Future of Life Institute
PO Box 454
Winchester, MA 01890

March 4, 2019
To: Defense Innovation Board, Science & Technology Subcommittee

Comments for DIB meeting on "The Ethical and Responsible Use of Artificial Intelligence for the Department of Defense (DoD)"

We appreciate the opportunity to provide public comments to the Defense Innovation Board's (DIB's) public listening session titled "The Ethical and Responsible Use of Artificial Intelligence (AI) for the Department of Defense (DoD)." Building upon the vision articulated in E.O. 13859, and the DoD's 2018 Artificial Intelligence Strategy (the 2018 AI Strategy), we believe the DIB's role in developing the "AI Principles for Defense" is a critical next step toward assuring the responsible and ethical use of AI. To that end, we are providing several practical summary recommendations for the DIB's consideration, and look forward to the opportunity to engage in a productive discourse in the future. We are aware that the DIB has a robust plan for continuing outreach and consultation during the development of these AI principles, and we would be happy to participate as desired. Please contact Jared Brown at jared@futureoflife.org for additional background information on these recommendations or to arrange further consultation.

1. **Adopt and translate the widely endorsed Asilomar AI Principles for the ethical and responsible use of AI by the DoD.** The 23 Asilomar AI Principles were developed by the Future of Life Institute in 2017 through a consultative process and have since been signed by more than 3,700 AI and robotics researchers and others. In August 2018, the Principles were also endorsed by the State of California. Several of the fundamental Asilomar AI Principles are highly relevant and important for the development and use of AI systems by the DoD. For example, the Principles state, "AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible," and that "If an AI system causes harm, it should be possible to ascertain why." In general, guiding principles for the use of AI in the military should include transparency, accountability, robustness, fairness, precaution, human dignity, and the common good.

2. **Maintain distinct directives on AI in weapons systems while creating broader DoD directives, principles, and other guidance that encompass the use of AI in non-weapon system applications.** It is advisable for DoD to develop overarching directives on the ethical and responsible use of AI in all manner of purposes across the national security enterprise, including those identified in the 2018 AI Strategy such as to streamline business operations and increase the safety of operating equipment. However, more specific directives, such as DoDD 3000.09 on Autonomy in Weapon

Future of Life Institute
PO Box 454
Winchester, MA 01890

Systems, should continue to exist and be reformed given the unique ethical considerations presented by within the Law of War and the extreme risk of unintended engagements. The more specific guidance on the use of AI in weapons systems should adhere to significantly higher standards for AI explainability and predictability, and take steps to counteract the ways in which automation could lower the threshold for military action by creating anonymity and psychological distance from conflict.

3. **Human judgment and control should always be preserved in the use of weapons systems, and DoD should advocate for this principle to be adopted internationally.** The future AI Principles for Defense must continue to ensure, as stipulated in DoDD 3000.09 on Autonomy in Weapon Systems, that commanders and operators can "exercise appropriate levels of human judgment over the use of force." Further, the DoD should advocate for the inclusion of this standard by international partners (e.g., within NATO) and by our near-peer adversaries.

4. **Prior to deployment, critical AI systems should be subject to rigorous verification and validation (V&V) and operational test and evaluation (T&E), including with adversarial examples, with the intent to manipulate the system into recommending unethical decisions.** It is essential that critical AI systems, such as those designed to assist the use of lethal weapon systems, be subject to rigorous testing with adversarial examples, perhaps through red teaming. For example, foreign combatants have been known to use civilian facilities, such as schools, to "shield" themselves from attack when firing long-distance munitions (e.g., rockets). An AI-system designed to support targeting acquisition of such combatants must be intentionally tested to try and provoke it to recommend unethical decisions, such as a recommendation to engage when collateral damage will be unacceptably high. V&V and T&E testing for AI systems should ensure reliability and alignment with human preferences, robustness against attack, protections from misuse, and close monitoring of the intersection of AI with other weapons systems such as nuclear control and command.

5. **Recognize the technical and other limitations of AI systems and identify unacceptable uses.** All existing AI systems are prone to adversarial attacks, bias, reward hacking, lack of explainability, and misuse, among other safety and ethical challenges. It is essential that the DoD exercise precaution in the integration of AI systems into military and national security processes. Particular attention should be paid to avoiding the use of "black box" or unexplainable systems in critical decision making. Steps should also be taken to prevent the use of AI to amplify the spread of

Future of Life Institute
PO Box 454
Winchester, MA 01890

disinformation and terrorist propaganda, as well as to support limitations on surveillance in order to protect the privacy and civil liberties of all Americans.

6. **DoD guidance on and safety measures for AI systems should be transparent and regularly communicated with the international community.** The 2018 AI Strategy appropriately emphasizes the importance of "promoting transparency in AI research" to "promote responsible behavior" and the need to advocate for "a global set of military AI guidelines." It is equally important for there to be universal transparency regarding DoD guidance on and safety measures for AI systems, especially as used in any weapons systems. Transparency about guidelines and doctrine would encourage other international actors to behave likewise and help prevent a "race to the bottom," a danger that could be exacerbated if weapons innovation becomes driven more directly by the software (rather than hardware) development timescale. By providing transparency about DoD's responsible and ethical approach to the development and deployment of AI, DoD would serve as a global and ethical leader.

7. **Civilian and military operators of critical AI systems should receive specialized training in machine ethics and on AI safety principles.** We are encouraged by the prominent inclusion of workforce training considerations in the 2018 AI Strategy. However, the unclassified summary of the Strategy does not specifically identify machine ethics or AI safety as part of this potential curriculum. As civilian and military personnel begin to more frequently interact with and receive support from AI systems, these operators must have an advanced understanding of machine ethics and AI safety principles in order to recognize potential unethical or irresponsible outcomes from the use of the AI system. Trained personnel should be able to recognize the limitations of AI technology and be cognizant of a human tendency to follow the guidance of machines, even when the software gives flawed or unethical suggestions. The training should be updated regularly, and operators should recertify their training frequently, as AI systems advance in complexity and the fields of machine ethics and AI safety evolve. Parallel support for research on the ethical and societal implications of AI in the military can also support ongoing improvements in this training.

8. **The DoD (e.g., the JAIC) should maintain a central unclassified and classified inventories of how, where, and for what purpose different AI systems are developed for national security purposes, including all National Mission Initiatives (NMIs) and Component Mission Initiatives (CMIs).** We have reservations regarding the desire articulated in the DoD's 2018 AI Strategy to enable "decentralized

Future of Life Institute
PO Box 454
Winchester, MA 01890

development and experimentation" at the "forward edge" in order to "scale and democratize access to AI." While well intentioned, overly decentralized development and experimentation may quickly lead to applications of AI systems for tasks they were not specifically designed for at the "forward edge." This can result in unintended, unethical, and unsafe outcomes. As briefly implied in the Strategy, such unintended outcomes could also occur as an emergent effect of the interaction of two or more AI systems, especially if one or more of those systems is being used in novel, unanticipated ways at the forward edge. To monitor and protect against these potential outcomes, the JAIC should maintain centralized inventories of developed AI systems. These inventories should be made available for independent oversight (e.g., DoD's Office of Inspector General (OIG) and Congressional committees) and should include information on the design and acceptable uses of all AI systems, ranging from those with relatively mundane purposes (e.g., CMIs involving specialized AI systems assisting with language translation for combatant commands) to the more consequential (e.g., NMIs involving specialized AI systems for cyberdefense and SIGINT analysis). These inventories should specify any and all exemptions from DoD guidance granted in the approval process for the AI system, NMI, or CMI. As research develops in AI safety and machine ethics, and DoD adopts new policy accordingly, these inventories will also facilitate the deployment of updates to all relevant AI systems to maintain proper ethical and responsible use.

9. **Any AI-related directives or other guidance should be required to be updated on a biannual basis at a minimum. An independent entity (e.g., DoD OIG or the DIB) should be given explicit authority to request reviews and potential updates to guidance on an as-needed basis.** Emerging technologies such as AI and machine learning often develop in unpredictable ways at an exponentially increasing speed. In acknowledgment of this fact, any directives, principles, or other guidance related to the ethical and responsible use of AI may become outdated quickly. Given a natural tendency to bureaucratic inertia, an independent entity should be able to order the review of guidance to address relevant changes in AI safety, machine ethics, or other research. Such reviews should acknowledge emerging international AI norms and principles and seek to align national guidance where possible.

10. **Robust public-private partnership, including engagement with diverse stakeholders and communities, should be prioritized.** Much of the development of AI is taking place in private and academic settings, while its use is already widespread. The DoD should support information sharing between sectors to help establish more reliable systems and prevent malicious use. Establishing opportunities for feedback from

Future of Life Institute
PO Box 454
Winchester, MA 01890

stakeholders and communities will additionally help protect the DoD from public backlash.

11. **Increase R&D spending on research into the comprehensive sociological, psychological, and political effects of using AI systems for national security, not just to how to improve the underlying AI technologies.** While increased spending on technical safety matters by DoD is extremely welcome, ensuring the eventual ethical and responsible use of AI also requires understanding the sociological, psychological, and political effects of using AI systems for various national security purposes. For example, as stated in the 2018 AI Strategy, it is often assumed, but not proven, that using certain AI technologies *may* "provide commanders more tools to protect non-combatants via increased situational awareness and enhanced decision support" to "reduce the risk of civilian casualties and other collateral damage." However, such a result does not depend solely on the technical capabilities and safety of the AI system. Rather, it also depends on understanding how using the AI systems ultimately influence: the individual behavior and decision-making of commanders and others using the AI systems (psychological research), the behavior of other combatants, non-combatants, institutions, and cultures interacting with the commander (sociological research), and the geopolitical responses that may result from the use these systems (political science research).