# AI Ethics

## Aff

### Extend – Inherency - DOD

#### The DOD has not committed enough resources to TEVV, despite support for AI

**Flournoy, Haines et al 2020 - former Under Secretary of Defense for Policy and Director of National Intelligence** [Michèle and Avril, Gabrielle Cheftiz Special Assistant to the Under Secretary of Defense for Policy October, “Building Trust through Testing Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems” https://cset. georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf Acc 6/23/22 JZ]

Machine learning applications can increase the speed and quality of human decision-making on the battlefield, enable human-machine teaming to maximize performance and minimize the risk to soldiers, and greatly improve the accuracy and speed of analysis that relies on very large data sets. ML can also strengthen the United States’ ability to defend its networks against cyberattacks at machine speeds and has the power to automate critical components of labor-intensive enterprise functions, such as predictive maintenance and personnel management. Advances in AI and machine learning are not the sole province of the United States, however. Indeed, U.S. global leadership in AI remains in doubt in the face of an aggressive Chinese challenge in the field. Numerous DOD and academic reports reflect on the need to invest more in AI T 2 research and development, train and recruit a skilled workforce, and promote an international environment supportive of American AI innovation—all while promoting safety, security, privacy, and ethical development and use. However, far too little attention is placed on the issue of trust, and especially testing, evaluation, verification, and validation (TEVV) of these systems. Building a robust testing and evaluation ecosystem is a critical component of harnessing this technology responsibly, reliably, and urgently. Failure to do so will mean falling behind.

### Extend – Harms - Interoperability

#### Diverse AI ethics undermine military interoperability because it prevents different nations militaries from cooperating. Coordinated standards improve cohesion and interoperability

van der Merwe, 2021 - Fellow Center for European Policy Analysis Defense Tech Initiative [Joanna, Feb 17, “NATO Leadership on Ethical AI is Key to Future Interoperability” https://cepa.org/nato-leadership-on-ethical-ai-is-key-to-future-interoperability/ BK]

If individual nations or groups are left to develop their own ethical principles without wider alignment to NATO, the result will be a number of AI-based systems with varying technical specifications based on the legal and policy decisions made by individual governments when answering the key questions. As has been demonstrated in areas such as facial recognition and policing algorithms, the assumptions made by those developing the tools and answering the key questions have a significant impact on the real-world functioning of the tool and societal acceptance of its ethics. The risk of tools failing to gain acceptance depends on the legal and ethical decisions made by governments. For the military, this may mean one state using an AI-based system that is seen as unacceptable by another, and in a joint operation one state fielding a system that cannot be used by another. Or worse yet, this could render a joint operation impossible. Without the ability to interoperate across NATO, the inability to effectively and efficiently respond to future threats would undermine the Alliance. The role of the private sector is another aspect of ethical AI development that has proved a challenge to governments and the transatlantic relationship. Within states, governments have struggled to adequately regulate Big Tech firms, which has led to these companies encroaching on government responsibilities to protect and uphold the public interest. This encroachment permeates all aspects of government, including defense and security. As Deputy Secretary of Defense Kathleen Hicks discussed during her confirmation hearings, the lack of competition is also a challenge to innovation in the private defense industry. This, along with a lack of regulation, feeds into the power imbalance between the sectors. Consequently, private sector companies building the AI and AI systems that are or will be deployed on the battlefield are deciding the ethics policies for themselves. The transatlantic partnership must focus on coordinating these core principles and systematic governance to ensure AI systems development aligns with the rule of law and democracy. In particular, this must ensure answering questions about human dignity, human control, and accountability. NATO is the ideal defense and security forum for this alignment. Given the US lead on adopting ethical principles for the entire DoD and the EU’s drive to assert checks and balances for private-sector tech companies, NATO remains the organization that can bring these two together and establishes the ethical bottom line. These will then ensure the diverging legal and ethical stances towards Big Tech do not lead to an interoperability barrier in the future. If developments surrounding the General Data Protection Regulation (GDPR) and the challenges it brought for U.S.-based, data-driven companies are any indication, a strong transatlantic led initiative is needed in order to ensure the same challenges do not hinder NATO. The solution to the challenge that ethical AI poses for the future of interoperability within NATO is for the Alliance to establish shared transatlantic ethical principles, informed by the US DoD, the EU, and others. Establishing these principles will not only strengthen transatlantic political relations; more technically, it will allow for the establishment of standardization agreements and inform training and education initiatives of the Alliance in the future.

### Extend – Harms - Instability

#### AI undermines human control of warfare – it speeds up war beyond our abilities to keep up.

Sanderson, 2019 - The Times arts correspondent [David, “Humans may lose control of war with AI weapons” The times. [https://go.gale.com/ps/i.do?p=GIC&u=gotitans&id=GALE|A601230754&v=2.1&it=r&sid=summon](https://go.gale.com/ps/i.do?p=GIC&u=gotitans&id=GALE%7CA601230754&v=2.1&it=r&sid=summon) LMSi]

The head of the armed forces has said that artificial intelligence weaponry is evolving so quickly that in the future humans may have limited control over how battles are fought. General Sir Nick Carter, chief of the defence staff, said that "technologies are beginning to question the ethical and moral basis on which we apply the rule of conflict", Speaking at the Cliveden Literary Festival, he said: "There's a debate we need to have about what does the future of warfare look like when you end up with a combination of manned, unmanned and autonomous technologies fighting that war for you ... We may not have the same control as we had in the past." At the same event, David Petraeus, the former head of the CIA, said that advances in AI would make it difficult to "keep the person in the loop". Sir Nick said that space and cyberspace were the new battlefields and that the advance of cyberwarfare meant that we were constantly "at war, just not in the way we defined it in the past". He also warned against thinking that the battle against Islamic State had been won. "We should all realise we are a very long way from dealing with that," he said. "People talked rather too early about beating the phenomenon that was Daesh in Syria and Iraq. "We now have Isis in west Africa, in South Africa, we will have Isis in east Africa quite soon, we have it in China, we have it in the Philippines and, of course, in Afghanistan."

### Extend – Harms - Dignity

#### AI weapons violate human dignity – they are not accountable and treat people as objects.

Amoroso and Tamburrini, 2021 - Prof of International Law at the University of Cagliari and Prof of Philosophy of Science and Technology at the University of Naples Federico [Daniele and Guglielmo, Feb Italian Journal of International Affairs “In Search of the ‘Human Element’: International Debates on Regulating Autonomous Weapons Systems” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/03932729.2020.1864995> TM]

Second, the Report contends that AWS are likely to cause an accountability gap, especially in connection with war crimes (Heyns 2013, paras. 75-81). One cannot exclude that AWS might make targeting decisions that, were they taken by human agents, would trigger individual criminal responsibility (Crootof 2016). Authors who oppose an AWS ban observe that, in such cases, the deploying officers would be held legally responsible under the doctrine of superior responsibility (see, for example, Margulies 2017). The rejoinder, however, is that deploying officers may cast an effective defence against criminal prosecution in terms of AWS complexities and their unexpected battlefield behaviours. The difficulty in predicting what these systems will actually do depends on both internal factors – their interacting components and functional modules – and external factors (for example, hostile battleground interactions with other artificial agents). Therefore, there may be cases in which the presence of that mental element required under international criminal law (ICL) to ascribe individual responsibility cannot be ascertained. Consequently, no one would be held criminally liable, even though the conduct in question would materially amount to an international crime (Liu 2016). This is hardly reconcilable with the accountability requirement for actions of military commanders and operators, as well as with the related principle of individual criminal responsibility under ICL. Third, it is maintained that AWS violate the principle of protection of human dignity, in that the latter dictates that decisions affecting the life and physical integrity of human beings involved in armed conflicts should be reserved for human operators (Heyns 2013, paras. 89-97). From this point of view, autonomous targeting is unacceptable because it “objectifies” human beings, reducing them to algorithmically processed “data points”, thereby systematically denying their inherent value as human beings (Moyes 2019, 6). Suppressing human life is ethically and legally justifiable only if it is based on human judgement, for only human decision-making offers a guarantee that the values at stake (human life, physical integrity and so on) can be fully appreciated (Heyns 2016).

### Extend - Harms – Authoritarian AI

#### Ethical values for AI are crucial to prevent Authoritarian abuse – AI can be misused for propaganda or government surveillance

Scharre, 2019 - Vice President and Director of Studies at CNAS [Paul, May 2019, Foreign affairs, “Killer apps: The real dangers of an AI Arms race”, https://www.foreignaffairs.com/articles/2019-04-16/killer-apps, acc 6/20/22, M.A.]

Harm from AI misuse isn’t hypothetical; it’s already here. Bots are regularly used to manipulate social media, amplifying some messages and suppressing others. Deepfakes, AI-generated fake videos, have been used in so-called revenge porn attacks, in which a person’s face is digitally grafted onto the body of a pornographic actor. These examples are only the start. Political campaigns will use AI-powered data analytics to target individuals with political propaganda tailored just for them. Companies will use the same analytics to design manipulative advertising. Digital thieves will use AI tools to create more effective phishing attacks. Bots will be able to convincingly impersonate humans online and over the phone by cloning a person’s voice with just a minute of audio. Any interaction that isn’t in person will become suspect. Security specialists have shown that it’s possible to hack into autonomous cars, disabling the steering and brakes. Just one person could conceivably hijack an entire fleet of vehicles with a few keystrokes, creating a traffic jam or launching a terrorist attack. AI’s power as a tool of repression is even more frightening. Authoritarian governments could use deepfakes to discredit dissidents, facial recognition to enable round-the-clock mass surveillance, and predictive analytics to identify potential troublemakers. China has already started down the road toward digital authoritarianism. It has begun a [massive repression campaign](https://www.hrw.org/report/2018/09/09/eradicating-ideological-viruses/chinas-campaign-repression-against-xinjiangs#;) against the Muslim Uighur population in Xinjiang Province. Many of the tools the government is using there are low tech, but it has also begun to use data analytics, facial recognition systems, and [predictive policing](https://www.wsj.com/articles/china-said-to-deploy-big-data-for-predictive-policing-in-xinjiang-1519719096) (the use of data to predict criminal activity). Vast networks of surveillance cameras are [linked](https://www.wsj.com/articles/the-all-seeing-surveillance-state-feared-in-the-west-is-a-reality-in-china-1498493020?mod=article_inline) up to algorithms that can [detect](https://www.wsj.com/articles/twelve-days-in-xinjiang-how-chinas-surveillance-state-overwhelms-daily-life-1513700355?mod=article_inline) anomalous public behavior, from improperly parked vehicles to people running where they are not allowed. The Chinese company Yuntian Lifei Technology [boasts](https://chinadigitaltimes.net/2018/10/company-touts-ai-powered-facial-recognition/) that its intelligent video surveillance system has been deployed in nearly 80 Chinese cities and has identified some 6,000 incidents related to “social governance.” Some of the ways in which Chinese authorities now use AI seem trivial, such as tracking how much toilet paper people use in public restrooms. Their proposed future uses are more sinister, such as monitoring patterns of electricity use for signs of suspicious activity.

#### AI Harms Disadvantaged Groups because “neutral” algorithms and big data reinforce existing discrimination

**Artificial Intelligence/Machine Learning Risk & Security Working Group, 2020** [Wharton, ‘Artificial Intelligence Risk & Governance’, <https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/>]

4. Interpretability and Discrimination Interpretability (presenting the AI system’s results in human understandable format), and discrimination (unfairly biased outcomes) are crucial concepts that factor into the risks associated with AI/ML systems used for certain use cases. In this section, we explore potential risks associated with discrimination and interpretability as they relate to certain applications of AI, e.g., loan approvals. 4.1 Discrimination in AI Depending on the use case, AI may potentially lead to discriminatory and/or unfairly biased outcomes if not implemented appropriately. Poor implementation may arise from biased data, the AI system itself not being properly trained or when there are alternate systems and data sources that could potentially be used to generate better outcomes for disadvantaged groups. Ultimately, the use of an AI system which may cause potentially unfair biased outcomes may lead to regulatory non-compliance issues, potential lawsuits and reputational risk. That said, these risks could be managed. There is even growing evidence that AI/ML systems could be harnessed to more effectively control for discriminatory outcomes. Existing Legal and Regulatory Frameworks Federal and state statutes prohibit discrimination in areas that impact our daily lives, including employment, housing, and lending, to name a few. By way of example, a potential impact in the use of AI for lending is described in greater detail below. The primary U.S. federal statutes that define illegal discrimination in lending are the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA); however, lenders are subject to many other federal regulations and state laws addressing fairness. Each statute defines types of “protected classes,” such as gender, race, or ethnicity, that a lender cannot legally disfavor. Generally speaking, three types of discrimination are recognized by federal banking regulators: overt discrimination, disparate treatment, and disparate impact when not supported by a legitimate business justification. Disparate treatment discrimination could occur when similarly situated individuals are treated differently based on a prohibited basis, but the treatment does not have to be motivated by prejudice or an intent to discriminate. In an AI context, this may potentially occur, for example, when a firm explicitly uses protected class status in an AI system used to underwrite creditworthiness. Disparate impact, on the other hand, occurs when a system includes features that lead to disproportionately unfavorable outcomes for a protected class. Importantly, evidence of disparate impact is almost always assessed independently of the accuracy and validity of the system. In other words, just because a given system is statistically sound does not mean that it is legally non-discriminatory. Such systems are generally not considered legally discriminatory if they and their constituent features could be demonstrated to meet a legitimate business need and where no less discriminatory alternative system or process could be identified that also meets those needs. Concerns over using and potentially amplifying implicitly biased data also arise in other contexts. For instance, the New York Department of Financial Security (NY DFS) discussed [7] the use of external consumer data and information sources in insurance underwriting, noting the potential of leveraging these sources to help establish lifestyle indicators that may inform the review of an application for insurance coverage. In doing so, however, NY DFS observed that such data may be inaccurate or unreliable, and its use may result in a significant detrimental impact to the insured. Similarly, in a speech [8] by Charles Randell, Chair of the UK Financial Conduct Authority, concerns over misuse of big data to inform potentially detrimental outcomes were raised, with a real-world example in the use of data mining credit card charges for services such as marriage counseling, and reducing cardholders’ credit limits on the basis of the correlation between marriage breakdown and debt default. The use and potential for misuse of big data is no longer a theoretical concern and should be considered in determining the types of data that may be used in developing AI/ML systems. We reference these legal and regulatory considerations to illustrate existing standards that already apply to many algorithmic activities of financial institutions, especially as they relate to unfairly biased outcomes. Data as a Cause of Discriminatory AI A host of factors may result in AI-related illegal discrimination. Input data may cause illegal discrimination if it identifies or closely proxies class membership, if it causes protected class members to experience less favorable outcomes, or if it is differentially predictive of the outcome for the protected class. Traditional data inputs, such as many credit bureau attributes, tend to be less likely to raise disparate impact concerns because they are generally thoroughly vetted and accepted for credit worthiness. They may also be differentially predictive if the system’s weights or coefficients do not properly account for class-specific idiosyncrasies. Non-traditional data, such as utility payment history, rental payments, or a person’s digital footprint (including social media posting), may generate heightened concerns relative to traditional data. From a fairness perspective, such data may have substantial merit, as its use has been shown to expand access to the financial system for unbanked or underserved populations that are often more likely to be members of some protected groups. However, such data use often raises coverage and accuracy concerns. Algorithms as a Cause of Discriminatory AI Algorithms themselves may result in discriminatory outcomes exacerbated by their complexity and opacity. Some of this concern arises from the fact that some machine learning algorithms create variable interactions and non-linear relationships that are too complex for humans to identify and review. These relationships have the potential to cause disparate treatment discrimination by creating proxies for protected class status. To some degree, these concerns have been lessened by advances in explainable AI techniques that allow additional insight into these complex relationships, which we address in Subsection 4.2 below. System misspecification may also cause discriminatory outcomes. Here, features may be independently predictive of both the outcome and protected class status, but the class effect is incorporated into the prediction. For example, suppose a credit system included whether a person tended to shop at a discount store. It is likely that such a variable would capture a measure of wealth, which may be a reasonable predictor of repayment, but may also unintentionally capture a race effect. In addition, if the store is more likely to be located in minority neighborhoods, then the system may further exacerbate this effect. That is, the variable may act as a proxy for the neighborhood, which in turn acts as a proxy for race. Importantly, this is not a problem that is unique to AI. In fact, to the extent that machine learning is more accurate than traditional methods, it may be more likely to identify such a relationship and remove the non-predictive race effect.

### Extend – Solvency – Human Control

#### Human control is necessary, because judgement is essential for complex military decisions.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

There are three types of machine learning algorithms.[45](javascript:;) All require human judgment. First, in “supervised learning,” the human tells the machine what to predict. Second, “unsupervised learning” requires judgment about what to classify and what to do with the classifications. Third, “reinforcement learning” requires advance specification of a reward function. A reward function assigns a numerical score to the perceived state of the world to enable a machine to maximize a goal. More complicated strategies may combine these approaches by establishing instrumental goals in pursuit of the main objective. In every case, a human ultimately codes the algorithm and defines the payoffs for the machine.

In economic terms, judgment is the specification of the utility function.[46](javascript:;) The preferences and valuations that determine utility are distinct from the strategies that maximize it. To take a trivial example, people who do not mind getting wet and dislike carrying umbrellas will not carry one, regardless of the weather forecast. People who dislike getting wet and do not mind carrying umbrellas might always have an umbrella in their bag. Others might carry an umbrella if the chance of rain is 75 percent but not if it is 25 percent. The prediction of rain is independent of preferences about getting wet or being prepared to get wet. Similarly, the AI variation on the notorious “trolley problem” poses an ethical dilemma about life-or-death choices. For example, should a self-driving car swerve to avoid running over four children at the risk of killing its human passenger? If the AI predicts even odds that someone will die either way, the car should swerve if all lives are equally valuable, but it should not swerve if the passenger's life is worth at least four times as much as that of a random child. This somewhat contrived dilemma understates the complexity of the judgment involved. Indeed, the ethical dilemmas of AI reinvigorate longstanding critiques of utilitarian reasoning. As Heather Roff points out, “We cannot speak about ethical AI because all AI is based on empirical observations; we cannot get an ‘ought’ from an ‘is.’ If we are clear eyed about how we build, design, and deploy AI, we will conclude that all of the normative questions surrounding its development and deployment are those that humans have posed for millennia.”[47](javascript:;) If the trolley problem seems far-fetched, consider the case of a self-driving Uber car that killed a cyclist in Tempe, Arizona.[48](javascript:;) The AI had predicted a low but nonzero probability that a human was in its path. The car was designed with a threshold for ignoring low-probability risks. The priority of not hitting humans was obvious enough. Yet, with an error tolerance set to zero, the car would not be able to drive. The question of where to set the tolerance was a judgment call. In this case, it appears that the prespecified judgment was tragically inappropriate for the context, but the prediction machine had absolutely no concept of what was at stake. A well-specified AI utility function has two characteristics. First, goals are clearly defined in advance. If designers cannot formally specify payoffs and priorities for all situations, then each prediction will require a customized judgment. This is often the case in medical applications.[49](javascript:;) When there are many possible situations, human judgment is often needed upon seeing the diagnosis. The judgment cannot be determined in advance because it would take too much time to specify all possible contingencies. Such dynamic or nuanced situations require, in effect, incomplete contracts that leave out complex, situation-specific details to be negotiated later.[50](javascript:;) Because all situations cannot be stipulated in advance, judgment is needed after seeing the prediction to interpret the spirit of the agreement. The military version of incomplete contracting is “mission command,” which specifies the military objective and rules of engagement but empowers local personnel to interpret guidance, coordinate support, and tailor operations as the situation develops.[51](javascript:;) The opposite of mission command, sometimes described as “task orders,” is more like a complete contract that tells a unit exactly what to do and how to do it. Standard operating procedures, doctrinal templates, and explicit protocols help to improve the predictability of operations by detailing instructions for operations and equipment handling. In turbulent environments with unpredictable adversaries, however, standardized task orders may be inappropriate. The greater the potential for uncertainty and accident in military operations, the greater the need for local commanders to exercise initiative and discretion. In Clausewitzian terms, “fog” on the battlefield and “friction” in the organization require commanders to exercise “genius,” which is “a power of judgment raised to a marvelous pitch of vision, which easily grasps and dismisses a thousand remote possibilities which an ordinary mind would labor to identify and wear itself out in so doing.”[52](javascript:;) The role of “genius” in mission command becomes particularly important, and particularly challenging, in modern combined arms warfare and multi-domain operations.[53](javascript:;) When all possible combinations of factors cannot possibly be specified in advance, personnel have to exercise creativity and initiative in the field. Modern military operations tend to mix elements of both styles by giving local commanders latitude in how they interpret, implement, and combine the tools, tactics, and procedures that have been standardized, institutionalized, and exercised in advance.

#### Increasing AI in the military will make human control More important, not less

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

Recent scholarship on artificial intelligence (AI) and international security focuses on the political and ethical consequences of replacing human warriors with machines. Yet AI is not a simple substitute for human decision-making. The advances in commercial machine learning that are reducing the costs of statistical prediction are simultaneously increasing the value of data (which enable prediction) and judgment (which determines why prediction matters). But these key complements—quality data and clear judgment—may not be present, or present to the same degree, in the uncertain and conflictual business of war. This has two important strategic implications. First, military organizations that adopt AI will tend to become more complex to accommodate the challenges of data and judgment across a variety of decision-making tasks. Second, data and judgment will tend to become attractive targets in strategic competition. As a result, conflicts involving AI complements are likely to unfold very differently than visions of AI substitution would suggest. Rather than rapid robotic wars and decisive shifts in military power, AI-enabled conflict will likely involve significant uncertainty, organizational friction, and chronic controversy. Greater military reliance on AI will therefore make the human element in war even more important, not less.

#### AI increases the importance of Human control, not technocracy – Human emotion and judgement become more essential.

Goldfarb and Lindsay, 2022 – Chair in AI and health care at the Univ of Toronto, Prof of Cybersecurity at the Georgia Tech [Avi, Jon, 2/25/22, <https://doi.org/10.1162/isec_a_00425>, “Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War” 6/18/22, LND]

We have defined judgment narrowly in economic terms as the specification of the utility function. The rich concept of judgment, however, deserves further analysis. Just as decision-making can be disaggregated into its components, judgment might also be disaggregated into the intellectual, emotional, and moral capacities that people need to determine what matters and why. Military judgment encompasses not only the Clausewitzian traits of courage, determination, and coup d'oeil, but also a capacity for fairness, empathy, and other elusive qualities. Some wartime situations merit ruthlessness, deviousness, and enmity, while others call for mercy, candor, and compassion. To these character traits must be added the engineering virtues of curiosity, creativity, and elegance insofar as personnel will have to reconfigure AI systems in the field. We expect that the general logic of complementarity will still apply at this more fine-grained level. Any future AI that is able to automate some aspects of judgment, therefore, will make other aspects even more valuable. Furthermore, the rich phenomenology of judgment, which AI makes more valuable, has important implications for professional military education. More technology should not mean more technocracy. On the contrary, personnel would be wise to engage more with the humanities and reflect on human virtues as militaries become more dependent on AI. In general, reliance on AI will tend to amplify the importance of human leadership and the moral aspects of war. In the end, we expect that more intensive human-machine teaming will result in judgment becoming more widely distributed in military organizations, while strategic competition will become more politically fraught. Whatever the future of automated warfare holds, humans will be a vital part of it.

#### AI systems need human control – only human correction can solve for unanticipated failures

Scharre, 2016 -- Vice President and Director of Studies at CNAS [Paul, Feb 2016, Center for New American Security, “Autonomous weapons and operational risk”, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS\_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515&focal=none, Acc 6/21/22, M. A.]

Maintaining effective control over autonomous systems Human operators will want to ensure that autonomous systems perform in a way consistent with their intentions. Because autonomous systems in many cases do not have real-time human supervision, maintaining effective control over the system has two components: 1. The ability of the human operator to accurately predict the autonomous system’s behavior in the environment in which it is being used. This includes its limitations and the conditions under which it will fail. This allows the human operator to employ the autonomous system only in situations where it will perform appropriately. 2. The ability of the human operator to undertake corrective action if/when the autonomous system fails to behave in accordance with the human operator’s intentions. A failure with an autonomous system is a loss of effective control—a situation in which the autonomous system no longer is behaving in accordance with human operator intention. Risk, in this context, refers to the risk of failure, both the probability and consequences of a loss of control: • The probability of failure is the likelihood of the system behaving in a manner inconsistent with human operator intentions in a particular environment. • The consequence of failure is the potential damage the autonomous system could do in that environment until such time as the human operator can undertake corrective action to bring the system back in line with human operator intentions or the autonomous system ceases operation. Autonomous systems can vary in the type of task they perform, their level of complexity, and degree of the human operator’s interaction with the system. As these aspects of the system change, the risks of employing an autonomous system change as well. The inherent hazard of a system depends on the task being performed Autonomous systems can perform a wide variety of tasks, from driving cars to regulating temperature or making toast. The inherent hazard of a system is the potential consequence if the autonomous system performs that task incorrectly. This depends on both the task being performed and its operating environment. The consequences of a failure with an autonomous car are far more potentially severe than a toaster failing to properly cook bread. The environment in which the system is operating is also a key component of the inherent hazard of the system. The hazard associated with an autonomous car driving on a closed-circuit track is much less severe than one driving through crowded city streets with pedestrians.

#### Increasing training and refining AI issues to incorporate more human controls decreases errors and increases trust – CBMs work

Atherton 2022 – Military Technology Journalist [Kelsey, 5/6/22, “Understanding the errors introduced by military AI applications”, <https://www.brookings.edu/techstream/understanding-the-errors-introduced-by-military-ai-applications/>, 6/18/22, LND]

Finding the right mix of trust between an autonomous machine and the human relying on it is a delicate balance, especially given the inevitability of error. Seventeen years after the Tornado shootdown, the automated features of the Patriot missile remain in place, but the way in which they are used has shifted. Air threats, such as aircraft, helicopters, and cruise missiles can now only be engaged in manual mode “to reduce the risk of fratricide,” as the U.S. Army’s manual for air and missile defense [outlines](https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN31339-FM_3-01-000-WEB-1.pdf). In manual mode, automated systems still detect and track targets, but it’s a human who makes the call about when and if to fire. But “for ballistic missiles and anti-radiation missiles,” like the kind the Patriot in Iraq assumed the Tornado was, “the operator has a choice of engaging in the automatic or manual mode,” though the manual notes that these “engagements are typically conducted in the automatic mode.” Defense researchers caution that human beings are not well-suited to monitoring autonomous systems in this way. “Problems can arise when the automated control system has been developed because it presumably can do the job better than a human operator, but the operator is left in to ‘monitor’ that the automated system is performing correctly and intervene when it is not,” the engineering psychologist John Hawley, who was involved in the U.S. Army’s efforts to study the 2003 friendly fire incidents, [wrote](https://www.cnas.org/publications/reports/patriot-wars) in a 2017 report. “Humans are very poor at meeting the monitoring and intervention demands imposed by supervisory control.”

### Extend – Solvency - Instability

#### Human control reduces the risk of AI accidents because they open channels of communication

Effoduh, 2021 – PhD candidate at Osgoode Hall Law School [Jake Okechukwu June 23 World Economic Forum, “Weapons powered by artificial intelligence pose a frontier risk and need to be regulated” <https://www.weforum.org/agenda/2021/06/the-accelerating-development-of-weapons-powered-by-artificial-risk-is-a-risk-to-humanity/#:~:text=are%20frontier%20risks%3F-,Frontier%20risks%20are%20low%2Dlikelihood%2C%20high%2Dimpact%20threats%20that,world%20as%20we%20know%20it>. BK]

Pathways to avert risks from L.A.W.S. While L.A.W.S. are another achievement for military intervention, it may be much safer to employ a “human-on-the-loop” approach in which people retain control over the weapons in a supervisory role. L.A.W.S. may not pose an apparent catastrophic risk like say COVID-19, cyberattacks or climate change, but we need to apply fresh perspectives to thinking about all these risks and how to avert them. We need more transparent channels for risk communication; to invest in institutions that oversee AI risk management and enact policies that check the speed at which military tech is developing. The United Nations can proscribe certain risk levels of military deployment and incentivize those whose deployment are compliant with IHL (e.g., like the Forum’s collaboration with the Smart Toy Awards to reward robot manufacturers that prioritize ethics in their creations). The focus should be less on perfecting the art of war and instead on learning to curbing , through technology, from the pervasive proclivity for waging war. In the end, this would be a win for IHL, scientific ethics and the human civilization.

#### Human control is key to prevent accidents with autonomous weapons – Empirical examples prove

Atherton 2022 – Military Technology Journalist [Kelsey, 5/6/22, “Understanding the errors introduced by military AI applications”, <https://www.brookings.edu/techstream/understanding-the-errors-introduced-by-military-ai-applications/>, 6/18/22, LND]

Automation is a compelling feature for an anti-air and, especially, for an anti-missile system. The calculations involved in shooting down aircraft and missiles are hard and require immediate translation of sensor information. Both interceptors and targets are traveling exceptionally fast. It’s the kind of task in which the involvement of a human introduces lag, slows down the process, and makes it less likely a missile is going to successfully shoot down an incoming projectile or aircraft. But human operators also serve an essential role: preventing accidental, incorrect shootdowns. And this requires a balance between human and machine decisionmaking that is difficult to achieve. When the Pentagon investigated the causes of the Tornado shootdown, as well as [two](https://www.latimes.com/archives/la-xpm-2003-apr-21-war-patriot21-story.html) [other](https://medium.com/war-is-boring/that-time-an-air-force-f-16-and-an-army-missile-battery-fought-each-other-bb89d7d03b7d) incidents of friendly fire involving Patriot systems, the missile system’s automated functions were identified as contributing factors in misidentifying friend as foe. U.S. Patriot batteries deployed to Iraq under the assumption that they would face heavy missile attacks, which would require the batteries to operate with a relative degree of autonomy in order to respond with sufficient speed. As a 2005 [report](https://dsb.cto.mil/reports/2000s/ADA435837.pdf) by the Defense Science Board Task Force on the Patriot system’s performance observed, operating autonomously required U.S. forces to trust that the automated features of the system were functioning properly. So when the assumptions underlying the decision to allow the Patriot system to autonomously identify and sometimes fire on targets no longer applied, the soldiers operating the system were not in a position to question what the weapon’s sensors were telling them. Had U.S. and coalition forces faced heavy missile attacks in the war, automating such defenses would have made more sense. Instead, U.S. and allied forces quickly established air superiority, enough to drastically shift the balance of what was in the sky. Instead of facing large amounts of incoming missiles, Patriot batteries were observing large numbers of allied planes operating in the sky above them and sometimes struggling to identify friend from foe. According to the Defense Science Board’s task force, the first 30 days of combat in Iraq saw nine ballistic missile attacks that Patriot batteries might have been expected to counter, compared to 41,000 aircraft sorties, amounting to a “4,000-to-1 friendly-to-enemy ratio.” Picking out the correct targets against the background of a large number of potential false positives proved highly challenging. In the case of the Tornado shootdown, automation—and the speed with which automated action was taken—was likely sufficient on its own to cause the tragedy, but it might have been prevented if other systems hadn’t failed. As the UK Ministry of Defence concluded in its [report](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/82817/maas03_02_tornado_zg710_22mar03.pdf) examining the incident, the battery culpable for the shootdown was without its communications suite, which was still in transit from the United States. Contact with battalion headquarters occurred through a radio relay with another battery equipped with voice and data links to headquarters. “The lack of communications equipment meant that the Patriot crew did not have access to the widest possible ‘picture’ of the airspace around them to build situational awareness,” the report found. Another system that failed and that might have prevented the shootdown was the identification-as-friend-or-foe system, a safety measure designed to avoid such deadly mistakes. That kind of information, transmitted securely and immediately, could have prevented an automated system from shooting down the jet. If the information was communicated to the human crew operating the Patriot battery, it would have been a signal to call off the attack. Tragically, the IFF transponder or the Patriot battery’s ability to receive such a signal failed. While it is tempting to focus on the automated features of the Patriot system when examining the shootdown—or autonomous and semi-autonomous systems more broadly—it is important to consider such weapons as part of broader systems. As policymakers consider how to evaluate the deployment of increasingly autonomous weapons and military systems, the complexity of such systems, the ways in which they might fail, and how human operators oversee them are key issues to consider. Failures in communication, identification, and fire-control can occur at different points of a chain of events, and it can be difficult to predict how failures will interact with one another and produce a potentially lethal outcome. The Defense Science Board’s examination of the Patriot concluded that future conflicts will likely be “more stressing” and involve “simultaneous missile and air defense engagements.” In such a scenario, “a protocol that allows more operator oversight and control of major system actions will be needed,” the task force argued.

### AT Human Control is Vague

#### Human control can be made meaningful – standards have been established that allow coherent discussions on Human Control

Amoroso and Tamburrini, 2021 - Prof of International Law at the University of Cagliari and Prof of Philosophy of Science and Technology at the University of Naples Federico [Daniele and Guglielmo, Feb Italian Journal of International Affairs “In Search of the ‘Human Element’: International Debates on Regulating Autonomous Weapons Systems” <https://www-tandfonline-com.proxy.lib.umich.edu/doi/full/10.1080/03932729.2020.1864995> TM]

How to make human control over weapons systems ‘meaningful’? The MHC formula made its appearance in AWS debates in a 2013 paper by Article36 commenting on the UK Ministry of Defence’s Joint Doctrine Note 2/2011 on unmanned systems. Taking issue with the paper’s interpretation of “human control”, which allowed weapons to carry out distinction and proportionality analyses without human intervention, Article36 made the case for retaining meaningful human control over individual attacks (Article36 2013, 2-4). In subsequent documents, it suggested that the notion of MHC has a twofold grounding: first, it is ethically and legally unacceptable to let machines apply (possibly lethal) force “without any human control whatsoever”; second, and more relevant here, not every form of human control is normatively satisfactory, since “a human simply pressing a ‘fire’ button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered ‘human control’ in a substantive sense” (Roff and Moyes 2016, 1). At the GGE, Article36 elaborated on the idea of human control with three basic claims. First, the human control issue should be addressed in the early stages of design, with a view to ensuring predictability, reliability and transparency of relevant technologies. Second, users should be given adequate information regarding weapons systems’ functioning and their specific deployment contexts. Finally, “timely human action” should be required for the weapon to initiate an attack, as well as “a capacity for timely intervention” to stop the system if there are extended time lags (Moyes 2016, 2).

### AT Industry Circumvents

#### Extending ethical guidelines to private companies is essential for effective governance

**Artificial Intelligence/Machine Learning Risk & Security Working Group, 2020** [Wharton, ‘Artificial Intelligence Risk & Governance’, <https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/>]

Most financial institutions follow a three-lines-of-defense model, which separates front line groups, which are generally accountable for business risks (the First Line), from other risk oversight and independent challenge groups (the Second Line) and assurance (the Third Line). AI governance frameworks should ensure that sufficient oversight, challenge, and assurance requirements are met in AI system development and utilization. Furthermore, as both the potential risks and regulations related to AI are evolving, the second and third lines of defense should, likewise, ensure they have adequate subject matter expertise to effectively challenge the first line in evaluating the proposed use and implementation of the AI systems, as outlined earlier in Section 2. Roles and Responsibilities Every organization is different with respect to their internal organizational structure and general roles and responsibilities. The roles/activities below provide some examples for organizations that are discussing roles and responsibilities with respect to AI implementations to consider. It is not intended to be an exhaustive or prescriptive list. Ethics Review Board An ethics review board may review AI projects in accordance with an organization’s ethical principles, e.g. AI deemed to be high risk. Center of Excellence A Center of Excellence (CoE) may provide a knowledge-sharing platform in an organization. Depending on the organization, a CoE could create a collective view and create and share best practices. Furthermore, the CoE could maintain engagement with the industry to share and learn best practices. Data Science Some organizations have mature Data Science practices. In addition to their assigned responsibilities, the Data Science team could manage AI system inventory and version control. ML Operations A ML operations team provisions data for analysis by the data science team. They may also create and maintain data sets for the purpose of training AI systems.

### AT Turkey Blocks Plan

#### Plan will not anger Turkey – they are already moving to develop safe and ethical AI – presidential Orders prove.

Ermis, 2021 – Lawyer from Galatasaray University[Fatih, September 2021, dataguidance.com, “Turkey: The first national AI strategy,” <https://www.dataguidance.com/opinion/turkey-first-national-ai-strategy>, 6/19/22 MD]

On 20 August 2021, Presidential Circular No. 2021/18 on the National Artificial Intelligence Strategy 2021-2025 ('the Strategy'), prepared in cooperation with the Digital Transformation Office of the Presidency of the Republic of Turkey ('DTO') and the Ministry of Industry and Technology, and with active participation of all relevant stakeholders, was published in the Official Gazette, immediately entering into effect upon its publication. Fatih Ermis, Can Taşdemir, and Meriç Yıldırım, from Ozbay & Okumus, analyse the Strategy and its potential impact on the development of AI in Turkey. The Strategy identifies six strategic priorities. Within the scope of these strategic priorities, 24 objectives and 119 measures have been determined. These aims and measures outline the actions that the implementing institutions will determine in detail afterwards. As in other countries, the Strategy will be revised and updated when the need arises during the implementation period. The Strategy will be implemented in a two-layered system. The first layer will provide coordination at the strategic level, while the second layer will coordinate at the administrative and technical levels. These strategies have been designed around the aspects of 'organizational competence', 'strategic consistency' and 'governance'. Furthermore, three core artificial intelligence ('AI') competencies, 'skills', 'data' and 'infrastructure', are at the Strategy's centre of focus. In addition, the Strategy envisages the following high-level objectives, to be achieved by 2025, which is the end of the implementation period of the Strategy,: the contribution of AI to Gross Domestic Product ('GDP') will be raised to 5%; employment in the field of AI in all public institutions and organisations, private sector, and universities will be increased to 50,000 workers; employment in the field of AI in central and local government public institutions and organisations will be increased to 1,000 people; the number of graduate-level diploma holders in AI will be increased from 1,218 to 10,000; AI applications developed by the local ecosystem will be prioritised in public procurement and commercialisation will be supported; active contribution will be made to the regulatory studies and standardisation processes of international organisations in the field of reliable and responsible AI and cross-border data sharing; and it will be ensured that Turkey is among the first 20 countries in the rankings in the international AI indices. Key values and principles The Strategy was prepared with the participation of local and central government institutions, private sector, international organisations, and non-governmental organisations, who contributed to the drawing up on a number of key principles upon which the Strategy is grounded. Respect for human rights, democracy, and rule of law: AI systems must give priority to universally recognised human rights, democratic values, and the rule of law. Supporting the environment: AI systems must support a sustainable environment and ecosystem. Ensuring diversity and inclusion: Social and cultural diversity must be ensured throughout the life cycle of AI systems. These systems cannot impose restrictions on people's lifestyles, thoughts, and beliefs. Living in peaceful, equitable, and interconnected societies: AI systems should aim to help contribute to human live in harmony with one other. Proportionality: AI systems can never pursue illegitimate purposes. Risk analysis must be made against potential damages and necessary measures must be taken to prevent occurrence of damage. Safety and security: AI systems must work reliably to ensure safety and security against possible damages which may occur on the environment and living beings. Fairness: AI systems should be designed in a manner that adheres to fundamental rights and freedoms and that benefits everyone equally. The special needs of disadvantaged groups in society must be considered and it must be ensured that the actions of AI systems do not discriminate against different demographic fractions. Privacy and data protection: AI systems should be designed in a such a way that they do not violate privacy and protection of personal data. Documents and information regarding the processing of personal date must always be open to inspection. Transparency and explainability: People have right to be informed about their decisions based on AI algorithms and to request information from relevant firms. Accountability: Individuals and organisations in the life cycle of AI systems are responsible for these systems running orderly and the applications of the principles enumerated. Data sovereignty: AI systems must fall into line with international and national regulations in the use of data throughout their life cycle.

### AT Cybernetics K

#### Human control is based on cybernetic thinking – the merging of human/machine teams

Kalpozos, 2020 - Prof of Law at Harvard [Ioannis 3-16-2020, , Leiden Journal of International Law, “Double elevation: Autonomous weapons and the search for an irreducible law of war", https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3545332 accessed on 6-18-2022]

3.2. Rising above oneself ‘We are in an arms race with ourselves – and we are winning’89 Technological evolution in war is not only about overcoming the enemy. It is also about overcoming one’s own imperfections in the wielding of violence. It is a process of progress, improvement, rationalization, optimization, ultimately the civilization of war-fighting. The role of this second elevation, which both facilitates and aims to justify the elevation above one’s enemy, is often underappreciated. I will highlight it in this section, complementing the historical narrative above and recognizing its influence on a certain view of the relationship between technology, war and international law. Elevation above oneself does not require asymmetrical relationships. The technological impetus of air power did not only serve the purpose of offense. It played a crucial role in the development of the relationship between human and machine for defensive purposes. The efforts to counter distancing and provide an effective defence against the German Luftwaffe and the early smart bomb technology of the V-1 and V-2 missiles significantly pushed forward artificial intelligence research.90 One such effort, led by Norbert Wiener, focused on the scientific articulation of human-machine interaction and the understanding of a pilot and his aircraft as a single unit, an integrated system, the behaviour of which could be predicted. While not successfully weaponized, the research led to Wiener’s theory of cybernetics,91 a widely influential theory for the scientific understanding of information, communication and the function of individuals in their socio-technical environment. Cybernetics is crucial for the evolution of human/machine merging, and the perception of self-improvement alongside the elevation above one’s adversary. Cybernetics is especially important for re-thinking law and agency in autonomous systems as it is, at the same time, based on a formalized understanding of information as the elementary unit for any sort of communication (human/human, human/machine, machine/human, or machine/machine) while having critical implications in relation to our understanding of agency and autonomy in human/machine systems. Therefore, it can be useful in appreciating that increasing autonomy and merged heteronomy are not opposites and that a ‘human-in-the-loop’ is not, by itself, the answer to the question of mechanization of judgement.92 For cybernetics, as Peter Galison has pointed out, the enemy, the German Luftwaffe, with its smart missiles and able pilots, is already perceived as hyper-rational, an advanced unit of human/machines, ‘a mechanized Enemy Other’.93 The distance already achieved by the enemy is an impetus for understanding them, through Wiener’s research, as a merged human/machine system. This perception of the enemy and the effort to predict their behaviour extends to and corresponds to the cybernetic perception of the world, and ourselves, as merged human/machine systems. The understanding of the enemy’s humanity as partial, as merged with a technical system, is reflected back to the view of oneself and it is emulated. Their distance becomes our distance; their elevation is the impetus for ours. A formalized system of information sharing and a merging in technological structures is the way forward. This is what cybernetics endeavoured to provide. Such impetus for self-improvement through military technology was applied to the creation of broader systems for the governance of war. In the Cold War, alongside the offensive asymmetry of Vietnam’s aspired automated battlefield, the period saw the creation of sophisticated human/machine systems for defensive purposes as well. The massive investment in the Semi-Automatic Ground Environment (SAGE) system in the 1950s – ‘the first large-scale, computerized command, control and communications system’ was aimed at ‘global oversight and instantaneous military response’.94 The identification of and response to the incoming threats would remain at a distance, achieved through the merging of human and machine surveillance power, in a complex and holistic system of artificial intelligence. In this mode of active defense, elevation above oneself and elevation above one’s enemy are seen as mutually reinforcing. The creation of distance and asymmetry in the elevation above one’s enemy envisions the conduct of war through an increasingly vertical relationship akin to governance.95 This entails qualities and aspirations associated with rational governance.96 Such qualities, like the rationalization and optimization of decision-making, span the range of war-making, from the level of planning and prioritizing targets (for example, on the production of ‘kill lists’97) to the level of the individual decision maker: the one pressing the button. The self-improvement through technology that puts one party in a position to govern through war is displayed in how that party governs through war, justifying its dominance.