Journal of Military Ethics

Routledge
Taylor & Francis Group

ISSN: 1502-7570 (Print) 1502-7589 (Online) Journal homepage: http://www.tandfonline.com/loi/smil20

“Trust but Verify”: The Difficulty of Trusting Autonomous Weapons Systems

Heather M. Roff & David Danks

To cite this article: Heather M. Roff & David Danks (2018) “Trust but Verify”: The Difficulty of Trusting Autonomous Weapons Systems, Journal of Military Ethics, 17:1, 2-20, DOI: 10.1080/15027570.2018.1481907

To link to this article: https://doi.org/10.1080/15027570.2018.1481907

Published online: 14 Jun 2018.

JOURNAL OF MILITARY ETHICS
2018, VOL. 17, NO. 1, 2–20
https://doi.org/10.1080/15027570.2018.1481907

"Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems

Heather M. Roff[a] and David Danks[b]

[a]Global Security Initiative, Arizona State University, Tempe, AZ, USA; [b]Departments of Philosophy & Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Autonomous weapons systems (AWSs) pose many challenges in complex battlefield environments. Previous discussions of them have largely focused on technological or policy issues. In contrast, we focus here on the challenge of trust in an AWS. One type of human trust depends only on judgments about the predictability or reliability of the trustee, and so are suitable for all manner of artifacts. However, AWSs that are worthy of the descriptor "autonomous" will not exhibit the required strong predictability in the complex, changing contexts of war. Instead, warfighters need to develop deeper, interpersonal trust that is grounded in understanding the values, beliefs, and dispositions of the AWS. Current acquisition, training, and deployment processes preclude the development of such trust, and so there are currently no routes for a warfighter to develop trust in an AWS. We thus survey three possible changes to current practices in order to facilitate the type of deep trust that is required for appropriate, ethical use of AWSs.

KEYWORDS

Autonomous weapons systems; trust; reliability; training and doctrine; autonomy

1. Introduction

The debate surrounding autonomous weapons systems (AWSs) is usually outward-looking. It addresses how such systems might have adverse impacts on non-combatants, adversaries, or broader social, political, legal, or military goals (Asaro 2012a, 2012b; Melzer 2013; Roff 2013, 2014; Sharkey 2008; Sparrow 2011). While such concerns are certainly important, they overlook inward-facing concerns, particularly how such systems, if developed and deployed, will affect those individuals they are supposed to help: the warfighters. Indeed, one primary rationale for autonomous systems is how they will enable militaries to fight better, both by facilitating maneuver in denied or contested environments as well as by enhancing force protection through the removal of the warfighter in some dangerous contexts. But this justification presupposes that AWSs will work as intended and, equally importantly, that the users and operators will *trust* these systems sufficiently to use them in combat.

Trust is critical for a warfighter in a hostile environment. She must trust that her comrades will provide covering fire, her commander's orders are appropriate given broader

CONTACT Heather M. Roff hroff@asu.edu; heather.roff@colorado.edu Global Security Initiative, Arizona State University, 4th Floor, 781 Terrace Road, Tempe, Arizona, 85287, USA.

strategic and tactical goals, a weapons manufacturer did not cut corners when making her gun, the others standing watch will not fall asleep, and so on. In all of these cases, we use the word "trust" as though it is a simple, binary notion: either one trusts or one doesn't. However, trust is a far more complex and multifaceted concept.

An individual may trust in another person, animal, or artifact; how they trust can vary in a fine-grained way along multiple and interacting dimensions. One dimension depends on reliability and/or predictability, usually in relation to the expectation of some behavior or action by an artifact or designed object. For example, the "trust" that one's car will start in the morning carries little more than an expectation that the car will work; it is, at its core, a morally neutral concept. A second dimension of trust depends on one's understanding of other people's values, preferences, expectations, constraints, and beliefs, where that understanding is associated with predictability but is importantly different from it. This dimension contributes to the more complex types of trust that one finds in rich, moralized interactions between humans. Both of these dimensions are continuums; we can trust in either way to varying degrees. Thus, even when we speak about trust in binary terms (present or absent) as a shorthand for more multifaceted statements about trust, matters can still be quite complicated because the threshold(s) required for someone to "trust" can vary in highly context-, goal-, and task-dependent ways.

The shortcomings of a simplistic, binary framework of "trust" are particularly apparent when thinking about the trust that warfighters can or should have in AWSs. For instance, the binary approach requires that warfighters either trust or not – full stop. Yet such a decision assumes that one either views AWSs as "mere tools," where reliability and predictability of behavior is sufficient to "trust" the system, or that an AWS is more akin to a moral agent with values and preferences, in which case the threshold for "trust" would be significantly higher. The reality, however, is that current-day and near-future AWSs do not fit neatly into either of these categories (mere tool or moral agent), and so the binary approach cannot provide action-guiding principles or practical advice about how to trust AWSs.

Without careful attention to the complexities and limits of trust in AWSs, we argue that states are likely to fail to calculate appropriate strategies, overestimate their operational capabilities, and create enduring vulnerabilities beyond lethal AWSs themselves. That is to say, a failure to understand how humans can or cannot trust AWSs has direct consequences for military strategy, the conduct of hostilities, and even long term economic impacts from investment and procurement choices. In this paper, we focus on two core questions: what kinds of trust are required for a warfighter to use AWSs in high-risk environments, and when (if ever) can those levels of trust be reached? We contend that quite high levels of trust – not mere predictability or reliability – are required in these environments (Section 2), but that there are significant barriers to achieving that trust with current practices, as well as with present and near-term AWS technology (Section 3). We then argue that technology alone cannot solve this problem, since AWSs become less capable of being trusted as they become more capable of undertaking elaborate or complex tasks (Section 4). That is, we have a paradox: improving the system's capabilities makes it less suitable for use in human–robot interactions or teams. Of course, our arguments depend on empirical details of human psychology and technological reality, and so they are open to revision as technology, policies, and practices change. In fact, we suggest several possible routes to establish the requisite levels of trust (Section 5), though we have serious concerns about each of them. But absent these or similar changes, AWSs ought not, and actually will not, be deployed in battlefield environments.

2. Varieties of autonomy

A key challenge here is the varieties of "autonomy" at play in AWSs. Many authors understand AWSs simply as those systems that can select, engage, and fire (or "select and attack") without the intervention of a human operator (e.g. the definitions provided in ICRC 2016 or USDoD 2012b, though we recognize that these are not the only approaches). While this is certainly useful for a general definition, it misses key features of what it means to "select" or "engage" as well as "attack" (Roff 2015, Forthcoming). For instance, if "select" means roughly "sense" or "detect," then large swaths of present-day weapons systems fall into these categories, ranging from land mines to advanced anti-ship missiles or ballistic missile defense systems. Indeed, this understanding collapses the distinction between "automatic" and "autonomous." This distinction is, however, an important one, and so we should reject arguments that assert that "autonomy can be considered as well-designed and highly capable automation" (USAFRL 2015).

Instead, we understand the "autonomy" of an AWS as residing on a multidimensional continuum characterized by the system performance on different tasks. That is, "autonomy" is about a system's *ability* to carry out a particular task assigned to it without the intervention of the human operator, not just the highest tier of "levels of automation" (Sheridan and Verplank 1978; see also Moray, Ingagaki, and Itoh 2000 for an overview of alternative taxonomies).[1] Moreover, because this is about ability, different capacities for autonomy can be interrelated and interdependent. For example, one might require a system to move from point A to point B by itself (i.e. autonomously), but that would require locomotion and, depending upon task, presumably planning and navigation as well. Its performance cannot be assessed simply on a one-shot basis, but instead on *how* it succeeds or fails in relation to its task. A system that performs very well through the use of navigation, homing, image recognition, guidance and positioning, etc., is, on par, more autonomous than one that performs very poorly, say by happenstance or only after getting lost a few times. Autonomy arises, we suggest, from *collections* of capacities and capabilities; it is not some single, standalone capability. In this paper, we focus on two related, but distinct, sets of capacities that an AWS could possess.

First, *planning-autonomy* is the ability to independently construct plans to realize a user's orders. Examples may be as simple as planning a navigation route with various way-points, or it may be more complex where a system needs to generate and identify a set of sub-goals to achieve an overall goal (determined by the user). This dimension of autonomy is clearly not a binary property, as planning-autonomous weapons systems (AWS$_P$) can differ radically in terms of the distance allowed between the order, the time it has to execute that plan, and the requirement to update plans as a function of time and distance. Complex goals related to targeting decisions to realize strategic or tactical goals would require much more capable autonomous planning than simple waypoint selection. Take, for example, a cruise missile that constructs a plan for the order "Remain 100 meters above the ground throughout flight." This requires little in the way of robust autonomy;[2] in contrast, a system given the order to "Clear this room of enemy soldiers" needs far more information and cognitive ability to plan a course of action, identify "enemy," "friendly," "neutral," and "protected" to accomplish this order. Such information could include a route to the room, as well as constraints regarding tactics, rules for engagement, and the laws of armed conflict.

Second, *learning-autonomy* is the ability to *adapt* to novel environments; learning-autonomous weapons systems (AWS_L) are those that exhibit the capability to learn underlying structures and relationships across multiple contexts, generalize from past experiences in various subtle ways, and adapt to rapidly changing, dynamic environments. As with AWS_P, AWS_L can differ widely in its capabilities. A simple sensor system might learn whether it is currently night or day, and thereby adjust its threshold for reporting anomalous sounds; a more sophisticated system might be capable of learning from prior experience that enemy soldiers have started wearing different types of clothing, or their patterns of life, and thereby update its target identification library.[3] As this last example suggests, there are clearly connections between planning- and learning-autonomy: systems with substantial degrees of one type of autonomy will often have significant degrees of the other. But they are importantly not identical, as we can have learning systems that cannot construct their own plans, and planning systems that cannot learn. Moreover, planning systems are now widely accepted and used – for example, the overall development of various precision-guided weapons systems, where these systems must be able to plan (and find) their way to various locations at great distances – but there is much more unease about learning systems (USAFRL 2015; Kester 2016). Our concerns about trust and AWSs are, however, largely independent of the planning vs. learning distinction in many places, and so we will use "AWS" to refer to either type of system, and specify the type only when it makes a difference.

For this paper, we focus on near-future systems that have significant degrees of planning- or learning-autonomy, but not full agency or human-like capabilities along either dimension. These near-future systems do not simply "react" in predefined ways to their environmental contexts, but also are not unbounded or unconstrained in their planning and learning. For example, they are able to construct a limited plan, but are unable to understand or invoke broader strategic or tactical factors that might render that same plan irrelevant, harmful, or unethical (Roff 2014). These AWSs are not operationally fielded yet, but the algorithms and capacities required for these systems are widely known and available.

As many commentators note, operators (and militaries) can only benefit from advances in autonomy if they are able to trust the systems they deploy (Hancock et al. 2013; USAFRL 2015; USDoD 2012a). Warfighters must be able to trust that the task they delegate to an AWS will not only be carried out, but that it will be carried out in the manner *intended* by the user. Without such "trust," commanders will (a) not field or deploy the AWS (nonuse); (b) field or deploy them in operationally inappropriate situations (abuse); or, most likely, (c) field them with other systems or human personnel in ways that incur significant costs in time, lives, and money (misuse). The questions then are what trust requires in relation to AWSs, and whether humans can possibly have such trust.

3. Trust and its importance with AWSs

Common usage of "trust" constructs it as though it were a single, binary notion, but it is significantly more complex and multidimensional. As we noted above, notions of predictability or reliability, such as through a performance standard, are commonly associated with "trusting" objects. Indeed, most discussions concerning human–robot-interaction or supervisory control, especially in military applications, focus on these dimensions (e.g. Beck, Dizndolet, and Pierce 2007; Chen and Joyner 2009; Cummings 2004; Cummings and Guerlain 2007; Lee and Moray 1992; Lee and See 2004; Parasuraman, Barnes, and Cosenzo 2007). To be sure, these studies take predictability and reliability as the sole bases of trust: users are claimed to experience a "decrease in trust" when they deem a system "unreliable" or when an experimenter manipulates the "predictability" of an outcome. To the extent that an AWS is unreliable, its use is *presumed* to be unethical (Klincewicz 2015).

Reliability and predictability are important dimensions of trust, not just for artifacts, but also for the expected behavior of other people. In particular, we often form this type of trust with strangers with whom we have little knowledge, evidence, or history. For instance, I "trust" that a random stranger on the street will not pinch me. Given general social practices, the desirability of cooperation, as well as the fact that I have never experienced a random nip from a stranger on the street, I generally trust that those I walk by will forgo from a friendly squeeze. Indeed, this minimum behavioral trust is well documented using game-theoretic contexts. Consider a standard Prisoners' Dilemma game: player A is commonly said to "trust" player B when A chooses to cooperate; contrarily, A "distrusts" B when she defects. This type of trust is focused solely on what the other (whether artifact or human) will do, all things considered and without deep knowledge of the other party.

Thus, trust based on reliability and/or predictability is often characterized by a pattern of behavior observed by the trustor: no particular psychological state or affect need be attributed to A, just as I can "trust" that my car will start or that my computer keys will work, without having any particular (occurrent) psychological state or affective response. Rather, *I need only have the (implicit) belief* that the recipient of my trust will behave or respond in a particular way (Chen and Barnes 2012). This type of trust arises and persists without having any detailed knowledge or understanding of the "inner workings" of the trustee, such as whether the trustee's intentions, desires, or beliefs, affect or theory of mind, is the same as mine, or in the case of an artifact, its design, structure, or past history. In a game theoretic context, A does not need to know *why* B will cooperate; she merely needs to be able to (accurately) *predict* that B will do so. Of course, prediction can be greatly aided by knowledge of the inner workings or causal structure of an artifact or human, especially when one is trying to "translate" or extrapolate from one context to another (Danks and London 2017; Kimmelman and London 2015). But such knowledge is rarely required for successful prediction. For example, many people find a plane's autopilot to be highly reliable,[4] even if they have no clue how it actually works, despite much work on ensuring an accurate interface for human and machine.

The second, and deeper or more moralized, notion of trust is found in interpersonal relationships and dependencies, and this usually depends on the trustor willingly accepting a measure of vulnerability because she has beliefs and expectations about the trustee that extend beyond mere predictions about what the trustee will (probably) do. (Deutsch 1958, 1973 are seminal works; Lewicki, Tomlinson, and Gillespie 2006 provide a recent review.) In these relationships, trust-relevant beliefs involve understanding, and perhaps even internalizing some of the values, preferences, and beliefs of the trustee. More colloquially, both the trustor and the trustee "share a mental model" of the world; in essence, the trustor knows roughly *what* the trustee will do, and also *why* she pursues that course of action. The "why" here is not strongly mechanistic: it does not require detailed knowledge of psychological processes and representation. Instead, it refers to knowing the values, principles, beliefs, and motives that guide the trustee's choices and actions. These mental models are exactly what we use in everyday life to direct and support our expectations, predictions, and explanations of those close to us. More critically, this type of trust *enables* the trustor to predict and respond to the trustee's choices and actions in novel or unforeseen situations that do not match ones they have previously (jointly) experienced. In other words, this deeper type of trust facilitates coordinated actions that are robust across a range of environments, problems, and challenges, including those never before encountered (Danks 2016).

For human warfighters participating in hostilities, both dimensions of trust are clearly at work (and required). The practice of modern warfare, including the use of objects and people, as well as the dependence upon processes, organizations, doctrine, and structure, mesh both kinds of trust together. On the predictability/reliability side, command structures provide a hierarchy whereby the status of rank denotes a particular capacity for leadership and responsibility, *eo ipso* and confers *trust*. A particular decision might not be predictable, but reliability of the commander's competence and the chain of command is critical for the success of a contemporary warfare and the command and control structure.[5] Moving from structures to technologies or artifacts, there are also explicit mechanisms and processes in place to ensure predictability and reliability of military objects. For instance, international law requires states to perform legal reviews of all new means and methods of war to ensure that they uphold the laws of war (ICRC 1977). These "Article 36" reviews look to whether a weapon's design or intended use may make it prohibited. Weapons that are indiscriminate, violate prohibitions on unnecessary suffering, or damage the environment, are *mala in se* and, as such, are not only illegitimate but not trustworthy.[6] Often weapons that cannot be accurately predicted – whether due to misfiring or targeting the wrong objects or persons – will likewise be deemed untrustworthy and most likely unlawful. Thus, reviews require that the weapons be tested, experimented, verified, and validated before one can be said to "trust" that they comply with the laws of war. In short, we must "trust" but "verify."

At the same time, interpersonal trust plays a key function for warfighters in hostile environments. These individuals rely not merely on their training, but also on their compatriots' attitudes, mental states, ideas, and beliefs, to achieve mission objectives and return home safely. Conflict is nothing if not a continual state of vulnerability, and so successful coordinated action requires interpersonal trust; that is, it requires "'positive' or 'confident' expectations about another party and a 'willingness to accept vulnerability' in the relationship, under conditions of interdependence and risk" (Lewicki, Tomlinson, and Gillespie 2006, 1014). Mere reliability and predictability are insufficient to enable appropriate trust and support between warfighters in these dynamic, frequently chaotic, high-risk, and hostile environments. Novel situations continually arise, and so group success will be achieved only when each individual, in her own vulnerable and uncertain state, accepts and understands *why* the other acts as she does. This kind of trust is particularly crucial for mission effectiveness and unit cohesion.

4. Challenges to developing trust in an AWS

Although there has been limited (if any) deployment of sophisticated AWSs in kinetic battlespaces, it is clear that they are on the very near-future horizon. "Machine assisted operations" will undoubtedly be utilized alongside human operations and through teaming. Speaking candidly at an awards ceremony in August of 2015, US Undersecretary of Defense Bob Work stated that "10 years from now if the first person through a breach isn't a friggin' robot, shame on us" (Pellerin 2015). The US's current "Third Offset Strategy" emphasizes human–machine teaming – Work calls this "Centaur Warfighting" as the machine and the human are "joined at the hip" (Freedberg 2015) – as a cornerstone of future combat scenarios. However, there are different kinds of loops at work. For instance, the US Deputy Chief of Naval Operations for Warfare Systems, Rear Admiral Michael Manazir, recently proposed that the "OODA Loop"[7] be divided up so that the machine Observes and Orients, and the human Decides and Acts (Kreisher 2016). In other contexts, The Netherlands has suggested that we view human control in the "wider loop" of various operations to allow machines to fulfill various tasks under existing command and control structures (The Netherlands 2016). This type of clean differentiation of tasks might be possible for information gathering and data processing operations, but is unlikely to be feasible when a system operates in a restricted or communications-denied environment. In these latter environments, the human "in the loop" will be a component in a larger *system of systems*, likely a heterogeneous mix of humans and AWSs.

As an aside, we have little reason to think that human warfighters will be removed entirely at any time in the near (if ever) future. Even quite robust AWSs are not *fully* autonomous agents in the moral or cognitive senses of the word; rather, they assume tasks for which they are currently superior (Chen, Barnes, and Harper-Sciarini 2013). Human warfighters remain necessary to play *at least* a "guidance" role to set mission goals and create larger strategic plans, despite the likelihood that even tactical decisions have strategic effects (Roff 2014). But this guidance role, including the amount of oversight, can only be fulfilled if the warfighters and commanders trust that the AWS will complete the tasks delegated to it.[8] More generally, for this human–machine collaboration to succeed, and a mission to be completed, the warfighters must trust the AWS as a team member.

Yet because a near-term AWS is unlike a mere tool and a well-understood friend or compatriot, humans will necessarily have uncertainty about how it will behave – along both dimensions. For predictability and reliability, there are enormous technological challenges in testing, validating, and verifying these systems (Danks and London 2017; USAFRL 2015). This, in turn, has wide ranging impacts on both trust development and risk analysis and allocation of responsibility and liability (KRHRW 2015; Roff 2013; Scharre 2016; Sparrow 2007). To be sure, systems with planning and learning capacities are not *necessarily* unpredictable; rather, the extent of their reliability and predictability depends on the details of the system, the environments in which it is deployed, the kind of orders it implements, the simulated and real-world environments it encounters, and whether the AWS is distributed across multiple machines (i.e. a swarm).[9]

AWSs are clearly much *less* predictable than typical non-autonomous munitions and weapons systems. Unlike warheads with a determinable blast radius, or planes that can be flight-tested, autonomous systems that can plan and/or learn in complex and novel environments will exhibit much greater variety of decisions. These are not minor or inconsequential decisions either, but rather, due to the nature of the system, often involve the use of lethal force.

What is more, we suggest that, to the extent that an AWS *is* highly predictable, people will not regard it as having significant autonomy. Rather, they will likely view such systems as "merely automatic." More importantly, highly predictable systems will not exhibit the types of self-direction that are thought to be characteristic of autonomy. Autonomy is valuable precisely when we do not know (and so cannot predict) the appropriate action in advance (Danks 2016). And even if we could somehow develop a system with "predictable autonomy" in the lab or a simulation, there is little reason to think that it will be sufficiently predictable in the field. Combat inevitably occurs in chaotic, rapidly changing environments, and so small over- or under-estimations of the system's abilities can produce numerous, significant prediction errors. Even if the system appears predictable to developers or testers, we have little reason to believe that *front-line* warfighters will have the experiences required to develop sufficient levels of trust on the reliability and predictability dimension.

One might object here that the warfighter is a capable human with much greater adaptive and cognitive abilities, and so she would be able to understand and thus trust the system's limits. While it is certainly true that humans are exceptionally adaptive, the empirical research that we have on human–robot-teaming suggests that humans often over- or under-estimate a machine's capabilities. Humans can suffer from "automation bias" where they accept a machine's decisions though it is incorrect (thus over-trusting), or they continually check-in or take-over various machine tasks (under-trusting) (Bartlett and Cooke 2015; Gorman, Cooke, and Winner 2006; Narayanan et al. 2015). Further, poor communication between adequately functioning robots and human team members also limits the ability of humans to trust in a machine's actions. Given that the types of situations where AWSs will be utilized are in poor or jammed communication environments, this does not bode well for developing even low levels of trust in reliability or predictability.

The problem is exacerbated when we turn to the dimension of interpersonal trust (i.e. trust based on understanding "why"). This type of trust requires the trustor to understand the values, beliefs, preferences, and motivations of the trustee. Obviously, one immediate concern is whether it is even coherent to talk about AWSs having these quasi-mental states or attributes, or whether those are the province of more sophisticated systems (or perhaps only animals or humans). For the sake of argument, let us suppose that there is a sense in which AWSs can be thought to have something like values, though presumably programed[10] by the system's developers. Even under this supposition, there is essentially no reason to think – given current development, testing, and acquisition practices in essentially all modern militaries – that any front-line warfighter will have any serious understanding of the values or conceptualizations of an AWS. While an AWS may be able to adequately "model" the human it is fighting alongside, the human is unlikely to have an adequate "mental model" of the machine.[11] Shared, joint mental models can be developed, but they require extensive training, even for fully human teams.[12]

Moreover, learned collaboration may have unintended negative side effects. Some proponents of AWSs claim that such systems will be "more moral" than human soldiers because they will not suffer from emotion or irrational beliefs (Anderson and Waxman

2012; Arkin 2009, 2014), and so will better uphold the letter of the law (Schmitt 2013). However, if a learning AWS comes to share a human team-member's mental model or emulate their behavior, then the AWS may in fact not behave any more morally. That is, the necessity of learning to collaborate may eliminate the claimed benefit of AWSs being more rational, and thereby more moral.

4.1 No simple solutions

Trust will not naturally emerge between human warfighters and near-future AWSs, and there are no simple solutions to this challenge. One might hope that we could solve each dimension of the trust problem separately, particular given the essentially linear development and acquisition process for non-autonomous weapons: they are designed, tested, verified, validated, and only then fielded with warfighters who have been trained. Essentially, the end-users receive a complete unit or system and training on how to use it. This linear process suggests the developers, testers, and validators should first focus on predictability and reliability; interpersonal trust can be developed second. However, we must solve for both dimensions of trust (to the extent possible) *simultaneously* for two distinct reasons. First, systems that can learn, evolve, and change are not predictable, at least in the supervisory control sense, when they are deployed to novel environments. Supervisory control (and accompanying notions of reliability, predictability, and testing) is built upon a manufacturing assumption, where the rote and predictable nature of automated robotics in simple uncomplicated environments is the standard. Warfighting, hostilities, and combat are the antithesis of this.[13] Second, the ability of human warfighters to develop deep, quasi-interpersonal trust depends on them understanding why the AWS responds as it does. The development of the AWS must thus have "end-user intelligibility" as a goal from the outset; it is not sufficient for the development and acquisitions processes to focus only on predictability (even that were even a clearly attainable goal).

More generally, one might hope that we could find a simple solution that resides in the AWS itself: perhaps there is some technological change or innovation that would enable trust in these systems. If so, then the solution is simply a matter of funding the appropriate research to discover this technology. However, this technology-centric approach will almost certainly fall short. The development of interpersonal trust requires that a trustor come to exhibit some degree of identification with, or internalization of, the preferences of the trustee (Lewicki, Tomlinson, and Gillespie 2006), and so there typically must be repeated interactions between the trustor and the trustee.[14] These repeated interactions provide indications about what the other will do and, more importantly, information about the other's values, preferences, or beliefs. These experiences allow us (humans) to perform our own risk calculations about what to internalize about the other agent's mindset and how vulnerable to make ourselves. Moreover, the social-cognitive aspects elicited by these interactions are precisely why we can move beyond mere predictability to the level of interpersonal trust required in novel, high-stakes situations with high degrees of uncertainty.

In fact, militaries go to enormous lengths to ensure that warfighters – certainly, members of the same unit – have exactly these shared (and known-to-be-shared) values and preferences. One major point of basic training is to achieve the kind of group identification that leads to shared values, beliefs, and expectations. Once a warfighter is beyond initial training, commanders are expected to provide a shared image (values, beliefs, goals) to their subordinates. Because each individual warfighter must identify with the group and share its values (to some extent), he or she can understand and trust the larger institutional and structural goals of the organization as well. For example, when a commander formulates a goal (sometimes referred to as an "operational concept" or even "commander's intent"), individuals are required to act in accordance with this goal and to have understanding of the values, beliefs, and reasons behind it so that they can achieve it in conditions of uncertainty, or without (necessarily) direct oversight. Use of AWSs, either in teams or in non-collaborative tasks, requires the same type of understanding and internalization by the human warfighter.

Now consider whether a purely technological shift can improve "identification" (in appropriate ways) with an AWS. If we have a learning AWS, then a system's "preferences" can change over time as it learns about its environment. A human interacting with a learning AWS must thus know more than just a list of "things that the AWS was programmed to value," given that the list may change as it learns. Moreover, those changes can be very un-human-like or, at very least, quite unexpected for a human. In this way, humans may have difficulties understanding and identifying the system's values as they may be non-human values (in certain ways), or human values that have changed through a non-human learning process. The battlespace is a dangerous place to be figuring out the preferences and values of a dynamically adapting weapon, so it is unsurprising that trust will be difficult to establish. In fact, in this type of human/robot weapon teaming, we face a dilemma: the extent to which the learning AWS actually *learns* and adapts to its environment in operationally effective ways will be *inversely* proportional to the extent to which the human team members can identify and internalize its values and preferences (and hence appropriately trust it). That is, AWS trust and AWS learning are mutually incompatible, or at least in significant tension.

Similarly, strong interpersonal-level trust in a planning AWS will depend on understanding or internalizing the system's methods of planning, its information acquisition mechanisms (since environmental information can change or constrain plans), and the values or constraints that are embedded in the planning algorithms. In present-day planning systems, this content – information, sensors, constraints – can be exceptionally dense and complex, and so correspondingly difficult for the human to process and understand. If we make an AWS_P even more sophisticated, then it will presumably find even more complex, surprising, or context-sensitive ways to complete its orders, and so become even harder for an individual to trust. Note that these concerns arise even if the system has no strong learning capabilities; it only needs to be responsive to its immediate environment, which will describe any worthwhile, autonomous planning system. A sufficiently complex planning capability can block the development of trust, even if that capability does not change over time. And as with AWS_L, improving the sophistication of the planning AWS_P, or otherwise hoping for some technological advance, gives no way to improve warfighters' trust in the system.

Whether autonomous in learning or planning, a "better" battlefield AWS will almost inevitably be *less* trustworthy from a human cognitive perspective precisely because the more advanced a system's capabilities are, the less human team members understand about the system's preferences. The rub, of course, is that we develop AWS technologies precisely so they can be used. As advocates point towards their many positive aspects, those benefits can only be (potentially) realized if they are actually deployed. Yet no commander will want to use a system that she cannot trust. Hence, a purely technological focus actually works *against* the adoption of the very technology being developed. We must instead look elsewhere for possible routes to developing strong bonds of trust between warfighters and AWSs.

5. Routes towards trust?

Rather than a purely technological solution, we should instead consider various changes in unit structure, training, and development pipelines. One option we do *not* consider is to significantly constrain the rules of use for AWSs, even though this strategy is currently employed with, for example, novel cyberweapons. Significant restrictions could certainly eliminate the trust challenge, by either ensuring that AWSs are used in only low-stakes situations (so trust is irrelevant) or removing substantial planning or learning capacities (so mere predictability is all the trust one could have). The first route, however, eliminates one of the putative advantages of AWSs, as these sophisticated systems are intended precisely for *high*-stakes situations where militaries require maneuverability in denied and contested environments. Moreover, given the large financial investments in AWSs, it is implausible to think that their use would be restricted to low-stakes situations. The second option clearly defeats the whole point of having an AWS: they are needed just when a non-adaptive or "dumb" system is insufficient. Given that constraints on rules of use are not a viable option, we instead consider three different, not mutually exclusive, routes to potentially build appropriate degrees of trust between soldiers and an AWS. We discuss these routes in roughly increasing order of cost, time, and difficulty, as well as likelihood of success.

First, we could attempt to leverage the phenomenon of "transitive trust." For example, if Alice trusts Bob, and Bob trusts Claire, then Alice will exhibit a degree of trust in Claire (assuming these bonds of trust are about similar domains, and everyone knows about them). Of course, Alice's trust in Claire will be attenuated in various ways. Nonetheless, some degree of Alice–Claire trust will arise by virtue of Bob's dual role as both trustee (of Alice) and trustor (of Claire). Organizational structures often depend on significant levels of transitive trust amongst their members, where one or both of the bonds of trust are grounded in the organizational role of one or more individuals. For example, suppose Sergeant Alice trusts Corporal Bob to carry out her orders, and Corporal Bob passes along part of those orders to Private Claire. Alice trusts Claire (again, in an attenuated sense) to carry out the relevant parts of the order, exactly because of the hierarchical roles that they each occupy. They understand the values of the organization and, as each are members of the greater structure, they presumably reason that each will act in accordance with them.

In the AWS case, we can potentially take advantage of transitive trust by focusing the trust-building efforts on a single warfighter who is designated as an "AWS liaison." Since the *other* soldiers presumably trust the AWS liaison, they should thereby have (attenuated) trust of the AWS, at least to the extent that the liaison trusts the AWS. Many warfighters in small units have specialized roles requiring additional training – medic, radio operator, and so forth – and the strategy would be to add "AWS liaison" to that list. That individual would require additional training focused on understanding the values, learning systems, and control mechanisms of the AWS, but this type of dedicated instruction is more feasible for an individual than for all individuals who may or may not come into operational contact with the system.

Essentially, the AWS liaison would need to combine a developer's understanding of the AWS with a warfighter's understanding of the complexity and dynamics of a high-stakes battlespace. This role would be analogous to warfighters who team with non-human animals. The United States Navy, for example, utilizes marine mammals for mine location and diver detection. These animals operate in open water, away from their trainers (unlike guard dogs and their handlers), and so present similar teaming challenges as might arise with humans and AWSs. Like marine mammal handlers, the AWS liaison would have to invest time and training (perhaps substantial amounts) to gain the knowledge and experience for full trust in the AWS.

This strategy has the advantage of being relatively straightforward to implement, as it leverages existing role-based infrastructure. Moreover, the analogy is better for humans. In particular, because we know that humans often over- and under-trust various types of autonomous systems and robotics, thinking of AWSs as more akin to animals, and not like humans or being human-like, will mitigate some of the human cognitive biases.

The disadvantages are, however, substantial. For each unit, there would be a single "point of failure," as the unit's (transitive) trust of the AWS depends entirely on a single individual having an appropriate degree of (non-transitive) trust of it. Anything that impairs the liaison's functioning would negatively affect the unit's ability to trust or to deploy the system. Additionally, the role of AWS liaison would likely not be simple or easy to train: relevant aspects of the technology will likely be highly classified, there will be difficulty in finding the right person to fill such a role, and one would presumably want this person to remain in the role for longer periods of time than the usual service commitment contracts.

A second strategy would be to provide whole units with multiple experiences with an AWS so that they can gradually build trust in it. For example, one could incorporate AWSs into basic training, so that warfighters can develop trust similarly to how they develop trust in one another. Alternately, one could embed an AWS with a unit in relatively low-stakes missions, perhaps with limited weapons functionality until the warfighters gain trust in it. In this latter example, one might not weaponize the AWS at first, but instead emplace targeting lasers so operators can gain some understanding of how the AWS evaluates threats, constructs plans, makes decisions, and so forth, without the risk of the AWS directly doing or risking harm. Indeed, presently the United States Marine Corps are training with non-armed robotic pack-mules, and this may be a good data point for further study (Seck 2016). Subsequent simulations and wargaming could also provide further experiences and interactions to help the warfighters come to have appropriate understandings of the AWS's values, preferences, and responses. Over the course of these experiences, the members of the unit would presumably develop trust in the AWS to some extent. Limited training and instruction could significantly increase the speed and quality of this trust construction by helping the warfighter to understand why the AWS is behaving in particular ways. In general, the success of this strategy would depend on the quality and generalizability of the shared AWS-unit experiences, and may vary depending upon unit.

This strategy has the advantage of preserving the current, relatively linear acquisition process. The strategy's costs are borne only when the systems are handed over to the operators for further "trust building" exercises. To the extent that this strategy succeeds, the resulting trust should be stronger than that developed via the first strategy because all members of the unit will have trust-building experiences. However, this trust-construction would require significantly more time, energy, and attention, as well as the ability and willingness to remove (or forgo deployment of) AWSs from a battlespace. This strategy is substantially more expensive in time and resources. Unit turnover also presents a challenge, as it might not be possible for all warfighters to have these shared (with the AWS) experiences. Most importantly, however, there are significant concerns about exactly what would be learned by the soldiers (and the systems) during these experiences. If an AWS is embedded in basic training, then its learning must be reduced or eliminated, else the AWS will learn or plan for the boot camp environment, rather than the (intended) battlefield environment. If the AWS is instead "tuned" to the battlespace, then the warfighters will have limited opportunities to see how it learns, plans, and adapts. If we place the AWS in low-stakes environments, then we must determine when it is reasonably safe to do so. We also must establish the relative priorities for the AWS – balancing its "normal" goals versus the "build trust with the unit soldiers" goal – as well as how those goals and constraints are satisfied. If the former is preferred, then trust-building will be negatively impacted; if the latter is preferred, then the AWS may act in inappropriate ways.

The third and final strategy would be to "close the loop" between the development lab and the battlespace by shifting radically away from the linear acquisitions process, so that *everyone* involved with an AWS – scientists, technicians, soldiers, commanders – has a shared understanding from the outset. Of course, each person would require a different degree of knowledge about different elements: the operators may not need to know what programing languages are used in the AWS, and the scientists may not need to know the exact situations in which systems will be deployed. But the strategy is to find mechanisms to establish sufficient shared understanding, so that warfighters would be able to understand how and why the AWS will behave in particular ways in novel situations, and thereby truly trust the AWS in a richer way. This type of understanding requires deeper knowledge than is codified in the standard rules of use for novel military technologies. This strategy would presumably involve such steps as bringing soldiers and commanders into the lab and scientists and technicians out into the field, so that each group can come to fully understand the values, learning, and control that are necessary for successful trust of the AWS.

This strategy would require substantial and qualitative shifts in development and acquisition processes, changes that go far beyond current attempts to engage in collaborative design between contractors and militaries, or provide limited fielding of a developmental system for testing purposes to obtain better feedback. We envision a radically different approach in which there is interaction between groups throughout every aspect of research, development, validation, and deployment. This potential solution would require the largest changes in current practices.

AWS design would be truly collaborative, and the training of the system and the warfighter would be simultaneous. In fact, this model is more akin to the education or rearing of a child rather than the engineering of a tool. And for that very reason, the present bureaucratic acquisition processes, rules, success metrics, and even the very idea of verifying a changing system cannot work. Even current proposals for acquisition reform – for example, in the United States[15] – do not extend as far as this strategy.

Rather than speeding development and deployment to get technologies onto the battlefield immediately, this type of change would require more time designing and understanding the weapons systems. The *overall* acquisition process would not necessarily take more time, but more time would be spent interacting with and developing the system holistically.

Indeed, these required changes in practices are the single biggest negative about this strategy, as it is simply impractical to achieve this type of shared understanding given current military practice, and it is implausible that current practices (including classification and compartmentalization of information) would change sufficiently to make this strategy feasible. Of course, the potential upside is the formation of exactly the high degrees of interpersonal trust required for successful use and deployment of AWSs. However, there is little reason to think that it would be feasible at the current time, despite some recent rhetoric about the importance of extensive training with robotic systems (Seck 2016). Moreover, even if significant changes were plausible, relatively little is known about exactly how disparate groups can and should work together to develop shared understanding of, and so widespread trust in, some technological object. One might hope that there would be analogous cases in other domains, such as the development and deployment of medical technologies, but there are few cases of successful trust-building in those processes either (Kimmelman and London 2015).

6. Conclusion

AWSs have been the focus of enormous debate and controversy, whether in academic, military, political, or public venues. Many of these discussions focus on their potential negative impacts on civilian populations – direct (e.g. erroneous targeting) and indirect (e.g. reducing psychological barriers to initiating hostilities) – and the possible lack of corresponding accountability or responsibility for an AWS's actions. While we agree that these are important issues, the present paper has instead looked at AWSs from the perspective of the warfighters. *Even if* all of those other issues of targeting and accountability could be resolved, we contend that soldiers would not have the required levels of trust in the AWS, and so would use it either inappropriately or not at all. That is, the warfighters themselves have pragmatic reasons not to use an AWS. Moreover, technological advances in learning and planning capabilities are (ironically) likely to make matters worse for the foreseeable future, not better. Improving the ability of an AWS to adapt to its environment and generate complex plans will likely worsen the ability of warfighters to understand, and thus to trust, the system. This level of trust in something autonomous, whether human, animal, or machine, requires more than mere predictability and reliability; it requires the trustor to understand *why* the trustee does what he, she, or it does. Achieving such trust in these systems is possible, but only with wholesale changes in training, doctrine, development, and acquisition. The suggestions that we have offered are first slices, but they do highlight the difficulties of establishing trust with weapons systems that are essentially agent-artifacts. The present military structure views weapons as tools and humans as agents. This structure is incapable of supporting trust in AWSs that can learn, plan, and adapt. War is a human activity, and trust in these systems is a necessary human element. Either the structure must change, or the weapons must be regarded with extreme caution.

Notes

1. We note two other sources of problems for the "10 levels of automation" approach (see Appendix Table 1). First, the levels do not reflect (smooth) variations in various dimensions. For example, there is a qualitative shift from level 1 to level 2 as the addition of the machine fundamentally changes the event. Second, many of the levels are distinguished by differences in the human, not the machine. For example, in levels 6–10, the system always has the same capability: it makes the decision to act. The only difference is interface design to allow the human to "veto" or intervene on the computer's decision. In short, there is no difference in how "automated" the *system* is. These levels are not about degrees of automation, but rather extent of human–computer interaction.
2. Many close-in weapons systems require only limited planning capabilities. For example, the Phalanx system is a close-in weapon system that searches, detects, and engages incoming rockets, mortars, missiles, and other craft without a human in the loop, but also without extended planning.
3. The US Defense Advanced Research Project Agency's present program "Target Recognition and Adaptation in Contested Environments (TRACE)" is a learning target identification system that uses deep neural networks to engage in rapid, real time target identification using synthetic aperture radar imagery. This particular system has anti-material capabilities, but extensions to anti-personnel roles are natural.
4. At least, assuming the relevant contextual features are reasonably stable, the system does not exhibit dynamic structural changes, and so on.
5. And in the other direction, command responsibility in modern militaries means that commanders will take every reasonable step to ensure the reliability of their subordinates' competence at their duties.
6. The use of such weapons would undermine the values associated with civilian protection, the laws of armed conflict, human rights, and the like, and thereby erode trust in the state that fields them, regardless of the reason. Of course, only about 26 of the world's almost 200 states claim to perform Article 36 weapons reviews.
7. The "OODA Loop" is the idea that military battles involve cycles of: Observing the battlespace; Orienting to the enemy through our ways of processing information; Deciding what action to take; and Acting. The developer of the idea, John Boyd, viewed combat as a series of simultaneous, iterated loops, and so interruption of an adversary's OODA loop would impair their ability to make effective combat decisions.
8. As Roff (2014) argues as well, there are important considerations to take into account when learning AWSs are used in combat because their tactical actions may have strategic effects.
9. Swarms will be predictable on the whole, as they are created and tested as a full and complete unit. However, each individual member of the swarm may not be predictable, and so we say that the swarm has emergent properties when all individual units are acting in concert. This may not seem entirely problematic if the swarm is a school of fish or a flock of birds, but in those instances the effects of one (presumably armed) AWS going astray or not behaving as intended can be tragic.
10. Or resulting from a learning algorithm that updates an initial value system given training data.
11. One might object that the developers do have appropriate knowledge, and so they can teach the warfighter. Even if the developers really did have that level of understanding, it is unclear how this could be translated to the front-line warfighter in a way that enables her to predict, control, and explain the AWS's functioning in novel, complex environments.
12. Interestingly, the United States Air Force Research Lab Report "Autonomous Horizons" (USAFRL 2015) notes that part of developing "trustworthy autonomy" would be to institute and support "airman-autonomy joint training." This would be a "mixed-initiative team training as part of any system development and deployment effort" to aid "in developing an understanding of common team objectives, the separate roles of the airman and the autonomous system, and the ways in which they are co-dependent" (23). The report is clear that

this kind of training should enable an airman to understand the limits of a system's operation, or when it may be approaching those limits due to particular observed behaviors. While we certainly agree with this assessment, it also shows that whatever system the AFRL has in mind, it is far more complex and cognitively capable along the learning and planning dimensions than any presently fielded weapons system.

13. One could instead require warfighters to constantly monitor autonomous systems in a supervisory control manner, but that would significantly increase their cognitive workload, potentially increasing risks associated with AWS use. Humans are incredibly poor at maintaining situational awareness when they are overloaded and over-taxed, and so these additional demands would likely have significant detrimental impacts on the warfighters who are supposed to be helped by the AWS (Kruij et al. 2012; Zhang et al. 2015).
14. Of course, repeated interactions are neither necessary nor sufficient. Interpersonal trust can arise for other grounds or reasons. Repeated interactions and histories with others are merely the typical route to build interpersonal trust.
15. The 2017 Defense Authorization Bill, for example, wants to cut at least three years off of acquisition cycles by eliminating the need to provide all funds for a new weapons system's life cycle. Moreover, there are initiatives afoot to institutionalize rapid acquisition and development.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Heather M. Roff is a Research Scientist at the Global Security Initiative at Arizona State University, a Senior Research Fellow at the Department of Politics and International Relations at the University of Oxford, and a Future of War and Cybersecurity Fellow at New America. Her research interests include the ethics, law, and policy of emerging military technologies, such as autonomous weapons, artificial intelligence, and cybersecurity. She publishes in a wide range of journals and policy outlets and is the author of *Global Justice, Kant and the Responsibility to Protect* (2012, Routledge).

David Danks is L.L. Thurstone Professor of Philosophy & Psychology, as well as Head of the Department of Philosophy, at Carnegie Mellon University (Pittsburgh, PA). His primary research interests are at the intersection of philosophy, cognitive science, and machine learning, including human-centric dimensions and impacts of autonomy in technological systems. He has published in a wide range of journals across multiple disciplines, and is the author of *Unifying the Mind: Cognitive Representations as Graphical Models* (2014, MIT Press).

References

Anderson, Kenneth, and Matthew C. Waxman. 2012. "Law and Ethics for Robot Soldiers." *Policy Review* 176. Accessed April 30, 2018. http://www.cfr.org/world/law-ethicsrobot-soldiers/9598p2.

Arkin, Ronald. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: Chapman and Hall/CRC Press.

Arkin, Ronald. 2014. "Lethal Autonomous Weapons Systems and the Plight of the Noncombatant." Paper at the United Nations informal meeting of experts at the convention on conventional weapons, May 14, 2014, Geneva, Switzerland. Accessed April 30, 2018. https://www.unog.ch/80256EDD006B8954/(httpAssets)/FD01CB0025020DDFC1257CD70060EA38/$file/Arkin_LAWS_technical_2014.pdf.

Asaro, Peter. 2012a. "How Just Could a Robot War Be?" In *Ethics of 21st Century Military Conflict*, edited by Erica L. Gaston and Patti Tamara Lenard, 257–269. New York: IDEBATE Press.

Asaro, Peter. 2012b. “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making.” *International Review of the Red Cross* 94 (886): 687–709.

Bartlett, Cade E., and Nancy J. Cooke. 2015. “Human-Robot Teaming in Urban Search and Rescue.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59 (1): 250–254.

Beck, Hall P., Mary T. Dizndolet, and Linda G. Pierce. 2007. “Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49 (3): 429–437.

Chen, Jessie Y. C., and Michael J. Barnes. 2012. “Supervisory Control of Multiple Robots: Effects of Imperfect Automation and Individual Difference.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54 (2): 157–174.

Chen, Jessie Y. C., Michael J. Barnes, and Michelle Harper-Sciarini. 2013. “Supervisory Control of Multiple Robots: Human-Performance Issues and User-Interface Design.” *Autonomous Systems* 2 (2): 119–138. Accessed April 30, 2018. https://www.arl.army.mil/www/pages/172/docs/Research@ARL_Autonomous_Systems_July_2013.pdf.

Chen, Jessie Y. C., and C. T. Joyner. 2009. “Concurrent Performance of Gunner’s and Robotics Operator’s Tasks in a Multitasking Environment.” *Military Psychology* 21 (1): 98–113.

Cummings, Missy L. 2004. “The Need for Command and Control Instant Message Adaptive Interfaces: Lessons Learned from Tactical Tomahawk Human-in-the-loop Simulations.” *Cyberpsychology and Behavior* 7 (6): 653–661.

Cummings, Missy L., and Stephanie Guerlain. 2007. “Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49 (1): 1–15.

Danks, David. 2016. “Finding Trust and Understanding in Autonomous Technologies.” *The Conversation*, December 30, 2016. Accessed April 30, 2018. http://theconversation.com/finding-trust-and-understanding-in-autonomous-technologies-70245.

Danks, David, and Alex John London. 2017. “Regulating Autonomous Systems: Beyond Standards.” *IEEE Intelligent Systems* 32 (1): 88–91.

Deutsch, Morton. 1958. “Trust and Suspicion.” *Journal of Conflict Resolution* 2: 265–279.

Deutsch, Morton. 1973. *The Resolution of Conflict*. New Haven: Yale University Press.

Freedberg, Sydney J. 2015. “Centaur Army: Bob Work, Robotics, & The Third Offset Strategy.” *Breaking Defense*, November 9, 2015. Accessed April 30, 2018. http://breakingdefense.com/2015/11/centaur-army-bob-work-robotics-the-third-offset-strategy/.

Gorman, Jamie C., Nancy J. Cooke, and Jennifer L. Winner. 2006. “Measuring Team Situation Awareness in Decentralized Command and Control Environments.” *Ergonomics* 49 (12–13): 1312–1325.

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. deVissar, and Raja Parasuraman. 2013. “A Meta-analysis of Factors Affecting Trust in Human-robot Interaction.” *U.S. Army Research Lab Autonomous Systems* 2 (2): 177–188.

ICRC (International Committee for the Red Cross). 1977. *Protocol Additional to the 1949 Geneva Conventions*. Accessed April 30, 2018. https://www.icrc.org/applic/ihl/ihl.nsf/Article.xsp?action=openDocument&documentId=FEB84E9C01DDC926C12563CD0051DAF7.

ICRC (International Committee for the Red Cross). 2016. *Autonomous Weapons: Decisions to Kill and Destroy are a Human Responsibility*, Statement of the ICRC, Meeting of Experts on Lethal Autonomous Weapons Systems United Nations Convention on Certain Conventional Weapons, Geneva, April 11–16, 2016. Accessed April 30, 2018. https://www.icrc.org/en/document/statement-icrc-lethal-autonomous-weapons-systems.

Kester, Leon. 2016. “Mapping Autonomy, Testimony to given to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems at the United Nations Convention on Certain Conventional Weapons,” April 11, 2016. Accessed April 30, 2018. http://www.unog.ch/80256EDD006B8954/(httpAssets)/D5BD627982BE28D8C1257F9200533ADF/$file/05+Leon+Kester_Mapping+autonomy.pdf.

Kimmelman, Jonathan, and Alex John London. 2015. “The Structure of Clinical Translation: Efficiency, Information, and Ethics.” *Hastings Center Report* 45: 27–39.

Klincewicz, Michal. 2015. “Autonomous Weapons Systems, the Frame Problem and Computer Security.” *Journal of Military Ethics* 14 (2): 162–176.

Kreisher, Otto. 2016. “Naval Aviation Goal is to Turn Kill Chain into ‘Kill Web’.” *SeaPower Magazine*, March 22. Accessed April 30, 2018. http://www.seapowermagazine.org/stories/20160322-manazir.html.

KRHRW (Killer Robots Human Rights Watch). 2015. *Mind the Gap: Accountability Gap in Lethal Autonomous Weapons Systems*, April 9. Accessed April 30, 2018. https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots.

Kruij, Geert, Miroslav Janicek, Shanker Keshavdas, Benoit Larochelle, Hendrik Zender, Nanja Smets, Tina Mioch, et al. 2012. “Experience in System Design for Human-Robot Teaming in Urban Search & Rescue.” In: *8th International Conference on Field and Service Robotics* (FSR 2012). Accessed April 30, 2018. https://pdfs.semanticscholar.org/58e2/3f277caca22b36d6db50da3b19ca638257d7.pdf.

Lee, John D., and Neville Moray. 1992. “Trust, Control Strategies and Allocation of Function in Human-Machine Systems.” *Ergonomics* 35 (10): 1243–1270.

Lee, John D., and Katrina A. See. 2004. “Trust in Automation: Designing for Appropriate Reliance.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1): 50–80.

Lewicki, Roy J., Edward C. Tomlinson, and Nicole Gillespie. 2006. “Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions.” *Journal of Management* 32 (6): 991–1022.

Melzer, Nils. 2013. “Human Rights Implications of the Usage of Drones and Unmanned Robots in Warfare.” *Directorate-General for External Policies of the European Union Directorate B Policy Department Study*. EXPO/B/DROI/2012/12, Accessed April 30, 2018. http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/410220/EXPO-DROI_ET(2013)410220_EN.pdf.

Moray, Neville, Toshiyuki Ingagaki, and Makoto Itoh. 2000. “Adaptive Automation, Trust, and Self-confidence in Fault Management of Time-critical Tasks.” *Journal of Experimental Psychology: Applied* 6 (1): 44–58.

Narayanan, Vignesh, Yu Zhang, Nathaniel Mendoza, and Subbarao Kambhampati. 2015. “Automated Planning for Peer-to-peer Teaming and its Evaluation in Remote Human-Robot Interaction.” In *Proceedings of the ACM/IEEE international conference on human robot interaction (HRI).*

Parasuraman, Raja, Michael J. Barnes, and Keryl Cosenzo. 2007. “Adaptive Automation for Human-Robot Teaming in Future Command and Control Systems.” *International C2 Journal* 1 (2): 43–68.

Pellerin, Cheryl. 2015. “Work: Human-Machine Teaming Represents Defense Technology Future.” *DoD News, Defense Media Activity*, November 8. Accessed April 30, 2018. http://www.defense.gov/News-Article-View/Article/628154/work-human-machine-teaming-represents-defense-technology-future.

Roff, Heather M. 2013. “Killing in War: Responsibility, Liability and Lethal Autonomous Robots.” In *Routledge Handbook for Ethics and War: Just War Theory in the 21st Century*, edited by Fritz Allhoff, Nicholas G. Evans and Adam Henschke, 352–364. London: Routledge.

Roff, Heather M. 2014. “The Strategic Robot Problem: Lethal Autonomous Weapons in War.” *Journal of Military Ethics* 13 (3): 211–227.

Roff, Heather M. 2015. “Autonomous or Semi-Autonomous Weapons? A Distinction Without a Difference?” *Huffington Post*, January 16. Accessed April 30, 2018. http://www.huffingtonpost.com/heather-roff/autonomous-or-semi-autono_b_6487268.html.

Roff, Heather M. Forthcoming. “An Ontology of Autonomy and Autonomous Weapons Systems.” In *The Ethics of Autonomous Weapons*, edited by Claire Finkelstein, Duncan MacIntosh, and Jens David Ohlin. Oxford: Oxford University Press.

Scharre, Paul. 2016. *Autonomous Weapons and Operational Risk*. Washington, DC: Center for a New American Security. Accessed April 30, 2018. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515.

Schmitt, Michael. 2013. “Autonomous Weapons Systems and International Humanitarian Law: A Reply to the Critics.” *Harvard National Security Journal Features*, February 5. Accessed April 30, 2018. http://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/.

Seck, Hope Hodge. 2016. "Corps Sees Hurdle in Getting Marines to Bond with Robotic Battle Buds." *Military.com*, May 18. Accessed April 30, 2018. http://www.military.com/daily-news/2016/05/18/corps-sees-hurdle-in-getting-marines-to-bond-robotic-battle-buds.html.

Sharkey, Noel. 2008. "Grounds for Discrimination: Autonomous Robot Weapons." *RUSI Defence Systems*, October 31: 86–89.

Sheridan, Thomas B., and William L. Verplank. 1978. *Human and Computer Control of Undersea Teleoperators*. Technical Report, Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77.

Sparrow, Robert. 2011. "Robotic Weapons and the Future of War." In *New Wars and New Soldiers: Military Ethics in the Contemporary World*, edited by Jessica Wolfendale and Paolo Tripodi, 117–133. Farnham: Ashgate Press.

The Netherlands. 2016. "Opening Statement at the General Debate, 3rd Informal Meeting of Experts on LAWS." United Nations Convention on Certain Conventional Weapons Informal Meeting of Experts. Accessed April 30, 2018. http://www.unog.ch/80256EDD006B8954/(httpAssets)/FC2E59B32F14D791C1257F920057CAE6/$file/2016_LAWS+MX_GeneralExchange_Statements_Netherlands.pdf.

USAFRL (United States Air Force Research Lab). 2015. *Autonomous Horizons: System Autonomy in the Air Force: A Path to the Future*. Washington, DC: Department of the Air Force.

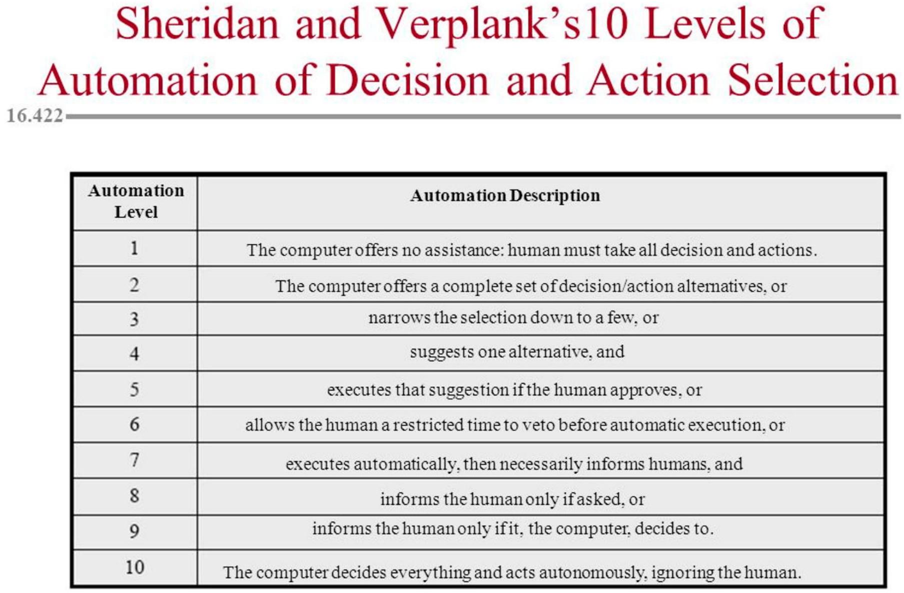
USDoD (United States Department of Defense). 2012a. *Task Force Report: The Role of Autonomy in DoD Systems*. Accessed April 30, 2018. https://www.acq.osd.mil/dsb/reports/2010s/AutonomyReport.pdf.

USDoD (United States Department of Defense). 2012b. *Department of Defense Directive 3000.09*. Accessed April 30, 2018. http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

Zhang, Yu, Vignesh Narayanan, Tathagata Chakraborti, and Subbarao Kambhampati. 2015. "A Human Factors Analysis of Proactive Support in Human-Robot Teaming." In *Proceedings of the 2015 international conference on intelligent robots and systems*. Accessed April 30, 2018. https://pdfs.semanticscholar.org/aac9/dd6200e4c3052c539fe6a9be0b3755415e5e.pdf.

Appendix

Table 1: Sheridan and Verplank and the citation.



Sheridan and Verplank's10 Levels of Automation of Decision and Action Selection

16.422

Automation Level	Automation Description
1	The computer offers no assistance: human must take all decision and actions.
2	The computer offers a complete set of decision/action alternatives, or
3	narrows the selection down to a few, or
4	suggests one alternative, and
5	executes that suggestion if the human approves, or
6	allows the human a restricted time to veto before automatic execution, or
7	executes automatically, then necessarily informs humans, and
8	informs the human only if asked, or
9	informs the human only if it, the computer, decides to.
10	The computer decides everything and acts autonomously, ignoring the human.