AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense

Supporting Document

Defense Innovation Board

Table of Contents

Prologue

The enduring mission of the United States Department of Defense (DoD), according to the 2018 National Defense Strategy, is "to provide combat-credible military forces needed to deter war and protect the security of our nation."[1] Maintaining a technological and military advantage over adversaries is central to this mission, but it is not sufficient. DoD also maintains a strong commitment to its set of values, both unique to the mission and Services, and also to the core democratic values of the United States.[2] Any advantage obtained at the expense of those values would therefore be self-defeating. DoD's enduring challenge is to retain a technological and military advantage while upholding and promoting democratic values, working with our allies, and contributing to a stable, peaceful international community.

DoD's development and use of artificial intelligence (AI) reflects this challenge. AI is expected to affect every corner of the Department and transform the character of war.[3] Maintaining a competitive advantage in AI is therefore essential to our national security. For these reasons, the Defense Innovation Board (DIB) recommends five AI ethics principles for adoption by DoD, which in shorthand are: **responsible, equitable, traceable, reliable, and governable**. These principles and a set of recommended actions in support of them are described in the primary document, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense.*[4]

The DIB proposes the adoption of this set of AI Ethics Principles to help guide, inform, and inculcate the ethical and responsible use of AI – in both combat and non-combat environments – by the Department to help maintain our technological and ethical advantage. After the initial tasking by the Department to undertake a study on AI Ethics Principles, the Board engaged in a 12-month transparent process involving two public listening sessions, three consultative sessions with groups of AI subject matter experts, two practical exercises with DoD leaders and personnel, 13 meetings of an informal DoD Working Group on Ethics & Principles that also included representatives from partner countries, and dozens of research-gathering interviews with academics, ethicists, lawyers,

[1] See National Defense Strategy of the United States
[2] See Department of Defense Core Values. U.S. Air Force Core Values, U.S. Army Core Values, U.S. Navy and Marine Corps Core Values, U.S. Coast Guard Core Values
[3] See Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity; and National Security Strategy of the United States of America, December 2017. United States Department of the Air Force. 2019 "The United States Air Force Artificial Intelligence Annex to the Department of Defense Artificial Intelligence Strategy".
[4] "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense," Defense Innovation Board, October 2019.

human rights experts, computer scientists, technologists, researchers, civil society leaders, philosophers, venture capitalists, business leaders, and DoD officials. The DIB also solicited public commentary both online and in person. We have detailed this process, along with the names of non-DoD experts, further in the Appendix.

The result is the ensuing set of principles, which the Board recommends for adoption by the Secretary of Defense. DoD is deeply committed to the legal, responsible development and ethical use of AI in all its potential applications, and it hopes to encourage other nations to make similar commitments.

AI Ethics Principles for DoD

The following principles represent the means to ensure ethical behavior as the Department develops and deploys AI. To that end, the Department should set the goal that its use of AI systems is:

1. **Responsible**. Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of DoD AI systems.
2. **Equitable**. DoD should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.
3. **Traceable**. DoD's AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.
4. **Reliable**. DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use.
5. **Governable**. DoD AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior.

This White Paper is organized into five chapters. Chapter One outlines the stated needs for a set of AI Ethics Principles and ethics guidance from existing strategic documents and law. It also addresses definitional approaches associated with AI and autonomy to clearly frame

how the DIB approaches key issues for DoD and its development and use of AI. Chapter Two provides the necessary grounding for a set of DoD AI Ethics Principles to ensure that they are coherently and consistently derived for DoD. We explain DoD's existing ethics framework that applies to all of DoD's technologies, AI included. In Chapter Three, we offer substantive and evidence-driven explanations of each of the five AI Ethics Principles. These principles are normative; they are intended to inform and guide action. However, we are mindful that depending upon context, some principles may override others, and for various AI use cases, these will apply differently. Chapter Four outlines our recommendations to the Department, while Chapter Five provides conclusions. We also provide a set of Appendices to aid readers by providing transparency and clarity about our process in developing these principles and providing high-level content about existing DoD processes.

Chapter 1: Why Should DoD Prioritize AI Ethics?

In 2018, the Department published its AI Strategy. In this public document, the Department noted it "will articulate its vision and guiding principles for using AI in a lawful and ethical manner to promote our values."[5] Noting especially the need for increased engagement with academia, private industry, and the international community to "advance AI ethics and safety in the military context," the Department stressed its commitments to the ethical and responsible development and deployment of AI. Supporting this document, the United States Air Force recently released its Artificial Intelligence Annex to the 2018 AI Strategy, noting "Artificial Intelligence is not the solution to every problem. Its adoption must be thoughtfully considered in accordance with our ethical, moral, and legal obligations to the Nation."[6]

The 2019 National Defense Authorization Act (NDAA) tasked a senior official within the Department with "the responsibility for the coordination of activities relating to the development and demonstration of artificial intelligence and machine learning for the Department."[7] Moreover, this senior official, the Director of the Joint Artificial Intelligence Center (JAIC), has a duty to "work with appropriate officials to develop appropriate ethical, legal and other policies for the Department governing the development and use of artificial intelligence enabled systems and technologies in operational situations."[8] The work of the Board, as a Federal Advisory Committee to the Secretary, has been in support of this NDAA mandate.

The Department has established and kept AI ethics and responsible development and use as a priority. Our work here is intended to aid the Department in its efforts.

A. Why this matters and why now

With the growing utility and generality of AI across all domains, as well as the Department's commitment to deploy AI across mission areas, ensuring responsible AI leadership is crucial. As the National Security Strategy (NSS) states:

> "To maintain our competitive advantage, the United States will prioritize emerging technologies critical to economic growth and security, such as data science, encryption,

[5] Department of Defense, AI Strategy op cit., p. 8.
[6] United States Department of the Air Force. 2019 "The United States Air Force Artificial Intelligence Annex to the Department of Defense Artificial Intelligence Strategy".
[7] National Defense Authorization Act. 2019. Sec. 238(b).
[8] Ibid, Sec. 238(c)2 §H

autonomous technologies, gene editing, new materials, nanotechnology, advanced computing technologies, and artificial intelligence."[9]

Maintaining not just technical competitive advantage, but also normative leadership is key. Thus, establishing a set of AI Ethics Principles for DoD is particularly timely, as:

- **Stakes are high.** AI is a powerful, emerging technology; there's much we do not know about the consequences of its application in various contexts or about the interaction and interoperability of such systems (including legacy and new systems). Additionally, with rising national security threats from near-peer adversaries, the need to maintain U.S. technological advantage to ensure U.S. national security and carry out the Department's mission is increasingly critical.
- **Contest of authoritarian versus democratic norms**. Norms governing how countries and their militaries develop and use AI are largely undeveloped. Near-term actions might affect whether they evolve to reflect democratic or authoritarian values.[10] DoD has a track record of participating in international institutions to develop international norms that comply with international humanitarian law; e.g. undersea warfare, air warfare, space, cyber, as well as helping to form a law governed international order.
- **Trust internal and external to DoD.** Building confidence and trust in AI within DoD to ensure responsible and effective adoption. Likewise, responsible research, innovation, and development can help to build external public trust and confidence in DoD uses of AI without decreasing the need to innovate and adopt AI where appropriate.
- **Opportunity for transparent and strengthened collaboration with academia and the private sector.** Historically, DoD had strong ties to the academic and commercial sectors to enable innovative work and groundbreaking applications. This model has faltered over the last thirty years. To redress this problem, DoD must update its research, funding, acquisition, training, and engagement strategies. This is not a one-size fits all fix: it needs acquisition reform; new models for small start-ups, such as increasing Small Business Innovation Research (SBIR) opportunities; more and new touchpoints for entrepreneurs not familiar with the Department, such as through SOFWERX, AFWERX, and the Defense Innovation Unit; and engaging in transparent and continual public discussion with academia and the private sector about its goals and intentions. In short, DoD must be transparent about its value commitments and funding approaches to AI.

[9] President of the United States. 2017. National Security Strategy of the United States, p. 20.
[10] Polyakova, Anna and Chris Meserole. 2019. "Exporting Digital Authoritarianism" Brookings Institution. However, we note that while these AI ethics issues are newly popular, they are not new, as they have been topics of discussion and research for decades. Ex: see IEEE's 1973 report, Forecasting and Assessing the Impact of Artificial Intelligence on Society. For an updated discussion of AI issues along these lines, see IEEE's 2019 report, *Ethically Aligned Design*, First Edition created by over 500 global academic and policy experts.

B. What is AI?

The FY2019 National Defense Authorization Act stipulates that within one year, the Secretary of Defense "shall delineate a definition of the term 'artificial intelligence' for use within the Department."[11] Furthermore, for working purposes, the Act maintains that AI includes:

(1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from its experience and improve performance when exposed to data sets.
(2) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication or physical action.
(3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
(4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.
(5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

However, each definition individually carries with it policy and ethical challenges, and attempting to incorporate them all yields an unwieldy definition of AI. If we define "AI" too narrowly or too broadly, we risk limiting the scope of AI capabilities or failing to specify the unique capacity that AI systems will have, respectively.

The NDAA, while not committed to any one of these five definitions, does explicitly favor "human-like" capacities, as well as human-like "thinking" or action (2, 3, 4, 5). This move, however, may be limiting in scope for AI capabilities, for some systems may require non-human like capabilities, such as sonar, various electromagnetic sensing abilities, and other capabilities that exceed human performance on a variety of tasks. Limiting AI to cognitive tasks (4) may also limit various uses of AI for physical task completion, where the repetitive physical task requires very little cognitive abilities. Additionally, "acting rationally" (5) may unjustifiably limit a definition of AI to prior arguments and assumptions about what constitutes rationality.[12] For example, classic Enlightenment thought takes a more robust approach to this, whereas economists take a more minimal approach bordering on means-ends reasoning. Yet, this is often contradicted by behavioral economics.

[11] See John S. McCain National Defense Authorization Act for Fiscal Year 2019.
[12] Roff, Heather M. 2019. "Artificial Intelligence: Power to the People" *Ethics and International Affairs*, 33, no. 2, pp. 127-140.

For (1), this definition appears to conflate autonomous systems and AI systems, while also leaving open questions pertaining to systems that are not dynamic in their design and architectures or non-learning systems. It also does not make any indications as to systems that may learn in a supervised setting, but then after development and training are "frozen" and not permitted to continually learn (i.e. life-long learning). Furthermore, the definition also seems to require AI to be performing tasks without "significant" human oversight, but this move seems to unnecessarily marry the definition to a rather contentious view of autonomy.

The 2018 Department of Defense Strategy on Artificial Intelligence defines AI as:

> "the ability of machines to perform tasks that normally require human intelligence – for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action – whether digitally or as the smart software behind autonomous physical systems."

Much like the NDAA's proffered approaches, the DoD AI Strategy also privileges "human intelligence" as the foundation or benchmark by which to measure non-natural computational systems,[13] though the types of tasks demarcated in the DoD AI Strategy are not solely in the domain of human experience or expertise. Additionally, as written, there is no difference between the objects in the final clause because software runs on digital information.

The general tasks of pattern recognition, learning, offering probabilistic predictions or estimations, or "taking action" are only a small subset of abilities that a variety of nonhuman organic entities, such as higher-order mammals but also distributed complex organic systems like bees and ants, can perform. Mammals, birds, insects, and reptiles all appear to be able to have varying capacities of planning (as a function of episodic memory), causal reasoning, and deceptive capabilities, while others display transitive inference, and teaching/imitation. The important point is that there are a variety of natural minds in the world that display various degrees of these same capabilities, and they are in no way limited to human beings, or that humans have the monopoly on them. The differences between human and animal cognition may be down to the ways in which their brains function, and as we have no full comprehension of how a machine mind functions, it would be premature to attempt to 1) emulate human brains in computational systems and 2) expect that the computational system will "think" like a human. There is danger in

[13] In subsequent reports, such as the Congressional Research Service's report "Artificial Intelligence and National Security," the same repetition of the NDAA's definitions are utilized.

anthropomorphizing the system, in both how humans interact with it, but also how we "expect" it to behave.[14]

The Board considers AI to be *a variety of information processing techniques and technologies used to perform a goal-oriented task and the means to reason in the pursuit of that task.* These techniques can include, but are not limited to, symbolic logic, expert systems, machine learning (ML), and hybrid systems. When referring to the wider range of considerations, we use the term AI; however, where we are specifically concerned with ML systems, we will refer to ML. Furthermore, we use the term "AI systems" to mean systems that may have an AI component within an overall system or a system of systems.

We use this definition because it comports with how DoD has viewed, developed, and deployed "AI" systems over the past 40 years. It permits us to make finer grained distinctions between legacy systems run on expert or hybrid systems and newer systems utilizing ML. The use of this term allows us to reinforce that this earlier and important AI work took place within the existing ethical frameworks that we will detail in a subsequent section.

We should also make clear that *AI is not the same thing as autonomy.* While some autonomous systems may utilize AI in their software architectures, this is not always the case. The interaction between AI and autonomy, even if it is not a weapon system, carries with it ethical considerations. Indeed, it is likely that most of these types of systems will have nothing to do with the application of lethal force, but will be used for maintenance and supply, battlefield medical aid and assistance, logistics, intelligence, surveillance and reconnaissance, and humanitarian and disaster relief operations. Various ethical dimensions may arise depending upon the system and its domain of use, and those will change depending upon context.

DoD's policy regarding autonomy in weapons systems comes from the 2012 DoD Directive (DoDD) 3000.09. There, DoD defines an autonomous weapons system (AWS) as:

> "a weapon system that, once activated, c*an select and engage targets without further intervention by a human operator*. This includes human-supervised autonomous weapons systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation."[15]

[14] Indeed, many of the calls for AI stem from a reason that "humans are fallible" or "humans under stress" cannot make adequate decisions, or that "humans aren't that smart." Therefore, it is odd, in some cases, to consider human intelligence as the benchmark for such systems. See the Affective Computing Chapter of IEEE's *Ethically Aligned Design*, First Edition. See also information on IEEE Standards Working Group P7008.
[15] Department of Defense. 2012. "Directive 3000.09." Updated Change 2017.

Here autonomy is limited to the ability of a system to act without direction and intervention by a human during target engagement. A human may watch or supervise a system, but that system can carry out a commander's intent and its task without any further guidance. *How* it does so is not specified; it is the behavior that is autonomous. Thus, DoDD 3000.09 does not explicitly address AI as such, but the Directive is broad enough to cover autonomous systems run on AI.[16]

Another approach to autonomy comes from the 2016 Defense Science Board's (DSB) "Summer Study on Autonomy." Here, the focus was on autonomy as such, and not its particular application in weapons systems. The DSB defines autonomy to be:

> "The capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation."[17]

Further, the Study also importantly notes that:

> "Autonomy results from the delegation of a decision to an authorized entity to take action within specific boundaries."[18]

What is crucial to note in the Directive's approach to autonomy is the requisite feature that a weapons system can independently "select and engage" a target. What ultimately constitutes "selection" as opposed to "detection" is open for discussion. However, this approach to defining autonomy is narrow.

If, however, one takes existing DoD policy as determining which *weapon* systems are autonomous and which are not, then this opens the aperture quite wide to include many systems that can function without intervention of a human operator, and those may or may not have AI as part of that system. The way in which the Directive circumscribes this wide aperture is that it excludes a wide variety of types of systems that would still ontologically fit the definition.[19] For instance, the Directive excludes "unarmed, unmanned platforms; unguided munitions; munitions manually guided by the operator (e.g., laser- or wire-guided munitions); mines; or unexploded explosive ordnance."[20] Thus systems like the U.S.

---

[16] During the Board's public consultation, we received comments that some believe autonomous weapons systems should be preemptively banned under international law. See Appendix Section III on the Law of War for further discussion.
[17] Defense Science Board. 2016. "Summer Study on Autonomy" p. 4.
[18] Ibid, p. 4.
[19] "Autonomous" if we take the Directive's definition of carrying out a function without human intervention. Note, however, that this minimal definition makes it difficult to sustain a meaningful difference between automatic and autonomous.
[20] DDoD 3000.09, 2(b).

Army Claymore mine, the U.S. Navy MK-60 CAPTOR mine, and the U.S. Navy Quickstrike Mine family (MKs 62, 63, and 65) are excluded. However, systems such as the U.S. Air Force CQM-121A Pave Tiger and the YGCM-121B Seek Spinner, both unmanned aerial vehicles with anti-radar munitions, would constitute autonomous weapons (per the Directive) without AI.[21] A deployed contemporary of Pave Tiger and Seek Spinner would be the Israeli Harpy and Israeli Harop anti-radar loitering munitions.

Moreover, autonomous systems, such as the U.S. Navy Sea Hunter or the U.S. Navy Orca, can possess AI. However, these two surface and subsurface submarine hunting systems are unarmed (and as such do not require the additional Senior Review and Approval as called for in DoDD 3000.09). The important thing to consider going forward is that however DoD integrates AI into autonomous systems, whether or not they are weapons systems, sharp ethical and technical distinctions between AI and autonomy may begin to blur, and the Department should consider the interaction between AI and autonomy to ensure that legal and ethical dimensions are considered and addressed.

As these developments unfold, we are mindful that while AI generates new ethical questions about DoD's operations, there are characteristics often cited about AI that in fact are not unique to AI. For example:

What is *not* different about AI

- **General-purpose technology.** Like electricity, the internal combustion engine, or the computer, AI can be applied in manifold ways across society. It is an enabling tool to achieve a particular outcome.
- **Subject to the same legal regimes**. Systems utilizing AI are still subject to the same legal regimes, such as international humanitarian law, international human rights law, and applicable treaties. DoD is also constrained by its authorities under Title 10 of the U.S. Code, the U.S. Constitution, as well as other statutory regulations.
- **Human responsibility and dignity.** Human beings are the developers and creators of AI technologies, and humans are the loci of responsibility for the actions of these systems. Enduring moral principles, such as human dignity, humanity, and respect still apply, as they are also principles that ground the Law of War and just war theory.[22]
- **Enabling automation**. AI, like technologies of the past century, facilitates automation in mission- and safety-critical environments. The basic functionality – if not the speed of operation – of many automated systems today, even when utilizing AI, resemble that of their industrial precursors, such as pre-World War I versions of autopilot systems on airplanes.

[21] Note however that both of these programs were cancelled and never saw full deployment. Pave Tiger was cancelled in 1984 and the Seek Spinner was cancelled in 1989. These are both unmanned aerial anti-radar munitions.
[22] See Chapter 2 and Appendix III for more information about Law of War.

- **Technological advantage.** AI, like all technologies, may provide a technological advantage to its possessor. However, there is nothing unique about AI, as such, that distinguishes it from other technologies that provide the advantage or offset capabilities.
- **Engineering fundamentals.** While AI is a new area of study and represents, to some, a new kind of engineering, its basic building blocks are rooted in the scientific method and fundamental engineering theory and practice. Assuring ML systems pose new challenges, but questions around safety, technical excellence, responsible construction, and unintended consequences are not new to the systems engineering field that is currently integrating AI, both in the public and other sectors.

What is *potentially* different about AI

While it is certainly true that escalation, attribution, and deterrence are challenges for past and present systems, AI systems might exacerbate these challenges. In this way, AI could present a "quantity has a quality all of its own" dimension, and the Department may want to keep a forward-looking approach to these enduring challenges and how they could change with increased use of AI, in particular ML:

- **Unintended escalation and speed**. Time pressures on the development and deployment of poorly understood systems could lead to unintended outcomes, such as emergent effects of tightly coupled systems, accidents, or unintended engagements leading to international instability (e.g., a "flash war").[23] How various AI systems couple with one another, and for which purposes they are used (decision aids, planning, targeting, etc.), may present escalation challenges if those systems are engaging adversary systems in greatly reduced time periods. Furthermore, increasing reliance on ML systems in cybersecurity, especially offensive cyber operations and potential hack-backs, increases this risk.
- **Attribution and deterrence.** States frequently are able to communicate their intentions through various forms of signaling, such as explicit threats, arms build-ups, or actual military exercises. For DoD AI applications, many of these assumptions may not hold true. Signaling may be difficult or misperceived at greater rates, knowledge of capabilities or intent is harder to discern, and the source of AI attacks at a distance and in digital form may be harder to attribute.

Nevertheless, we do see AI carrying with it particular characteristics, especially for DoD, that require further consideration. For instance:

[23] For nuclear risks, see: Geist, Edward and Andrew Lohn. 108. "How Might Artificial Intelligence Affect the Risk of International War" Rand Corporation. Boulain, Vincent. 2019. "The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk" SIPRI. Center for a New American Security, "Artificial Intelligence and International Stability." Danzig, Richard. 2018. "Technology Roulette: Managing Loss of Control as Militaries Pursue Technological Superiority." Center for a New American Security.

What *is* different about AI

In some respects, *AI is different in degree, rather than kind*. Some of the aspects noted below exist for other technologies but carry with them ethical ambiguities or risks that their technological forerunners do not. Notwithstanding, there are direct benefits from using AI that other systems cannot achieve. These benefits include increased situational awareness, higher quality data for decision-making, increased speed for decision making and planning, streamlining business processes, and increasing safety and reliability of DoD equipment through predictive maintenance and advances in material sciences. What counts as a benefit or a risk, however, depends on the maturity of the technology, its testing and evaluation, the context of its use, and various design features, such as user interfaces. Therefore, what is new about AI is the following:

- **Augments or replaces human judgment**. AI is an information-processing technology that can augment or substitute for human cognition and abilities.[24] For moral and legal purposes, the use of force requires human deliberation, and it is DoD policy that autonomous weapons systems – with or without AI – will ensure appropriate human judgment and not replace it. However, for other AI use cases, such as the use of ML for personnel decisions, logistics planning, or predictive maintenance, the system may replace as well as augment human judgment in various aspects of a task.
- **Safety challenges**. For many AI techniques, including but not limited to ML, ensuring systems are safe, secure, and robust presents unique challenges.[25] For legacy systems, robust Test and Evaluation (T&E) and Verification and Validation (V&V) processes are well established, both mathematically as well as institutionally. However, for newer forms of ML, for example, T&E and V&V face serious challenges because there are open research questions within the field of AI about how best to achieve these. Additionally, for ML systems that learn over their lifetime, challenges remain for continual certification that these systems do not learn behaviors outside of their intended use and parameters. For multiple agent systems, as well as for interacting AI systems, the ability to model complexity and emergent behaviors is not well understood.

[24] For applications involving autonomous weapons systems, DoD Directive 3000.09 requires that all autonomous systems enable appropriate levels of human judgment and do not replace human judgment. DoD Law of War Manual states that the Law of War imposes "obligations on persons" and do "not impose obligations on the weapons themselves." In addition, any use of an autonomous weapon system with AI-enabled capabilities must comply with the Law of War, and all legal obligations (proportionality calculations, determination of a military objective, feasible precaution, and distinction) apply to human beings, such as commanders and operators. See: United States Department of Defense, Law of War Manual (June 2016), esp. 6.5.9.3. Also: See: Launchbury, John. 2016. "A DARPA Perspective on Artificial Intelligence." Launchbury notes "AI is a programmed entity to process information." United States Government Accountability Office. 2018. "Artificial Intelligence: Emerging Opportunities, Challenges, and Implications for Policy and Research", esp. page 5.

[25] See: Executive Office of the President, National Science and Technology Council. 2016. "The National Artificial Intelligence Research and Development Strategic Plan." The Networking and Information Technology Research and Development Subcommittee noted that emergent behavior in online learning systems, the predictability of system behavior may be difficult, while also noting that complex and uncertain environments raise many safety concerns for exhaustive testing.

- **Speed/scale/breadth of applications**. As AI is a computational and information technology, AI today can be quickly and cheaply deployed across a wide range of tasks at scale that was not previously possible. Earlier applications of AI (sometimes referred to as "machine intelligence" or just "expert systems") lacked significant computational power, were brittle, and required massive amounts of labor to hand-code knowledge representation. Thus, they were costly both in time and effort. However, the rapid decrease in costs of computation and storage, the availability of orders of magnitude more computational power, the increasing ease of amassing large amounts of data, and the ability of ML systems to learn models and representations without time-intensive labor are rapid drivers for adoption of AI. As ML progresses without needing access to large amounts of data to learn, these costs will also decrease substantially.
- **Uncertainty regarding speed and direction of progress**. Given the momentous progress in ML and other techniques that companies and universities are making, the direction and speed of progress in AI is highly uncertain.[26] The novel integration of ML into different industries and the burgeoning types of new jobs and careers that AI is spawning reveal a lack of clarity around what the AI field will look like, even a decade from now. In addition, the speed of progress is also difficult to determine due to a strong culture within the AI community of open sourcing code, data, and algorithms, allowing significantly more researchers and analysts to contribute to the field in a decentralized fashion.[27] Given the financial and reputational incentive to duplicate AI and ML systems quickly and at scale, AI researchers continue to make technical progress without having fully resolved the challenges regarding the brittleness and vulnerability of these systems.
- **Role of the private sector**. Since the early 1980s, much innovation in computing, including AI, has occurred within the private sector, outside the conventional defense industrial base.[28] This is the first time in recent history that neither DoD nor the traditional defense companies it works with controls or maintains favorable access to the advances of computing and AI for both for civilian and for military relevant technologies. Thus, the drivers and the investments from the private sector, especially in AI technologies, are for business/enterprise products and not necessarily military applications. Additionally, the ownership of intellectual property, massive amounts of data, and the requisite talent base to create AI systems remains with the private sector.
- **Diffusion and access**. Depending upon the application and hardware needs, AI systems can be easily produced, distributed, and accessible to states and non-state actors because the technology is more diffuse than ever before. Traditional regulatory mechanisms, such as the

[26] Geroski, P.A. 2000. "Models of Technology Diffusion" Research Policy, Vol. 29, pp. 603-625. Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemedy, Barbara Grosz and Zoe Bauer, "The AI Index 2018 Annual Report", AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA, December 2018.

[27] Stuart Armstrong, Kaj Sotala & Seán S. Ó hÉigeartaigh (2014) "The errors, insights and lessons of famous AI predictions – and what they mean for the future," *Journal of Experimental & Theoretical Artificial Intelligence*, 26:3, 317-342, DOI: 10.1080/0952813X.2014.895105

[28] As early as 1987 the Defense Science Board noted this fact. See "Report to the Defense Science Board Task Force on Military Software", esp. page 24. For an updated and more current view of this, see the Defense Innovation Board Software Acquisition and Practices (SWAP) Study.

International Traffic in Arms Regulations, the Export Administration Regulations, and the Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, may have difficulty regulating the proliferation of AI due to diffusion. With increasing utilization of open-source software, open-source datasets, publication of open access scholarship, and the availability and access to inexpensive large amounts of computational power on cloud systems, AI systems can be created, trained, and deployed to anyone with access to the internet.

- **Data and algorithmic bias**. Unlike other technologies, access to and necessity of massive datasets are presently required for most ML models.[29] The outsized role played by massive data brings with it various questions about the quality, fidelity, and provenance of the data. Additionally, ethical questions arise around who has access to and ownership of the data, whether data includes personally identifiable information (PII), what protections are in order for that data, and what sorts of power relations these engender.[30] Finally, because of the reliance on data, ML systems are also vulnerable to data poisoning and model inversion attacks. Such adversarial attacks on ML require new technical solutions and countermeasures. For systems relying on deep learning in particular, training data may have undesired biases inherent in its provenance and training procedures may introduce further bias, yet those training process and the attending outputs are presented as factually or even causally objective.[31] [32] While non-ML learning systems have depended on data in the past and are susceptible to biased inputs, the speed of learning of which ML is capable outpaces that of non-ML learning systems in such a way that it raises new ethical concerns about deploying algorithms in practical settings.
- **Types of mistakes.** While AI systems may generally perform better by providing a user with an ability to rapidly reach consistency and standardization of tasks, any mistakes made by the system may be very "dumb" when compared to human mistakes, and it may not be clear exactly what caused these mistakes. For example, a small child would recognize that a turtle and a rifle are different objects, but image recognition classifiers can still incorrectly identify similarly distinct sets of items for various technical reasons.[33] This same notion is found in debates around self-driving cars that highlight how thousands of lives could be saved, but some traffic accident deaths would happen because the car made the kind of mistake (e.g. identifying a stop sign as a speed limit sign) that a human never would.

What makes DoD different

Compared to other organizations and political bodies that have created AI Ethics Principles, DoD is different in meaningful ways:

[29] There is increasing research on the ability to train ML models on small datasets. However, this is an open research problem and for the present, big data is required to train most models.
[30] See the Personal Data and Individual Agency Chapter of IEEE's *Ethically Aligned Design*, First Edition. See also information on IEEE Standards Working Groups P7002, P7003, P7004, P7005, P7006, P7012.
[31] Op. cit. GAO report 2018.
[32] See information on IEEE Standards Working Group P7003 and IEEE's Ethics Certification (ECPAIS) program.
[33] See more on the turtle vs. rifle computer vision case study here.

- **Non-commercial.** DoD does not make commercial products. While DoD may utilize commercial off the shelf (COTS) products for some tasks, such as more "back office use," or modify existing commercial products for Department use, DoD is not a commercial entity. While commercial industry may solve some AI problems that are useful for DoD purposes, DoD will require other applications that are not dual-use.
- **Scale**. DoD is larger than any US company in terms of personnel, equipment, etc. Thus when incorporating AI across the enterprise, the technology may need to scale in orders of magnitude larger than current private sector efforts.
- **Access to Data**. DoD has access to a wide range of data, including PII, such as physical and mental health records, financial records, travel histories, information on family and friends, and other pertinent information needed for security clearances. The need for protections on privacy for DoD warfighters and government employees is even greater than in other organizations. The ways in which DoD may use its data is also different than commercial industries or even other government agencies due to national security needs.
- **Battlespace stakes.** Battlespace context has different features and stakes than other contexts. The most obvious is that DoD operations have lethal implications, both for combatant and civilian populations. Although other safety-critical industries, such as commercial aviation or construction, include conduct that has life or death consequences, none of these industries involve deaths or injuries occurring on behalf of an entire nation. Additionally, to prove superior in a battlespace, DoD may require systems that enable ruses and deception. In these instances, while DoD tries to *create* consistency and standardization to facilitate predictability, AI systems may excel by *avoiding* predictability (e.g., via adversarial learning that provides very unexpected outputs or actions) to better achieve mission objectives against an adversary.
- **Mission-driven.** While a company vies for greater market share within its respective industry to maximize financial returns for its shareholders and investors, DoD aims to gain global influence to ensure that other countries embrace the values of openness and rule of law over authoritarianism and illiberalism to maximize security for the American people both at home and abroad. With its unique mission to deter war and as the most advanced military in the world able to project power further than that of any other country, DoD plays a unique leading role among its international allies in shaping global norms around military conduct in the service of the American people.
- **Restricted and classified developments**. DoD engages in classified research, development, and deployment of technologies for purposes of maintaining national security. While commercial industry closely protects and holds its intellectual property, the security implications are different, as disclosure of classified information could result in grave damage to national security.
- **Deterrence and escalation risk**. DoD, unlike other federal agencies, let alone commercial actors, academia, or civil society, must put forth combat-credible forces to deter war. This is central to DoD's stated mission. However, failing to deter armed conflict can lead to the perception among DoD's competitors that the U.S. military is weak, thereby incentivizing competitors to take actions that will kill American citizens or those of allied nations. In

addition, deterrence failures could lead to potentially catastrophic destabilizing and escalatory dynamics that do not exist outside of the military sphere.

- **Potential for norm development.** DoD has the potential to help shape the norms of AI design, development, and use by States, as well as in civilian spheres. As international law is formed by and for States, particularly customary international law, how DoD uses AI has norm-generating effects for the society of states.

C. What DoD is Doing to Establish an Ethical AI Culture

DoD's "enduring mission is to provide combat-credible military forces needed to deter war and protect the security of our nation."[34] As such, DoD seeks to responsibly integrate and leverage AI across all domains and mission areas, as well as business administration, cybersecurity, decision support, personnel, maintenance and supply, logistics, healthcare, and humanitarian programs. Notably, many AI use cases are non-lethal in nature. From making battery fuel cells more efficient to predicting kidney disease in our veterans to managing fraud in supply chain management, AI has myriad applications throughout the Department.

DoD is mission-oriented, and to complete its mission, it requires access to cutting edge technologies to support its warfighters at home and abroad. These technologies, however, are only one component to fulfilling its mission. To ensure the safety of its personnel, to comply with the Law of War, and to maintain an exquisite professional force, DoD maintains and abides by myriad processes, procedures, rules, and laws to guide its work. These are buttressed by DoD's strong commitment to the following values: leadership, professionalism, and technical knowledge through the dedication to duty, integrity, ethics, honor, courage, and loyalty.[35] As DoD utilizes AI in its mission, these values ground, inform, and sustain the AI Ethics Principles.

As DoD continues to comply with existing policies, processes, and procedures, as well as to create new opportunities for responsible research and innovation in AI, there are several cases where DoD is beginning to or already engaging in activities that comport with the calls from the DoD AI Strategy and the AI Ethics Principles enumerated here. This list is not exhaustive, but highlights DoD and Service Component's initiatives:

[34] See National Defense Strategy of the United States, p. 1

[35] Service values: Air Force: integrity first, service before self, excellence in all we do; Army: loyalty, duty, respect, selfless service, honor, integrity and personal courage; Coast guard: honor, respect, devotion to duty; Marine Corps: honor, courage, and commitment; Navy: honor, courage and commitment; Joint Staff: integrity, competence, physical courage, moral courage, teamwork.

- The Joint AI Center (JAIC) announced plans to hire an ethicist to expand and improve the JAIC's current ethical assessments and to help guide DoD's development and deployment of AI. This individual will ensure that ethics is taken into consideration throughout the life cycle of an AI application.
- The Air Force (USAF) announced in September 2019 its AI Annex to the AI DoD Strategy. The Annex identifies a specific focus area and line of effort devoted to engaging in "dialogue on the ethical, moral and legal implications of employing AI in military operations in concert with the Joint AI Center."[36] This line of effort is devoted to increasing the transparency and cooperation, with a commitment to "staying engaged, informed, and accountable through our [USAF] relationships stemming from AFRL [Air Force Research Lab], academia, and various consortia."[37]
- The Army has also created an AI Task Force and appointed an officer in charge of examining the ethical implications of the Army's present and portended use of AI. Such projects include thinking through the ethical implications for using AI for personnel matters, as well as considering responsible design and development of its Integrated Visual Augmentation System (IVAS) for dismounted soldiers.
- The Army Judge Advocate General School has also added additional instruction for military lawyers with the inclusion of an AI Ethics & Principled Counsel course, and the Naval War College recently launched a new graduate certificate course in ethics and emerging military technology, with AI as one of its core focus areas.
- The Defense Research Project Agency (DARPA) has partnered with the Institute for Defense Analysis (IDA) to study the legal, moral and ethical implications of autonomy. The Urban Reconnaissance through Supervised Autonomy (URSA) project looks to derive technical requirements for autonomy from law, morality, and ethical first principles, particularly distinction, proportionality, necessity, and humanity, to enable commanders and operators to make legally and ethically compliant judgments more rapidly.[38]
- URSA, while different in focus, rounds out some of DARPA's existing work on explainable AI (XAI). XAI seeks to produce more explainable ML models while not sacrificing in efficiency and accuracy, with the goal of enabling end users to appropriately understand, and ultimately trust, AI systems. Although not the primary focus of XAI, its progress and popularity to date has substantial ethical implications. For example, XAI may be unique in how technical outputs take on morally important roles for human judgement. That is, the role of explanation for end users gives users particular inputs for their own judgments, whereby those

[36] Specifically Focus Area 5 and Line of Effort 5.
[37] Ibid.
[38] Urban Reconnaissance through Supervised Autonomy.

inputs may act as justifying or excusing reasons. The information provided by an AI system thus supports human decision making in a morally relevant sense.

These efforts are real world cases of executing the DoD AI Strategy's focus on AI safety and ethics. As noted, the Strategy explicitly calls for six lines of effort to do so: develop AI Ethics Principles for defense; investing in research and development for resilient, robust, reliable, and secure AI; continuing to fund research to understand and explain AI-driven decisions and actions; promoting transparency in AI research; advocating for a global set of military AI guidelines; and using AI to reduce the risk of civilian casualties and other collateral damage.[39]

[39] DoD AI Strategy, pp. 15-16.

Chapter 2: Existing DoD Ethics Frameworks and Values

AI Ethics Principles for DoD are intended to guide research, design, development, and use of AI by and for DoD. However, there is an existing ethics framework for *all* DoD technologies, and it provides grounding for the proposed AI Ethics Principles here. This existing ethics foundation is meant to guide the development and deployment of weapons and the use of lethal force, but the absence of a formalized ethics framework in non-combat situations means DoD also relies on direction from its existing values to undergird its non-combat activities. Non-combat aspects of DoD include intelligence, surveillance, and reconnaissance (ISR); command and control; logistics; predictive maintenance; humanitarian operations; force management; personnel training, and administrative and "back office" applications. Thus, it is important to note that while DoD's existing ethics framework applies to AI, it is not unique to AI.

These well-established ethical frameworks and values guide DoD in how it makes and executes decisions. Each of these is reflected through various statements, policy documents, and existing legal obligations. Formal accords include the Law of War and existing international treaties, while numerous DoD-wide memoranda from Secretaries of Defense highlight the intrinsic importance of ethical behavior across the armed services, which in turn forms the basis for the trust and confidence of the American people in DoD.[40] [41] The following describes this framework:

A. DoD's Conduct is Based on Deep-rooted Values

DoD is committed to a core set of values: leadership, professionalism, and technical knowledge through the dedication to duty, integrity, ethics, honor, courage, and loyalty.[42] Internally, these values ground and inform DoD policy, doctrine, training, techniques, practices, and procedures. Externally, they shape and guide how DoD articulates and develops international norms for its allies and the international community.

[40] Law of War refers to a body of international law that is adapted to warfare and provides a well-established framework to address the legality of conduct in the context of armed conflict.
[41] See memo from Secretary Mark Esper and memo from former Secretary James Mattis.
[42] Service values: Air Force: integrity first, service before self, excellence in all we do; Army: loyalty, duty, respect, selfless service, honor, integrity and personal courage; Coast guard: honor, respect, devotion to duty; Marine Corps: honor, courage, and commitment; Navy: honor, courage and commitment; Joint Staff: integrity, competence, physical courage, moral courage, teamwork. See Department of Defense Core Values, U.S. Air Force Core Values, U.S. Army Core Values, U.S. Navy and Marine Corps Core Values, U.S. Coast Guard Core Values.

Important for any set of principles, there first needs to be a coherent grounding in a set of normative values. We require this normative grounding for not only logical consistency, but also for justification for the selection of this particular set of principles over others. As our purpose is directed toward DoD, we can circumscribe our values to those of the Department. Furthermore, DoD is also committed to a set of core democratic values, e.g., human dignity, individual rights (including privacy), and the rule of law (including international law and applicable treaties, such as law of war and international human rights law).[43] These values, coupled with the mission-oriented nature of the Department, support at a more general level, the more specific set of AI Ethics Principles.

Since the Department espouses the values of leadership and professionalism, it also considers the ways by which it will uphold those values. Practically, this requires a shared understanding of commitments and obligations within DoD. For the direct purposes related to AI, this will at least require a common vocabulary and understanding of how DoD views AI. This can enable DoD to be purposeful in how it trains, educates, and grows its leaders.

Yet, leadership is also externally facing. In this sense, DoD also provides leadership through the articulation and development of international norms for its allies and the international community. Such norms can include how AI will be developed and used, and whether there ought to be any regulation on particular applications. This latter consideration requires considerable efforts on thoughtful development of military technologies (not just AI), as well as foresight into how future conflicts may arise and be fought, so that the development of those technologies complies with the values and obligations of DoD.

B. DoD Adheres to the Rule of Law

DoD is committed to compliance with existing U.S. law and the Law of War, especially the principles of necessity, humanity, and honor, from which principles of precaution, distinction, and proportionality arise.[44] This includes upholding and defending the Constitution, the laws of war, and relevant international treaty obligations.

The Department has a long tradition of care and deliberation, including ethical considerations, before implementing new means and methods of war, as well as upholding

[43] All DoD personnel are required to pledge to uphold and to defend the US Constitution.

[44] *Military necessity* justifies the use of all measures needed to defeat the enemy as quickly and efficiently as possible that are not prohibited by the Law of War; *humanity* forbids the infliction of suffering, injury, or destruction unnecessary to accomplish a legitimate military purpose; *proportionality* means that, even where one is justified in acting, one must not act in a way that is unreasonable or excessive; *distinction* obliges parties to a conflict to distinguish principally between the armed forces and the civilian population, and between unprotected and protected objects; *honor* demands a certain amount of fairness in offense and defense and a certain mutual respect between opposing military forces

domestic law. To this end, viewing ethical considerations as a late-stage constraint on innovation, and thus military effectiveness, is historically and practically false. DoD views ethical considerations as an inseparable part of research, design, and deployment for DoD AI systems, which in turn increase military effectiveness and operational legitimacy. While perhaps not explicitly labeled as "ethics," the Department's stringency with regulatory compliance, but also its strong commitment to testing and evaluation, are noteworthy.

These ethical considerations are supported by DoD's deep commitment to compliance with the laws of war. For all DoD systems, the Department must use these systems in a manner consistent with the laws of war and other applicable bodies of law. This holds equally true for AI systems.

For example, pursuant to DoD Directive 2311.01E (DoD Law of War Program), it is existing DoD policy that "members of the DoD Components comply with the law of war during all armed conflicts, however such conflicts are characterized, and in all other military operations."[45] The Law of War is:

> "That part of international law that regulates the conduct of armed hostilities. It is often called the 'law of armed conflict.' The law of war encompasses all international law for the conduct of hostilities binding on the United States or its individual citizens, including treaties and international agreements to which the United States is a party, and applicable customary international law."[46]

To comply with the obligations under LOW, DoD has a robust system to prevent violations of the LOW, as well as to enable prompt reporting of incidents, investigating authorities and capabilities, and corrective actions. DoD also publishes the DoD Law of War Manual, and qualified legal advisers provide legal guidance and advice to DoD Components to ensure compliance "during planning and execution of exercises and operations."

As noted in the Law of War Manual, "the law of war is part of who we are," meaning that the values established by this set of treaties, agreements, and behaviors are fundamental to the character of DoD. It further explains, "the law of war is part of our military heritage, and obeying it is the right thing to do."[47] Of course, there are prudential reasons to obey these laws, such as for good order, discipline, and establishing a just peace, but there are also ethical reasons for doing so. The mainstay principles of the LOW - necessity, precaution, distinction, proportionality, and humanity - are there to protect people who are not taking

[45] DoD Directive 2311.01E (2011). See Section 4.1.
[46] Ibid. See also Appendix III in this report.
[47] DoD Law of War Manual (2016). See p. ii.

part in hostilities or who are rendered unable to participate in combat. In short, the LOW is about civilian protection.

Accordingly, DoD will use AI systems in support of the LOW's purposes, including protecting combatants and noncombatants from unnecessary suffering, and assisting military commanders in ensuring the disciplined and efficient use of military force. Through the use of AI systems, DoD can enhance implementing, complying with, and enforcing the LOW.

The Department recognizes that along with the opportunities AI systems affords comes challenges. The Department should continue to assess the processes by which weapons and weapons systems are acquired, procured, fielded, and subjected to legal review and whether and how the introduction of AI systems may warrant modification to these processes. Similarly, the Department should consider whether adjustments to training, practices, procedures, doctrine, and other programs may more fully ensure that all members of DoD Components have an appropriate understanding of the relevant technologies within their organizations.

Additionally, AI systems may provide capabilities and challenges in the conduct of intelligence activities. It is DoD policy that:

> "All Defense intelligence activities will be conducted in accordance with the applicable laws, Executive orders, and Presidential directives, and governed by procedures issued by the Secretary of Defense and, where appropriate, approved by the Attorney General in accordance with E.O. 12333."[48]

The use of AI for DoD intelligence activities may minimize potential intrusiveness of collection techniques. At the same time, AI capabilities may facilitate the collection on US persons, and the combination of many sources of large amounts of data on US persons (such as through incidental collection) may be easier, potentially prompting broader collection activities. DoD should seek to balance this incentive with its longstanding commitment to collect only information that is reasonably necessary to conduct its mission in a lawful manner.

[48] DoD Manual 5240.01 Procedures Governing the Conduct of DoD Intelligence Activities (2016). See Section 1.2. See also Executive Order 12333 (1981)

C. DoD is Committed to People

Ethics constitute a set of action-guiding principles, so human deliberation and subsequent human decisions and actions are at the center of ethical and legal evaluations.

A commitment to ethics is a commitment to people, and any principles offered need to be practical and adaptable. This means that principles should be practically relevant and inform future guidelines or relevant directives. Additionally, to be adaptable means that principles need to be flexible and generalizable and should be able to lead to more granular guidelines and recommendations.

That a commitment to ethics is a commitment to people, shows us plainly that human deliberations – and subsequent human decisions and actions – form the center of ethical and legal evaluation. When it comes to AI, we see that a commitment to human agency demands that AI systems are designed, developed, and deployed in a *human-centered* way. Thus, AI should be used to augment human performance, enhance human judgment, and extend moral agency.

D. Users' Trust in DoD Systems Stems from a Well-organized Ecosystem

DoD is committed to providing personnel the requisite doctrine, training, techniques, practices, and procedures for all systems to ensure that DoD personnel possess an appropriate understanding of a system to engender trust and other properties necessary for the human to remain involved and responsible.

Human operators and commanders must be informed and educated on how a system operates, how the system is supposed to perform, and how to use systems safely and effectively to achieve mission objectives. Additionally, this approach complements DoD's existing legal requirements. Thus, being informed requires that the appropriate programs be put into place. Commanders, operators, and legal advisers need to have appropriate understanding of the technologies under their authorities.

E. An Ethical Culture of Safety and Precision Engineering Drives DoD's Technological Progress

DoD's longstanding track record of safety and technical excellence in complex systems engineering underpins any future AI projects that the Department undertakes. For mission-

and safety-critical systems utilizing AI, the Department should continue to push for the same safety and technical excellence it has achieved with other systems.[49]

Over the last half-century, DoD has invested hundreds of billions of dollars to ensure the safety and reliability of its weapons systems and platforms, as well as create more precise and accurate weapons to reduce civilian casualties and protect civilian infrastructure while achieving military objectives. Additionally, DoD continually encourages changes in how it trains its personnel to uphold these standards and use these tools responsibly.

An additional example is noteworthy: Since their launch, U.S. nuclear-powered warships have safely sailed for more than five decades without a single reactor accident or release of radioactivity that damaged human health or marine life. For more than 162 million miles, nuclear reactors have safely steamed on nuclear power, amassing over 6,900 reactor-years of safe operation.[50]

We do not highlight this example to make the case that DoD should apply AI to its nuclear enterprise. Rather, we highlight the efforts to create a culture of safety and precision that fully represents the standard that DoD has established for complex systems engineering. It is a critical foundation for DoD as it enhances its ethical culture around new and emerging technologies like AI.

[49] For a description of DoD's Test and Evaluation (T&E) and Verification and Validation (V&V) procedures and how they would be utilized or modified for AI systems, see Appendix IV of this report.
[50] National Nuclear Security Administration and Department of the Navy factsheet, "United States Naval Nuclear Propulsion Program." September 2017.

Chapter 3: AI Ethics Principles for DoD

We reaffirm that the use of AI must take place within the context of these existing DoD ethics principles. Building on this foundation, we propose the following principles that are more specific to AI, and note that they apply both to combat and non-combat systems. AI is a rapidly developing field, but it is still new. No organization that currently fields AI systems or espouses AI ethics principles can claim to have solved all the challenges embedded in the following principles, but the Department's goal should be that its AI systems are:

1. Responsible. Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of DoD AI systems.

Humans are the subjects of legal rights and obligations, and as such, they are the entities that are responsible under the law. AI systems are tools, and they have no legal or moral agency that accords them rights, duties, privileges, claims, powers, immunities, or status.[51] However, the use of AI systems to perform various tasks means that the lines of accountability for the decision to design, develop, and deploy such systems need to be clear to maintain human responsibility. With increasing reliance on AI systems, where system components may be machine learned, it may be increasingly difficult to estimate when a system is acting outside of its domain of use or is failing. In these instances, responsibility mechanisms will become increasingly important.

While it is certainly true that some human decision makers may be more responsible in some instances than in others, there is a system of distributed responsibility at work for all DoD systems, AI included. We can think of this as a nested set.[52] The first layer of responsibility lies with those persons with authorities for and over the design, requirements definition, development, acquisition, testing, evaluation, and training for any DoD system, including AI ones. DoD has robust mechanisms in place to demarcate who has authorities for each of these activities, and in many instances multiple or redundant authorities.

[51] Hohfeld, Wesley Newcomb. 1913. "Some Fundamental Legal Conceptions as Applied to Judicial Reasoning" The Yale Law Journal, No. 1: 16-59.

[52] For discussion on the nested set approach, see: Roff, Heather M. and Richard Moyes. 2016. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons" Briefing Paper for the Delegates at the Convention on Certain Conventional Weapons Informal Meeting of Experts on Lethal Autonomous Weapons Systems. Roff, Heather M. 2016. "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons" Briefing Paper for the Delegates at the Review Conference on the Convention on Certain Conventional Weapons.

Nevertheless, particular attention ought to be paid to the often overlapping authorities in these instances to ensure effective communication between persons with authority and a feeling of "ownership" for this set of activities. Often tracing responsibility for systems' behaviors is only performed once something has gone wrong, when it is too late. So, rather than wait for an incident to occur, it behooves the Department to foster clearer communication and coordination from vendors, contractors, program managers, DoD Components, and DoD leadership from the outset with responsibility in mind.

For AI systems involved in the conduct of hostilities, a second layer concerns responsibility mechanisms for actions taken by decision makers during hostilities. As discussed, the Law of War is the bedrock for much of the guidance and requirements for commanders and warfighters in armed conflict. However, rules of engagement, commander's intent, as well as doctrine also play important roles for responsibility mechanisms in armed conflict.

For AI systems deployed during armed conflict, accountability requires a robust Command and Control (C2) architecture that demarcates responsibilities for human commanders and operators. For DoD, C2 is "the exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission." To effectuate C2, robust and reliable communication systems are required; as such, "the facilities, equipment, communications, procedures, and personnel essential for a commander to plan, direct, and control operations of assigned and attached forces pursuant to the missions assigned" comprise a C2 system.[53]

For commanders and operators to be held responsible under the requisite C2 system, they require appropriate information on a system's behavior, relevant training, and intelligence and situational awareness. In short, responsibility here requires certain epistemic thresholds. We can think of this along the lines of classic theories of consent. Consent requires that individuals satisfy three criteria: that an action be voluntary, intentional, and informed.[54] While military commanders may order operators to use a particular AI system, the commander's choice would ultimately be the ground for voluntary use. Likewise, her decision would be intentional - that is not accidental or arbitrary - and informed. She would have the requisite information given the circumstances ruling at the time and exercise her judgment in accordance with the laws of war, rules of engagement, and other pertinent information.

The third layer of responsibility pertains to remediation mechanisms for actions after hostilities have ended. This can be in two forms. First is internal to DoD, and this involves

[53] Department of Defense Dictionary of Military and Associated Terms. July 2019.
[54] Simmons, A. John. 1981. *Moral Principles and Political Obligations* (Princeton University Press).

holding warfighters to account for any alleged violations of the Law of War or the Uniform Code of Military Justice. The second is in external to DoD and involves the doctrine of State Responsibility.

State responsibility arises from the simple fact that States are the principal bearers of international obligations and the holders of particular international rights, such as sovereignty. Acts of any entity or organ of a State deemed wrongful, according to international law, are attributable to the State itself. As DoD is an organ of the State, any internationally wrongful acts would then be attributed to the United States as such.

In addition, issues concerning indemnification may arise. In the U.S., some defense suppliers are indemnified against liability for harm to third parties arising from the use of those products or services by agencies or departments of the U.S. government. While standard practice for other contractual goods and services, this practice may incentivize some defense contractors to engage in more risk acceptant development. Thus, it is worth considering how indemnification will affect AI development and procurement and may stress responsibility. Nevertheless, the indemnification does not detract from the general point that, if DoD uses AI systems in a manner that is internationally wrongful, then State responsibility is still an operating body of law.[55]

Ultimately, human responsibility involves making appropriate judgments, for which DoD has already set a precedent in DoD 3000.09: "[a]utonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force." While we have argued that AI and autonomy are not the same thing, the requirement that systems be designed so that human decision makers can exercise appropriate levels of human judgement is a standard to continue using.

As the U.S. Delegation to the Convention on Certain Conventional Weapons explained to the Member States in 2016 with regard to what "appropriate human judgment" means in the context of Lethal Autonomous Weapons Systems (LAWS):

> "Some may criticize "appropriate" as an overly vague standard. But we chose "appropriate" because there is no "one-size-fits-all" standard for the correct level of human judgment to be exercised over the use of force with autonomous and semi-autonomous weapon systems, including potential LAWS. Rather, as a general matter, autonomous and semi-autonomous weapon systems vary greatly depending on their intended use and context. In particular, the level of human judgment over the use of force that is appropriate will vary depending on factors, including, the type of functions performed by the weapon system;

[55] Crootof, Rebecca. 2016. "War Torts: Accountability for Autonomous Weapons" 164 *U. Pa. L. Rev.* 1347.

> the interaction between the operator and the weapon system, including the weapon's control measures; particular aspects of the weapon system's operating environment (for example, accounting for the proximity of civilians), the expected fluidity of or changes to the weapon system's operational parameters, the type of risk incurred, and the weapon system's particular mission objective. In addition, engineers and scientists will continue to develop technological innovations, which also counsels for a flexible policy standard that allows for an assessment of the appropriate level of human judgment for specific new technologies."[56]

In a similar manner, the breadth and scope of AI use cases throughout the Department will also vary greatly and technological solutions and innovations will continue to develop. Better data for decision makers can enable higher quality decision making, and intuitive user-interfaces can increase situational awareness and the speed of decision-making. This is true for AI systems utilized for personnel, healthcare, communication, command and control, and logistics.

Requiring the ability for humans to exercise levels of appropriate judgment over the use of force also entails that human decision makers need available options over whether to use force, the proportionate amount of force, and whether to de-escalate or walk back from using force. In this latter situation, decision makers require "off-ramps" from escalatory dynamics.

Off-ramps are usually political or diplomatic courses of action that can ease rising tensions in an adversarial and escalatory situation. Such off-ramps usually require time to formulate, communicate, and undertake. However, situations where AI systems interact with adversary systems may be exceedingly fast and may strain decision makers' abilities to formulate desired off-ramps.

While some AI physical systems may not have this rapid escalatory dynamic, other AI digital systems may. To this end, the Department should consider what types of mitigation strategies and technological requirements to put in place for its AI systems that foreseeably have a risk of unintentional escalation. For instance, in the stock market, the Security Exchange Commission emplaced "circuit breakers" or "collars" to halt trading on exchanges or securities when prices hit pre-set thresholds. In DoD's case, these may include limitations on the types or amounts of force particular systems are authorized to use, the decoupling of various AI cyber systems from one another, or layered authorizations for various operations.[57] The point, however, is that DoD ought to consider various

[56] United States Delegation to the Convention on Certain Conventional Weapons Informal Meeting of Experts on Lethal Autonomous Weapons. 2016. Statement on "Appropriate Levels of Human Judgment".

[57] Recently, two scholars associated with the US Air Force argued for a revamping of US nuclear command, control and communication (NC3), based on AI and akin to the Russian "dead-hand" system, Perimeter. This suggestion has

technological options to increase a decision maker's ability to de-escalate and find an appropriate off-ramp.

2. Equitable. DoD should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.

We wish to acknowledge that the term "fairness" is often cited in the AI community as pertinent to this principle. Indeed, we use the term below. The reason we do not word this principle as such stems from the DoD mantra that fights should not be fair, as DoD aims create the conditions to maintain an unfair advantage over any potential adversaries, thereby increasing the likelihood of deterring conflict from the outset. By intentionally seeking an unfair or asymmetric advantage, DoD can reduce the risk of armed conflict. Additionally, though this principle mentions combat and non-combat AI systems, it might end up applying to the latter more than the former, as DoD already intends to inflict harm on adversaries, when legal and ethical. Doing so with AI does not change the framework under which DoD engages in armed conflict. However, applying AI to certain non-combat situations could cause new types of unintended harm – unrelated to life or death – to DoD personnel.[58]

Benefits of AI and ML technologies are that they enable low-cost, enterprise level, and rapid decision-making capabilities. As DoD will utilize such applications across the Department, from personnel decisions to predictive maintenance to combat operations, it is incumbent upon the Department to ensure that those applications do not inequitably or unfairly discriminate or engender unjustified biased outcomes.[59] Equitability, as we conceptualize it here, has two related dimensions: First, for some AI use cases, privacy protections will be a needed feature. Second, not all bias is bad.[60]

---

been met with serious skepticism by academics and senior DoD leadership. See: Lowther, Adam and Curtis McGiffin. 2019. "America Needs a 'Dead Hand'" *War on the Rocks*, 16 August. Field, Matt. 2019. "Strangelove Redux: US Experts Propose Having AI Control Nuclear Weapons" *Bulletin of the Atomic Scientists*, 30 August. Former Deputy Secretary of Defense Robert Work and the Director of the Joint AI Center, Lieutenant General Jack Shanahan, also publicly stated at a recent event at the Johns Hopkins Applied Physics Laboratory on Assuring AI that they personally felt that such systems should not be deployed. Event: The Future of Humans and Machines: Assuring Artificial Intelligence, August 29, 2019.

[58] For example, applying ML to the Services' personnel management systems to select service members for promotion more effectively could rely on biased historical data sets that might not lead to the intended change in the status quo. This result would be harmful to service members who, according to DoD leaders who approved this approach, are supposed to benefit from ML-driven promotion decisions and subsequently assume higher-ranking roles in which they could better contribute to maintaining DoD's competitive advantage.

[59] For a sampling of papers on addressing fairness in machine learning, see Google, Machine Learning Fairness paper archive, as well as the publications of Microsoft's Fairness, Accountability, Transparency, and Ethics in AI group's publication page.

[60] Danks, David and Alex John London. 2017. "Algorithmic Bias in Autonomous Systems" *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence:* 4691-4697.

The relationship between privacy and fairness is complex, and particular design choices and assumptions about both ought to be considered at the outset and made explicit.[61] For some applications, keeping not merely PII protected but also veiled can increase the values associated with procedural fairness, where "fairness" is the result of an agreed upon process (e.g. due process of law).[62] That is, in some instances, it is desirable to treat similar cases similarly without reference to other demographic pieces of information. The opposite may also hold true: that for protected categories of individuals, global solutions to procedural fairness may be insufficient and differential considerations will be overriding.[63]

An example may be worth noting is that in 2016, an Idaho judge ruled in K.W. v. Armstrong that Idaho's Medicaid program for adults with developmental disabilities "arbitrarily deprives participants of their property rights and hence violates due process" by utilizing an automated decision system to allocate disability benefits.[64] The use of the automated system's formula for resource allocation was not made public to its recipients, and upon examination, the "tool" Idaho used had data flaws, disproportionate results for different populations, and statistical errors.[65]

The reliance on inappropriate historical data for the Idaho Medicaid tool is noteworthy, as many AI decision-aids or tools will undoubtedly rely on exactly this type of information. For the Department, use of such historical data in personnel decisions, such as billeting or promotions, will likely draw from such sources. It is therefore incumbent that equitability act as a guiding principle in the development and deployment of such systems, and that DoD performs appropriate checks for unwanted bias.[66]

In applying AI to prediction, attention must also be paid to upholding a principle of equitability because for predictive models, personnel may not have actually undertaken any overtly prohibited, undesired, or unlawful activity, but the predictive model's outputs may cause superiors or others to treat personnel unfairly. As many of these types of

---

[61] Dwork, Cynthia. Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel. (2011). "Fairness Through Awareness"

[62] Rawls, John. 1971. *A Theory of Justice* (Belknap Press of Harvard University Press).

[63] Ekstrand, Michael, Rezvan Joshaghani and Hoda Mehrpouyan. (2018). "Privacy for All: Ensuring Fair and Equitable Privacy Protections" *Proceedings of Machine Learning Research*, Vol. 81: 1-13.

[64] *K.W. et. al. v. Richard Armstrong.* Case No. 14-35296. See also: ACLU (2016) "Ruling Mandates Important Protections for Due Process Rights of Idahoans with Developmental Disabilities" ; Stanley, Jay (2017).

[65] "Pitfalls of Artificial Intelligence Decisionmaking Highlighted in Idaho ACLU Case" . While it is uncertain that this system was AI and not merely a mathematical formula, the potential for similar effects remains.

[66] Amazon recently had to end an experimental ML based recruitment model because it disproportionately favored men over women, and this was surmised to be the fault of the underlying data. The model was trained on candidate applications over a 10-year period in which most applications came from men. See: Dastin, Jeffery. (2018). "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women" *Reuters*. Many open source platforms have free access to bias mitigation tools. See GitHub IBM/AIF 360 for instance.

projects are in pilot form, attention and care should be continuously taken not only to ensure the appropriate data, architectures, and privacy considerations are utilized, but also that the metrics for success adequately reflect values of equitability as well.[67]

While many DoD personnel consent to having their data collected, such as through disclosure through their clearance forms, job duties, communications, and personal contacts, a principle of equitability and reasonable privacy still exists for them, though reasonable is more tightly circumscribed here than for others. For applications that utilize non-DoD personnel data, the privacy requirements ought to comply with existing U.S. domestic law.

As DoD utilizes AI and ML across its mission areas, it should ensure appropriate data sources and provenance for applications that have the potential to violate a principle of equitability, include explicit and transparent notifications about sundown clauses with regard to personal data, as well as provide appropriate training for those utilizing such decision-aides, appropriate safeguards, such as employing zero-trust architectures, to limit security breaches and access to personally identifiable information,[68] and mechanisms for redress for personnel who wish to dispute particular evaluations made by these tools.

Some applications will be permissibly and justifiably biased because the intent of the designer is to weigh particular parameters more significantly than others to achieve an optimal outcome.[69] Specifically, DoD should have AI systems that are appropriately biased to target certain adversarial combatants more successfully and minimize any pernicious impact on civilians, non-combatants, or other individuals who should not be targeted.

3. Traceable. DoD's AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.[70]

The DIB's 2019 Software and Acquisition and Practices Study (SWAP) argues that DoD needs to adopt the "the same modern tools, systems, environments, and collaboration

[67] Tucker, Patrick. (2019). "The US Military Is Creating the Future of Employee Monitoring" *Defense One*. Kimmons, Sean. (2019). "Army Leaders Discuss Benefits, Challenges with AI Systems" *US Army*.
[68] Koerner, Brendan. (2016). "Inside the Cyberattack that Shocked the US Government" *Wired Magazine*.
[69] At a more basic level, all systems will have some form of bias implicit in them due to the choices of metrics, data, algorithms, deployment scenarios, etc. In this respect, there is no value-neutrality in any AI system, and continual emphasis on complete objectivity is misleading. There is a vast literature in philosophy of science to this fact, especially regarding inductive approaches. Cf: Douglas, Heather. 2000. "Inductive Risk and Values in Science" *Philosophy of Science*, Vol. 67, no. 4: 559-579.
[70] We use "traceable" as the overarching principle here to align with the Organisation for Economic Cooperation and Development's Principles on AI, which the U.S. approved, along with 42 other countries, in May 2019.

resources that commercial industry has adopted" to ensure readiness and superiority.[71] Additionally, the SWAP Study notes:

> "With the introduction of new technologies like ML and AI and the ever-increasing interdependence among networked heterogeneous systems, software complexity will continue to increase logarithmically. DoD needs to continuously invest in new development tools and environments including simulation environments, modeling, automated testing, and validation tools. DoD must invest in research and development (R&D) into new technologies and methodologies for software development to help the Department keep up with the ever-growing complexity of defense systems."[72]

The traceability of AI systems is thus an important part of maintaining the best practices and standards.

How might DoD ensure AI traceability? There are several potential avenues to pursue at this time; however, undoubtedly there will be changes in the future as technology and standards progress. For our purposes, we can identify two important phases for DoD: development and deployment. At the development phases, design methodology, relevant design documents, and data sources ought to be provided to the appropriate DoD stakeholder. For example, as datasets are increasingly important for ML systems, understanding not only the provenance of the data, but also questions pertaining to the motivation, the composition, and the collection of data are key. Recent progress on "Datasheets for Datasets" or even "model cards" for ML systems are seen as instituting a best practice for developing and deploying responsible AI.[73]

In short, Datasheets for Datasets are directed toward two stakeholder groups: dataset creators and dataset consumers. As DoD can be either or both of these groups at the same time, it is important for DoD to have familiarity with the objectives of each group. For creators, Datasheets could note a variety of aspects, including but not limited to: the process of writing down assumptions about the data or collection process; the kinds of data; the population or sampling group of data; the timeframe of collection; the suggested sunset of data use or retainment; the preprocessing, cleaning and labeling of data; the potential uses and off label uses of the data.[74] For users, having detailed access to this type of information, as well as any other relevant information to enable monitoring (such as requirements, design specifications, model information), will be increasingly important to

[71] Defense Innovation Board. 2019. Software Acquisition and Practices Study, p. 7.
[72] Ibid, p. 7.
[73] Gebru, Timnit et. al. 2019. "Datasheets for Datasets" arXiv. Microsoft, Google, and IBM have all started pilot projects related to this approach.
[74] Ibid.

ensure appropriate and responsible use. This type of tracking can also help to mitigate data dependencies.[75]

For deployment, traceability is equally important. First, from the perspective of the AI or ML system, being able to probe a system throughout its life cycle is increasingly important, especially if that system relies on streaming data or "online learning." Given the inability to trace all of the data for online learning systems, one approach that may provide a trail of system performance may be to automatically probe such systems with hypothetical cases to query a system to check for bias, brittleness, or potential distributional shift.[76] The results of such probes can provide a sufficient history of system operation.

Another approach may be to provide after-action reviews of a system after a deployment. To do so, however, will require greater attention paid to the data storage requirements for the data emanating from these systems. For systems deployed at the tactical edge and have small weight and power limitations, research may be needed on the tradeoffs between after-action reviews and data storage on a system.

Second, from the perspective of program managers, commanders, and operators, the creation of appropriate logs of user access and authorities may be increasingly important. Some systems may require not just reviews of user access, but also records of use and for what purpose. This requirement can mitigate harms related to off-label use of an AI system, as well as reinforce the principle of responsibility.[77] In short, DoD will need to rethink how it traces its AI systems, who has access to particular datasets and models, and whether those individuals are reusing them for other application areas. This will also become increasingly important as advances in transfer learning become more apparent.

4. Reliable. DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use.

DoD has a long-established history of test and evaluation (T&E) and verification and validation (V&V) of its systems. In this respect, AI systems ought not be any different. In

[75] Scully, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison. 2015. "Hidden Technical Debt in Machine Learning Systems" *NIPS Proceedings*.
[76] Etzioni, Oren and Michael Li. 2019. "High Stakes AI Decisions Need to Be Automatically Audited" *Wired Magazine*. Distributional shift is where a learning system continually interacting with its environment learns a pattern of unwanted behavior because its deployment environment is different than its training environment. See: Amodei, et. al. 2016. "Concrete Problems in AI Safety" arXiv.
[77] This can be seen as a problem of "undeclared consumers" where a systems output is then consumed by other systems, in classic software engineering this is a form of "visibility debt." Cf: Scully, D., et. al., op. cit.

high-risk areas, such as with nuclear weapons, there also exist additional programs for authorization, safety, and reliability (i.e. nuclear surety). As AI will be employed across a wide variety of applications – from enterprise systems to weapons systems – it is crucial that individual AI systems and interacting AI systems are assured appropriately. While AI systems that DoD intends to field for less risky purposes may not warrant the same level of rigorous testing, all AI systems need to be reliable with respect to their safety, security, and robustness continually and across their life cycles and domains of use. We can think, then, of reliable AI systems as those which appropriately, safely, and robustly act within their domain.

However, some AI systems, such as online ML systems, pose T&E and V&V challenges. For some ML systems, such as those that are non-deterministic, nonlinear, high-dimensional, probabilistic, and continually learning, traditional T&E and V&V techniques are insufficient. In part, this is because defining the domain of use can be incredibly difficult. For these types of systems, new research and standards for T&E and V&V are required.

V&V for ML models should take into consideration data validation and model selection.[78] Verification for these types of ML systems will also have to be done at runtime, but this entails onboard verification and will look distinctly different than prior V&V approaches in the Department and may have substantial policy implications for systems that require weapons reviews.[79]

Even ML systems that are "frozen" before deployment (offline learning) require novel techniques for V&V, though it is possible to have some form of verification.[80] Acknowledging some of the current difficulties with ML verification is pressing and increasingly important if DoD wants to push new data to an existing model to update its learning and ultimately create a new model in a rapid and iterative fashion, especially for systems deployed in theater. Without sufficient automatic data validation, as well as model and parameter selection validation, DoD risks aggregating errors across ML applications.

ML models utilized for predictive analytics, say for decision aides that are continually receiving new data inputs, is one instance where particular care ought to be taken. As there may be no ground truth available to verify the model, there may be nothing to compare a given output (e.g. label, classification, likelihood, etc.). Additionally, because predictive analytical systems may be continually receiving new data, verifying the new data inputs

[78] Pullum, Laura, et. al. "Mathematically Rigorous Verification & Validation of Scientific Machine Learning". Breck, Eric. et. al. 2019. "Data Validation for Machine Learning" arXiv.
[79] For new and novel work on nonlinear verification, see: Qin, Chongli, et. al. 2019. "Verification of Non-Linear Specifications for Neural Networks" arXiv.
[80] Van Wesel, Perry and Alwyn E. Goodloe. 2017. "Challenges in the Verification of Reinforcement Learning Algorithms" NASA/TM-2017-219628.

may also be infeasible. Ensuring the reliability of these systems would also seem to require specific and objective criteria/metrics to evaluate the performance of the system, being mindful of various sub-optimal behaviors that may emerge. These might include being stuck in unwanted or undesired optimums or mis-specifying a goal or value-function.

T&E also poses unique challenges for AI systems, particularly ML systems. While it is certainly true that DoD has a very well established T&E enterprise, one that consists of not only DoD level authorities, but also Service Component entities, existing techniques and approaches to T&E will be increasingly difficult with ML systems, software complexity, and human-machine interaction. Existing test ranges may be insufficient, aging, or inadequate for testing new systems, and the classic approach to T&E by formulating a T&E Master Plan (TEMP) will require a more flexible and adaptable approach, along the line of DevOps, to ensure sequential T&E grounded in subject matter expertise, throughout the life cycle of a system.[81] This life cycle approach to T&E may pose challenges for appropriate or adequate cost evaluations for the sustainment of some AI systems.

There is also the risk of interactions of AI systems. While some AI applications will be stand-alone solutions, many of the Department's efforts include layering AI solutions. Robustness and interoperability, for example, must now also be considered for AI systems of systems. While designers and developers might functionally decompose the operations and activities of individual systems, DoD should take care during T&E and V&V to adequately consider the overarching AI system of systems, including the interaction of subordinate, layered systems, and identification of and solutions to failure in one or more of the subsystems. This may in fact be impossible, given the inability to test, model, or simulate such a large state space, as well as adequately test all components in dynamic, unpredictable, and unstructured environments with high fidelity.[82]

There are many challenges on the horizon for reliable, assured AI. Of particular concern are nonlinear models, complex adaptive systems and attending emergent properties, self-learning systems, and a range of vulnerabilities, such as model inversions, adversarial

[81] Defense Business Board. 2016. "Best Practices for the Business of Test and Evaluation: Recommendations on the Test and Evaluation Enterprise to Improve Management and Effect Process Improvements" DBB FY17-01. For a response to the DBB report see: Gilmore, Michael. 2017. "Memorandum for Chairman, Defense Business Board". For clarity on current U.S. test ranges, there are 23 components to the Major Range and Test Facility Base (MRTFB) across the United States. The U.S. also has International Test and Evaluation (IT&E) bilateral efforts with 11 countries: Australia, Canada, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Sweden, and the United Kingdom. Beginning in 2015, the U.S. along with Australia, Canada, New Zealand, and the United Kingdom established the Multinational Test & Evaluation Program (MTEP), as well. See: United States Director of Operational Test and Evaluation. 2018. "Director, Operational Test and Evaluation FY 2018 Annual Report" pp. 11-12

[82] See: Zacharias, Greg. 2019. "Emerging Technologies: Test & Evaluation Implications" 10, April.

attacks (adversarial images, generative adversarial networks finding vulnerabilities previously unforeseen), and issues of distributional shift.

However, greater focus and study on AI assurance can go far to help ameliorate these issues. As noted, it is crucial to document assumptions for traceability, including data, context, domain of use, and interoperability with new or legacy systems. The assumptions about the operating environment and the properties and constraints in a platform (such as sensors, etc.), in the data and the algorithm must all be made explicit for traceability as well as for T&E and V&V. Assumptions may also interact in uncertain ways, so appropriate testing of systems is crucial to ensure that anomalous inputs do not drastically change the behavior of the system.

Explainability of ML models may also be also another route toward verification. There may be a need for explainable AI – the ability of an AI solution to explain the logic behind a recommendation or action. Indeed, DARPA's initiative on explainable AI demonstrates DoD's commitment to this field of research. Current AI decision support systems have demonstrated the ability for AI to model countless options that can result in recommendations that are effective but appear unorthodox to humans. The ability to understand the logic behind recommendations, at least in the near term, is critical to developing trust in AI solutions. However, it is important to clarify that the level of explainability or assurance should depend on the nature of the mission. If no one's life, well-being, or property is at risk, then perhaps less or even no assurance is required. Conversely, if an AI system is meant to be used in a mission-critical scenario with lives at stake, then a higher level of assurance should be required.

Finally, greater reliance on modeling and simulation (M&S), adaptive red teaming, novel test designs (such as new designs of experiments, rapid and sequential test designs), and appropriate metrics for ML systems are crucial areas to address in the pursuit of AI reliability.[83]

5. Governable. DoD AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior.[84] [85]

[83] Air Force Science Advisory Board. 2017. "Adapting Air Force Test and Evaluation to Emerging System Needs." Tate, David. 2019. "A Framework for Thinking About the Challenges of TEV&V of Autonomy" IDA D-10872. Institute for Defense Analysis. Also, thanks to David Sparrow, Institute for Defense Analysis.
[84] While "harm" in Principle 2 refers to harm to persons, "harm" here may also refer to harm to DoD's technical infrastructure, communications, or ability to conduct operations or make decisions.
[85] We recognize that AI systems may act or react differently than conventional automated or non-automated systems, to the point where they might exhibit undesirable behavior, such as AI on AI adversarial or other "flash crash" evolution. It is worth discussing whether DoD should design an AI system that can be disabled completely,

AI systems will fail to achieve their stated function, and it is foreseeable that when they fail, they may fail in surprising ways. As DoD will utilize myriad AI systems throughout the Department and Service Components, careful attention ought to be paid to the specifics of each system, ensuring that all reasonable measures to minimize unintended harm or disruption of a faulty or failing AI system. This is increasingly important for safety-critical systems, especially those utilizing AI in an open world (or "in the wild"), where the real-world complexity and dynamic changes in an environment offer many opportunities for a system to fail.[86] Thus, correctly designing and engineering AI systems to be governable when they move outside of their domain of use is crucial.

Reasonable measures in system design, development, and deployment will vary depending upon the particular AI system and its domain of use. However, careful attention should be paid to the following areas: ontologies of failures, system characteristics, and architecture design for isolation or ability to safely interrupt, where appropriate.

Creating a robust ontology of potential failures is important not only for designers and developers of AI systems, but also for users. Thus, understanding the type, scope, number, and scale of harm for potential failure modalities is increasingly important. Failures due to adversarial attack or manipulation, latent variables or un-modeled influences on a system, the ability of the system to modify itself, and failures arising from human-machine collaborations are a few examples.[87]

Understanding the characteristics of specific AI systems will also help to understand various failure modalities and help to ensure responsible design and deployment. Recent work from scholars at Cambridge University outlines what they deem the "safety relevant AI characteristics" of any given system.[88] Such characteristics include internal features; effects of the external environment on the system; impact of the external system on the environment; various types of computational systems, distribution, and interaction with other systems; and capabilities for supervision, amongst others.[89] In this respect, providing taxonomies of various safety-relevant features of a system complements work on various

---

but given existing non-AI systems today that DoD might use but cannot turn off or call back after a certain point, setting a higher bar for AI systems could constrain the Department. However, concerns over AI systems learning and evolving to disobey commands are not unreasonable. This is why we chose to frame the principle as written.

[86] Dietterich, Thomas and Eric Horvitz. 2015. "Rise of Concerns about AI: Reflections and Directions" Communications of the ACM, Vol. 58, No. 10: 38-40.

[87] For a thorough consideration of some of these issues, see: Eric Horvitz. 2016. "Reflections on Safety and Artificial Intelligence" Exploratory Technical Workshop on Safety and Control for AI, Carnegie Mellon University.

[88] Hernández-Orallo, José, Fernando Martínez-Plumed, Shahar Avin, Seán Ó hÉigeartaigh. 2019. "Surveying Safety Relevant AI Characteristics" Proceedings of 1st AAAI's Workshop on Artificial Intelligence Safety

[89] Ibid.

failure ontologies and technical AI safety problems that are yet unresolved within the AI research community.[90]

Architecture design for safe interruptibility, isolation, and self-monitoring are also vital as DoD deploys AI throughout its mission areas. Depending upon the task, as well as the mission objectives and risk calculation, AI systems may work with little human supervision. Due to the scale of interactions, time, and cost, humans cannot be "in the loop" all the time. This is not to say that humans have no supervisory capacity; humans are the responsible agents for AI systems. However, DoD ought to consider the task and role of the AI system under development and deployment, how that system interacts with other legacy or new systems, and what sorts of risks may result from unintended interactions with adversary systems, such as escalatory dynamics.

System architectures should be designed, where appropriate, to permit safe interruptibility. For complex systems and systems of systems, engineering for isolation or graceful failure is prudent. For other systems, particularly ML systems that continually learn, the ability to safely interrupt without causing suboptimal learning is no easy task. However, recent work on safe interruptibility for reinforcement learning agents shows that they can learn to behave optimally despite interruptions from operators.[91] This work also extends to multi-agent and dynamic environments, and is therefore of particular interest to DoD applications in open real world deployments.[92]

However, additional research on dynamic systems reasoning in uncertain conditions is also needed. Depending on the significance or risk level of a particular task, AI systems need to know when they do not know something, and they need to be able to detect when they shift outside of their intended domain of use.[93] In some cases, this may entail a needed "reach-back" to a human operator for guidance; in other cases, it may require a system to halt its actions. The particularities of the system will drive design and development, but we suggest taking a considered effort in thinking through the various kinds of failure modalities, and the technological solutions required to mitigate against risks.

Governability is important because operators and users of AI systems should understand the potential consequences of deploying the system or system of systems to its full extent,

[90] See: Ortega, Pedro, Vishal Maini and DeepMind Safety Team. 2018. "Building Safe Artificial Intelligence: Specificiation, Robustness, and Assurance." *Medium*.
[91] Orseau, Laurent and Stuart Armstrong. 2016. "Safely Interruptible Agents" in Uncertainty in Artificial Intelligence: 32nd Conference, ed. Alexander Ihler and Dominik Janzig.
[92] El Mhamdi, El Mahdi, Rachid Guerraoui, Hadrien Hendrikx, Alexandre Maurer. 2017. "Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning." arXiv.
[93] If AI is being used for high-risk purposes in less-mature application areas, it is important that AI systems abide by this recommendation. However, in low-risk situations and mature application areas, this approach is less important.

which may lead to unintended harmful outcomes. In these cases, DoD should not use that system because it does not achieve mission objectives in an ethical or responsible manner.

Chapter 4: Recommendations

In support of the deliberation and implementation of its ensuing AI Ethics Principles, the Defense Innovation Board has identified useful work that can aid in the articulation and adoption of said principles. The following twelve recommendations will support these efforts:

1. **Formalize these principles via official DoD channels.** The Joint AI Center should recommend to the Secretary of Defense the proper communications and policy issuances to ensure the lasting nature of these AI ethics principles.

2. **Establish a DoD-wide AI Steering Committee.** The Deputy Secretary of Defense should establish a senior-level committee reporting to him/her with the responsibility for ensuring that oversight and execution of the DoD AI Strategy and that the Department's AI projects are consistent with the DoD's AI Ethics Principles. Upholding AI ethics principles requires DoD to integrate them into many underlying aspects of decision-making, from a conceptual level such as DOTMLPF[94] to more tangible AI-related areas like data sharing, cloud computing, human capital, and IT policies.

3. **Cultivate and grow the field of AI engineering.** The Office of the Under Secretary for Research and Engineering (OUSD(R&E)) and the Service Labs should support the growth and maturation of the discipline of AI engineering by: building on sound engineering practices that DoD has long fostered, engaging the broader AI research community more extensively, providing specific opportunities for early-career researchers, and adapting the Department's legacy of safety and responsibility to the field of AI to integrate AI technology into larger complex engineered systems.

4. **Enhance DoD training and workforce programs.** Each Service, Combatant Command, defense agency, defense field activity, and component within the Office of the Secretary of Defense should establish programs for training and education that are relevant to their respective DoD personnel in AI-related skills and knowledge.[95] Various AI training programs should be made widely available, from junior personnel to AI engineers to senior leaders, and should leverage existing digital content combined with tailored instruction from leaders and experts.[96] It is imperative that junior officers, enlisted service members, and civilians are exposed to AI in their training and education early in their careers, and that DoD provides opportunities for continued learning throughout their careers through formal professional military education and practical application.

[94] A DoD term that refers to Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, and Facilities.
[95] See DoD AI Strategy on p. 14 ("Providing comprehensive AI training and cultivating workforce talent").
[96] Ibid.

5. **Invest in research on novel security aspects of AI.** The Office of the Under Secretary for Policy and the Office of Net Assessment should invest in understanding new approaches to competition and deterrence in an age of AI, particularly when it is coupled with other fields such as cybersecurity, quantum computing, information operations, or biotechnology. Areas for increased focus include AI competitive and escalatory dynamics, avoiding dangerous proliferation, effects on strategic stability, options for deterrence, and opportunities for positive-sum commitments between nations.

6. **Invest in research to bolster reproducibility**. OUSD(R&E) should invest in research that improves the reproducibility of AI systems. The challenges that the AI community is experiencing in this area provides an opportunity for DoD to contribute to understanding how complicated AI models work.[97] This effort will also help address the so-called "black box" problem with AI.[98]

7. **Define reliability benchmarks.** OUSD(R&E) should explore how best to craft appropriate benchmarks for measuring the performance of AI systems, including relative to human performance.

8. **Strengthen AI test and evaluation techniques.** Under the leadership of the Office of Developmental Test & Evaluation (ODT&E), DoD should use or improve existing DoD test, evaluation, verification, and validation procedures, and, where necessary, create new infrastructure for AI systems. These procedures should follow the software-driven guidelines for T&E detailed in the DIB Software Acquisition and Practices (SWAP) Study.[99] [100]

9. **Develop a risk management methodology.** The JAIC should create a taxonomy of DoD uses of AI based on their ethical, safety, and legal risk considerations.[101] This taxonomy should encourage and incentivize the rapid adoption of mature technologies in low-risk applications, and emphasize and prioritize greater precaution and scrutiny in applications that are less mature and/or could lead to more significant adverse consequences.

---

[97] Numerous prominent AI researchers, including those associated with NeurIPS, the community's most well-known conference, have recently begun tackling the technical and financial obstacles inherent in the challenge of AI system reproducibility.

[98] The "black box" problem refers to the inability of humans to understand how AI systems reach a particular conclusion, due to the many hidden or inexplicable ways that algorithms evaluate various inputs, often leading to a lack of trust in the AI system.

[99] See DIB SWAP Study.

[100] For more details about DoD's existing test and evaluation capabilities for AI and recommendations to improve them, please see Appendix IV of this report.

[101] DARPA supported a 2014 National Academy of Sciences study that resulted in a report, *Emerging and Readily Available Technologies and National Security: A Framework for Addressing Ethical, Legal, and Societal Issues*, which recommended a risk assessment and mitigation framework addressing ethical, legal, and societal issues posed by research into emerging technologies for national security purposes.

10. **Ensure proper implementation of AI ethics principles.** The JAIC should assess appropriate implementation of these principles and any related directives as part of the governance and oversight review required by Section 238 of the 2019 National Defense Authorization Act or other future instructions.

11. **Expand on research around understanding how to implement AI ethics principles.** OUSD(R&E), in conjunction with the Services' research offices, should form a Multidisciplinary University Research Initiative (MURI) project on AI safety, security, and robustness. This MURI should serve as a starting point for continuous fundamental and academic research in these areas.[102] [103]

12. **Convene an annual conference on AI safety, security, and robustness**. In light of the rapidly evolving nature of the broad field of AI, the JAIC should convene an annual conference that examines the ethics embedded in AI safety, security, and robustness, involving a diverse array of internal and external voices.

[102] See DoD AI Strategy on p. 15 ("Investing in research and development for resilient, robust, reliable, and secure AI."). National AI Strategic Plan, 2016, esp. Strategy 4 ("Ensure the Safety and Security of AI Systems").
[103] Supplementary research should also harness the existing capabilities of DoD Federally Funded Research and Development Centers (FFRDCs) and University Affiliated Research Centers (UARCs).

Chapter 5: Conclusion

The DIB has attempted to provide an objective and authoritative accounting of the AI Ethics Principles for DoD it has proposed in the primary document, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, as well as a set of recommendations to help enable the implementation of those or other principles. We are mindful that this is a complex space, with many important dimensions. What is crucial, however, is that the discussion around AI ethics and technological advancement and innovation start to happen simultaneously. Ethics cannot be "bolted on" after a widget is built or considered only once a deployed process unfolds, and policy cannot wait for scientists and engineers to figure out particular technology problems. Rather, there must be an integrated, iterative development of technology with ethics, law, and policy considerations happening alongside technological development. We have not offered exhaustive or comprehensive lists here, but have attempted to highlight crucial elements, policies, and technological and ethical challenges. We hope this will also be the beginning of a broader conversation and effort both within and beyond the Department around how AI can help achieve important objectives in accordance with our values and the rule of law.

The AI Ethics Principles offered here, along with our further recommendations, are not designed to gloss over contentious issues or to restrict the Department's capabilities. In our three years of researching issues in technology and defense, we have found the Department of Defense to be a deeply ethical organization, not because of any single document it may publish, but because of the women and men who make an ongoing commitment to live and work – and sometimes to fight and die – by deeply held beliefs. These values must be the subject of open discussion and critical thinking to remain relevant and true. While the field of AI is new and continuously evolving, the Department's commitment to the laws of the United States, the Law of War, and international humanitarian law is enduring. We offer these recommendations with the hope that they contribute to the important discussion the Department must have on interpreting existing commitments in the context of emerging technologies such as AI.

Appendix I: Definitions

- **Algorithm**: a method or set of rules or instructions to be followed in calculations or other problem solving operations, particularly by a computer.
- **Algorithmic Bias:** Systematic bias in an AI system's outputs. Can be due to biased input or training data, a statistically biased estimator in the algorithm, off-label use, incorrect assumptions, or misinterpretation.
- **Autonomy:**
    - a) DDoD 3000.09 autonomous weapon system definition: "a weapon system that, once activated, c*an select and engage targets without further intervention by a human operator*. This includes human-supervised autonomous weapons systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation."
    - b) Defense Science Board definition: ""The capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation."
- **Artificial Intelligence**: an umbrella term that covers a variety of information processing techniques and information communication technologies utilized to perform a goal-oriented task and possesses the requisite means to pursue and complete that task.
- **Deep Learning:** Multiple layers of neural networks stacked "deep".
- **Machine Learning:** the capability of machines to learn from data without being explicitly programmed.
- **Neural Network: (AKA: Connectionist System or Artificial Neural Network):** a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Typically organized in layers of interconnected "nodes" where data inputs are observed in the input layer, then communicated to and processed in one or more hidden layers, to finally link to an output layer.
- **Online Learning:** Machine learning systems that learn and continue to learn on dynamic inputs in real time.
- **Offline Learning**: Machine learning systems that have learned their approximate target functions or policies after initial training phase and no longer learn or are "frozen."
- **Regression Testing:** Statistical software testing to re-run function and non-functional tests to ensure that previously developed and tested software still performs after a change.

- **Reinforcement Learning:** Machine learning system where software agents learn to take actions in an environment through the requirement to maximize some notion of cumulative reward (often discounted for future rewards) through episodic training.
- **Supervised Learning:** Machine learning that learns a function that maps inputs to outputs based on known input-output pairs from labeled data in a training sample.
- **Test & Evaluation:** The process by which a system or components are compared against the requirements and specifications through testing. Results are evaluated to assess progress of design, performance, supportability, etc. *Testing* is a program or procedure designed to measure characteristics of an entity under identified conditions. *Evaluation* is the determination and substantiated judgment of risk associated with the significance, worth, or quality of capabilities or limitations of an entity, components, integrated system, or participant in a system-of-systems, using criteria established by systems engineers or users.[104]
- **Transfer Learning:** Machine Learning method where a model developed for one task is applied to another, often related, task.
- **Unsupervised Learning:** Machine learning that learns the underlying structure or distribution of unlabeled input data.
- **Validation:** The assessment of a planned or delivered system to meet sponsor's operational need in the most realistic environment achievable.
- **Verification**: The process of assessing how well a system meets a specification requirement.

[104] https://www.dau.edu/guidebooks/Shared%20Documents/Chapter%208%20Test%20and%20Evaluation.pdf ; original citation: https://uscode.house.gov/view.xhtml?req=%28title:10%20section:196%20edition:prelim%29%20OR%20%28granuleid:USC-prelim-title10-section196%29&f=treesort&edition=prelim&num=0&jumpTo=true

Appendix II: Principles Development Process

In July 2018, the Department asked the DIB to undertake an effort to help catalyze a dialogue and set of consultations to establish a set of AI Ethics Principles for Defense, especially while the adoption of this technology is at a nascent stage. The DIB is well-suited to enjoin business, academic, and non-profit perspectives with their accumulated insights into defense, and to engage a diverse array of stakeholders in our society who have important views to bring to this discussion. To that end, the DIB has sought to make this process as robust, inclusive, and transparent as practical. DIB Chairman, Eric Schmidt, tasked the Science and Technology Subcommittee to take on this line of effort.

In August 2018, the Department formed an informal, internal DoD Principles & Ethics Working Group that included international partners and met monthly to assist the DIB in gathering information and promoting cooperation. This group consisted of roughly 25 different organizations within DoD, with all Service Components represented, as well as the Office of the Under Secretary for Research & Engineering, Office of the Under Secretary for Acquisition & Sustainment, Office of the Under Secretary for Policy, Office of the General Counsel, the Office of the Chief Information Officer, the Joint AI Center, the Joint Staff, all the Services (including their respective Judge Advocate General Corps), the National Security Commission on AI, MITRE Corporation, Institute for Defense Analysis, CNA, and other Components. This effort aided us in mapping the internal ecosystem of stakeholders in DoD working on AI, as well as ensuring representative viewpoints were shared and heard.

In addition, more than 20 General Officers, Flag Officers, and civilian members of the Senior Executive Service participated in a session designed to "cross-examine" the draft principles to assess their viability and applicability within DoD. These leaders' comments illuminated potential operational challenges that the principles might unintentionally exacerbate, helping us refine our recommendations and craft them in a way that will preserve the ethical use of AI without compromising mission effectiveness.

The initial stage of the DIB AI Ethics Principles Project was to identify a representative set of stakeholders, taking special care to include all of the different perspectives noted above, but also to be diverse in their academic backgrounds, the levels of support (or lack thereof) for DoD, and to include gender diversity. These experts spanned computer science, AI, design, ethics, law, sociology, robotics, political science, arms control/disarmament, and operational military expertise. From this stage, invitations were sent to the identified experts where they were able to choose one of the three planned consultative session to attend. Each of these sessions was to have an accompanying public listening session where any member of the public could submit comments to the attending DIB members for their

consideration. Members of the public who could not attend in person could also submit comments through the DIB website.

The first consultative session was held at Harvard University on January 22, 2019. Due to the federal government shutdown at this time, the corresponding public listening session had to be cancelled. As the DIB is a federal advisory committee, it is required to publish its public meetings in the Federal Registry, and it was unable to do so for this meeting, as the Office of the Federal Registry was closed due to the shutdown. Thus, the first meeting held only the consultative session.

The second set of sessions were held at Carnegie Mellon University. The closed part was held on March 13, 2019, while the public listening session took place on March 14, 2019. For those interested in reading the public comments to the Board members, transcripts for the public session are available here.

The third set of sessions were held at Stanford University. The public listening session took place on April 25, 2019. The two sets of consultative sessions were held on April 26, 2019. Unlike the previous two sessions, this session had an over-subscription for attendees and we felt it was more productive to break the session into two to better enable discussion. Transcripts for the public session at Stanford are also available here.

After extensive discussions during the consultative session, reviewing the public written submissions and the public comments provided at the open sessions, as well as partaking in ongoing meetings with the informal internal DoD Principles & Ethics Working Group, the DIB Science and Technology subcommittee, along with DIB Staff and Dr. Heather Roff, the external Special Governmental Expert attached to the DIB for this project, began developing the AI Ethics Principles and this document.

Dr. Roff, Board members, and DIB staff undertook a mapping of all the existing AI Ethics Principles projects offered to date, as well as conducted various literature reviews on many of the topics traversed in this White Paper. Writing was undertaken by Dr. Roff and the subcommittee members. Additionally, we sought external peer reviews of this document by subject matter experts.

With many thanks, the following individuals participated in the formal consultative sessions. We are very grateful for their unique perspectives, insights and experiences on this topic, as well as taking the time out of their schedules to travel from near and far to provide their expertise. These individuals and their organizations do not necessarily endorse the principles or any other product that ultimately results from the DIB's expert consultations.

Participants

- General John Allen (Ret.), President, Brookings Institution
- Dr. Dario Amodei, Research Director, OpenAI
- Dr. Genevieve Bell, Director, Director of the Autonomy, Agency and Assurance (3A) Institute, Australian National University, and Senior Fellow, Intel Corporation
- Mr. Jack Clark, Policy Director, OpenAI
- Dr. Paul Cohen, Dean, School of Computing and Information, University of Pittsburgh
- Mr. August Cole, Nonresident Senior Fellow, Scowcroft Center for Strategy and Security, Atlantic Council
- Ms. Rebecca Crootof, Clinical Lecturer and Executive Director of the Information Society Project, Yale Law School
- Dr. David Danks, Professor of Philosophy and Psychology, Carnegie Mellon University
- Dr. Richard Danzig, Board Trustee, RAND, and Board Member, Center for a New American Security
- Dr. Neil Davison, Policy Adviser, Legal Division Arms Unit, International Committee of the Red Cross
- Dr. Ed Felten, Director, Center for Information Technology Policy, and Professor of Computer Science, Princeton University
- Dr. Mary Gray, Senior Researcher, Microsoft Research, and Fellow, Harvard's Berkman Klein Center for Internet and Society
- Mr. Andrew Grotto, International Security Fellow, Center for International Security and Cooperation; and Research Fellow, Hoover Institution, Stanford University
- Dr. Martial Hebert, Head of Robotics Institute, Carnegie Mellon University
- Ms. Evanna Hu, CEO, Omelas
- Dr. Joi Ito, Director, MIT Media Lab
- Dr. Sheila Jasanoff, Professor of Science and Technology Studies, Harvard Kennedy School
- Dr. Colin Kahl, Co-Director, Center for International Security and Cooperation, Freeman Spogli Institute for International Studies, Stanford University
- Dr. Michael Klare, Senior Visiting Fellow, Arms Control Association
- Dr. Mykel Kochenderfer, Co-Director, AI Safety Center, and Assistant Professor of Aeronautics and Astronautics, Stanford University
- Ms. Marta Kosmyna, Silicon Valley Lead, Campaign to Stop Killer Robots
- Ms. Mira Lane, Head of Design and Ethics, Microsoft
- Dr. Seth Lazar, Professor of Philosophy, Australian National University
- Dr. Yann LeCun, Chief AI Scientist, Facebook
- Dr. Fei-Fei Li, Professor of Computer Science, Stanford University
- Mr. Frank Long, Associate Product Manager, Google
- Dr. Bill Mark, President, Information and Computing Sciences Division, SRI International

- Mr. Chris Martin, Director R+D Intelligent + Secure IoT, Bosch
- Dr. Michael McFaul, Director, Freeman Spogli Institute for International Studies, Stanford University
- Dr. Paul Nielsen, CEO, Software Engineering Institute, Carnegie Mellon University
- Dr. Lisa Parker, Professor and Director, Center for Bioethics and Health Law, University of Pittsburgh
- Dr. Rob Reich, Faculty Director, Center for Ethics in Society, and Professor of Political Science, Stanford University
- Ms. Dawn Rucker, Principal, Rucker Group
- Dr. Tuomas Sandholm, Professor of Computer Science, Carnegie Mellon University
- Mr. Michael Sellitto, Deputy Director, Institute for Human-Centered Artificial Intelligence, Stanford University
- Dr. David Sparrow, Institute for Defense Analysis
- Dr. Lucy Suchman, Professor of Sociology, Lancaster University
- Mr. Jonathan Zittrain, Professor of International Law, Harvard Law School; Professor, Harvard Kennedy School; and Professor of Computer Science, Harvard School of Engineering and Applied Sciences

Several additional experts participated in these sessions but requested that their names be withheld. We thank them for their participation and candor.

Additionally, the following individuals provided their input to the DIB on this initiative via stakeholder interview sessions. Like the participants listed above, they too do not necessarily endorse the principles or any other product that ultimately results from the DIB's expert consultations. Particular thanks also go to the Johns Hopkins Applied Physics Laboratory and the MITRE Corporation for their valuable support in this project.

- Mr. Jared Brown, Senior Advisor for Government Affairs, Future of Life Institute
- Mr. Miles Brundage, Research Scientist (Policy), OpenAI
- Dr. Charina Chou, Global Policy Lead for Emerging Technologies, Google
- Dr. Lorrie Cranor, Associate Department Head, Engineering and Public Policy; FORE Systems Professor, Engineering & Public Policy, and School of Computer Science
- Dr. James Crawford, Founder and CEO, Orbital Insight
- Mr. Jeffrey Ding, Researcher, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford
- Dr. Ann Drobnis, Director, Computing Community Consortium, Computer Research Association
- Dr. Baruch Fischhoff, Professor, Institute for Politics and Strategy, and the Departments of Social and Decision Sciences and Engineering and Public Policy, Carnegie Mellon University
- Dr. Jodi Forlizzi, Director, Human-Computer Interaction Institute, Carnegie Mellon University
- Dr. Matt Gee, CEO, BrightHive

- Dr. Yolanda Gil, President, Association for the Advancement of Artificial Intelligence (AAAI), and Research Professor of Computer Science and Spatial Sciences, University of Southern California
- Mr. Mina Hanna, Chair, IEEE-USA Artificial Intelligence and Autonomous Systems Policy Committee
- Dr. Mark Hill, Professor of Computer Science, University of Wisconsin
- Ms. Natalie Evans Harris, Head of Strategic Initiatives, BrightHive
- Dr. Michael Horowitz, Professor of Political Science, University of Pennsylvania
- Ms. Christine Fox, Assistant Director, Policy and Analysis, Johns Hopkins University Applied Physics Laboratory
- Mr. Christopher Jenks, Assistant Law Professor, Southern Methodist University
- Mr. Andrew Kim, Manager of Government Affairs and Public Policy, Google
- Dr. Lydia Kostopoulos, Member, IEEE-USA Artificial Intelligence and Autonomous Systems Policy Committee
- Dr. Larry Lewis, Director, Center for Autonomy and Artificial Intelligence, CNA
- Mr. Ashley Llorens, Chief of Intelligent Systems Center, Johns Hopkins University Applied Physics Laboratory
- Dr. Alex London, Professor of Ethics and Philosophy, Carnegie Mellon University
- Dr. Mark MacCarthy, Senior Fellow, Institute for Technology Law and Policy, Georgetown Law
- Dr. Jason Matheny, Director, Center for Security and Emerging Technology, Georgetown University
- Mr. Brendan McCord, President, Tulco Labs
- Dr. Chris Meserole, Fellow in Foreign Policy, Brookings Institution
- Mr. Michael Page, Research Fellow, Center for Security and Emerging Technology, Georgetown University
- Ms. Lindsey Sheppard, Associate Fellow, International Security Program, Center for Strategic and International Studies
- Dr. David Sparrow, Research Staff, Institute for Defense Analysis
- Dr. Molly Steenson, Senior Associate Dean for Research, College of Fine Arts; Associate Professor of Ethics & Computational Technologies; and Associate Professor, School of Design, Carnegie Mellon University
- Mr. Craig Ulsh, Project Leader, MITRE Corporation
- Mr. Steve Welby, Executive Director, Institute of Electrical and Electronics Engineers (IEEE)
- Hon Robert Work, Distinguished Senior Fellow, Center for a New American Security

Appendix III: Law of War

The Law of War is particularly significant, as it is an internationally recognized legal guide for the conduct of all armed forces. It is important to note that underpinning the Law of War is the U.S. Constitution, to which all DoD personnel are required to pledge to uphold and to defend. Additionally, Title 10 of the U.S. Code, published by the Office of the Law Revision Counsel of the House of Representatives, describes in substantial detail the legal role of the armed forces. U.S. military use of AI across a broad range of activities could raise diverse issues to consider, including Law of War issues. Thus, it is crucial to have a firm grasp of Law of War and current DoD commitments.

Background on Law of War

The Law of War is a body of international law specially adapted to war. For the United States, this body of law includes treaties the United States has accepted, such as the 1949 Geneva Conventions, and customary international law, which results from the general and consistent practice of States done out of a sense of legal obligation.[105]

Different parts of the Law of War address different situations. One part of the Law of War, called *jus ad bellum* (the just resort to war), governs the resort to force. The Law of War that regulates the conduct of hostilities is called *jus in bello* (justice in war).

The Law of War has very well-developed rules that apply to "international armed conflicts," in other words, conflicts between opposing States. The Law of War also provides more basic and fundamental protections in the context of "non-international armed conflicts," which are all other types of conflicts, such as military operations against terrorist groups.

Existing Law of War rules can apply when new technologies, such as AI, are used in armed conflict.

Existing Law of War rules can regulate uses of AI in armed conflict. As DoD has explained in the context of cyber operations:

The Law of War affirmatively anticipates technological innovation and contemplates that its existing rules will apply to such innovation, including cyber operations. Law of War rules may apply to new technologies because the rules often are not framed in terms of specific technological means. For example, the rules on conducting attacks do not depend on what type of weapon is used to conduct the attack. Thus, cyber operations may be subject to a variety of Law of War rules depending on the rule and the nature of the cyber operation. For example, if the physical consequences of a cyber attack constitute the kind of physical damage that would be caused by

[105] DoD Law of War Manual [§§ 1.3, 1.8].

dropping a bomb or firing a missile, that cyber attack would equally be subject to the same rules that apply to attacks using bombs or missiles.[106]

A similar point may be made with respect to the use of AI in military operations. When AI is used to enhance military effectiveness in activities that are subject to existing rules, such as conducting attacks or detention operations, those rules continue to apply, notwithstanding the addition of AI. For example, if AI functions were added to weapons, such weapons would be reviewed to ensure consistency with existing legal requirements, such as the requirement that the weapon not be calculated to cause unnecessary suffering or be inherently indiscriminate.[107]

As an initial matter, it is important to understand that it is not prohibited under the Law of War to use tools, such as AI, to aid in decision-making. For example, as discussed in Section 5.4.3 of the DoD Law of War Manual:

In making the judgments that are required by the Law of War rules governing attacks, persons may rely on information obtained from other sources, including human intelligence or other sources of information. For example, in a long-distance attack, a commander may rely on information obtained from aerial reconnaissance and intelligence units in determining whether to conduct an attack.[108]

Aids to decision-making can be especially important in war because "[d]uring war, information is often limited and unreliable."[109] In addition, under the Law of War, commanders and other decision-makers must make decisions in good faith and based on the information available to them at the time. The use of AI to support command decision-making is consistent with Law of War obligations, including the duty to take feasible precautions to reduce the risk of harm to the civilian population and other protected persons or objects.[110]

In understanding how the Law of War applies to the use of AI in armed conflict, it may also be useful to consider that Law of War rules apply to persons. As the DoD Law of War Manual explains in reference to the use of autonomy in weapon systems:

The Law of War rules on conducting attacks (such as the rules relating to discrimination and proportionality) impose obligations on persons. These rules do not impose obligations on the weapons themselves; of course, an inanimate object could not assume an "obligation" in any event. Thus, it is not the case that the Law of War requires that a weapon determine whether its target is a military objective. Similarly, the Law of War does not require that a weapon make other legal determinations such as whether an attack may be expected to result in incidental harm that is excessive in relation to the concrete and direct military advantage expected to be gained. The Law of War does not require weapons to make legal determinations, even if the weapon (e.g., through

[106] DoD (n 41) p. 1056.
[107] DoD (n 41) § 6.4.1.
[108] DoD (n 41) p. 241.
[109] DoD (n 41) § 1.4.2.2.
[110] DoD (n 41) § 5.3. and 5.2.3.2.

computers, software, and sensors) may be characterized as capable of making factual determinations, such as whether to fire the weapon or to select and engage a target. Rudimentary autonomous weapons, such as mines, have been employed for many years, and there has never been a requirement that such weapons themselves determine that legal requirements are met.

Rather, it is *persons* who must comply with the Law of War. For example, persons may not use inherently indiscriminate weapons. In addition, in the situation in which a person is using a weapon that selects and engages targets autonomously, that person must refrain from using that weapon where it is expected to result in incidental harm that is excessive in relation to the concrete and direct military advantage expected to be gained. In addition, the obligation on the person using the weapon to take feasible precautions in order to reduce the risk of civilian casualties may be more significant when the person uses weapon systems with more sophisticated autonomous functions. For example, such feasible precautions a person is obligated to take may include monitoring the operation of the weapon system or programming or building mechanisms for the weapon to deactivate automatically after a certain period of time.[111]

As explained above, even if human beings are not making all the specific judgments or determinations related to military action, the Law of War imposes general affirmative duties on persons, such as the duty to take feasible precautions for the protection of the civilian population, that may be important to consider.

Fundamental principles provide the foundation for the Law of War.

Certain fundamental principles serve as the foundation of Law of War: military necessity, humanity, proportionality, distinction, and honor:

- *Military necessity* justifies the use of all measures needed to defeat the enemy as quickly and efficiently as possible that are not prohibited by the Law of War.
- *Humanity* forbids the infliction of suffering, injury, or destruction unnecessary to accomplish a legitimate military purpose.
- *Proportionality* means that, even where one is justified in acting, one must not act in a way that is unreasonable or excessive.
- *Distinction* obliges parties to a conflict to distinguish principally between the armed forces and the civilian population, and between unprotected and protected objects.
- *Honor* demands a certain amount of fairness in offense and defense and a certain mutual respect between opposing military forces.

These principles: (1) help practitioners interpret and apply specific treaty or customary rules; (2) provide a general guide for conduct during war when no specific rule applies; and (3) work as interdependent and reinforcing parts of a coherent system.[112] These principles may also help in

[111] DoD (n 41) p. 395. Italicized emphasis added
[112] DoD (n 41) § 2.1.2.

answering novel ethical or policy questions presented by the use of emerging technologies, such as AI, in armed conflict. As a potential frame for AI, the U.S. noted in an August 2018 working paper regarding autonomous weapons:

40. For example, if the use of a new technology advances the universal values inherent in the Law of War, such as the protection of civilians, then the development or use of this technology is likely to be more ethical than refraining from such use.

41. The following questions might be useful to consider in assessing whether to develop or deploy an emerging technology in the area of lethal autonomous weapons systems:

(a) Does military necessity justify developing or using this new technology?

(b) Under the principle of humanity, does the use of this new technology reduce unnecessary suffering?

(c) Are there ways this new technology can enhance the ability to distinguish between civilians and combatants?

(d) Under the principle of proportionality, has sufficient care been taken to avoid creating unreasonable or excessive incidental effects?

(e) Under the principle of the honor, does the use of this technology respect and avoid undermining the existing Law of War rules? [113]

Background on DoD's Implementation of the Law of War

History

The U.S. armed forces have a long tradition of compliance with the Law of War. As noted in the Foreword to the DoD Law of War Manual:

The Law of War is part of who we are. George Washington, as Commander in Chief of the Continental Army, agreed with his British adversary that the Revolutionary War would be "carried on agreeable to the rules which humanity formed" and "to prevent or punish every breach of the rules of war within the sphere of our respective commands." During the Civil War, President Lincoln approved a set of "Instructions for the Government of the Armies of the United States in the Field," which inspired other countries to adopt similar codes for their armed forces, and which served as a template for international codifications of the Law of War.

After World War II, U.S. military lawyers, trying thousands of defendants before military commissions did, in the words of Justice Robert Jackson, "stay the hand of vengeance and voluntarily submit their captive enemies to the judgment of law" in "one of the most significant tributes that Power has ever paid to Reason." Reflecting on this distinctive history, one Chairman of the Joint Chiefs of Staff observed that "[t]he laws of war have a peculiarly American cast." And it is

[113] See U.S. Working Paper, "Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems." 28 August 2018, Group of Governmental Experts, Geneva.

also true that the laws of war have shaped the U.S. armed forces as much as they have shaped any other armed force in the world.[114]

Current DoD issuances that implement Law of War requirements

DoD implements Law of War requirements through a variety of issuances. DoD Directive 2311.01E, DoD Law of War Program, establishes DoD policy to "comply with the Law of War during all armed conflicts, however such conflicts are characterized, and in all other military operations."[115] The DoD Law of War Program includes requirements for training, regulations and procedures, reporting of incidents involving alleged violations, investigations and reviews of incidents, and appropriate corrective actions.

Issued by the DoD General Counsel, the DoD Law of War Manual (June 2015, Updated Dec. 2016) serves as the authoritative statement on the Law of War within the Department of Defense. Over many decades, DoD components have often issued publications on the Law of War for their personnel, such as manuals, pamphlets, or instructional guides.

A variety of DoD issuances contain provisions that reflect Law of War requirements or support implementation of the Law of War. For example, DoD Directive 2310.01, DoD Detainee Program, reflects requirements for humane treatment of detainees as provided in the 1949 Geneva Conventions.[116] DoD Directive 5000.01, The Defense Acquisition Program, includes a requirement for the legal review of the intended acquisition of weapons or weapons systems to ensure consistency with U.S. international obligations, including Law of War requirements.[117] DoD Directive 3000.09, Autonomy in Weapon Systems, reflects requirements to ensure legal reviews of autonomous and semi-autonomous weapon systems, including preliminary and final legal review of certain systems raising novel issues done in coordination with the DoD General Counsel.[118]

In military operations, rules of engagement (ROE) issued by the command often implement Law of War requirements, and ROE are reviewed for consistency with the Law of War.

Role of DoD lawyers

More than 10,000 military and civilian lawyers within DoD advise on legal compliance with regard to the entire range of DoD activities, including the Law of War.

Military lawyers train DoD personnel on Law of War requirements, for example, by providing additional Law of War instruction prior to a deployment of forces abroad. Lawyers for a Component DoD organization advise on the issuance of plans, policies, regulations, and procedures to ensure

[114] DoD (n 41) p. 17.
[115] See DoD Directive 2311.01E.
[116] Ibid.
[117] See DoD 5000.01 ¶E1.1.15,
[118] Note that although a special policy review procedure may be waived in cases of urgent operational need, the requirement for legal review is not waivable: DoD (n 15) Enclosure 3, ¶2.

consistency with Law of War requirements. Lawyers review the acquisition or procurement of weapons. Lawyers help administer programs to report alleged violations of the Law of War through the chain of command and also advise on investigations into alleged incidents and on accountability actions, such as commanders' decisions to take action under the Uniform Code of Military Justice. Lawyers also advise commanders on Law of War issues during military operations.

DoD lawyers also participate in U.S. efforts to engage internationally on Law of War issues (or "international humanitarian law" issues, to use a term for the Law of War also frequently used outside of DoD) including concerning how the Law of War applies to emerging technologies. For example, in 1999, DoD lawyers authored an assessment of legal issues in cyber operations. Similarly, the 2015 DoD Law of War Manual reflects work done in 2012 in connection with DoD Directive 3000.09, Autonomy in Weapon Systems, to elaborate on how the Law of War applies to the use of autonomous functions in weapon systems.[119] In 2017 and 2018, DoD lawyers, in coordination with Department of State lawyers, drafted four working papers for the Group of Governmental Experts on Emerging Technologies in the area of lethal autonomous weapons systems, which has been convened by the High Contracting Parties to the Convention on Certain Conventional Weapons: i) a paper discussing characteristics of lethal autonomous weapons systems;[120] ii) a paper discussing the Law of War and autonomy in weapon systems;[121] iii) a paper discussing the potential humanitarian benefits of emerging technologies in the area of lethal autonomous weapons systems;[122] and iv) a paper discussing human-machine interaction in the development, deployment, and use of emerging technologies in the area of lethal autonomous weapons systems.[123]

[119] DoD (n 48).
[120] See U.S. Working Paper, "Characteristics of Lethal Autonomous Weapons Systems." 10 November 2017, Group of Governmental Experts, Geneva.
[121] See U.S. Working Paper, "Autonomy in Weapons Systems." 10 November 2017, Group of Governmental Experts, Geneva.
[122] See U.S. Working Paper, "Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems." 28 March 2018, Group of Governmental Experts, Geneva.
[123] U.S. Working Paper (n 50).

Appendix IV: Defense Acquisition / Test and Evaluation Process

A. Introduction

Adoption of AI requires that DoD test it to ensure it is satisfactory for deployment and use. DoD has a long history of investing in rigorous test and evaluation (T&E) and verification and validation (V&V) procedures, and will use and tailor them to AI, as the Department has done with previous emerging technologies. Just as DoD has a strong legal, ethical, and policy framework to build on as AI adoption increases across the Department, DoD's existing T&E and V&V infrastructure constitutes a robust foundation on which future modifications for AI will be developed.

B. Basic overview of the role of T&E in the DoD acquisition process

We seek here to explain, in broad terms, the life cycle of a potential AI system from the identification of a need, to development, deployment and disposal. This short introduction is high level, and it is grossly oversimplified for the sake of brevity. Additionally, this overview is of *existing* DoD Acquisitions policy. DoD is, however, leveraging recommendations from the Defense Innovation Board's software acquisition report, and it is developing new policy and guidance for a distinct software pathway as part of an overarching reform of the DoD Instruction 5000.X acquisition policies. These reforms are intended to bring DoD in line with industry standards, streamline acquisitions and more rapidly bring capabilities to the Department. Also note, this appendix excludes Commercial Off-the-Shelf items (COTS), as there are different regulatory rules and requirements for COTS items.[124] The appendix also does not address Internal Use Software requirements for the Department.[125]

[124] See: Defense Acquisition University discussion of COTS here: https://www.dau.edu/cop/e3/Pages/Topics/Commercial%20Off-The-Shelf%20COTS.aspx . For the FY2017 proposed changes to procurement of COTS items, see: Cassidy, Susan B. and Alexis N. Dyschkant. 2018. "Takeaways from DoD's Proposed Changes to Commercial Item Contracting" *Covington*. 13 Sept. Also see: Department of Defense. Frequently Asked Questions Commercial and Nondevelopmental Items.

[125] Department of Defense. 2019. DoD Instruction 5000.76 "Accountability and Management of Internal Use Software (IUS)" 2 March 2017, updated 7 June 2019.

The existing guidelines for acquisition within DoD are outlined in DoD Instruction 5000.02.[126] However, depending upon the designated program category, there are different statutes that govern different categories and can be quite complex; we cannot enter into this here.[127] This process is well established, even if byzantine in application. To briefly explain what a generic acquisition cycle looks like, we can reference 5000.02:

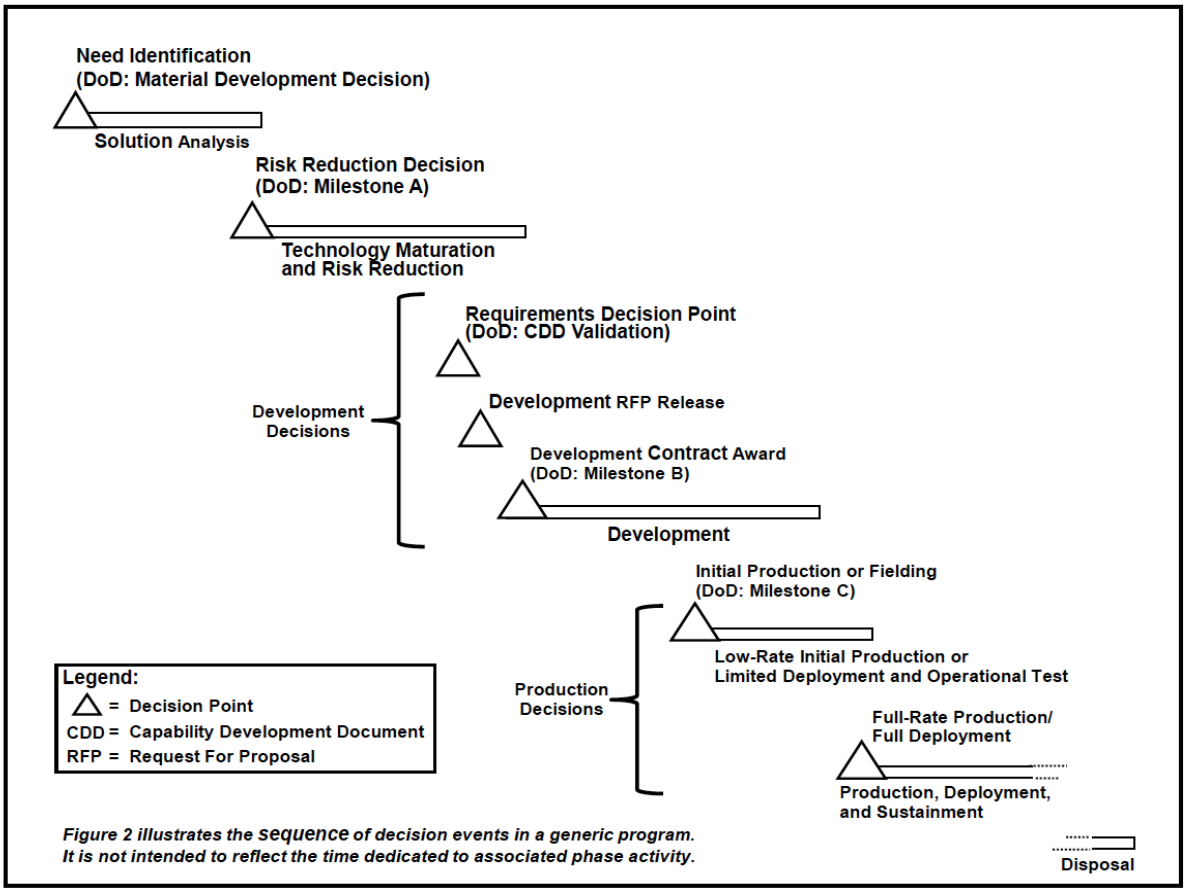

Figure 1: Generic Acquisition Phases and Decision Points, DOD Instruction 5000.02, Operation of the Defense Acquisition System, Section 5.c.(2) (a) January 2015,USD(AT&L)

This generic cycle is complicated in various ways depending upon the technology under development. For instance, various phases change or have different decision points depending upon whether the technology is hardware intensive, "unique" software intensive, incrementally deployed software intensive, accelerated acquisition, or a hybrid acquisition.

[126] Department of Defense. 2017. Instruction 5000.02 "Operation of the Defense Acquisition System" 7 January 2015, updated 10 August 2017.
[127] For further information regarding these categories, see DoDI 5000.02.

Given that AI systems can be almost any one of these programs (except perhaps hardware intensive), how the specific acquisition and T&E programs will specifically work is particular to the application. Additionally, the length of each phase will depend upon the necessary builds and the T&E requirements, which include Developmental Test and Evaluation (DT&E), Operational Test and Evaluation (OT&E) and potentially Live Fire Test and Evaluation (LFT&E).

To see the difference, see below for Unique Software Intensive Programs, which are characterized by "complex, usually defense unique, software program that will not be fully deployed until several software builds have been completed."[128]

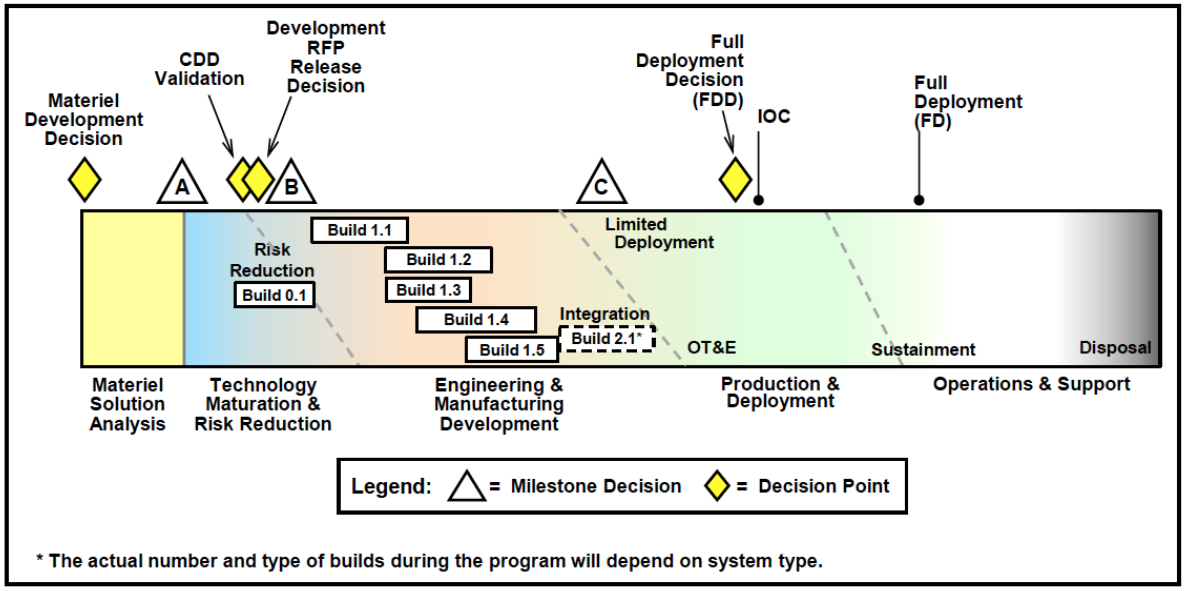

Figure 2: Model 2: Defense Unique Software Intensive Program. DOD Instruction 5000.02, Operation of the Defense Acquisition System, Section 5.c.(3) (c) January 2015,USD(AT&L)

From the initial identification of a need, through a capabilities requirements process or Initial Capabilities Document (ICD), the first decision point arises: a Material Development Decision. This is based on the requirements documents provided and a completion of an Analysis of Alternatives Study Guidance and Study Plan. This prompts the first phase of Material Solution Analysis (MSA), which is not as of yet a formal acquisition program (which can be delayed until Milestones B or C depending upon the technology and statutory requirements).

[128] Ibid, p. 12.

For AI systems, it is likely that the ICD will provide indicators or design documentation that calls for the specific needs of AI or even more specifically ML.

For the MSA, plans for development, testing and evaluation are required. This entails a need for knowledge of data sources, quality, the characteristics of a system (such as human-machine teaming or human-machine interaction), the required infrastructure of the system, the permissible and impermissible actions of the system, and an understanding of the capabilities for various missions. This ends with Milestone A.
For the Technology Maturation and Risk Reduction phase, T&E is already underway, at least in formal development and planning if not within one build. The purpose of this phase is to reduce cost risks for program development, as well as analyze technology maturation, and can be accompanied by tech demonstrations, competitive prototyping, as well as initial planning for sustainment and life cycle costs.[129] These also include testing for vulnerabilities, data needs and size, weight and power (SWAP) constraints.

Additionally, initial validation occurs at this phase to ensure that requirements are "technically achievable, affordable, and testable, and that requirements trades are fully informed by systems engineering trade-off analyses completed by the Program Manager or the DoD Component."[130]

Phase III, the Engineering and Manufacturing Development Phase (EMD), begins with a Request for Proposal (RFP) for the EMD. The RFP is issued to ensure risk reduction and assess rigorously whether the program is affordable, technologically sound, and executable. If RFPs are successful, then Milestone B is accomplished and provides authorization to begin EMD, the formal initiation of an acquisition program.

EMD Phase's purpose is to "develop, build, and test a product to verify that all operational and derived requirements have been met, and to support production or deployment decisions."[131] In short, much of the hard work on hardware and software engineering takes place during this phase. Moreover, additional DT&E occurs throughout this phase, as does initial OT&E. These early OT&E events are undertaken by Service Component test organizations, and as noted in 5000.02, should be planned in conjunction with one another.

Finally, during EMD, Program Managers complete designs for production support elements, looking to incorporate performance tests and sustainment plans that do not exceed the affordability caps. Depending upon the system under development, various other actors

[129] Ibid, pp. 21-23.
[130] Ibid, p. 23.
[131] Ibid, p. 27.

may also be involved during this phase, including Lead System Integrators and the need for advanced purchase of production items.

Given that many production and sustainment plans and choices are made during the EMD, Program Managers and various other authorities ought to strive for "concurrency" between the beginning of production stages and the completion of DT&E. EMD is completed when the Phase's purposes noted above are completed, and for software components they have sustainment processes in place and functioning. Completion is when Milestone C and limited deployment decisions are made, and a Production and Development RFP will be released.

Once a system enters into the Production and Deployment (P&D) phase, the goal is to deliver the requirements-compliant systems to the particular organization for fielded use by operational units. The phase consists of several different activities, such as limited deployment, OT&E, and the decision to engage in full-rate production or full-deployment. For software systems, limited deployment supports OT&E, and in some cases will provide the required operational testing metrics. OT&E in this phase is intended to be as realistic as possible for the system. In short, realistic environments are crucial.

After a full-rate production decision or full deployment decision, the cycle enters into its final stage: Operations & Support. As noted in 5000.02, "the purpose of the O&S Phase is to execute the product support strategy, satisfy material readiness and operational support performance requirements, and sustain the system over its life cycle (to include disposal)."[132] This phase is divided between Sustainment and Disposal.

Given that many acquired systems can be operational for decades, the best cost estimations, as well as the need for replacement components, needs to be specified correctly. Included here are license rights, tools, equipment and facilities needed for sustainment and maintenance. Upon the end of life cycle, the system is demilitarized and disposed.

All this said and as noted in the opening of this section, the Department is undergoing policy revisions to its software acquisition processes. It will soon shift to an Adaptive Acquisition Framework (see Figure 3 below). This framework is comprised of various pathways for acquisition, where programs are tailored to the particular characteristics and risks of the capability being acquired.

The software pathway underscores the critical role of automated testing in the software process, as well as iterative and incremental software development methodologies and a

[132] Ibid, 31.

DevSecOps approach. Moreover, complex systems may utilize multiple pathways during the acquisition cycle.

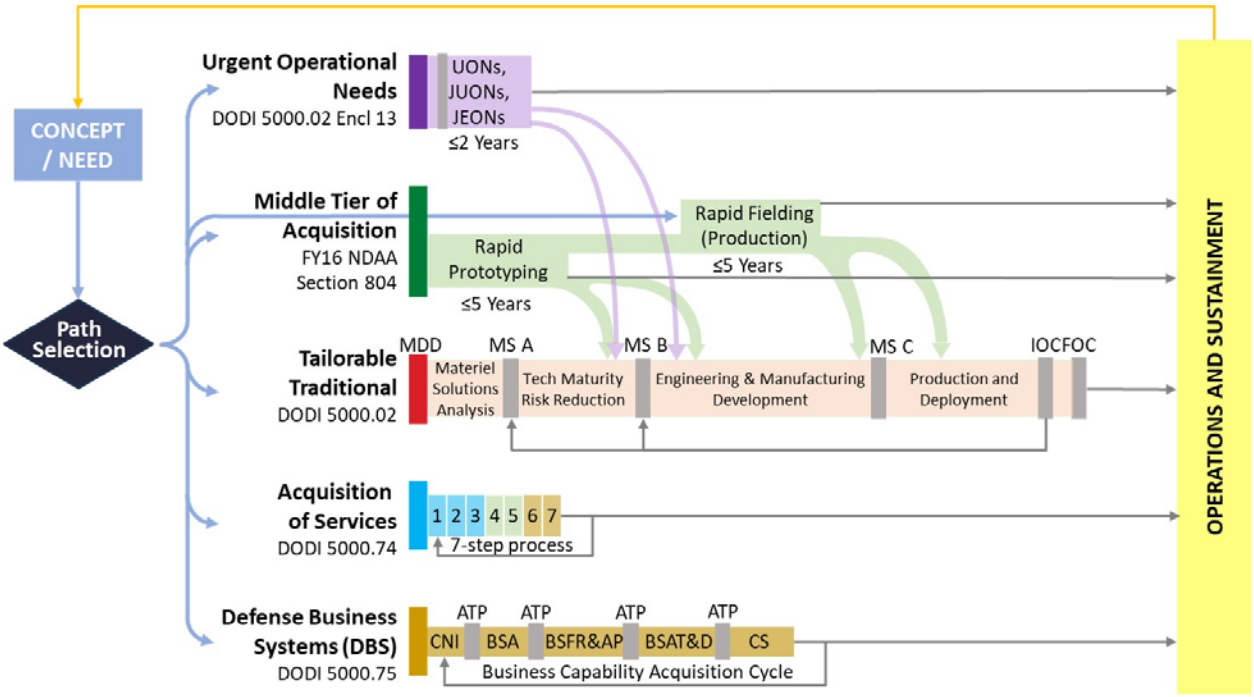

Figure 3: Adaptive Acquisition Framework

C. The Current State of Test & Evaluation for Artificial Intelligence in DoD[133]

Background

AI will have a significant impact on military operations and DoD must be prepared to adopt this technology within a clear ethical framework. As DoD adopts an ethics framework, T&E of AI systems is a critical enabler for that outcome.

The abundance of national security policies and directives that emerged in the past few years complicates and enables U.S. AI pursuits. The Defense Science Board (DSB) 2016 Summer Study on Autonomy identified the need to ensure trust and trustworthiness in AI

[133] The DIB asked the MITRE Corporation to conduct a high-level assessment of the current state of AI T&E within DoD and to offer recommendations for acceleration of AI T&E adoption. Section C of Appendix IV reflect MITRE's findings.

systems. The Study stated that DoD has a far-reaching responsibility "not only for the operator and the Commander, but also for designers, testers, policy and lawmakers, and the American public."[134] The DSB made recommendations to the T&E community, including the need to "establish a new paradigm for T&E of autonomous systems that encompasses the entire system lifecycle."[135]

In its 2018 DoD Artificial Intelligence Strategy, DoD highlighted the need to provide leadership in the areas of military ethics in AI safety. Specifically, the AI Strategy stressed the importance of "pioneering approaches for AI T&E, verification, and validation."[136]

Moreover, the 2019 National Defense Authorization Act (NDAA) tasked a senior DoD official with "the responsibility for the coordination of activities relating to the development and demonstration of artificial intelligence and machine learning (ML) for the Department."[137] The NDAA also mandated that this official, the Director of the JAIC, "work with appropriate officials to develop appropriate ethical, legal and other policies for the Department governing the development and use of artificial intelligence enabled systems and technologies in operational situations."[138]

Toward AI T&E Processes

The T&E process is a critical component of DoD's ability to acquire systems that deliver value to warfighters.[139] T&E provides evidence to users and decision makers that the acquired system is assured to operate as expected when deployed. The DoD community has years of experience developing program T&E requirements as well as the necessary hardware and software support tools. However, these current T&E processes and tools were largely designed without consideration of AI technologies.[140]

Since AI systems are software, they inherit the same T&E structure that software developers use, such as IEEE Std 1012.[141] IEEE Std 1012 provides a rigorous structure for the software T&E process. However, AI systems will be more difficult to evaluate than traditional software due to their heavy use of data and stochastic operational behavior.

[134] Defense Science Board (n 17).
[135] Ibid.
[136] Department of Defense (n 5).
[137] Congress (n 11).
[138] Ibid.
[139] DAU Guidebook, Chapter 8: Test and Evaluation.
[140] Defense Science Board (n 17) p. 30
[141] IEEE Standards Association, "Standard for System, Software, and Hardware Verification and Validation 1012-2016," New York: IEEE, 2016.

The assurance that AI systems will function as expected will require a holistic approach that tests and evaluates the AI software across its entire lifecycle: from data capture to operational deployment to sustainment. This process blurs the lines between the traditional Development T&E (DT&E) and Operational T&E (OT&E) approaches to software systems testing. There are several compelling reasons for evolving current practices:

1. The commercial sector has converged on several guidelines to test and evaluate its ML models. Industry adheres to a Continuous Integration Continuous Development (CICD) process,[142] [143] where multiple models are trained, tested, and deployed in live environments.
2. Testing of complex software systems is known to be computationally challenging because of increasing reliability requirements.[144] Adding AI-enabled software components will increase complexity as the components are "entangled" in a complex manner not seen in traditional software.[145] Real-time testing of AI software components will not cover all operational edge cases. Consequently, advances in modeling and simulation (M&S) will be required to establish operational envelopes in many DoD applications.
3. AI capabilities will require large, relevant data sets (both labeled and un-labeled) used to train ML-based software as well as test other types of AI decision-making algorithms. In many DoD use cases, these data requirements will be difficult to satisfy, as DoD does not have control over all aspects of its operational environments.

DoD could follow a feasible path toward assured AI systems adapted from the commercial sector, which would decrease the initial cost of T&E tool development. However, as discussed above, some use cases will require new investments to satisfy DoD's unique needs.

Comparing DoD and Commercial AI T&E Needs

Commercial organizations have spent the better part of a decade developing cloud computing and data harvesting platforms that facilitate their analytic pipelines. For example, Netflix, Amazon, Google, and Facebook apply mature ML algorithms to build recommendation systems. Two key contributors to the commercial sector's successful

[142] https://aws.amazon.com/getting-started/projects/set-up-ci-cd-pipeline/
[143] https://cloud.google.com/docs/ci-cd/
[144] The Infeasibility of Quantifying the Reliability of Life-Critical Real-Time Software, Butler and Finelli, IEEE Transactions on Software Engineering, January 1993.
[145] Software Engineering for Machine Learning: A Case Study, Amershi et al., CSE-SEIP '10 Proceedings of the 41st International Conference on Software Engineering, pp. 291-300. (2019).

application of AI technologies are: 1) centralized control over infrastructure and 2) development and deployment into low-risk environments. To develop and test mission-critical AI capabilities, DoD will require mission-ready infrastructure and testing tools to build assurance cases for its AI systems before operational use.

The commercial sector has also spent some ten years developing integrated data analytics pipelines over cloud computing infrastructures. Since commercial firms have closer control (from sensors to databases) over their information streams, their AI development and test teams can rapidly analyze, develop, and deploy solutions. The commercial sector does not confront challenges to the degree that DoD does with its infrastructure (e.g., legacy computing systems, multiple classification networks). Consequently, commercial solutions cannot be inserted unaltered into DoD environments to capture and curate operationally relevant data that is useful for technology development and T&E. Even if the infrastructure problem were solved, the diverse types of data collected for DoD missions present a challenge in terms of knowledge organization. Establishing data standards among DoD and other U.S. government (USG) agencies would enable better AI T&E methodologies.

When possible, DoD should repurpose commercial T&E methods and tools for use in the DoD enterprise. For example, Azure and AWS provide secure environments and tooling that DoD could leverage in the short term.[146,147] However, as noted, industry infrastructure and innovation will not apply in all DoD use cases. For example, live A/B testing is a common industry T&E approach on new ML models, but such a process would expose warfighters to unnecessary risk. The commercial autonomous vehicle community is pushing the boundaries of how T&E could function for safety-critical AI systems. This nascent work leverages new techniques in synthetic data generation and M&S. DoD could learn from this application domain what functions well (or does not) when deploying AI systems in higher risk situations.

Current State of DoD AI T&E

Recent commercial and academic advances in AI provide new options to help DoD meet some of its mission needs. Two technology areas being studied extensively are object classification, for example, by Project Maven,[148] and AI assistants, which provide situational awareness and recommendations to warfighters.[149] Since these two technology areas have

[146] https://aws.amazon.com/blogs/publicsector/announcing-the-new-aws-secret-region/
[147] https://azure.microsoft.com/en-us/blog/announcing-new-azure-government-capabilities-for-classified-mission-critical-workloads/
[148] https://www.defense.gov/Newsroom/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/
[149] https://www.armytimes.com/news/your-army/2018/07/26/want-siri-or-alexa-ready-for-tactical-ops-this-army-command-is-working-on-it

received the most attention, organizations have more efforts underway to develop methods and tools that support T&E. However, many developing AI-enabled capabilities have minimal or no T&E support tools. As DoD funds new capabilities, it will also need T&E tools that align with the technology development process.

Typically, technology development and acquisition are measured via Technology Readiness Levels (TRLs):[150]

- TRL 1–3: Basic research and establishing proof of concept. Service laboratories and OSD science and technology (S&T) organizations fund basic science exploration of AI, starting at TRL 1, and ending with lab-level proof of concept at TRL 3.
- TRL 4–6: Technology demonstration to prototype. Technologies at these TRLs in DoD are also typically handled by Service labs and Office of the Secretary of Defense (OSD) S&T organizations. However, this is also the point in the development cycle at which a program office or other acquisition authority might assume responsibility for the effort. By TRL 6, a subsystem or system prototype has been demonstrated in an operationally relevant environment.
- TRL 7–9: Engineering development to launch and operation. The final levels of development take a prototype technology and engineer it into a deployable capability.

Several DoD organizations have started to develop AI test tools. The following (non-exhaustive) list provides a sample of T&E process and tool development activities focused primarily on OSD investment:

- **Test Resource Management Center (TRMC):** TRMC is housed within the Office of the Assistant Secretary of Defense for Research and Engineering. It is congressionally mandated to provide the Major Range and Test Facility Base (MRTFB) with the technology required to test Programs of Record. The TRMC has augmented its T&E/S&T portfolio to include AI in the Autonomy and Artificial Intelligence Test (AAIT) Test Technology Area (TTA). This portfolio focuses on maturing technology, specifically from TRL 3 to TRL 6. In addition to its S&T investment, the TRMC advances technology from TRL 6 to TRL 9 in its Central T&E Investment Program (CTEIP).

- **Joint AI Center (JAIC):** The JAIC is a newly established center under the DoD Chief Information Officer (CIO) that is mandated to accelerate the adoption of AI

[150] Héder, Mihály. "From NASA to EU: The evolution of the TRL scale in Public Sector Innovation." *The Innovation Journal* 22.2 (2017): 1-23.

throughout the Department.[151] During FY19, it established several National Mission Initiatives (NMIs) to develop new AI-enabled capabilities. The JAIC designated AI T&E as one of its priorities and is investing in the creation of tools for the development and operational test of AI systems. The JAIC is also establishing the Joint Common Foundation (JCF), an infrastructure environment designed specifically for training, testing, and transitioning AI technologies intended to be available for use by all the Services.[152] As part of its investments over the next few years, JAIC is also incorporating red teaming and adversarial AI into its capabilities to meet the increasing demand for those technologies both as standalone technology and for T&E purposes.

- **Autonomy Community of Interest (COI) T&E Validation and Verification (V&V) Group:** OSD created the Autonomy COI to advance autonomous systems by assessing S&T investments, gaps, and opportunities, and initiating enabling technology development. Autonomous systems software uses AI technologies; thus, the COI's Test, Evaluation, Validation, and Verification (TEVV) process addresses the technological need to test and evaluate everything from algorithms to multi-vehicle teams.

- **Defense Advanced Research Projects Agency (DARPA):** DARPA, OSD's primary S&T organization, created the Explainable AI Program[153] to produce more explainable models and facilitate user understanding of and confidence in AI systems. The AI Next Program, focused primarily on the development of the next generation of AI/ML algorithms, will likely have T&E components, particularly with respect to the adversarial use of AI.[154]

Investments in AI T&E are not limited to OSD. The Services have their own AI programs, many of which include test components. The Services have also established multiple organizations and developed strategies that address AI testing needs. Examples include the Army AI Task Force,[155] the Navy Center for Applied Research in Artificial Intelligence located at the Naval Research Laboratory ,[156] and the Air Force Artificial Intelligence Annex to the DoD AI Strategy.[157]

[151] https://dodcio.defense.gov/About-DoD-CIO/Organization/JAIC/
[152] https://www.fedscoop.com/jaic-fiscal-2020-jack-shanahan/
[153] https://www.darpa.mil/program/explainable-artificial-intelligence
[154] https://www.darpa.mil/work-with-us/ai-next-campaign
[155] https://www.army.mil/article/225642/ai_task_force_taking_giant_leaps_forward
[156] https://www.nrl.navy.mil/itd/aic/
[157] https://www.af.mil/News/Article-Display/Article/1959225/air-force-releases-2019-artificial-intelligence-strategy/

Addressing technologies at TRLs 1–6, the Service labs investigate T&E of AI as part of their efforts to create new AI technologies. Service AI developers will need to execute T&E activities that ensure that the AI technology developed meets program requirements.

U.S. Special Operations Command (USSOCOM) is exploring industry-aligned paradigms for performing AI T&E. For example, USSOCOM is experimenting with commercial-inspired CICD processes, wherein capability developers work in a tight loop with operators. This tightly coupled developer-user process facilitates rapid T&E that has informed an alternative method to increase the speed and assurance of AI systems. Furthermore, USSOCOM elements are investigating ways to change the Command's workforce to represent multiple levels of AI technology knowledge: from aware leaders to technology developers. The Army and Air Force have also started exploring these structural changes through their AI acceleration engagements.

Consideration of Human-AI Interaction

Many of the deployed AI capabilities that receive the most attention are "black box" models, such as deep neural networks. Modern neural networks are complex pattern recognition models that contain millions of parameters. When such systems make incorrect predictions, only an expert can perform deeper analysis of the parameters to determine their validity. DoD T&E teams will not have the same level of expertise as the model builders and will therefore require tools that augment their understanding of the models' behavior.

DARPA's recent Explainable AI Program provided a funding boost that advances such explainability methods, which produce approximate models that allow non-experts to gain some insight into the model. However, these methods do not provide intuitive detail about the learned model. The program has alternative approaches to develop more transparent models that still exhibit good performance but are more readily understood by operators. Combining these transparent models with post-hoc explanations would facilitate the T&E measurement process and provide new avenues to assure AI systems. The DoD DT&E and OT&E communities should maintain awareness of OSD investments in interpretable AI technologies.

Observations on DoD's AI T&E Capabilities

The current DoD infrastructure and processes are not yet equipped to handle the rapidly changing environment of AI and the associated testing at an enterprise level. Some organizations, particularly in OSD, are starting to change that landscape, but they cannot by themselves effect the alterations needed to create an AI-enabled fighting force.

DoD lacks AI T&E tools for validation of AI/ML models. A primary issue is the gap (often called the "valley of death") between late S&T efforts and delivery of capability to the warfighter (i.e., many efforts that reach TRL 6 have trouble navigating the DoD acquisition system to reach TRL 9). Several factors contribute to this problem, including:

1. **Lack of investment in AI T&E tools:** Standard software pipelines and tools may be used for many aspects of verification but will not be applicable to validating models (e.g. through explainability of interpretability). An exception is that TRMC has started to establish efforts in S&T for AI T&E for TRLs 3–6.
2. **Few early S&T investments:** With some exceptions, DoD organizations have made relatively few investments in the development of TRL 1–3 AI S&T tools. Within OSD, the DARPA Explainable AI Program represents a notable exception, though DARPA does not have a program focused on test of AI.
3. **Service and OSD coordination:** The Services comprise the users who require AI solutions for their problems and have the program offices for acquisition. OSD is working to establish centralized AI development and AI T&E solutions. This means that the Services and OSD need to coordinate to avoid duplicated application of DoD resources in the establishment of AI T&E solutions.

AI T&E development is in its early stages, so DoD has time to create the bridges that cross this "valley" and create infrastructure using established OSD organizations. The notable exception is that it seems no OSD organization has a program dedicated to developing TRL 1–3 test technology for AI, although DoD can address this shortcoming with strategic investment.

<u>Recommendations to Accelerate Adoption</u>

The recommendations included in this section represent a concise response to the gaps discovered through the preliminary analysis summarized in this paper. All recommendations focus on changes at the OSD organization level, as a comprehensive survey of the Service AI T&E efforts is out of scope for the MITRE effort. The list below presents three major recommendations with more detailed sub-recommendations:

1. Establish Memoranda of Agreement (MOA) between the appropriate OSD organizations to create lanes for AI T&E development.
    1.1. A full TRL 1-9 path is needed within OSD for AI T&E tool development, with separate organizations given the mandate and funding for development and acquisition.

1.2. In working to take technologies from TRL 7 to TRL 9, OSD should establish formal relationships with the Service test commands and program offices to cross the technology "valley of death" for AI T&E. This should include working with the program offices to establish tool requirements that filter down to TRL 3–4 tool S&T efforts.

2. Increase investment in the early science of AI T&E
    2.1. OSD should provide funding for a TRL 1–3 program (e.g., under DARPA or TRMC) to study the science of AI test for validation (e.g., explainability and interpretability).
    2.2. OSD should fund the development of ground truth sets for operationally relevant DoD problems. This effort should be executed in conjunction with the Service program offices.
    2.3. OSD should mature the test technology for non-classification (e.g., AI assistants and red teaming) ML systems.
3. Leverage relevant commercial sector infrastructure and practices.
    3.1. The JAIC has made initial investments in establishing the JCF, but the infrastructure is not ready for full-scale deployment and use by OSD and the Services. OSD should ensure that the relevant organizations (e.g., JAIC, TRMC, DOT&E, etc.) have the resources they need to develop a modern environment for centralized AI T&E.
    3.2. OSD should provide resources and coordinate among the Service program offices/test commands, OSD/Service T&E organizations (e.g., JAIC, the MRTFB, etc.), and industry partners to establish data standards and practices for testing of AI.
    3.3. OSD should mandate that developed test tools conform to the industry standard CICD processes that have proven successful in the commercial world.

Appendix V: Overview of Other AI Ethics Principles

Over the past several years, increasing attention has been paid to generating AI Ethics Principles in one sort or another. The public sector, including the OECD, European Commission, the United Kingdom House of Lords, and ministries or groups from the governments of Germany, France, Australia, Canada, Singapore, and Dubai, have all formulated AI ethics or governance documents. The private sector, such as Microsoft, Google, OpenAI, the Partnership on AI, and IBM, has also provided their own approaches to AI ethics and governance. Likewise civil society groups, professional societies and other multi-stakeholder groups of experts and academics continue to issue reports and guiding principles for ethical AI. Some of these groups include the Future of Life Institute's Asilomar Principles, the IEEE, Amnesty International's Toronto Declaration, the Montreal Declaration, and many more. The closest set of "guiding principles" that may have relevance for the DoD AI Ethics Principles outlined in this paper are those put forward by the 2019 United Nations Convention on Certain Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons. Those principles are affirmed by the States Parties at the 2019 August meeting in Geneva. However, those directly relate to weapon systems and emerging weapon systems, and do not consider the use of AI more broadly by states' militaries.

As these activities gain in prominence, more sets of principles continue to emerge from all corners of governments, industries, civil society and even academics. Indeed the proliferation of AI Ethics Principles projects has now garnered a need for mapping all of the existing approaches. Recently, Harvard's Berkman-Klein Center announced it's new project "Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches" to highlight commonalities and gaps.[158] They find commonalities around nine different categories in these documents: human rights, promotion of human values, professional responsibility, human control of technology, fairness and non-discrimination, transparency and explainability, safety and security, accountability and privacy. More recently, researchers at ETH-Zurich conducted an analysis of whether there is a "global agreement" on what constitutes ethical AI. They found that there is broad consensus around 5 principles: transparency, justice and fairness, non-maleficence, and responsibility and privacy.[159] However, there was substantial disagreement about what the content of each principle contains and how each was interpreted. In effect, this finding leaves open

[158] Fjeld, Jessica. Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Joshua Feldman and Sally Kagay. 2019. "Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches."
[159] Jobin, Anna. Marcello Ienca and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines" *Nature Machine Intelligence*, 2 September.

questions as to terminological bandwagoning, cultural or organizational relativism, or even mere self-interest.

Each set of principles will undoubtedly reflect the particular values of the organization or group from which they come. But the myriad of principles suggests that societies across the globe are concerned with the responsible and ethical use of AI technologies in all facets of people's lives.

A set of AI Ethics Principles for DoD must likewise reflect the values and commitments of the Department. In many respects, there are overlaps and commonalities with some of the other proffered approaches. However, DoD differs in many substantive ways that we have highlighted here. What is noteworthy when canvassing the plethora of available AI Ethics Principles documents is that there is no other military in the world that has offered its approach to ethical design, development, and deployment of AI systems.[160] In this respect, DoD is leading in this space, showing its commitments to ethics and law.

[160] Some countries, such as France, have announced AI strategies for national security that include ethics as a part of the strategy, but have not developed an overarching ethical framework as the foundation of military decision-making or behavior related to AI. Other countries, such as Germany, have mentioned AI as part of an ethical set of principles they have developed, but have focused on data, not AI specifically, as the locus of the principles, or have concentrated on applying these principles to society in general, not national security.