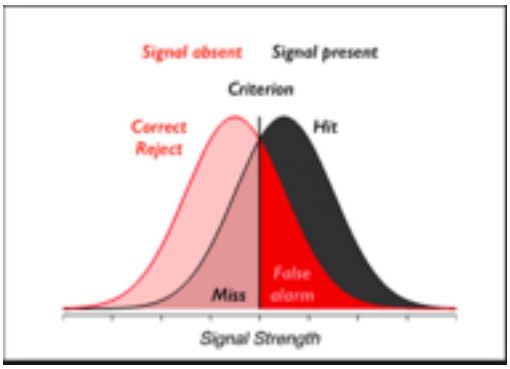
Dr. Jacquelyn Burkell: 'Trust' is fundamentally a psychological concept, and considering it through this lens promotes a better understanding of how humans might interact with machines. Trust is a belief or an attitude held by a 'trustor' toward a 'trustee,' usually with respect to the performance of a particular task or with respect to the outcome of a decision-making process. The trust held by a trustor toward a trustee can be broken down into two types: affective and cognitive. Affective trust is based in emotion, is non-evaluative, and tends to develop between people (and perhaps between people and intelligent technologies) over varying time frames, sometimes quite quickly based on "feeling". Cognitive trust is evaluative and based on performance, and may be placed in individual people or things, but also in organizations and groups, *i.e.*, those tasked with producing systems upon which we may rely.

Cognitive trust develops over time and through experience; it is enhanced when the trustee is observed to perform reliably and effectively. In this way, trustworthiness in the form of reliability can enhance a trust attitude, with increasing reliance as the result: given repeated experiences of effective performance, a stronger trust attitude is assigned to the person or thing (the trustee), and the trustor will develop an enhanced belief that the trustee will behave as expected.

We rely on something or someone when we delegate tasks or decision-making to them (or it, in the case of technology). Trust is almost certainly a necessary condition for reliance–that is, you cannot rely on something or someone you consider to be untrustworthy with respect to the issue at hand. At the same time, trust does not necessarily and immediately lead to reliance; in fact, it is very likely (especially in conditions where the outcomes in question are significant) that reliance requires both initial trust (possibly affectively based) and repeated direct experience of reliable behaviour. Reliability, on the other hand, describes the performance of the trustee. Reliability tends to foster trust: in general, the more reliable the performance of the trustee, the greater the trust held by the trustor.

There is an important relationship among the concepts of trust, reliability, and reliance. Reliable performance by a trustee will tend to create trust in the trustor. As trust increases, so too may reliance–we can easily begin to rely on a demonstrably effective delegate to perform tasks for us. This, in turn, can result in a lack of control but, also, to a loss of capacity, since we are no longer performing the task and thus no longer perfecting our own performance. Ultimately, we can come to believe (perhaps accurately) that the task in question will be better performed by the machine to which it is delegated.

Describing human relationships with machines through the psychological construct of trust is not a perfect fit. We are borrowing concepts meant to describe a relationship between two humans. However, this application to the human-machine relationship is not wholly incongruent because interactions with machines automatically engage social responses that are typically the result of human interaction. These innate human responses are subsequently extremely difficult to disengage. Thus, humans interacting with machines act and respond very much like humans interacting with other humans, including in the area of trust.



What does it mean to trust 'appropriately' – that is, neither to 'overtrust' nor 'undertrust'? The answer cannot be to get it 'right' every time – that is, to let the machine do the job when it gets it 'right', and to take over when the machine gets is 'wrong' – because that would require prescience that none can possess. In fact, the question of whether to trust the (fallible) machine do its work (whatever that work may be) in each individual case is analogous to a *signal detection* problem: in each individual situation the decision maker has to trust or not trust the system (declare 'signal present' or 'signal absent' in signal detection terms) in the face of inconclusive evidence regarding how the system will perform. The issue with signal detection is that even the *best* criterion, or cutoff, for determining when to trust and when not to trust will result in some 'misses' (accurate performance when the system was not trusted) and some 'false alarms' (inaccurate performance when the system *was* trusted). This analysis makes evident that 'overtrust' and 'undertrust' cannot be defined in terms of the calibration of trust to the outcome of each and every decision or action that a system—in other words, individual failures of appropriate trust cannot be taken as evidence of 'overtrusting' or 'undertrusting'.

Instead, we have to think of appropriately calibrated trust (neither 'undertrust' nor 'overtrust') as a function of overall system performance – that is, aggregate performance across multiple individual tasks. All else (including the significance of the task) being equal, appropriately calibrated trust would have the result that a less reliable system–a system that performs less accurately and more variably–would be more closely monitored, while a more reliable system would attract less close monitoring. This implicates human control–whether MHC or 'appropriate human involvement–in the calibration of trust. Critically, we will need to accept that calibration failures are inevitable. Even with optimally calibrated trust (and given a less than perfectly performing machine), we *will* sometimes delegate to a machine that fails in its task (or

vice versa)–because if we were able to *predict* these failures we would either improve trust calibration or improve system performance.

There are methods that can be used to improve trust calibration: perhaps most critically, systems can signal their level of certainty in a decision or action. This extra information, in addition to (and preceding) system action or decision, can signal specific instances when trust should be moderated, and greater human oversight or control should be affected.

Lightning Round – Experts' "Big Questions" Presentations and Discussions

Expert 1: Measuring trust as part of MHC is important. There are multiple lenses through which trust might be considered. If engineers measure reliability objectively, trust will remain esoteric and subjective. This problem is compounded by attempting to translate engineering knowledge to later stages of AWS development, such as legal review. Designers and lawyers undertake their various considerations differently: coherently predicting weaknesses or potential failures in an automated weapon system is difficult because of the apparently inconsistent legal or engineering understandings of reliability and trust. Being unable to fully predict how AWS will function further challenges legal review, because legal review must be completed before deployment of AWS, yet it is only through deployment that AWS is likely to encounter scenarios and environments that will challenge its design. This creates a paradox.

Two Big Questions to consider:

1. What role can trust play in establishing MHC over weapons? How could it be observed, tested or verified?
2. How does the concept of trust in existing weapons systems differ between engineers, reviewers (acquisition/Article 36[6]), commanders and operators?

Expert 2: AWS questions are related to elements of Accident Theory, originally proposed by Charles Perrow. He posits that the tightly coupled interactions that are a necessary facet of high risk technologies (nuclear, chemical, genetic, *etc*.) make it almost impossible to predict the exact point of failure when things go wrong. We must accept that AWS are a similarly high risk technology and that they are therefore inherently untrustworthy. This problem is compounded by machine deep learning and the theory of emergence, which holds that AI arrives at decisions and conclusions beyond the scope of its original programming and in a manner that is inscrutable to human operators. Despite these serious concerns, there are also practical problems to be solved in AWS procurement and interoperability.

Two Big Questions to consider:

1. Canada typically fights in alliances/coalitions. How will the interoperability required between military forces be affected by artificial intelligence (AI)?

---

[6] International Committee of the Red Cross (ICRC), *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, of 8 June 1977.

2. We [Canada] cannot even procure inflatable boats; how does the government think it can procure AI?

Expert 3: There are many different permutations of MHC being proposed at the international level, but it is not explicitly clear what they all mean. On the one hand, the technology-neutral nature of this term is beneficial since it leaves room for flexible development. On the other, an open-ended term runs the risk of being interpreted too broadly. It will be problematic if the prioritization of human control in operations begins to eclipse other requirements under IHL. Moreover, it is necessary that human involvement remains meaningful, rather than as a rubber stamp, offering nothing more than a place in which liability may be located when a system fails.

Any application of MHC will require striking the proper balance between over and undertrust. Thought must be given to how tendencies to over or undertrust, and trust relationships between man and machine, might be exploited by bad actors. Failing to do so might render MHC more of a weakness than a strength.

Two Big Questions to consider:

1. Consider different versions of the general proposition "*X* human *X*". One is *meaningful* human *control*. Another is *appropriate* human *judgment*. What are we really concerned about losing/preserving according to this proposition, as AI systems multiply?
2. Sometimes humans trust AI too little, sometimes too much. When are these traits problematic? How are they likely to be abused by bad actors?

Expert 4: The collective international handwringing around AWS and MHC does not appear to consider cyber operations as part of the same debate, despite the fact that cyberattacks can have potentially devastating kinetic effects. If the international community is seeking a unified policy on AWS and MHC, then cyber must be incorporated.

Whether AWS are cyber or not, humanity has vast experience in dealing with untrustworthy and unpredictable agents: other humans. Humanity deals with this trust gap by insisting on accountability, and by developing norms and systems of enforcement. No such system exists for AI—yet. Examining existing systems can help inform how we might approach these same issues for AI.

Two Big Questions to consider:

1. Why is "meaningful human control" rarely insisted upon for cyberops, even though it can be autonomous too? What does it mean for AWS policy?
2. Can trust exist if an autonomous system might be unpredictable? If not, how do we ever trust humans who are also unpredictable?

Expert Interaction: The invited experts were offered a chance to discuss each other's proposed questions, as well as respond to questions from participants. This preliminary discussion was intended as a precursor to a vote on which questions would form the basis of the morning breakout sessions.

Some of the experts debated whether a unified policy of AWS and MHC, including cyber, is truly necessary. Concerns were raised that a policy encapsulating all forms of weaponized AI would be too broad and introduce vagueness. Some parties believed that it would be best to tailor policy to more specific visions of weapons systems. It was suggested MHC could be developed as a standalone concept for technology policy and AWS, before determining how it might apply to the cyber realm. Nonetheless, the experts agreed that 'cyber' should be kept in mind as part of the greater AWS debate.

The experts briefly discussed the relationship between trust, trustworthiness and MHC. They all agreed that the exploration of trust/trustworthiness could help stakeholders to articulate the many concepts subsumed within the overarching concept of MHC. This discussion mirrored Dr. Millar's proposition that MHC could be a proxy for trust. The experts suggested that MHC, as a governing concept, might eventually be eliminated once appropriate trust, i.e. a level of trust that is justified, is established with AWS.

Morning Breakout Sessions – Discussion Summaries

Big Question 1

Expert 3: Consider different versions of the general proposition "*X* human *X*". One is *meaningful* human *control*. Another is *appropriate* human *judgment*. What are we really concerned about losing/preserving according to this proposition, as AI systems multiply?

Expert 3 asked the participants to consider two additional questions as a way of getting at the overall Big Question:

1) Is the introduction of MHC merely an attempt to preserve a certain conservatism in how we conduct military operations?
2) Can MHC be understood as an attempt to preserve a level of comfort, since autonomy has not yet fully found its way into our everyday lives?

MHC provides a conceptual path towards acceptable legal and moral uses of AWS. Participants suggested that if AWS were to function reliably *vis-à-vis* IHL 99% of the time, the debate around the legality requirement for MHC would not exist. Given the current legal state of affairs, assuming AWS made decisions in violation of IHL, MHC would provide just enough of a human element to allow our current legal system to accommodate the technology. This is especially so given the nascent state of AWS, the absence of Article 36 frameworks for testing and verification and issues of accountability. Participants recognized that existing legal frameworks are meant to address human actors because humans are a known quantity in a world of rapidly advancing technology, even if imperfect. MHC may provide for a more understandable and predictable interface with machines, while also addressing concerns of comfort and accountability.

As a moral precondition, many participants believed that a degree of humanity would be lost if AWS were to make lethal decisions without MHC. Participants agreed that the decision to kill is best made by a human. Some participants expressed fear that a degree of mercy would be lost, as this is an attribute that could be shown by a combatant, but not by an AWS.

MHC in all aspects of the design and deployment of AWS is not a panacea. Participants expressed concern that human biases could negatively affect AI programming, resulting in AWS that reflect certain social and cultural prejudices. At the same time, AI development might also require MHC as a means of combatting this risk.

Opinions varied regarding the quality of AI decision-making as compared to human decision-making. Some prioritized the emotion and creativity of human decision-making, which they argued better facilitate proper moral and legal decisions to be made in battlespace, on a case-by-case basis. Conversely, some participants placed greater faith in the unemotional nature of AWS, believing this to be an especially useful feature when tasks are dull, dirty, and dangerous.

This discussion highlighted the need to work from common definitions and understandings of exactly what we mean by "autonomous" weapons systems, as well as by MHC. It became clear that participants were sometimes conflating the concepts of "automated" (*i.e.*, systems that contain elements or subsystems that are automated but not full end-to-end automation) and "autonomous" (*i.e.*, systems that have full end-to-end automation of all features, most controversially targeting and firing) systems. Furthermore, participants debated what exactly is included in the concept of MHC, and whether that means "ultimate control" including the ability to override AWS decision-making.

Conversations coalesced around the notion that MHC could preserve accountability and responsibility in the event of grave breaches of IHL. This might increase public trust at both an operational and institutional level as well as increasing the trustworthiness of the overall system.

Big Question 2

Expert 1: How does the concept of trust in existing weapons systems differ between engineers, reviewers (acquisition/Article 36), commanders, and operators?

Trust means different things to different stakeholders within the AWS debate.

Trust is an integral part of the military chain of command. Participants (some of whom were current or former military members) described the command chain as 'personality driven' and, therefore, subjective: some commanders are more willing to engage with operators and systems at "lower levels," while others "stay above the fray." (editors: these terms were not clearly defined). Similarly, some commanders were described as more willing to trust subordinates while others might micromanage. These differing styles of leadership reflect the varying levels of trust that exist within chains of command. This applies to both affective and cognitive trust.

Other participants disagreed with some of the claims made in the prior expert panels, arguing that trust, insofar as it is an aspect of engineering and programming, is more a

reflection of reliability than of affect. They suggested that reliability in current systems is simpler to gauge, making their trustworthiness more easily established. Increased autonomy and complexity, they thought, may lead to diminished trust in systems because machine decision-making can result in unanticipated and/or inscrutable outcomes and behaviours. It was suggested that concerns about trust might be mitigated by some form of meaningful human control or involvement.

Some participants felt that AWS stakeholders will develop the trust necessary for their given tasks, whether in development, procurement, *etc*. But since these parties do not exist in a vacuum, it was suggested that it would be necessary to transfer trust along the chain to the end user. Establishing shared standards against which reliability can be measured is vital. Participants with military experience tended to believe that operators are more willing to trust a system based on demonstrable reliability.

The idea that weapons procurement should involve wider stakeholder engagement also emerged in this discussion. For example, participants suggested that vulnerable populations that have previously been the target of military operations should be consulted as new weapons are developed. Their potential unfamiliarity with how weapons are used and how targets are selected could be a source of anxiety and psychological strain. The example of armed drones was offered. Participants seemed to agree that wide-ranging, multifaceted stakeholder involvement could increase transparency which, in turn, could increase trust, insofar as those considerations find their way into the system design, at all levels.

Big Question 3

Expert 4 Question # 2: Can trust exist if an autonomous system might be unpredictable? If not, how do we ever trust humans who are also unpredictable?

Despite the framing question, the conversation focused on cyber operations and the similarities to and differences from AWS.

Cyber and autonomous weapons are used differently; the use of the former tends to be virtual with possible kinetic effects, while the use of the latter normally involves technologically-guided deployment of kinetic weapons. Going forward, AI will play a vital role in both forms of warfare. Participants agreed that trust in each form of weapon is compromised by issues of unpredictability and inscrutability.

While some headway has been made in regulating cyber warfare at the international level due to the *Tallinn Manual*,[7] no such progress has been made regarding AWS. Participants conceded that this is likely a result of the fact that cyber is already a clear and present danger, while AWS remains mostly in the theoretical sphere.

Finally, the discussion turned to human soldiers and the possible ways in which they could be enhanced in the future, potentially blurring the line between human combatants and weapons. Participants posited that such enhancements could reach a threshold at which Article 36 review would be triggered. This notion prompted a consideration of whether trust might still be an issue when the line between human and weapon is blurred.

Afternoon Breakout Sessions – Discussion Summaries

Dr. Burkell: In the context of AWS, what is the relationship between "meaningful human control", "trust", and "trustworthiness"? Are "trust/trustworthiness" necessary or sufficient for MHC, or vice versa?

One important discussion point is that trust and/or trustworthiness are not explicitly necessary for meaningful human control. Some felt that control may still be meaningful even in the absence of trust—perhaps more so due to an operator's wariness of AI decision making. However, most agreed that the two concepts seem to be intertwined.

The view was expressed that MHC could influence the trustworthiness of AWS operations at the macro level, since humans and machines working together could be more reliable than either working alone. This could lead to greater societal acceptance of AWS, as citizens may be willing to trust that their government and military are conducting legally and morally sound operations. Trust at this level includes notions of accountability, as MHC would allow for responsibility to be located in a human actor in the event of a mishap.

Most participants held that trust and trustworthiness are both relevant to MHC. These concepts are linked, but they have different implications for MHC in AWS operations. Some participants expressed that trustworthiness is essential for MHC, but over- or undertrust can undermine it. Some believed that trustworthiness is necessary for a system to be used effectively. Without it, they thought, the human exercising MHC would be forced to second-guess all of the AWS' decisions and there would be no

[7] Michael N. Schmitt, ed, *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge: Cambridge University Press, 2013).

advantage to having automated the system. As trustworthiness and reliability increase over time and use, humans will delegate more to autonomous systems. Increasing anthropomorphization of machines interacting with humans could cause humans to place affective trust in autonomous systems. As one workshop participant explained it, once trust is established (or perceived as having been established), the human operator tasked with MHC is less likely to scrutinize the system's decision-making. This might undermine the meaningfulness of their control. The tension between trustworthiness and trust was thought to be conceptually similar to Dr. Millar's exploration of undertrust and overtrust. Finding the appropriate balance is imperative.

Finally, participants observed that the appropriate balance between trustworthiness and trust in MHC would likely vary depending on what exactly has been delegated to AWS. If AWS is engaged in lethal operations, then trust might be lower and the perceived need for MHC higher. This is connected to the apparent trustworthiness and reliability of the given system in similar circumstances.

Dr. Millar: When considering the many stakeholders necessary for the complex operation of AWS, where are the concepts of trust and trustworthiness most critical to maintain?

- What would be the consequence of NOT having trust with respect to one of these stakeholders?
- How does trust function with each of the stakeholders?
- What does it mean for each stakeholder to trust the system? (List the stakeholders and ask for more input. Think of the two notions of trust: (affective versus epistemological.)

Participants considered that the possible stakeholders in AWS are wide-ranging and might include: operators, procurement officers, suppliers, non-governmental organizations (NGOs), international organizations, engineers, combatants, defence contractors, and civilians (who sometimes become 'collateral damage').

It was noted that the nature of the relationship between these various stakeholders has not yet been determined in the context of AWS. However, it was suggested that stakeholders must establish a relationship of trust between the defence industry and the military end users/operators. Inherent in this relationship are notions of legality, review, and demonstrable reliability. Governments procuring AWS could benefit from close collaboration with industry, since the bulk of AI research and development is occurring in the private sector. It was acknowledged that governments must be able to trust that the systems they procure and ultimately deploy will be able to meet requisite national and international legal standards. The parties to this relationship must also

recognize that there may be a difference in the degree of trust required for government and industry. Governments and operators are the entities likely to be called upon to answer for AWS related mishaps, unless legal liability for industry is explicitly established. Trust must not be blind. Parties must collaborate to create shared standards that meet their collective needs.

Participants pointed out that trust relationships also exist between a military chain of command and its operators, due to accountability models wherein operators bear a higher legal burden than the chain of command. An operator has no choice but to rely on her chain of command's assurances of legality, trusting that a specific weapon meets legal requirements. Relatedly, those higher in the chain of command must trust that the operator will deploy the weapon in accordance with the law. This relationship is simpler when considering conventional weapons, since their legality has been established. The *meaningfulness* of human control takes on greater importance for the individual operators if they must answer for breaches of law.

Participants also considered that trust between stakeholders does not take into account the varying positions on AWS held by each individual stakeholder. It was pointed out that some stakeholders are less likely to trust industry. This disparity in trust may have impacted the AWS debate. This is another example of an area in which specific definitions could help move the debate forward. Establishing a common conceptual and definitional framework would increase trust that the many stakeholders are indeed debating about similar ideas.

Dr. Kerr: Reflecting on examples of other autonomous systems that replace humans with technology (*e.g.*, caregiving, surgery, driving), under what conditions are we justified in trusting the technology and why? In what circumstances would that trust justify delegating a task to that technology? Provide specific examples to illustrate the point.

There are many near-term technologies in development to which humans appear willing to delegate certain tasks and decisions, including autonomous cars, robotic caregivers, and surgical robots.

Participants reiterated that the AWS debate has made clear that the notion of delegating lethal decision-making does not feel "right". Technologies such as surgical robots were distinguished by participants insofar as they are ostensibly meant to benefit society and tend to generate less uneasiness. It was pointed out that as the risk of possible harm increases, people become less comfortable with the idea of delegating tasks to autonomous systems even if such systems are aimed at achieving beneficial ends.

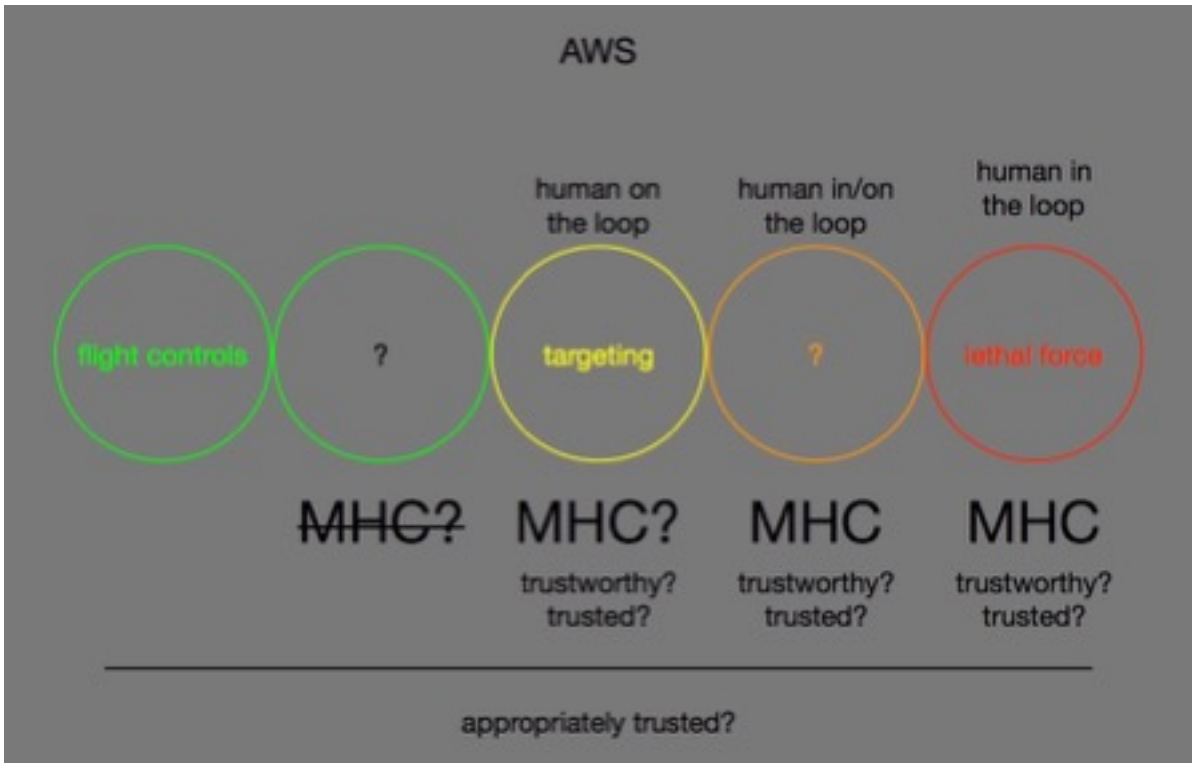
Another discussion focused on the fact that the nature of trust and the appropriateness of delegation are also impacted by the ability to make a choice with respect to the use of the technology. It was observed that many autonomous technologies currently in use require the user to willingly engage with them. Individuals who are skeptical of the merits of autonomous technologies can choose to opt out. This is not the case in the context of AWS. It is unlikely that an individual operator will have the ability to refuse or exercise individual discretion to use a weapon; combatants and noncombatants have absolutely no say in how weapons are used against them.

Participants generally held that trustworthiness is a necessary precondition when considering the appropriateness of delegation. Functionally, trustworthiness comes from demonstrated reliability. Some people were of the view that this allows individuals to satisfy themselves (lethal decision-making aside) that their decision to delegate a task is consistent with morality. Others held firm to that fact that trustworthiness is insufficient, since they believe that the moral decision to kill is not permissibly delegated to machines regardless of their trustworthiness in adhering to IHL or rules of engagement.

Discussion of Policy Implications

Part 1 – AWS System Diagram: The organizing experts began by presenting the group with a System Diagram, visually representing a hypothetical and abstracted AWS with varying degrees of MHC identified at various points in the overall system. Participants were also given a handout with questions to guide the discussion.[8]

[8] See Appendix III.



Participants were asked the following (with reference to the diagram above):

1. In your opinion, what key sub-tasks (bubbles) have we missed, if any, that are worthy of discussion?
2. In your opinion, what considerations (red flags?) arise with respect to each of these tasks when considering automation?
3. In order for you to be able to trust automation of each sub-task, what would have to be in place? How would you address it?

Question 1: Participants proposed that the "Targeting" bubble was an oversimplification of that step, since it did not explicitly account for the IHL-imposed obligations of distinction and proportionality. Satisfaction of these obligations is a legal precondition to launch an attack. Participants agreed that the obligation of distinction is easier to achieve technologically than other IHL obligations, because it is a more objective, binary decision. On the other hand, the proportionality requirement was seen as a more subjective determination, requiring a holistic analysis of many contextual factors. It was thought that the differing nature of these IHL obligations impacts the nature of the relationship between human and machine, particularly with respect to delegation and MHC.

Both proportionality and distinction are vital aspects of IHL, but distinction is a mandatory first condition that acts as a gateway to further action. If a combatant, or

AWS, is unable to accurately determine whether a person or object is a valid military objective, then proportionality is moot, since any further attack would be illegal.

Some participants thought that it is reasonable to project that humans may eventually privilege the superior computing and sensory power of AI. If overtrust in AWS were to occur, some participants questioned, would human operators likely be removed from battlespace altogether? Over-trust in such circumstances could result in immoral or illegal determinations. Additional Protocol I[9] of the Geneva Conventions holds combatants to a standard of reasonableness. Some participants wondered how this standard is applied and whether it might have to be reevaluated in light of AWS and MHC.

Question 2: Participants agreed that proportionality presents the greatest challenge for AWS. There was a general perception that the subjective nature of the proportionality determination is at odds with automation and the MHC standard. It was thought by most that, for the time being, human input remains necessary to synthesize data for assessing the appropriateness of anticipated collateral damage.

Many participants strongly supported the idea that targeting and kill decisions necessarily require a human in the loop.

With respect to human control, the functional difference between MHC and "appropriate human involvement" remains unclear, except that most participants saw the latter quite clearly as a lower standard. Once again, this discussion highlighted the importance of definitions; the difference between "control" and "involvement" needs to be delineated. This is especially true if human involvement/MHC is implemented as a means of addressing issues of accountability.

Participants expressed concern that individuals including roboticists, programmers, and engineers could be held accountable for AWS actions, even if they were not the humans exercising meaningful control. Although MHC and accountability are important, scapegoating is not a viable answer to the so-called "accountability gap". Many agreed that a nation that operationalizes AWS must clearly define chains of accountability.

Question 3: Some participants believed that the ability to create affective trust between human and machine in a MHC context could lead to increased performance of both entities: the optimal outcome for machine-human partnerships would be for humans to trust a machine that performs reliably and accurately. Misplaced trust–that is, trust

---

[9] *Supra* note 3 at Article 57.

placed in a machine that was *not* performing reliably and accurately–was recognized as a potentially problematic outcome. It is critical, most participants felt, that trust be appropriately calibrated to the trustworthiness (typically reliability) of the machine. This calibration is necessary for not only the proper performance of an AWS with MHC, but also for larger notions of trust. Some thought that machines designed to engender affective trust could prove problematic insofar as they might destabilize traditional notions of trust, as it is an inherently emotional type of trust placed in a system devoid of emotion.

Part 2 – Final Discussion Questions:

1. What do you feel are some relevant policy issues related to all of this?
2. Would it be useful to adopt or avoid certain terminology for reasons expressed in today's workshop?
3. Do any key knowledge gaps jump out at you that might prevent the development of concrete policy?
4. Would it be useful to recommend the formation of specific working groups to further investigate particular issues?
5. Are there any areas for coordination either within the Government of Canada or internationally that seem useful, important or essential?

Participants focused their discussion on Question 2. They debated which concept was better: MHC or appropriate human involvement. Most participants agreed that, irrespective of the selected phrase, the vital consideration is the human element of the equation.

With this in mind, Dr. Millar's introductory question was once again put to the participants: is MHC a proxy for trust? Participants suggested that a better way to frame this question is, "What would it take for you to trust AWS? Is MHC sufficient?" Many thought that reliability in AWS is a prerequisite for trust. If a system is sufficiently reliable and trustworthy, then MHC becomes less of a focal point. Some participants were of the view that this does not apply in the case of projecting lethal force that might morally require human participation of some sort or other. However, in the absence of reliability, participants generally agreed that MHC becomes necessary. The legal versus moral necessity of MHC differs based on stakeholder perspective. Some feel that MHC is required on both moral and legal grounds, while others think that MHC is required on one ground or the other only.

Differences in opinion amongst the participants illustrated the need to integrate a wide range of stakeholders into future policy debates, since AWS and its related moral and legal quandaries implicate many facets of society. For these reasons, the debate around

AWS was seen to trigger emotional responses in the participants regardless of their stakeholder position. Whether or not the rise of AWS is inevitable remains to be seen. Regardless, it is imperative that governments, militaries, industry, and civil society remain engaged in order to ensure that any possible future systems are the result of honest consideration and engagement. It was generally held that it is not necessary or desirable to give up our collective moral values in the process of delegating authority to AWS.

PART III: Policy Considerations for Further Exploration

The discussions during *Targeting Trust* touched on almost all areas of the AWS debate. However, participants and experts alike appeared to continually return to a few key themes. We conclude this report by enumerating four themes with the aim of briefly highlighting related insights that will be useful for policy purposes when considering MHC going forward.

1. Clearly defined concepts and terminology

The workshop made clear the importance of shared concepts and terminology. There seems to be a general consensus within the global AWS debate that *some* form of human oversight/control in AWS operations is desirable. Opinions on various conceptions of MHC and which should be the prevailing concept tended to differ based on stakeholder perspective. For some, mere human involvement may be sufficient; for others, meaningful control and veto power is necessary; for others still, these were only a baseline for regulation.

The Government of Canada, in its 2017 Defence Policy, *Strong, Secure, Engaged*, stated that it is "committed to maintaining *appropriate human involvement* in the use of military capabilities that exert lethal force."[10] Canada should fully articulate what this concept means and how it will impact the operationalization of AWS. This definition will significantly affect all facets of AWS, from development through procurement, standards and testing, training, and eventually to deployment in battlespace. Clear initial articulation of these concepts and consistency in the use of terms will add greater certainty for the many engaged stakeholders, while simultaneously laying the foundation for trust at an institutional level.

2. *X trusts Y to do Z* – The Tripartite Nature of Trust

This simple expression of a trust-based relationship belies the complexity of trust in the context of the delegation of decision-making to autonomous systems. When X, Y, or both, are no longer human, it becomes far more challenging to determine the appropriate trust requirements prior to the permissible delegation of decision-making to an automated system. The substitution of machines for humans without any human oversight would affect the way that trust is perceived and how trustworthiness is measured, though any calculus will necessarily involve considerations of machine

[10] Canada's Defence Policy, Strong, Secure, Engaged, 2017 at 73, accessed at: http://dgpaapp.forces.gc.ca/en/canada-defence-policy/docs/canada-defence-policy-report.pdf. [emphasis added]

reliability. In the absence of meeting the threshold of reliability, it appears that some degree of human involvement or control will remain part of the equation.

It is also necessary to consider this ternary relationship at an institutional and societal level. Thought must be given as to the nature of public trust in public or private sector organizations charged with deciding whether to delegate certain tasks and decision-making to autonomous systems.

3. Transparency as a means of enhancing trust

Transparency is often viewed as a way of generating trust. In the context of AWS and AI, transparency must be considered beyond its usual sense of opacity or lack of secrecy. In the machine learning context, where algorithms are able to recognize patterns invisible to human detection, transparency must be understood not in terms of the visibility of the code itself but, rather, by the scrutability of AI decision-making. Human operators will be more likely to trust AWS if they are able to understand how a system arrives at decisions. However, it is important to recognize that the scrutability of such systems is, on its own, no indicator of the system's trustworthiness. Scrutability will offer humans one basis for trust and should be considered an essential design feature.

Transparency in machine decision-making understood in this sense will also increase trust at the institutional level. It will help to demystify machine decision-making while simultaneously enhancing accountability frameworks. Still, it is no panacea; it is a necessary condition but not a sufficient one.

4. The Promise of Trustworthiness and the Peril of Trust

Trustworthiness and trust are distinct yet intertwined concepts, with significant implications for MHC in AWS operations. While trustworthiness (established through reliability) is an essential element of MHC, trust can threaten to undermine the meaningfulness of human control. Thought must be given, not only to the tension between these two concepts, but also regarding how to achieve balance between them.

Attention paid to this balance will impact all facets of AWS development and use. An understanding of concepts of trust and trustworthiness will allow for the development of AWS that contemplate MHC from the outset. This will enhance the effectiveness of human-machine interactions. Such an approach is likely superior to the development of AWS to which human control must be added at a later date. This approach will also help ensure that the role of the human is more than a rubber stamp.

Conclusion

The authors of this report will utilize the core themes enumerated in the previous section and other knowledge produced at the workshop to guide their continuing work on the important conceptual distinction between trust and trustworthiness, and the role of each in better understanding the regulatory and policy issues associated with AWS.

We would like to end by expressing our gratitude to the *Defence Engagement Program* for the opportunity to develop this workshop, thereby helping us to refine our ideas about the nature and role of trust in AWS Policy Development. We are also grateful to the 40 workshop participants from various sectors of government and the academy for their enthusiastic participation and shared wisdom, which is sure to inform our future scholarship in the field.

Finally, we wish to offer special thanks to Brittainy Cerqua, Ida Mahmoudi, and Dustin Moores for their outstanding substantive contributions to the workshop, and their excellent research assistance, transcription, and assistance with workshop logistics.

Appendix I

Targeting Trust Invitation

Targeting Trust: A Workshop on the Nature and Role of Trust in Autonomous Weapon Systems (AWS) Policy Development

What: You are invited to attend a one-day intensive workshop focused on developing policy options that ensure trustworthy Autonomous Weapon Systems (AWS). Participants will develop a better understanding of "meaningful human control" and "appropriate levels of human judgment" by investigating differing conceptions of these standards based on stakeholder categories and perceived trustworthiness.

This IS NOT your typical workshop. Instead of podium presentations, the workshop will consist of a series of carefully curated and moderated group discussions. Leveraging the knowledge and perspectives of several international subject matter experts, the discussions aim at provoking broad interaction amongst all workshop participants to ensure key stakeholder engagement. The workshop will be held under Chatham House Rules.

Who: Attendees will be participants from the defence community, including DND, Defence Research and Development Canada, JAG, Strategic Joint Staff, ADM POL, and Global Affairs Canada.

When: October 5, 2017

Where: University of Ottawa, Alex Trebek Hall, 157 Séraphin-Marion Private, Ottawa, ON K1N

Why: Workshop participants will have the opportunity to engage directly with internationally renowned experts to enhance their knowledge and understanding of the AWS policy, and will have direct involvement in the development of a Policy Report sponsored by the Department of National Defence's (DND) Defence Engagement Program.

Background

As weapons are increasingly automated, policy makers will need to determine the level of continued human involvement necessary in the target and kill decisions that AWS are meant to expedite. Developing standards that maintain "meaningful human control" or "appropriate levels of human judgment" are typically understood as preconditions for the permissible use of AWS in battlespace. From the perspective of many, such standards are necessary to instill appropriate trust in the deployment of AWS.

The *Targeting Trust Workshop* brings together a group of invited AWS experts from academia, the Department of National Defence, Defence Research and Development Canada, and the Government of Canada, that includes policy scholars, experts on the law and ethics of AI and robotics, and decision-makers. The group will investigate how perceptions of trust in AWS can impact Canada's policy choices, and vice-versa. Workshop participants will investigate whether and to what extent differently framed requirements impact the trust/trustworthiness of AWS and how the perceived trustworthiness of AWS impacts the framing of requirements such as "meaningful human control" and "appropriate levels of human judgment."

To date, only a handful of scholars, activists, and stakeholders have linked the study of AWS policy directly to the concept of trust.[1] And yet, as Roff and Danks correctly state, "a failure to understand how humans can or cannot trust autonomous weapons systems has direct consequences for military strategy, the conduct of hostilities, and even long term economic impacts from investment and procurement choices."[2]

Canada has the opportunity to provide leadership on the world stage by commenting on these important questions. This workshop seeks to spur policy considerations necessary for the development of trustworthy autonomous military systems. Improved domestic expertise will help measure Canada's current and future policy, improve procurement and deployment decisions, and inform dual-use strategies in the Canadian context. This workshop is an important step toward refining Canada's approach to, and participation in, international discussions to determine standards and best practices for AWS at the international level.

Confirmed List of Expert Participants include:

Expert 1: name and biography withheld (the workshop was held under the Chatham House Rule).

Dr. Jacquelyn Burkell: Dr. Burkell is Assistant Dean of Research in the Faculty of Information and Media Studies at the University of Western Ontario. Her research focuses on the influence of cognition on information use, including examinations of information credibility assessment, the use of information in decision making, and human-computer interface design. She is the recent co-recipient of a SSHRC grant to develop ethical and legal frameworks for the governance of machines with artificial intelligence.

Expert 2: name and biography withheld (the workshop was held under the Chatham House Rule).

Expert 3: name and biography withheld (the workshop was held under the Chatham House Rule).

[1] Heather M Roff & David Danks, "The Necessity and Limits of Trust in Autonomous Weapons Systems" (2017, forthcoming) J Military Ethics; Thomas W Simpson, "Robots, Trust and War" (2011) 24:3 Philosophy & Technology 325; Brent J Steele & Eric A Heinze, "From Smart to Autonomous Weapons: Confounding Territoriality and Moral Agency" in Caron E Gentry & Amy E Eckert, eds, *The Future of Just War: New Critical Essays* (Athens: University of Georgia Press, 2014) 98 at 104.
[2] *Ibid.*

Dr. Ian Kerr: Dr. Kerr is a Full Professor and holds the Canada Research Chair in Ethics, Law and Technology at the Univeristy of Ottawa, Faculty of Law. He is a pioneer in the field of robotics and AI law and policy. He is an expert on the law and policy of military robots, driverless cars, and robots and AI in the health sector. Dr. Kerr is a co-recipient (with Dr. Burkell) of a SSHRC grant to develop ethical and legal frameworks for the governance of machines with AI, including autonomous military robots.

Expert 4: name and biography withheld (the workshop was held under the Chatham House Rule).

Dr. Jason Millar: Dr. Millar is a Postdoctoral Fellow at Stanford University's McCoy Family Center for Ethics in Society. His research focuses on examining trust/trustworthiness and responsibility in the context of automation technologies, including AWS, as well as on developing design and governance frameworks to help engineers, lawyers, and policymakers incorporate ethical analyses of automation technologies into their work.

Appendix II

***Targeting Trust* Agenda**

Targeting Trust: A Workshop on the Nature and Role of Trust in Autonomous Weapon Systems (AWS) Policy Development

October 5, 2017

Workshop Location

University of Ottawa, Alex Trebek Hall, 157 Séraphin-Marion Private, Ottawa, ON K1N

* Breakfast, lunch, coffee and snacks will be provided.

Workshop Agenda

Day at A Glance

1	08:30– 9:00	Breakfast
2	09:00 – 10:00	Workshop Introduction: Primer & Provocations
3	10:00 – 11:00	Lightning Round (Big Questions)
4	11:00 – 11:30	Morning Break – Coffee and Voting
5	11:30 – 12:30	Morning Breakout Session
6	12:30 – 13:30	Lunch
7	13:30 – 15:00	Afternoon Breakout Session
8	15:00 – 15:30	Afternoon Networking Break
9	15:30 – 16:30	Policy Implications
10	16:30 – 17:00	Wrap-Up

Details

1. Workshop Introduction: Primer & Provocations:

9:00 – 10:00: Welcoming remarks, followed by brief primers from the workshop organizers.

2. Lightning Round (Big Questions)

10:00 – 11:00: Each of our invited experts is invited to "pitch" two Big Questions related to the workshop themes. All participants will have a chance to vote on which questions will be the subject of discussion for the morning breakout session.

3. Morning Break – Coffee and Voting

11:00 – 11:10: Each participant has three green stickers to place beside up to three questions. Each sticker counts as a "vote" for that question to form the basis of further discussion during the Morning Breakout Session.

11:10 – 11:20: Covfefe.

11:20 – 11:30: Each participant must "sign up" for a morning breakout group.

4. Morning Breakout Session

11:30 – 12:30: Goals**:** Find your group. Refine the question. Start identifying knowledge gaps that would prevent you from answering the question. Identify potential information that moves the yardstick towards answering the question. Focus on trust, trustworthiness and MHC during all of this.

5. Lunch

12:30 – 13:30: Lunch details TBA.

6. Afternoon Breakout Session

13:30 –15:00: Workshop organizers will pose three new questions for consideration, followed by a guided investigation of those questions in breakout groups.

7. Afternoon Networking Break

15:00 – 15:30: Mingle.

8. Policy Implications

15:30 – 16:30: We will begin translating the day's discussions into clear policy implications. *All participants are asked to contribute to this **important** discussion.*

9. Wrap-up

16:30 – 17:00: Closing remarks.

Appendix III

Morning Presentation Slides

Trust and Meaningful Human Control in Autonomous Weapons Systems

Jason Millar

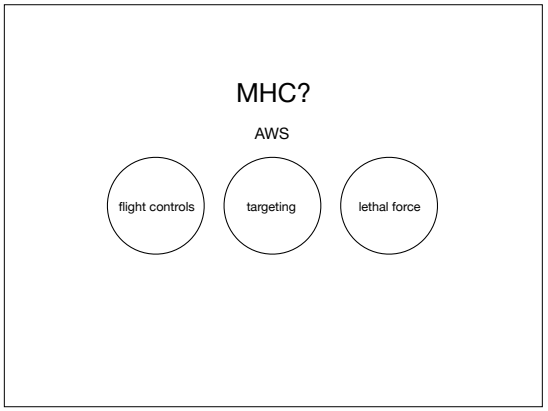
trust

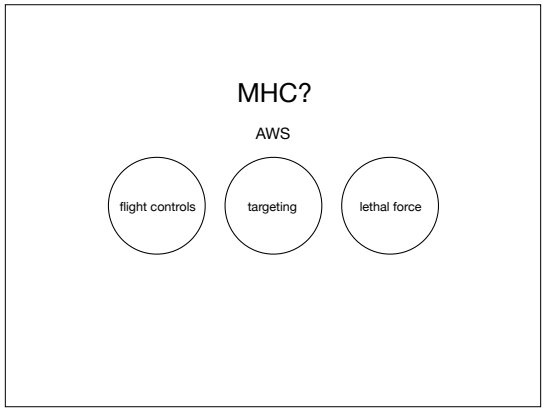
trustworthiness

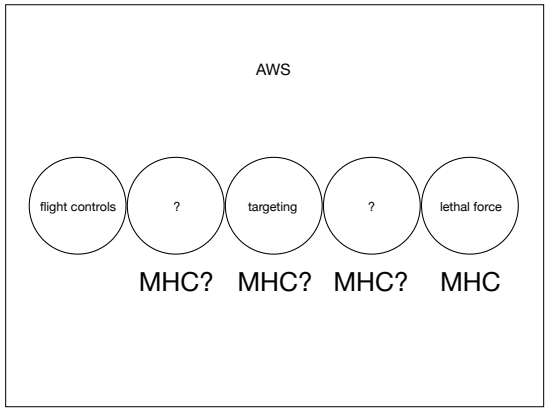
moral vs. conventional trust(worthiness)

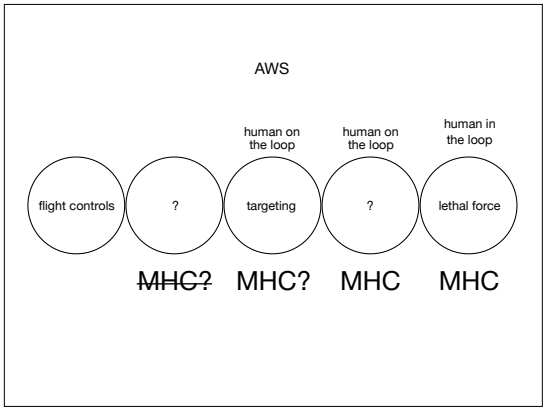
over/under trust

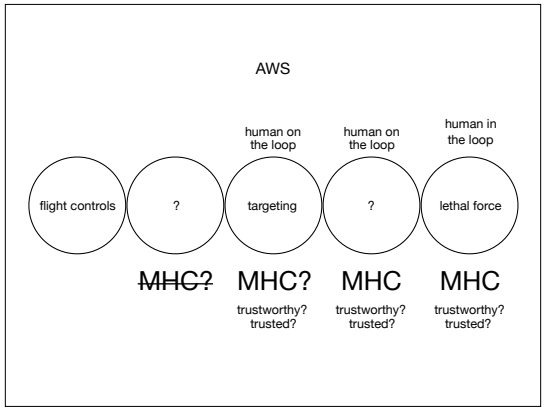
meaningful human control (MHC)

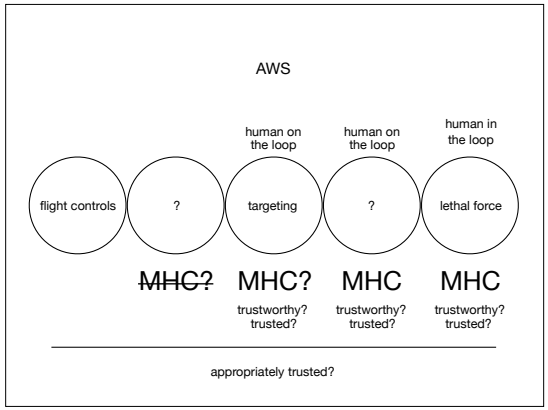

MHC?
AWS
flight controls
targeting
lethal force


MHC?
AWS
flight controls
targeting
lethal force


AWS
flight controls
?
targeting
?
lethal force
MHC?
MHC?
MHC?
MHC


AWS
human on the loop
human on the loop
human in the loop
flight controls
?
targeting
?
lethal force
~~MHC?~~
MHC?
MHC
MHC


AWS
human on the loop
human on the loop
human in the loop
flight controls
?
targeting
?
lethal force
~~MHC?~~
MHC?
MHC
MHC
trustworthy?
trusted?
trustworthy?
trusted?
trustworthy?
trusted?


AWS
human on the loop
human on the loop
human in the loop
flight controls
?
targeting
?
lethal force
~~MHC?~~
MHC?
MHC
MHC
trustworthy?
trusted?
trustworthy?
trusted?
trustworthy?
trusted?
appropriately trusted?

keep these concepts and relationships in mind while discussing the issues today

Trust in Machines {and, or, vs} Humans

Ian Kerr

X trusts Y to do Z

can we trust machines ?

building reliable machines

building reliable machines

^

intelligent

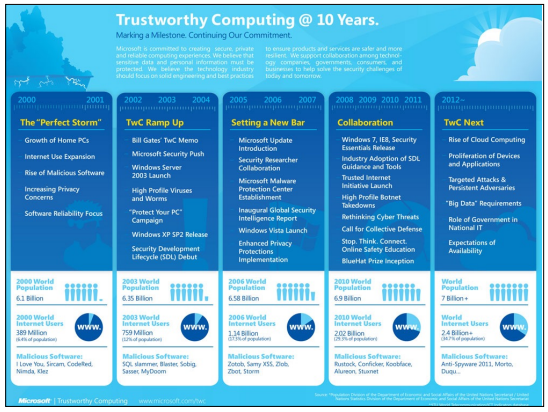
ML

building reliable humans

the problem of unreliable people

trustworthy computing
vs
trusted computing

trustworthy computing (microsoft 2002)



trusted systems (stefik, 1997)

TRUSTED SYSTEMS

Devices that enforce machine-readable rights to use the work of a musician or author may create secure ways to publish over the Internet

by Mark Stefik

Everyday experience with computers has led many people to believe that anything digital is ripe for copying—computer programs, digital books, newspapers, music and video. Some digital-age pundits have gone so far as to proclaim that the ease of duplicating data heralds an end to copyright: information "wants to be free," they assert. It is impossible to thwart the spread of information, so the argument goes. Anything that can be reduced to bits can be copied.

This provocative notion undermines the dream behind the creation of the Internet: the possibility of universal access in a digital age—where any author's work could be available to anyone, anywhere, anytime. The experience of most people, however, is not that the Net contains great works and crucial research information. Instead most of what is there is perceived to be of low value.



JIM BOZ

The root of the problem is that authors and publishers cannot make a living giving away their work. It now takes only a few keystrokes to copy a paragraph, an entire magazine, a book or even a life's work. Uncontrolled copying has shifted the balance in the social contract between creators and consumers of digital works to the extent that most publishers and authors do not release their best work in digital form.

Behind the scenes, however, technology is altering the balance again. Over the past few years, several companies, including Folio, IBM, Intertrust, NetRights, Xerox and Wave Systems, have developed software and hardware that enable a publisher to specify terms and conditions for digital works and to control how they can be used. Some legal scholars believe the change is so dramatic that publishers will be left with too much power, undercutting the rights and needs of consumers and librarians.

Yet consumers' needs can be served

Copyright 1997 Scientific American, Inc.

MHC

feature vs. bug

aws

Trust in Machines {and, or, vs} Humans

Ian Kerr

Trust and Technology

Jacquelyn Burkell

Trust

Reliance

Reliability

Trust → **Reliance**

Trust → **Reliance**

↖

Reliability

Dispositional Trust

Affective Trust
Cognitive Trust

Basis of Cognitive Trust

Situational, Specific, Dynamic

Trusting humans
Trusting Machines

Trusting appropriately

Trustworthiness

Trust → **Reliance**

↖

Reliability

Level of trust
Level of trustworthiness

DECISION to rely (or not)

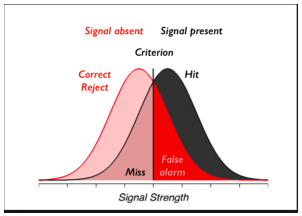
‘go’ or ‘no go’
reject or concur

signal detection problem

	‘Trust’	‘Don’t Trust’
System is accurate	✔	✗
System is inaccurate	✗	✔

	‘Trust’	‘Don’t Trust’
System is accurate	Hit	False Alarm
System is inaccurate	Miss	Correct Rejection

	‘Trust’	‘Don’t Trust’
System is accurate		Under-trust
System is inaccurate	Over-trust	



Catch-22

Trust and Technology

Lightning Round (Big Qs)

Experts 1, 2, 3 and 4

1. What role can trust play in establishing meaningful human control over weapons? How could it be observed, tested or verified?

(Expert 1)

2. How does the concept of trust in existing weapons systems differ between engineers, reviewers (acquisition/ Art36), commanders, & operators?

(Expert 1)

3. Canada typically fights in alliances/coalitions so how will required interoperability between military forces be affected by AI?

(Expert 2)

4. We can't even procure inflatable boats, how does the government think it can procure AI?

(Expert 2)

5. Meaningful human control. Appropriate human judgment. X human X. What're we really concerned about losing/preserving as AI systems multiply?

(Expert 3)

6. Sometimes humans trust AI too little, sometimes too much. When are these traits problematic? How are they likely to be abused by bad actors?

(Expert 3)

7. Why is “meaningful human control” rarely insisted for *cyberops*, even though it can be autonomous too? What does it mean for AWS policy?

(Expert 4)

8. Can trust exist if an autonomous system might be unpredictable? If not, how do we ever trust humans who are also unpredictable?

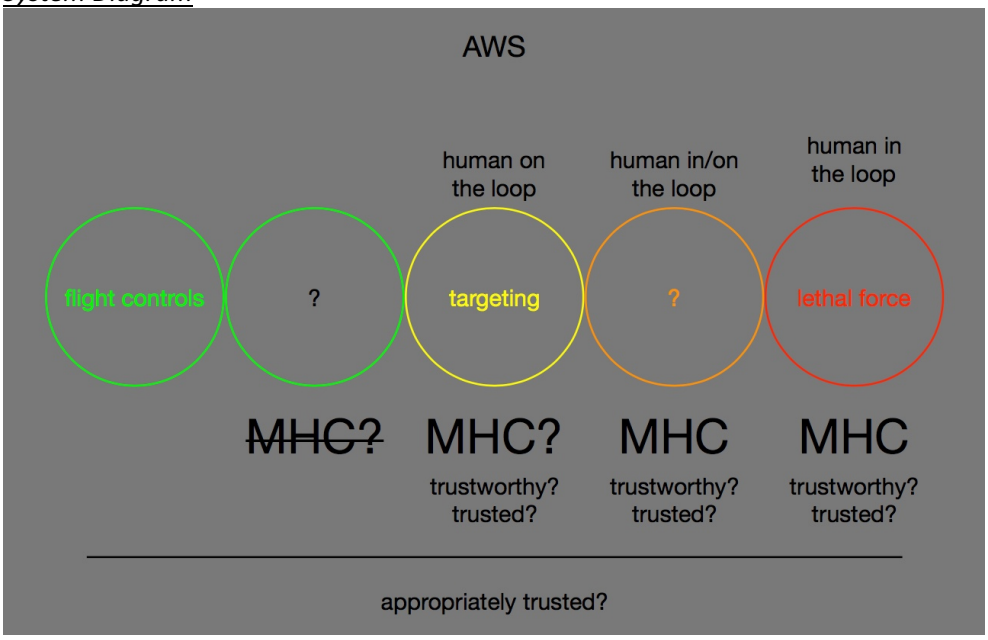
(Expert 4)

Appendix IV

Policy Implications Handout

Policy Implications

System Diagram



Guiding Questions

Based on the system diagram (above):

1. In your opinion, what sub-tasks (bubbles) have we missed, if any, that are worthy of discussion?
2. In your opinion, what considerations (red flags?) arise with respect to each of these sub-tasks when considering automation?
3. In order for you to be able to trust the automation of each sub-task, what would have to be in place? How would you address it?

Discussion Questions

In this session, we will discuss potential policy directions, such as:
Feel free to propose potential policy initiatives, such as:

1. What do you feel are some relevant policy issues related to all of this?
2. Would it be useful to adopt or avoid certain terminology for reasons expressed in today's workshop?
3. Do any key knowledge gaps jump out at you that might prevent the development of concrete policy?
4. Would it be useful to recommend the formation of specific working groups to further investigate particular issues?
5. Are there any areas for coordination either within the Government of Canada or internationally that seem useful, important or essential?

Remember: Be Creative!