AI-ENABLED WEAPONS AND JUST PREPARATION FOR WAR

JOVANA DAVIDOVIC
MILTON REGAN

babl

STOCKDALE
Return with Honor

EXECUTIVE SUMMARY

Just preparation for war (*jus ante bellum*) requires that the development and deployment of weapons be done in such a way so as to minimize unjust resort to war and unjust fighting in war. This report examines what *jus ante bellum* requires of states with respect to the development and deployment of AI-enabled weapon systems.

The development and use of AI-enabled weapons gives rise to distinctive risks for our ability to wage just wars. To minimize such risks, we ought to engage in rigorous testing, evaluation, validation and verification (TEVV) of AI-enabled weapons. Such testing should be a) cradle to grave, and b) modular and principled, and should be followed by c) gradual fielding in d) clearly defined operational envelopes, with e) appropriate explainability. In addition to adjusting TEVV to bette suit risk mitigation during the development of AI-enabled weapons, countries should also seek to mitigate the AI-fueled security dilemma, which might arise from an AI arms race. To do so countries should invest in AI-weapon specific confidence building measures as well as a careful language of war preparedness.

INTRODUCTION

No potential application of artificial intelligence (AI) has prompted more urgent ethical concern than the prospect of AI-enabled weapons. The fear is that machines will precipitate war without human intervention and take human life during conflict without regard for innocent civilians. Framed in terms of the centuries-old just war tradition, AI-enabled weapons pose the risk of violating the principles of *jus ad bellum*, which govern when it is ethically justifiable to resort to war, and *jus in bello*, which determine when it is ethically justifiable to take life during war.

These concerns are appropriate and significant. It is important to appreciate, however, that whether a state resorts to war and how a state conducts it can be shaped by how it engages in preparation for the possibility of war. Certain forms of military preparedness may make war less or more likely, for instance, or may enable or foreclose ways in which a state fights war. Recognition of this has led to an emerging strand of just war thought known as *jus ante bellum;* a strand of just war theory that asks: what ethical principles should guide how a state prepares for war?

As Harry van der Linden puts it, *jus ante bellum* asks "whether the military preparation of the country is such that it is conducive to the country resorting to force only when justice is on its side, as well as to executing and concluding war in the just manner. How should we prepare for the possibility of military conflicts so that wars will be only justly initiated, executed, and concluded?"[i] In other words, just preparation for war minimizes the chances that a state will resort to unjust wars, or that it will fight in war using unjust means, such as targeting civilians, mistreating prisoners of war, and using inhumane methods.

This report examines what *jus ante bellum* requires of states with respect to the development and deployment of weapons that incorporate AI, or AI-enabled weapons.

In other words, the report examines the key risks to just preparation for war that the development of AI-weapons can present. While precisely defining an AI-enabled weapon can be a challenge, for the purposes of this report we define it as a weapon that for its functioning utilizes machine learning algorithms, which may include algorithms developed through deep learning techniques. This would include, for instance, a weapon that uses AI for object recognition to inform the targeting process, a weapon that uses AI to identify incoming missiles and potentially engages them or a decision-support system that provides advice on potential tactical choices available to a unit leader.

While much has been said in literature on whether or not developing weapons and having a standing military can be justified in the first place, we agree with Cecile Fabre that maintaining a standing army that is prepared to wage war if need be is morally justified because it enables a state to protect persons from violent infringements of their fundamental rights.[ii] With that said, and in light of increasing attention by several states to the potential for incorporating AI into weapon systems, we believe that a state may be justified in investing in the development of AI-enabled weapons.[iii] Development of AI-enabled weapons by potential adversaries need not necessarily be met by a comparable response, but it is plausible to believe that this may be one motivation. Furthermore, the potential value of AI in, for example, minimizing civilian casualties makes it worth considering whether and when AI-enabled weapons can contribute to fighting just wars.

AI-weapons development also carries some significant risks. For example, premature deployment of AI-enabled weapon systems, and the deployment of systems with an inappropriate delegation of authority between machines and humans, can increase the risk of violations of *jus ad bellum* and *jus in bello*. With regard to *jus ad bellum*, preparation that leads other states to fear overwhelming attacks increases the likelihood that states will use these systems preemptively rather than defensively. In addition, systems that rely on automation in situations that require human judgment create the risk that low-level crises will escalate to armed conflict. With respect to *jus in bello*, systems that have not been

adequately tested for safety and reliability, and in which there is an inappropriate balance of machine and human involvement, heighten the risk of harm to innocent persons.

Because of those significant ethical risks, *jus ante bellum* requires that a state:

(1) not deploy AI-enabled weapons without rigorous testing, evaluation, verification, and validation (TEVV), which should include due consideration of the appropriate delegation of tasks between machines and humans;

(2) not engage in development of these weapons in ways that trigger insecurity in other states that leads them to deploy AI-enabled systems without engaging in this analysis.[iv]

The next section elaborates on the key risks of AI-enabled weapons systems. The following section discusses the critical role of rigorous TEVV in reducing these risks, which is the basis for the claim that *jus ante bellum* requires engaging in this process before deploying AI-enabled weapons systems. The article then discusses how even a state that conducts rigorous TEVV may develop AI weapons systems in ways that generates a security dilemma in which other states have incentives to hasten deployment of AI-enabled weapon systems without sufficient testing and review. The section argues that *jus ante bellum* requires states to attempt to take steps to minimize these risks, and suggests how they might do so. A concluding section summarizes the importance of subjecting development of AI-enabled weapons to *jus ante bellum* analysis.

RISKS OF AI-ENABLED WEAPONS

Any complex system whose components are tightly coupled is subject to what is called "normal accidents."[v] Such accidents in facilities such as nuclear power plants, passenger airline operations, hydroelectric systems, and some military operations can lead to large-scale unintended harms to innocent people.[vi] Infliction of these harms by military systems also create the risk of escalating tensions that may lead to conflict, as other states respond to what they regard as aggression by another state. An accidental nuclear strike, for instance, could trigger a retaliatory strike, with large-scale annihilation of innocent populations as the result.

AI-enabled weapons will feature tightly coupled human-machine interactions that culminate in the use of lethal force. AI systems are based on models with million or sometimes billions of statistical parameters, which are learned automatically and are often very challenging for humans to interpret. And as Michael Horowitz suggests "[e]ven if commanders understand how the systems are supposed to work, and deploy them appropriately, the complexity of the programming raises the prospect of unintended behaviour and accidents."[vii] Furthermore, these complex systems will be competing on the battlefield, which can "further raise the degree of uncertainty about their behaviour and the risk of accidents." AI-enabled weapons systems therefore pose the same risk that other complex systems do of normal accidents with potentially grave consequences. AI-enabled weapons also carry distinctive risks because of their incorporation of artificial intelligence. First, testing systems for reliability and safety is a major challenge because it is impossible fully to replicate before deployment the unpredictable conditions that characterize warfare. The data on which a system is trained therefore can never completely represent the conditions under which it will operate. Second, related to this, systems can be vulnerable to adversarial attacks that distort the information they receive. Warfare, by definition, involves conflict between adversaries, so the risk of this seems especially high.

Third, systems at this point tend to be brittle, in the sense that they are not able to function effectively outside the specific set of circumstances for which they are trained. The consequences they cause when they operate in such an environment will be difficult to predict and could result in significant harm to innocent persons or one's own force. Fourth, it can be challenging for human operators to determine when a system has ventured into this situation and therefore when it is necessary to disable it. In addition, a system may not be able to provide an explanation of its analysis and recommendations in terms that are comprehensible to a human operator. This opacity can make it difficult for humans to exercise effective judgment, thus creating the risk of deferring to AI without possession of full situational awareness.

These features (brittleness, opacity, hackability, etc.) of AI-enabled weapon systems could increase risk of violations of *jus ad bellum* and *jus in bello*.

With respect to *ad bellum*, states could field systems that are less flexible than conventional weapons and lack sensitive contextual awareness of likely human intentions. "This brittleness of machine decision-making may particularly be challenging in a pre-conflict crisis situations, where tensions among nations run high," and contextual human judgment can be crucial in lessening the risk of escalation.[viii] Furthermore, even if a system performs as intended, adversaries may not know whether its behavior reflects human intention. This ambiguity may lead to escalation of conflict if states assume that they must ascribe hostile intention to an adversary in order to protect themselves. Finally, systems' "intrinsic vulnerability to unexpected interactions or operational accidents raises the specter of inadvertent escalation into a 'flash war' between autonomous military systems, similar to the algorithmic flash crashes already observed in the financial sector."[ix] All these scenarios pose the risk that states will resort to war without just cause, or meeting other requirements of the *jus ad bellum*.

With respect to *in bello* violations, delegation of some tasks to machines could mean that "minor tactical missteps or accidents that are part and parcel of military operations in the chaos and fog of war, including fratricide, civilian casualties, and poor military judgment, could spiral out of control and reach catastrophic proportions before humans have time to intervene."[x] If so, mechanisms and processes for deliberation and reconsideration based on human judgment would be unavailable, which could mean intensification of warfare resulting in greater suffering and larger loss of innocent life. This risk would be exacerbated by the interaction between and among competing AI-enabled systems that could result in a cycle of attacks and counterattacks at a speed that humans could not control.

These risks underscore the crucial importance of rigorous pre-deployment review of AI-enabled weapons. The next section discusses what states must do to conduct rigorous testing, evaluations, validation and verification (TEVV), while the following section discusses how they might take steps that avoid triggering a security dilemma.

TESTING, EVALUATION, VALIDATION AND VERIFICATION

Deployment of AI-enabled weapons that have not been rigorously tested for safety and reliability increases the risk of unjust resort to war and of harm to innocent persons. To avoid this, deployment should be preceded by a high quality, rigorous process known as of testing, evaluation, verification, and validation (TEVV). This process, drawn from systems engineering, is designed to assess the future performance of new technology and the risks that it may pose.[xi] The TEVV process in general terms seeks to provide assurance that technology will work as expected, which generates what Heather Roff and David Danks call predictability-based trust.[xii] Because a weapon can cause significant harm, however, TEVV of weapon systems also aims to provide what Roff and Danks call values-based trust: confidence that a weapon will operate in a way that is consistent with certain ethical demands, such as the principle of not targeting civilians.

Roff and Danks observe that the paradigm of values-based trust is interpersonal relationships, in which trust reflects confidence that another person will act ethically in unpredictable future situation because we know the "values, principles, beliefs and motives that guide the trustee's choices and actions."[xiii] Such understanding is based on repeated interactions with another in which we "perform our own risk calculations about what to internalize about the other agent's mindset and how vulnerable to make ourselves."[xiv] In human-machine teams, however, the values of the machine may be inscrutable to the human. Roff and Danks argue that AI-enabled weapons present us with a fundamental tension: the extent that an AI-enabled weapon "actually learns and adapts to its environment in operationally ways will be inversely proportional to the extent to which the human team members can identify and internalize its values and preferences."[xv] The more advanced an AI-enabled weapon system, the greater the need to have not only the reliability-based trust that we have in tools, but the kind of values-based trust that we have in other human beings.

A weapon system TEVV thus must seek to generate the right kind of calibrated trust in commanders who decide to deploy the weapon system and operators who use it. [xvi] Trust is calibrated when "the warfighters' operational reliance aligns with the system performance for the context."[xvii] Trust is the right kind when it grounded both in predictability and conformity with values. [xviii] As the discussion below describes, there are features of AI-enabled weapons that can make it difficult for the TEVV process to generate this trust in such weapons.

CHALLENGES

Appropriate Unit of Analysis

Fitness for Purpose

Generalization from Testing

Unpredictable Failures

Piecemeal Approach to Development

Dynamic and Open Systems

CHALLENGES

AI-enabled weapons present distinctive challenges for the TEVV process because of their complexity, opacity, and brittleness. [xix] We discuss these challenges below, and suggest how TEVV should respond to them in order to satisfy the *jus ante bellum* requirement to provide assurances of safety, operational effectiveness, and conformity with values and in order to provide grounding for calibrated trust, of the kind discussed above. The challenges to TEVV for AI-enabled weapons include: appropriate unit of analysis, fitness for purpose, generalization from testing, unpredictable failures, piecemeal approach to development and unique challenges posed by open systems. We address each one of these in turn.

Appropriate Unit of Analysis

It is harder to define the appropriate unit of analysis for the TEVV process for an AI-enabled weapon than for a conventional one. First, the same algorithm can be utilized across a range of weapon system applications. An object recognition algorithm, for instance, might be used to enable autonomous navigation in a tank, or to inform targeting by distinguishing between objects that are weapons and those that are not.

Second, algorithms often function within a system of systems. This refers to the fact that several different algorithms may each provide inputs for one another. Thus, for instance, referring to the object recognition algorithm above, this algorithm could provide input to a decision support algorithm that collates and processes inputs from other algorithms and presents alternative courses of action to a commander.

Third, the ability of an algorithm to function properly greatly depends on the interaction between humans and machines. With respect to a weapon, this involves the way in which an algorithm's outputs are presented to a commander or operator and how the human is to use these outputs in deciding how to proceed. This means that it may be necessary to test AI-enabled weapons with humans who have different types of training in order to provide assurance of the weapon's performance.[xx]

All this means that it is not always easy to identify and circumscribe the appropriate unit of analysis for a test or a risk assessment. For instance, should we test an algorithm or a particular application? How often should testing be repeated and in which configurations?

Fitness for purpose

The ability of TEVV to provide assurance that an AI-enabled weapon will perform as it should requires assessing its performance in the environment in which it will be used and developing it with a clear purpose in mind and a clear and correct model of the environment in which it is to be used. This requires the availability of training data that accurately simulates that environment. Each deployment of a weapon in a new environment therefore may require a new TEVV. More fundamentally, this requires a definition of what constitutes a "new" environment for which a previous assessment cannot provide assurance of performance. The consequences of using the system in the wrong context - when the algorithm is not fit for purpose are significant, but what counts as a 'wrong' context is often difficult to discern.

Generalization from testing

Related to the above issue, generalizing and extrapolating from test results is extraordinarily challenging for many AI-enabled weapons systems because of the exceptional difficulty in anticipating all the conditions under which these weapons will operate. It is true that conventional weapons present a similar obstacle to some extent -- we can test only a fraction of the settings in which a weapon may operate. AI-enabled weapons, however, perform extremely complex tasks, they do so in radically unpredictable environments, and they provide "non-deterministic, dynamic responses to those environments."[xi] All of this makes the range of potential scenarios to test immense, if not infinite. Compared with conventional weapons, we can generalize with much less confidence about performance across varied environments.[xxii]

Flournoy, et.al. discuss this challenge as one of "brittleness," observing that the "traditional TEVV approach is not well suited for ML/DL [machine learning/deep learning]" because, "ML/DL system performance is difficult to characterize and bind, and the brittleness of such systems means they will require regular system updates and testing."[xxiii] Thus, in addition to the inability to predict all potential scenarios and to extrapolate from testing data, TEVV for AI-enabled weapons requires ongoing adjustment to provide assurance of performance in new operational environments. This means that TEVV cannot simply be done once and for all, prior to deployment, but will need to be continued after a weapon is deployed.

Unpredictable failures and opacity

Brittleness also refers to the fact that failures of AI-enabled weapons are harder to predict and more difficult to understand than is the case with conventional weapons. This complicates the ability of TEVV to anticipate how AI systems will perform in various operational environments, and to identify those settings to which its use should be confined. Deep learning techniques and systems of systems are particularly likely to present these challenges.

This leads some observers to suggest that explainability may be a prerequisite for adequate TEVV, and thus for deploying a weapon.[xxiv] This will require clarifying the meaning of this concept (explainability), since what needs to be explained may be different for each of the several parties involved in the development and deployment of AI-enabled weapons. The opacity of AI-enabled systems also may mean that TEVV will influence certification schemes for such weapons, such as requiring machine learning experience for operators of some high-risk systems.[xxv] It also may require moving away from insistence on complete risk avoidance and precise risk quantification toward acceptance of the risk of failure, with a focus on ensuring that a system fails "gracefully" in ways that do not cause harm or jeopardize the larger operation in which it is deployed.[xvi]

Piecemeal approach to development

A further complexity of AI-enabled weapons systems is that they often are not customized or built by the DoD, but might come from small or large private enterprises in response to a particular request from the Pentagon. Unlike traditional weapon systems, AI-enabled weapon systems (and specifically their AI components) are more likely to be assembled piecemeal from a variety of sources.[xvii] This is because much AI development is occurring in the private sector, and because AI is often utilized to solve for specific problems. It is more likely that an AI-enabled weapon is going to come not from a single weapons developer, as traditional weapon systems did, but from a range of sources.

This affects how some AI-enabled weapons can be tested. For example, for ML systems, verification and validation require unprecedented and publicly unavailable testing data that is fit for a specific purpose. A major obstacle for building and meaningfully testing AI-enabled weapons thus is access to large data sets that are appropriately built and maintained for such testing.[xviii]

Dynamic and open AI systems

Finally, some machine learning solutions to war-fighting problems will result in open/dynamic systems (sometimes called online learning systems) – that is, models that take in data from the environment and update the model as it is being used. That presents an obvious problem for the TEVV process, as each new algorithm (ML model) based on such learning in a sense may effectively become a new weapon. A robust TEVV process needs to make clear at what level of change in the model the weapon is sufficiently different from the previous one so as to trigger the TEVV process anew.

ADAPTING TEVV TO AI-ENABLED WEAPONS

Integrated, Cradle to Grave

Modular

Integrated with Legal Review

Gradual Deployment

Purpose-driven Transparency and Explainability Tools

Defining Operational Envelope

Enabling Certification Schemes

Undertaken in Various Configurations

ADAPTING TEVV TO AI-ENABLED WEAPONS

Given various ways in which AI-enabled weapons are different from conventional ones, the TEVV process needs to be adapted to address the challenges Ai weapons present. We focus here on recommendations that will serve the requirements of *jus ante bellum* to assure safety, precision, and accuracy, to avoid unjust resort to war, and to ensure that AI-enabled weapons are used in accordance with the *jus in bello*. The following recommendations are meant to address the challenges discussed in the previous section.

Ongoing and Integrated Cradle to Grave TEVV

TEVV process for an AI-enabled weapon should be ongoing throughout the life cycle of the weapon.[xix] This will enhance the ability to anticipate problems as the weapon encounters new circumstances and improve the ability to set limits on the operational environments in which it may be used. Both benefits will increase the likelihood that a weapon is developed and used in ways that meet the requirements of *jus ante bellum*. Ongoing TEVV also has the potential to address concerns about transparency and explainability.

Principled and Modular Approach to TEVV

Clear rules should be laid out for TEVV to identify what types of changes to application, use, context, or training data require or trigger a new TEVV. In other words, it should be clear when one or more components of an AI-enabled weapon need to be subjected to a new portion of the TEVV process or a new process altogether. We might call this "modular" TEVV: not every change to any part of an AI system requires a completely new TEVV, but clear rules should indicate what aspects of TEVV need to be repeated under which circumstances. Also it should be clear when a weapon requires an entirely new TEVV from the ground up. This is because a weapon's performance depends heavily on the operational environment, and new operational environments, or updates to a weapon, will require at least some new testing and evaluation.

A robust TEVV process needs not only to assess performance in appropriate operational environments, but also to define those environments, often in collaboration with those who are developing or integrating AI into weapon systems. Defining such operational environments for AI-enabled weapons is much harder than for conventional weapons. It may be, for example, that an object recognition model works very well in one climate or geographic region, but not in another, for unanticipated reasons. Thus, both new training data, and potentially new operational environments should trigger the requirement that some or all elements of the TEVV process be conducted anew.

Integration of TEVV and Legal Weapons Review

Article 36 of Additional Protocol I to the Geneva Conventions requires review of a weapon to determine if its use can comply with international law requirements of discrimination, humanity, and respect for the environment. Some legal scholars have argued that this review should be parallel to and incorporated into the TEVV process. As Vestern and Rossi suggest, although testing and technical assessment traditionally have been conducted prior to legal weapons review, and provide evidence for such review, it may be necessary to incorporate legal requirements into the technical specifications of AI-enabled weapons.[xxx] This means legal review would need to be undertaken in tandem with the TEVV process.[xxxi] This seems appropriate in order for the TEVV process to help ensure compliance with *jus ante bellum* requirements and *jus in bello* requirements. Under this approach, operational circumstances that trigger the need for additional TEVV may also trigger the need for additional legal weapons review.

Better Testing Data and Gradual Deployment

Many algorithms relevant to weapons systems, such as object recognition or decision augmentation algorithms, are trained and validated in simulated environments. Simulation-based testing data, however, will often be inadequate when the risks of deploying a weapon are especially high. In these cases, data sets based on actual conditions are preferable because they can increase commanders' and operators' ability to trust a system in high-risk operational environments.

It thus will be crucial in many cases that an AI-enabled weapon be tested in actual rather than simulated environments, and that such systems be deployed only gradually. "[A] strategy of graded autonomy (slowly stepping up the permitted risks of unsupervised tasks, as with medical residents) and limited capability fielding (only initially certifying and enabling a subset of existing capabilities for fielding) could allow the services to get at least some useful functionality into warfighters' hands while continuing the T&E process for features with a higher evidentiary burden."[xxxii]

Differing Purposes of Transparency and Explainability

The transparency and explainability required to conduct the TEVV process may differ from the transparency and explainability necessary for its operation. TEVV, for instance, requires transparency and explainability that enables, among other things, the defining of an operating envelope, that is, the set of conditions under which we expect the system to perform in expected ways. An operator, by contrast, may require transparency and explainability that enables an informed judgment about the appropriate level of reliance on machine outputs as the basis for choosing a responsible course of action.

This reflects the fact that transparency and explainability can serve various purposes. We therefore may need different types of transparency and explainability for different purposes. First, for TEVV, transparency can enable diagnosis: knowing why the system is exhibiting undesired behavior is the first step toward fixing it.[xxxiii] Second, it can assist prediction: being able to forecast how the system will behave in given circumstances is essential to effectively deploying it.[xxxiv] Third, transparency and explainability is important in bounding the system. This means understanding the limits of dependable performance in order to formulate tactics, techniques, and procedures for using the system, as well as identifying when monitoring the state of a system during its operation may be the only way to avoid undesirable behavior.[xxxv]

One way to assure the type of transparency that TEVV might require is to have systems for recording metadata. Wojton, et al. for example suggest that "If systems are recording data about their own decisions and internal processing, then stakeholders, including developers, testers, and even users, can gain more transparency into the system."[xxxvi] With respect to TEVV, this might be combined with "safety middleware or disabled functionality to execute what some call 'shadow testing,' where the complex system makes decisions about what it would do in the current situation without being allowed to implement or execute those actions."[xxxvii] Such shadow testing could also provide meaningful updated data from the operational environment as well as the equivalent of counterfactual explanations for certain behaviors that can be useful for the TEVV process and operators.

Defining the operating envelope with an eye on alternatives to AI

TEVV is traditionally primarily about safety and accuracy. As we have seen when it comes to AI-enabled weapons, however, TEVV must include assessing operational environments and defining the operating envelope. This will allow operators and commanders to know whether and when to trust the weapon system, as well as to inform them about the potential risks.

TEVV is meant to do this: provide insights into variable performance in a range of operational environments (the so-called operating envelope) – for all systems, whether legacy or AI-enabled. When it comes to AI-enabled weapons system, however, defining the operating envelope is significantly more complicated. TEVV therefore has a more significant role to play in defining appropriate operational environments for the use of an AI-enabled weapon.

TEVV can also provide some insight into the initial decision whether a particular algorithm should be used instead of a human or non-AI alternative. There may be times when an AI-enabled system that works relatively well in a given context will not be better than a human. This means that TEVV ought not to assess the safety and precision of a weapon in vacuum, but with an eye to the likely benefits and risks of reliance on humans or machines for similar functions in varied operational environments. TEVV can base such assessments on how well different systems would achieve the goals of a weapon, taking into account performance and risks from its use. In other words, TEVV should not simply assess the safety and precision of a weapon in isolation, but in comparison with available alternatives for similar functions in different operational environments.

TEVV should drive certification schemes

The iterative process used in TEVV can help guide appropriate training, skills and certifications of operators. For example, the US Joint AI Center proposed including four types of testing: algorithmic testing, human-machine testing, systems integration testing, and operational testing with real users in real scenarios.[xxxviii] The human-machine testing and the operational testing provide evidence not just for the evaluation of the weapon, but for how best a weapon should incorporate and present machine outputs in order to augment human judgment in the decision-making process. While TEVV has always played a role in US certification schemes for operators, the training content that can emerge from TEVV of AI-enabled weapons may well be significantly greater.

TEVV should be undertaken in various configurations of systems and people

As described above, AI-enabled weapons are often systems of systems – that is, chains of algorithms with one algorithm's output serving as input for another. In such cases it may not be possible or desirable to test only one algorithm at a time.[xxxix] This suggests that ML algorithms will need to be tested in various configurations, operating with, alongside, and/or in a chain with several other ML models/algorithms. Similarly, some scholars have argued that rigorous testing should focus on testing various configurations of both systems and humans.[xl]

In conclusion, a TEVV process for AI-enabled weapons must test not only for safety and reliability, but provide meaningful insights regarding appropriate human-machine interaction as well as compliance with relevant values. As we have suggested, this may mean that a robust TEVV process must incorporate criteria that are involved in a legal weapons review. In the ways we have described above, a TEVV process that is sensitive to the unique challenges of AI-enabled weapons can meet the requirements of the *jus ante bellum*. As the next section discusses, however, this alone will be insufficient to meet these requirements if a state develops AI-enabled weapons in ways that create the risk of a security dilemma.

AI-FUELED SECURITY DILEMMA

Uncertainty About Capabilities

Uncertainty About Intentions

Perception of Acute Threat

AI-FUELED SECURITY DILEMMA

A security dilemma exists when one state's investment in military capabilities prompts other states to increase their own investments because they perceive that the first state's actions make them less secure. Even if every state's investment is only for defensive purposes, uncertainty of other states about this intention can lead to increasing militarization and states' perception of increasing insecurity. This in turn increases the risk of violent conflict.

Two factors may be especially important in determining whether a security dilemma arises. The first is whether states perceive that the offense or the defense has the advantage. As Robert Jervis puts it, "[W]hen we say that the offense has the advantage, we simply mean that it is easier to destroy the others' [forces] and take its territory than it is to defend one's own."[xli] When a state believes this is the case, it is likely to conclude that it "cannot afford to wait until there is unambiguous evidence that the other is building new weapons because the war may be over before it can get arms to its forces."[xlii] This can fuel a security dilemma because each state believes that other states' weapons development poses a threat to its security.

A second factor is the ease of distinguishing development of offensive and defensive weapons. The ability to differentiate between the two allows non-aggressive states "to behave in ways that are clearly different from those of aggressors. In this situation, states can effectively signal their intentions by the type of weapons that they develop."[xliii] One state's investment therefore need not make other states feel more vulnerable, reducing the likelihood of a security dilemma.

The risks and unpredictability of AI-enabled weapon systems may naturally lead states to refrain from deploying them until there is assurance of their safety and reliability. As the discussion below describes, however, various features of AI may increase the likelihood of a security dilemma that creates incentives for states to develop and deploy AI enable weapons as soon as possible without a rigorous TEVV process to provide such assurance.[xliv]

Uncertainty about Capabilities

First, AI-enabled weapon systems will not be directly observable in the way that conventional weapons are. Whether a system is enabled by AI depends not upon its visible physical characteristics but the software that guides its operation. Two weapons that are identical in appearance therefore may have dramatically different capabilities, with differing roles for machines and human operators. This means that it is likely to be extremely difficult for one state to determine the AI enabled weapon capabilities of another. This opacity surrounding AI capabilities could lead states to adopt worst case assumptions about the threat posed by other states with AI-enabled weapons.

Second, the dynamic rate of AI innovation means that even if it were possible to make an assessment of a state's AI-enabled capabilities at one point, this assessment may soon be outdated. States are likely to stay abreast of AI research and development, and to seek continuously to incorporate new capabilities into their systems. This "uncertainty of measuring relative progress in AI research and its military applications" can make it difficult for states to have a stable understanding of the balance of power.[xlv]

Third, AI is not itself a weapon but a technology that can be put to a variety of uses. Horowitz & Kahn observe: "As an enabling technology with many discrete applications, the amorphous quality of AI exacerbates uncertainty over how its integration into existing platforms and doctrine will change the character of warfare."[xlvi] A state therefore faces a considerable challenge in attempting fully to comprehend all the ways in which other states may be incorporating AI into their military operations. This means that states' capabilities may be especially opaque to one another.

Fourth, at least in the near term, states have little experience with the use of AI-enabled weapons that could provide a shared understanding of their capabilities, limitations, and risks. This is in contrast to conventional weapon systems and, notably, to nuclear weapons. An understanding of the devastating impacts of nuclear weapons provided a basis during the Cold War for the United States the Soviet Union to take steps intended to reduce risk that such weapons would ever be used.

There is not yet comparable clarity about the capabilities and risks of AI-enabled weapons that could lessen the likelihood of a security dilemma.

All these features of AI-enabled weapon systems are likely to mean that it will be extremely difficult to distinguish between offensive and defensive weapons. States therefore may be likely to believe that other states' development of such weapons threaten their security and that they need to develop and deploy AI enabled systems as soon as possible, including those that are for offensive purposes – which in turn will trigger other states' sense of insecurity.

Uncertainty about Intentions

Certain features of AI also may make it especially difficult for states to discern one another's intentions with regard to using the AI-enabled capabilities that they have. First, the difficulty of distinguishing between offensive and defensive systems can make it difficult for a state to signal its benign intentions through its choice of the AI-enabled systems in which it invests.

Second, there may be significant limits to how transparent states are willing to be about their AI-enabled capabilities, because transparency would involve disclosure of highly sensitive software. This software, rather than the platform in which it is used, could be a significant source of competitive advantage that a state would not want to reveal. This creates challenges for any attempt to use inspection and verification as a way of reducing uncertainties and the risk of misperception about capabilities.

Perception of Acute Threat

Uncertainties about both capabilities and intentions with regard to AI enabled weapons could well lead states as a prudential matter to overestimate other states' capabilities and to assume that they have aggressive intentions. The nature of AI-enabled weapons may intensify this sense of threat because of the perceived decisive advantage of operating at machine speed compared to a "remotely controlled, 'slower' adversarial system."[xlvii]

A state may feel especially vulnerable because it fears that another state's use of such weapons against it would inflict such damage that it would prevent it from defending itself or retaliating. As Lieutenant General Jack Shanahan, the first Director of the Joint AI Center, declared: "What I don't want to see is a future where our potential adversaries have a fully AI-enabled force and we do not… I don't have the time luxury of hours or days to make decisions. It may be seconds and microseconds where A.I. can be used."[xlviii]

Under these circumstances, states are likely to believe that the balance of military capabilities favors the offense, which can make a preemptive strike seem advantageous. They therefore may hasten to develop and deploy offensive AI-enabled weapons in order to develop this capability. As Altmann & Sauer note, "Destabilisation becomes a particular concern when qualitatively new technologies promising clear military advantages seem close at hand."[xlix] If "the situation is seen as urgent. . . there are compelling incentives for accelerating the development of technology and incorporating it into militaries."[l]

Furthermore, to the extent that use of an AI enabled system involves less risk to military forces than a conventional weapon, a state's risk calculus may make a preemptive strike seem more appealing. This will be especially likely if a state fears that suffering an attack from such a system will result in its decisive defeat because of the potency of AI. In this way, the features of AI could make it more likely that the result of a security dilemma is the emergence of a conflict.

Finally, as Horowitz and Scharre suggest, a state's fear that AI-enabled weapons could quickly disable command and control capabilities could lead a state to develop weapons that automatically fire without human intervention upon warning of an impending attack.[li] Experience from the Cold War indicates the risk of such pre-programmed responses, when humans determined that an apparent warning of an imminent nuclear attack was the product of a technological failure, and prevented nuclear war by overriding the system. AI-enabled weapons therefore could place pressure on mechanisms that are designed to control the risk of escalation.

Conclusion

At its core, the security dilemma reflects states' difficulties in determining the extent to which other states pose a military threat to them. The greater the uncertainty about other states' capabilities and intentions, the greater risk of this dilemma. This risk may be exacerbated by indistinguishability between offensive and defensive weapons, and by the perception that the balance of power favors offensive use of military capabilities. The perceived decisive advantage of operations conducted at machine speed beyond the ability of humans to respond can make the fear of state insecurity especially acute, and foster the conviction that the balance of power favors the offense. The result may be that states believe it is necessary to deploy AI-enabled weapons as soon as possible without rigorous TEVV, and to substitute machines for humans as much as possible. Just development of AI-enabled weapons systems therefore requires that states engage in such development in ways that minimize the likelihood of these risks.

AVOIDING THE SECURITY DILEMMA

The Language of Military Preparedness

Confidence Building Measures

AVOIDING THE SECURITY DILEMMA

What might states do to minimize the risk that development of military AI systems will generate a security dilemma that could risk harmful deployment of such systems? One important step is to avoid using language likely to trigger a sense of insecurity on the part of other states. A second is to explore ways to reduce the uncertainties that give rise to the dilemma, and to adopt measures that can build trust among states about how such systems will be used.

The Language of Military Preparedness

On the one hand, a state needs to signal to other states that it has a military that will enable it to defend itself effectively if it is attacked, and win any conflict that occurs. On the other hand, it needs to communicate that it is not taking steps to give its military such a distinctive advantage that other states may feel so threatened that they disregard ethical constraints on the use of military force.

Regarding the development of AI-enabled weapons, state therefore should avoid characterizing its systems as providing it with an unprecedented decisive military advantage over other states. Given the potential of AI to conduct operations at machine speed, this may well trigger an intense sense of insecurity on the part of other states. This insecurity in turn could lead them to hasten development and deployment of AI systems without sufficient considered deliberation.

Language that can create the same risk is the public declaration that states are engaged in an "AI arms race." Unfortunately, there is no shortage of such language.[lii] "A 2019 survey of AI experts from technical and policy-oriented fields, for instance, indicated that an overwhelming majority of respondents predicted an AI arms race, however defined, in the next 15 years.[liii] A risk of framing the situation in this way is that it suggests that states need to invest in developing and deploying AI-enabled weapons as soon as possible if they want to be secure.[liv]

Another potentially problematic approach is to suggest that future warfare inevitably will be conducted primarily by machines. In this scenario, humans are relegated to the role of what General John Allen and Amir Husain describe as "providing broad, high-level inputs while machines do the planning, executing, and adapting to the reality of the mission and take on the burden of thousands of individual decisions with no additional input."[lv] Allen and Husain define this as "hyperwar," a fundamental transformation in warfare in which "human decision making is almost entirely absent from the observe-orient-decide-act (OODA) loop, as responses become "near instantaneous.""[lvi]

Describing this state of affairs as inevitable is likely to heighten states' sense that they need to develop as quickly as possible systems in which current human tasks are performed by machines. On this assumption, carefully deliberating about and identifying where such substitutions should take place in the system could risk leaving a state at a significant military disadvantage.

Just development of AI-enabled weapons thus requires that a state avoid language about such weapons that heightens states' insecurity. Aside from avoiding the use of such language, a state may take affirmative steps to reduce other states' fears about its development of AI enabled weapon systems. As the next section describes, one rubric under which such steps can be described is "confidence building measures."

Confidence-Building Measures

In the military context, confidence-building measures (CBM) are designed to reduce states' suspicion of one another through the exchange of information about capabilities and intentions, and establishment of some agreement on how military operations will be conducted.[lvii] Marie-France Desjardin's study of CBMs concludes that "[i]ncreasing transparency in military matters lies at the core of the confidence building approach. . . Secrecy breeds suspicions, and when states do not communicate, other is a lack of information about other states' military capabilities or activities, officials tend to make worst-case analyses."[lviii]

Some observers maintain that pursuing such measures may help reduce uncertainty about state capabilities and intentions that can fuel a security dilemma.[lix] Our discussion below draws on this literature in suggesting what measures might serve this purpose.

CBMs gained particular prominence during the Cold War as a way of reducing the likelihood that misinterpretation of capabilities and intentions could lead to nuclear war. One example was the Open Skies agreement, under which the United States and the Soviet Union agreed to permit aerial surveillance to establish their missile capabilities. Another is the Incidents at Sea Agreement that regulated the movement of US and Soviet naval vessels, and established means to communicate the presence of submarines and surface naval movements. A third example is the creation of a hotline for communication between top US and Soviet Union leadership after the Cuban Missile Crisis. In addition, the 1972 Anti-Ballistic Missile Treaty imposed certain limitations on nuclear weapon capabilities. Finally, NATO and the Soviet Union agreed to notify one another of military exercises above a certain threshold to reduce the risks of escalatory responses.

Desjardins cautions that CBMs can be difficult to negotiate and may not necessarily provide the benefits that parties seek. She says they typically have three main weaknesses: "the ambiguous nature of the level of obligation, the vague formulation of many stipulations and the absence of verification provisions."[lx] In addition to these caveats about CBMs in general, the conditions regarding AI-enabled weapons may not be completely comparable to those that provided incentives for CBMs during the Cold War.

First, both during the Cold War sides were keenly aware of the destructive power of nuclear weapons because of the bombs that had devastated Hiroshima and Nagasaki. This created a common interest in avoiding mutual annihilation. By contrast, the impacts of AI enabled weapons systems are unclear at this point because they have not been widely deployed, particularly for offensive purposes. States thus do not have a shared understanding of the exact nature of the risks that they pose. Second, weapons' capabilities during the Cold War generally were discernible from physical observation.

This made it possible to use surveillance to assess such capabilities, and to engage in verification of compliance with arms control agreements. By contrast, as described above, the capabilities of AI-enabled weapons are not readily observable, but are contained in software that is states may well be reluctant to disclose. As Michael Horowitz and his co-authors have noted, this opacity can give rise to a third challenge: it may be difficult to verify whether harm caused by an AI-enabled system was accidental or intentional.

Finally, states were the parties who engaged in the development of nuclear weapons during the Cold War, and they relied on highly centralized systems to control their use. By contrast, the developments in AI are generated to a significant degree in the private sector, the technology is widely available, and the uses to which it can be put are manifold. As Altmann & Sauer suggest, "While the development of AWS [automated weapon systems] clearly presents a challenge to less technologically advanced actors, obtaining AWS with some degree of military capability is a feasible goal for any country already developing, for example, remotely controlled armed UAVs."[lxi]

These features of AI-enabled weapons suggest caution in assuming that we can rely on approaches during the Cold War to reduce risks from the use of these systems. At the same time, Kenneth Payne argues, there are some general similarities that could provide at least some incentives to develop measures that are tailored to AI-enabled systems. First, he says, "nuclear weapons and AI are both highly technical scientific developments, requiring coordinated expertise."[lxii]

Second, Payne says that "the 'revolution' is concentrated in a few states, and the research involves a degree of secrecy which, coupled with the inherent technicalities, constrains public debate."[lxiii] The implication of this is that the widespread availability of algorithms does not automatically translate into the widespread possession of AI-enabled weapons. As Matthijis Maas argues:

"In practice, cutting-edge AI still requires very large (and rapidly increasing) amounts of computational power. Tacit knowledge possessed by experienced researchers proved a critical if underappreciated brake on the proliferation of nuclear weapons (and will likely play a similar role in limiting the rapid diffusion of military AI capabilities that actually offer strategically meaningful performance improvements (over humans; or against rival systems)."[lxiv]

Finally, advancements in technology depend not simply on the technology itself, but organizational structures and processes that are effectively designed to capitalize on it.[lxv] Maas argues, "All of this suggests that the set of leading state parties which must be brought in line with governance is not much larger for military AI, than it was for nuclear weapons."[lxvi]

With these considerations in mind, there may be some steps that a state could take to reduce the concerns that underlie the security dilemma, and thereby engage in ethically responsible development of AI-enabled weapon systems.

One step is for a state to publicly announce that it is committed to ensuring that deployment of these systems is consistent with ethical principles and legal requirements, and that there is assurance of their reliability and safety. The US Defense Innovation Board, for instance, has released AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense, which have been adopted by the Department (DoD). [lxvii] These signal to other states that the US military will develop and deploy AI systems only after careful review to ensure that they can be used ethically.

DoD has followed this with guidance on how to implement these principles across the Department through "the continuous identification, evaluation, and mitigation of risks, including risks from inaction or opportunity costs, across the entire product life cycle and well beyond deployment."[lxviii] Similarly, DoD is adapting both its TEVV and weapons review process to conduct assessments of AI-enabled systems. These steps will involve additional financial costs, as well as some delay in deploying systems. Commitment to them therefore can serve as a "costly signal" to other states that they will not be disadvantaged by likewise committing to use AI-enabled weapons only after rigorous review.

A second step could be to work to develop common definitions and shared understanding among states of core concepts that are relevant to the safety, reliability, impact, performance, and risks of AI enabled weapon systems. As Horowitz & Kahn suggest, "Knowledge about the material facts, that is, the technical and organizational features of military-relevant AI applications, is the foundation on which to build understandings about the risks themselves and the means to reduce them."[lxix] Differences in how concepts are described can reflect differences in how they are understood, which can increase the likelihood of misunderstandings that could lead to tensions. Common definitions can enable states to analyze the likely performance of systems under different conditions, and the ways in which their uses could inflict harm and create misunderstandings that could lead to escalation of conflict. As Andrew Imbrie & Elsa Kania observe, "The field of AI today is relatively globalized, but there appear to be some discrepancies emerging in technical and doctrinal concepts."[lxx]

A third measure could be to encourage information sharing and the development of communication channels among states. Some degree of transparency about TEVV, for instance, could involve public release of general information about the process for assessment of military AI-enabled systems without disclosing their specific technical features. This would be similar to the US approach to weapons review, which involves disclosing the process but not the review of particular weapons, in an effort to encourage other states to conduct reviews.

States might also share information on how to establish parameters that limit the domain in which a system can operate without human supervision, and how safely to shut it down if it begins to pose risks by operating beyond that domain. They might use certain joint projects to share information and conduct research that increased knowledge about issues concerns of common concern. During the Cold War, for instance, the Apollo-Soyuz Test Project involved collaboration between US and Soviet scientists and engineers for the first international human space flight.[lxxi]

There could be some risk to a state from sharing such information, since it may enhance the ability of adversaries to deploy effective and reliable systems that they could use to threaten the sharing state's security. As Imbrie an Kania put it, "On the one hand, collaboration in AI safety and security can reduce the risks of accident and strategic miscalculations among great powers. On the other hand, such collaboration may improve the reliability of machine learning techniques and therefore enable strategic competitors to deploy AI/ML-enabled military systems more quickly and effectively."[lxxii] A state therefore would need to decide how to weigh the security risk of an adversary's improved AI capabilities compared to the risk of an adversary and other states deploying unsafe and unreliable AI systems in ethically problematic ways. One way to address this could be encouraging ongoing exchanges among technical experts, members of the private sector, and academics, which could provide informal channels for sharing information that do not require official state involvement.

The measures described above could also help build confidence by serving as the impetus for a fourth step, which is establishing common norms and codes of conduct about the deployment and use of AI-enabled systems. In addition to the other purposes they might serve, the public commitment to ethical use of such systems and the willingness to share information described might, for instance, gradually generate a norm that states are expected to engage in ethical assessment and rigorous TEVV of military AI systems, and a weapons review of AI enabled weapon systems, before they deploy them. This could gradually generate a norm that states are expected to engage in rigorous pre-deployment review of AI-enabled weapons.

Over time, states might bolster these measures by taking a fifth step, which is providing for some degree of inspection and verification. As described above, there are particular challenges in subjecting AI-enabled weapons to such a regime. One measure to address this could be for states to share the general characteristics of an AI-enabled weapon without revealing all its training data or other components that they fear would compromise security. Another might be to permit outside parties to observe the operation of the system without disclosing its algorithms.

Horowitz and Scharre suggest that in this case "the outward behavior of the system would be observable, even if its code is not." As with information sharing in general, states would need to assess the relative security benefits of verification and of limiting other states' access to their systems.

Finally, states might work to develop "rules of the road" for the conduct of AI-enabled military operations and perhaps "red lines" that establish limits on their use. An International Autonomous Incidents Agreement, for instance, similar to the Incidents at Sea Agreement, could provide rules to govern and deconflict the interaction of military forces operating with a high degree of autonomy. States also could agree to declare some geographic areas off limits to autonomous systems because of their risk of unanticipated interactions.This could be to avoid unintended escalation in a contested region (e.g., a demilitarized zone), or because a region is near civilian objects (e.g., a commercial airliner flight path). Aside from territorial restrictions, states might also agree that AI-enabled systems not be used for crucial functions related to nuclear weapons.[lxxiii]

In conclusion, we have argued that rigorous state testing and review of AI-enabled weapon systems is a crucial condition for meeting jus ante bellum. In addition, states must engage in such development in a way that minimizes the risk of a security dilemma that could lead to incentives to deploy these systems without testing and review. States attempting to satisfy this requirement will need to balance competing ethical considerations. On the one hand, measured description of a state's capabilities, and transparency about those systems and standards for testing their safety and reliability, can mitigate the risk of creating a security dilemma. On the other hand, a state has an ethical responsibility to protect its population.It therefore will want to describe its capabilities in a way that discourages attack, and to limit how transparent it is about its AI-enabled systems. There is no formula to guide states in how best to navigate this tension. The important point, however, is that jus ante bellum insists that they attempt to do so in good faith. It requires, in other words, that they focus on "how to foster responsible competition" in AI-enabled systems.[lxxiv]

REFERENCES

[i] Harry van der Linden, "Just Military Preparedness: A New Category of Just War Theory," Paper Presented at the Department of Philosophy at Michigan State University (2010): 1-24, 7. https://digitalcommons.butler.edu/facsch_papers/1073.

[ii] Cecile Fabre, "War, Duties to Protect, and Military Abolitionism," *Ethics & International Affairs*, 35 no. 3 (2021), 395-406.

[iii] For a discussion of the current state of such efforts, see Vincent Boulanin & Maaike Verbruggen, "Mapping the Development of Autonomy in Weapon Systems," *Stockholm International Peace Research Institute* (2018), https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems.

[iv] Note about the term 'TEVV': Various sources have the 'VV' as validation and verification (Flournoy, et.al.) and others (NSCAI and DoD AI Strategy documents) as verification and validation. Here we use VV to mean verification and validation partly, because we see sources such as NSCAI as authoritative in the U.S. context. and partly because validation is the last step in the process in which machine learning models are built and tested. We thank Joe Chapa for this clarification.

[v] Charles Perrow. *Normal Accidents*; *Living with High-risk Technology*, Princeton University Press, 1984.

[vi] Scott D. Sagan et.al, "Leanring from Normal Accidents," *Organization and Environment*, 17:1, 2004: Matthijs Maas, "Regulating for Normal AI Accidents," Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.

[vii] Michael Horowitz, "When speed kills" in *Emerging Technologies and International Stability*, ed. Sechser, et.al., Routledge, 2022.

[viii] Michael Horowitz & Paul Scharre, "AI and International Stability," *CNAS*, 2021 https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures

[xix] Paul Scharre, *Army of None*, National Geographic Books, 2018.

[x] Horowitz & Scharre, 2021.

[xi] One formulation of the process, for instance, characterizes it in general terms as test and evaluation, defined as "a structured and recursive process that helps manage and/or reduce risk in defining, developing, acquiring, fielding, using, and supporting new capabilities." In the defense setting, the Department of Defense Instruction on Test and Evaluation (T&E) says, "The fundamental purpose of T&E is to enable the DoD to acquire systems that support the warfighter in accomplishing their mission. "A more detailed approach, however, differentiates between verification and validation. Verification seeks to ensure that the technology meets the specifications that a prospective user has provided, while validation assesses whether those specifications will meet the goals of the user. As Hand and Khan put it, "[V]alidation is sorting out that you are answering the right question, and verification is ensuring that you find the right answer to that question." One prominent report, for instance, reverses the last two steps and describes the process as testing, evaluation validation, and verification. Michèle A. Flournoy, Avril Haines, & Gabrielle Chefitz, "Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems," October 2020' https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf. Defense Acquisition University, Test and Evaluation. https://www.dau.edu/training/career-development/test-eval. DOD INSTRUCTION 5000.89 TEST AND EVALUATION §3.1(a). David J. Hand & Shakeel Khan, Validating and Verifying AI Systems. Patterns (N Y). 2020 Jun 12;1(3). https://pubmed.ncbi.nlm.nih.gov/33205105/.

[xii] Roff and Danks, "Trust but Verify; The Difficulty of Trusting Autonomous Weapons systems," *Journal of Military Ethics* 17(1). 2-20.

[xiii] Id. 7.

[xiv] Id. 10.

[xv] Id. 11.

[xvi] Flournoy, et.al., 3. "The ultimate goal of any TEVV system should be to build trust—with a commander who is responsible for deploying a system and an operator who will decide whether to delegate a task to such system—by providing relevant, easily understandable data to inform decision-making."

[xvii] Jane Pinelis, SERC talks: Progress in TEVV, https://www.youtube.com/watch?v=1eSKngsJvvo

[xviii] Roff and Danks, "Trust, but Verify."

[xix] DoD directive 3000.09, TEVV enclosure 2, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

[xx] Flournoy et.al. It might also need this to help define appropriate operational environments for setting a particular weapon in autonomous vs. semi-autonomous mode (e.g. Aegis system)
[xi] Wojton H., Porter D., Dennis J. (2021). Test and Evaluation of AI-Enabled and Autonomous Systems: A Literature Review. Alexandria, VA: Institute for Defense Analysis. Available online at: Wojton et al., 4.
[xxii] Pinelis, SERC presetation.
[xxiii] Flournoy, et.al. 7.
[xxiv] Id..
[xxv] This might be especially the case when teaching "common mistakes" is less useful and understanding the way that the system works is necessary for anticipating/recognizing when the system is failing to operate as expected.
[xxvi] Pinelis, SERC presentation.
[xxvii] Interview notes; Chief Responsible AI officer for Air Force,
[xxviii] Interview notes: TEVV director of CDAO.
[xxix] Flournoy, et.al.
[xxx] Vestner & Rossi, Legal Review of War Algorithms, USNWC, 547. https://digital-commons.usnwc.edu/ils/vol97/iss1/26/
[xxxi] Id.
[xxxii] Wojton, 20.
[xxxiii] Haugh, B., Sparrow, D., & Tate, D. (2018). The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems. 3-3.
[xxxiv] Id.
[xxxv] Id.
[xxxvi] Wojton, 20.
[xxxvii] Id.
[xxxviii] Pinelis SERC
[xxxix] Flournoy et. al.
[xl] Id.
[xli] Robert Jervis, "Cooperation under the Security Dilemma," World Politics 30, no. 2 (January 1978): 167–214.
[xlii] Id.
[xliii] Id.

[xliv]The discussion in this section draws in part on thoughtful analysis in Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability," Survival 59(5) (2017), 117–142; Kareem Ayoub & Kenneth Payne (2016). "Strategy in the Age of Artificial Intelligence", Journal of Strategic Studies, 39:5-6, 793-819; Ingvild Bode & Hendrik Huelss, "Autonomous weapons systems and changing norms in international relations, Review of International Studies, (2018), 44(3), 393–413; Carter Bowman, "Could AI-Proliferation Be The Next Nuclear Crisis?" (2020), Broad Street Humanities Review, 2, 1-15; Chatham House Report, M. L. Cummings, Heather M. Roff, Kenneth Cukier, Jacob Parakilas, and Hannah Bryce, (2018), Artificial Intelligence and International Affairs: Disruption Anticipated; Denise Garcia, " Lethal Artificial Intelligence and Change: The Future of International Peace and Security, International Studies Review (2018) 20, 334–341; Michael Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," Texas National Security Review: 1(3) (May 2018); Michael Horowitz, "Do Emerging Military Technologies Matter for International Politics?" Annu. Rev. Political Sci. 2020. 23:385–400; Michael C. Horowitz (2019) "When speed kills: Lethal autonomous weapon systems, deterrence and stability," Journal of Strategic Studies, 42:6, 764-788; "How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons," Contemporary Security Policy, 40:3, 285-311; Kenneth Payne, "Artificial Intelligence: A Revolution in Strategic Affairs?" Survival 60(5) (2018); Matthjis M. Mass (2019),

[1xlv Imbrie & Kania 10.

[xlvi] Id. 539.

[xlvii] Altmann & Sauer, 119.

[xlviii] "Lt. Gen. Jack Shanahan Media Briefing on A.I.-Related Initiatives within the Department of Defense." Department of Defense, 30 August 2019.

[xlix] Id. 121.

[l] Id. 12.

[li] Horowitz & Scharre.

[lii] See, e.g., Edward Moore Geist, "It's already too late to stop the AI arms race—We must manage it instead," *Bulletin of the Atomic Scientists* 72, no. 5 (2016): 318-321; Jean-Marc Rickli, Artificial Intelligence and the Future of Warfare, in *World Economic Forum*, The Global Risks Report 2017 (12th ed.)("An arms race in autonomous weapons systems is very likely in the near future"). https://www.weforum.org/reports/the-global-risks-report-2017.

[liii] Christine Carpenter and Casey Mahoney, "How Emerging Technologies Are Rewiring the Global Order: Fall 2019 Colloquium Report," University of Pennsylvania Perry World House (2019), https://global.upenn.edu/perryworldhouse/fall-2019-global-ordercolloquium-report-and-thought-pieces.

[liv] Paul Scharre, "Debunking the AI Arms Race Theory," Texas Nat. Sec. Rev. 4(3) (2021).

[lv] John Hussain Aim, "On Hyperwar," *U.S. Naval Institute Proceedings*, July 2017, Vol. 143, Issue 7.

[lvi] Allen & Husain.

[lvii] The discussion in this section draws on Marie-France Desjardin, Rethinking Confidence Building Measures (2014).

[lviii] Desjardin, 21.

[lix] See, e,g, Michael C. Horowitz, Lauren Kahn, & Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" Foreign Policy Research Institute (2020), 528-543; Michael C. Horowitz & Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures," Center for a New American Security (2021); Andrew Imbrie & Elsa Kania, "AI Safety, Security, and Stability Among Great Powers," Center for Security and Emerging Technology (2019).

[lx] Desjardins 38.

[lxi] Id. 125

[lxii] Payne, Kenneth. Artificial Intelligence: A Revolution in Strategic Affairs?, 15.

[lxiii] Id.

[lxiv] Maas, 290. Computing power: Amodei & Hernandez, 2018; Hwang, 2018. Tacit knowledge: MacKenzie & Spinardi (1995).Availability of algorithms deceptive: Ayoub & Payne (2016).

[lxv] Horowitz, 2018a, p. 4

[lxvi] Maas, 290.

[lxvii] Defense Innovation Board, AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense; US Department of Defense, DOD Adopts Ethical Principles for Artificial Intelligence, Feb. 24, 2020. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[lxviii] U.S. DEPARTMENT OF DEFENSE RESPONSIBLE ARTIFICIAL INTELLIGENCE STRATEGY AND IMPLEMENTATION PATHWAY 16 (June 2022).

[lxix] 541

[lxx] Imbrie & Kania, 8.

[lxxi] Imbrie and Kania

[lxxii] Imbrie & Kania 4.

[lxxiii] James Johnson (2020) "Deterrence in the age of artificial intelligence & autonomy: a paradigm shift in nuclear deterrence theory and practice?," Defense & Security Analysis, 36:4, 422-448; Matthijs M. Maas (2019)

[lxxiv] Roff, 97.

ABOUT THE AUTHORS

Jovana Davidovic is an Associate Professor in Philosophy at the University of Iowa; with a secondary appointment at the Law School and the Center for Human Rights. She is also a Senior Research Fellow at the Stockdale Center for Ethical Leadership at the United States Naval Academy and a Chief Ethics Office for BABL AI. Her academic work focuses on military ethics, international law and AI-enabled weapons systems. Her advisory work focuses on risk assessment for AI and automated decision systems.

Milton Regan is McDevitt Professor of Jurisprudence, and the Director of the Center on the Legal Profession, and Co-Director of the Center on National Security and the Law at Georgetown University Law Center, as well as Senior Fellow at the Stockdale Center for Ethical Leadership at the U.S. Naval Academy. His work focuses on international law, national security, international human rights, and legal and military ethics. Mitt also serves on the Institute for Defense Analyses committee that is advisory to the Defense Advanced Research Projects Agency (DARPA) on incorporating assessment of legal, moral, and ethical considerations into research projects for the agency.