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

Defense Innovation Board

I. Purpose

The leadership of the Department of Defense (DoD) tasked the Defense Innovation Board (DIB) with proposing Artificial Intelligence (AI) Ethics Principles for DoD for the design, development, and deployment of AI for both combat and non-combat purposes. Building upon the foundation of DoD's existing ethical, legal, and policy frameworks and responsive to the complexities of the rapidly evolving field of AI, the Board sought to develop principles consistent with the Department's mission to deter war and ensure the country's security. This document summarizes the DIB's project and includes a brief background; an outline of enduring DoD ethics principles that transcend AI; a set of proposed AI Ethics Principles; and a set of recommendations to facilitate the Department's adoption of these principles and advance the wider aim of promoting AI safety, security, and robustness. The DIB's complete report includes detailed explanations and addresses the wider historical, policy, and theoretical context for these recommendations. It is available at innovation.defense.gov/ai.

The DIB is an independent federal advisory committee that provides advice and recommendations to DoD senior leaders; it does not speak for DoD. The purpose of this report is an earnest attempt to provide an opening for a thought-provoking dialogue internally to Department and externally in our wider society. The Department has the sole responsibility to determine how best to proceed with the recommendations made in this report.

II. Background

Why should DoD prioritize AI ethics? AI is transforming our society and affecting the ways in which we do business, interact socially, and conduct war.[1] In many respects, the field of AI is in its adolescence. Recent rapid advances in computing have enabled progress on AI applications that have for decades been only theoretical. Nevertheless, practical applications of AI are often brittle and the discipline of AI development is evolving, leaving the norms of AI use inchoate. Globally, the public sector, private industry, academia, and civil society are engaging in ongoing debates over the promise, peril, and appropriate uses of AI. National security is a crucial facet of these debates. Now is the time, at this early stage of the resurgence of interest in AI, to hold serious discussions about norms of AI development and use in a military context – long *before* there has been an incident.

Our adversaries and competitors have recognized the transformative potential of AI and are investing heavily in it by modernizing their forces while actively engaging in provocative activities around the globe. China has forcefully and publicly committed to becoming the global leader in AI by 2030, and it is spending billions of dollars to gain advantage.[2] Russia is likewise investing heavily in AI applications and testing those

[1] 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.

[2] Savage, Luiza Ch., and Nancy Scola. "'We are being outspent. We are being outpaced': Is America ceding the future of AI to China?" Politico. 18 July 2019.

systems in live combat scenarios.[3] The overwhelming reality for the future of DoD is clear to Lieutenant General Jack Shanahan, Director of the Joint AI Center (JAIC): "What I don't want to see is a future where our potential adversaries have a fully AI-enabled force and we do not… I don't have the time luxury of hours or days to make decisions. It may be seconds and microseconds where A.I. can be used."[4]

The 2018 National Defense Strategy (NDS) calls for greater investments in AI and autonomy to provide the United States with competitive military advantages.[5] The DoD's AI Strategy, aligned to the NDS, asserts that DoD is committed to harnessing the potential of AI to "transform all functions of the Department positively, thereby supporting and protecting service members, safeguarding U.S. citizens, defending allies and partners, and improving the affordability, effectiveness and speed of our operations."[6] The AI Strategy further notes that it "will articulate its vision and guiding principles for using AI in a lawful and ethical manner to promote our values."[7] Stressing 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 underscored its commitments to the ethical and responsible development and deployment of AI. *"Leading in military ethics and AI safety"* is indeed one of the five pillars of the strategy.

DoD is not the first organization to recognize the importance of producing ethics principles for the development and use of AI, but the DIB has observed that many existing sets of such principles have generated more questions than answers about the limits of how AI might be used. Within the high-stakes domain of national security, it is important to note that the U.S. finds itself in a technological competition with authoritarian powers that are pursuing AI applications in ways inconsistent with the legal, ethical, and moral norms expected by democratic countries. Our aim is to ground the principles offered here in DoD's longstanding ethics framework – one that has withstood the advent and deployment of emerging military-specific or dual-use technologies over decades and reflects our democratic norms and values.

However, we acknowledge that AI's unique characteristics and fragilities require new ways to address its potential unintended negative consequences.[8] In the national security arena, analysis of unanticipated behavior is key when considering whether to field emerging technology. The uncertainty around unintended consequences is not unique to AI; it is and has always been relevant to all technical engineering fields. For example, humans built bridges and buildings and manipulated energy and physical materials before the respective

[3] Konaev, Margarita, and Samuel Bendett. "Russian AI-enabled combat: coming to a city near you?" War on the Rocks. 31 July 2019.
[4] See "Lt. Gen. Jack Shanahan Media Briefing on A.I.-Related Initiatives within the Department of Defense." Department of Defense, 30 August 2019.
[5] National Defense Strategy of the United States, 2018.
[6] Department of Defense, Summary of the 2018 Department of Defense Artificial Intelligence: Harnessing AI to Advance Our Security and Prosperity. p. 4. (Hereafter "AI Strategy").
[7] Department of Defense (n 6) 8.
[8] See Supporting Document for detailed descriptions of how AI presents different kinds of ethical challenges than other technologies do, as well as for further in-depth discussion of the principles.

fields of civil and chemical engineering crystallized as formal disciplines, leading to many unforeseen accidents.[9] Today, despite the lack of agreed-upon ways to use AI that maximize societal benefit and curtail unintended consequences, "humans are proceeding with the building of societal-scale, inference-and-decision-making systems that involve machines, humans and the environment."[10] AI ethics principles should therefore enrich discussions about how to advance the still-nascent field of AI in safe and responsible ways. There is a parallel here in the way engineering disciplines like civil and chemical engineering developed their cultures of ethical behavior. They did so by identifying and upholding obligations of technical excellence on the part of their practitioners.

Given the ongoing global debates over when and under what circumstances employing AI in a national security context is appropriate, it is essential to note that DoD has a duty to the American people and its allies to preserve its strategic and technological advantage over competitors and adversaries who would use AI for purposes inconsistent with the Department's values. We intend for the ensuing principles to serve as a guide for this effort, while DoD continues its timeless commitment to lawful and responsible behavior; builds on its existing ethical foundation by translating and adapting ethics to the field of AI; helps shape new international norms on AI use; and ensures we simultaneously capitalize on the technology's benefits while mitigating its potential harms.

A Diversity of Views. To aid the Department in this challenge, DoD leadership tasked the DIB to engage a broad set of audiences to deliver recommendations on possible AI ethics principles for DoD and how those principles might be integrated into the existing ethics framework under which the Department executes its mission.

The DIB conducted a 15-month study designed to be robust, inclusive, and transparent. The process involved collecting public commentary both online and in person; holding two public listening sessions at major universities; and facilitating three expert roundtable discussions with dozens of subject matter experts in academia, industry, civil society, and the Department. Roundtable participants included Turing Award-winning AI researchers, retired four-star generals, human rights attorneys, political theorists, arms control activists, tech entrepreneurs, and others.[11] Additionally, the Department formed an informal DoD Principles and Ethics Working Group, including government officials from close partner nations, to assist the DIB in gathering information and promoting cooperation. The DIB also held a classified "red team" session and a Table Top Exercise to pressure-test the principles in realistic policy scenarios and against current intelligence about potential applications of AI in warfare. After thoughtful consideration of the input of more than 100 internal and external experts, reflecting a wide range of perspectives and almost 200 pages of submitted public comments,[12] the DIB developed this set of proposed AI Ethics Principles and accompanying recommendations for consideration by the Secretary of Defense. These principles – which are intended to be specific to AI – nest

[9] See Dr. Michael Jordan's view of AI as a potentially new engineering discipline.
[10] Ibid.
[11] A complete list of participants (who approved the inclusion of their name) is listed in an appendix of the corresponding Supporting Document and on the DIB website.
[12] See links to and videos of public comments on DIB website.

within the context of the existing ethical, legal, and policy frameworks the Department uses to guide its activities.

Defining AI. Artificial intelligence is an extremely broad discipline, defined in many different ways for many different purposes.[13] For clarity and to guide this project, we use the term to mean *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.* When referring to the wider range of considerations, we use the term artificial intelligence (AI); however, where we specifically address machine learning (ML) systems, we refer to ML. Furthermore, we use the term "AI system" to mean systems that have an AI component within an overall system or a system of systems.[14]

We use this definition of AI 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 and newer ones such as those using ML. The use of this term allows us to reinforce that the earlier and important AI work achieved by DoD took place within DoD's existing ethics framework outlined below.

We also distinguish and make clear that *AI is not the same thing as autonomy.* While some autonomous systems may use AI in their software architectures, this is not always the case. For example, DoD Directive 3000.09 addresses autonomy in weapons systems, but it neither addresses AI as such nor AI capabilities not pertaining to weapon systems.[15]

Finally, AI is neither inherently positive nor negative.[16] It is an enabling capability, akin to electricity, the internal combustion engine, or computers, and as such, it is the decisions of human beings that will determine whether AI will advance or undermine our efforts to make the world safer and more prosperous.

[13] The 2018 DoD Strategy on AI defines AI to be: "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." Our definition does not preclude the DoD AI Strategy's approach, but allows for a wider range of AI applications that do not require a human intelligence benchmark. We have an extended discussion of the various definitions for AI in several policy documents in the Supporting Document.

[14] Some documents prefer the phrase, "AI-enabled systems," but for our purposes, an AI system or an AI-enabled system would be the same.

[15] See DoD Directive 3000.09, which defines an autonomous weapon system as "a weapon system that, once activated, can 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."

[16] This is not to say that technology, and thus AI, is value-neutral. Technological artifacts, like AI systems, reflect the values of their human designers, developers, and users, as well as the societies in which they reside and make decisions.

III. Existing DoD Ethics Frameworks and Values

AI is and will be an essential capability across DoD in non-combat and combat functions. Indeed, AI constitutes just one of the many technologies used by the Department and presents testing and fielding challenges similar to other large, technically complex systems that DoD has deployed safely and successfully. In all of these cases, the values-based framework under which DoD and the Services operate[17] and the legal constructs under which DoD and the U.S. civil society operate, including the U.S. Constitution, Title 10 of the U.S. Code, and other applicable laws, provide the foundation on which any AI ethics principles must function. The proposed AI-specific principles outlined below arise from *existing and widely accepted* ethical and legal commitments.

This well-established ethical framework and its accompanying values guide DoD in how it makes and executes decisions. Evidence for this 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 importance of ethical behavior across the armed services. In isolation and taken together, this body of evidence shows that DoD's ethical framework reflects the values and principles of the American people and the U.S. Constitution.[18] [19]

Of particular importance is DoD's commitment to uphold the Law of War, as it is an internationally recognized legal guide for the conduct of all armed forces.[20] For the U.S., this body of law includes treaties the U.S. has accepted, such as the 1949 Geneva Conventions; customary international law, which results from the general and consistent practice of States done out of a sense of legal obligation; and the DoD Law of War Manual.[21]

Existing Law of War rules can apply when new technologies are used in armed conflict.[22] For example, the 2015 DoD Law of War Manual reflects work done in 2012 in connection with DoD Directive 3000.09, to elaborate on how the Law of War applies to the use of autonomous functions in weapon systems.[23] The fundamental principles of the Law of War provide a general guide for conduct during war, where no more specific rule applies, and thus provide a framework to consider novel legal and ethical issues posed by emerging technologies, like AI. For example, if AI was 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

[17] See Department of Defense Core Values, US Air Force Core Values, US Army Core Values, US Navy and Marine Corps Core Values, and US Coast Guard Core Values.
[18] 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.
[19] See memo from Secretary Mark Esper and memo from former Secretary James Mattis.
[20] While the Law of War is an important guide for DoD, it does not apply to all situations in which the Department might apply AI. We describe these situations in more depth in the Supporting Document.
[21] See DoD Law of War Manual.
[22] These rules are based on five fundamental principles that serve as the foundation of the Law of War: military necessity, humanity, proportionality, distinction, and honor.
[23] DoD (n 18) 395.

indiscriminate. Additionally, under the Law of War, commanders and other decision-makers must make decisions in good faith and based on the information available to them and the circumstances ruling 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 and objects.[24]

DoD has robust processes to implement the Law of War, including training, regulations and procedures, reporting of incidents involving alleged violations, investigations and reviews of incidents, and appropriate corrective actions.[25] To complement and facilitate these actions, DoD has invested hundreds of billions of dollars in the last half-century to ensure the safety and reliability of its weapons systems and platforms to create more precise and accurate weapons that reduce civilian casualties and protect civilian infrastructure while still 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.[26]

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 technically complicated efforts, like the development and deployment of AI.[27]

IV. AI Ethics Principles for DoD

We reaffirm that the use of AI must take place within the context of the existing DoD ethical framework. Building on this foundation, we propose the following principles, which are more specific to AI, and note that they apply to both combat and non-combat systems. AI is a rapidly developing field, and no organization that currently develops or fields AI systems or espouses AI ethics principles can claim to have solved all the challenges embedded in the following principles. However, the Department should set the goal that its use of AI systems is:

[24] DoD (n 18) § 5.2.3.2 and 5.3.

[25] See DoD Directive 2311.01E.

[26] National Nuclear Security Administration and Department of the Navy factsheet, "United States Naval Nuclear Propulsion Program." September 2017.

[27] For a deeper description of all aspects of the Department's existing ethical framework, please see the Supporting Document.

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.

V. Recommendations

In the course of its work developing these proposed AI Ethics Principles, the DIB has identified useful actions that can aid in the articulation and implementation of these proposed principles. DoD will ultimately determine the exact principles it wishes to adopt, but regardless of the exact nature of the principles approved, the following twelve recommendations will support that effort:

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[28] to more tangible AI-related areas like data sharing, cloud computing, human capital, and IT policies.

[28] A term that refers to DoD Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, and Facilities.

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, Office of the Secretary of Defense Component, defense agency, and defense field activity should establish programs for training and education that are relevant to their respective DoD personnel in AI-related skills and knowledge.[29] 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.[30] 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.

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.[31] This effort will also help address the so-called "black box" problem with AI.[32]

[29] See DoD AI Strategy on p. 14 ("Providing comprehensive AI training and cultivating workforce talent").

[30] Ibid.

[31] 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.

[32] 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.

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.[33] [34]

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.[35] 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.

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 research into 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.[36]

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.

[33] See SWAP Study.

[34] 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.

[35] 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.

[36] See DoD AI Strategy on p. 15 ("Investing in research and development for resilient, robust, reliable, and secure AI."). National AI R&D Strategic Plan: 2019 Update, esp. Strategy 1 ("Make Long-Term Investments in AI Research") and Strategy 4 ("Ensure the Safety and Security of AI Systems").

VI. Conclusion

These principles are designed neither to gloss over contentious issues nor restrict the Department's capabilities. Rather, we intend that these principles should enable AI systems and operations that are aligned with DoD's mission to deter war and protect our nation. Further, these principles are consistent with existing policy frameworks, the Law of War, domestic law, such as Title 10 of the U.S. Code, and enduring ethical norms that reflect democratic values. The DIB proposes these AI Ethics Principles for consideration by DoD, including recommendations for their implementation. After all, ethics is not merely a collection of ideas as much as it is a set of purposeful activities and ongoing inquiries.

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 evolving, the Department's commitment to the laws of the United States, the Law of War, and democratic values 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.