Public Comments Ahead of April 25, 2019 Defense Innovation Board Listening Session

Table of Contents

1) **David Kerley, Big Kahuna Technologies LLC**

Moral behavior of AI(s) is unlikely to be achieved without a causal dynamic model capability. Will the DoD adopt a mandate and pursue the necessary research and development required to achieve AI solutions that have access to or incorporate integral models of causality (cause and effect) such that a moral basis of behavior can be trained, monitored, and explained for any deployed system?

2) **Glenn Keselman, Private Sector**

What steps is the DoD taking to create an ethical standard for the rules of war and AI?

3) **Paul Losiewicz, Defense Technical Information Center**

Autonomous Agent-based cybersecurity for military platforms entails clearly delineating joint AI-Human responsibilities in military operations. Furthermore it requires significant testing and M&S of the interfaces that will be employed to ensure that Tactics, Techniques and Procedures can be followed. NATO IST -152 has developed a reference architecture for future R&D on this topic, but the cognitive interactions described to date have not included a formalism for a joint deontic logic usable by Agents as well as Human operators. This is an area in which CSIAC would be happy to assist, with the Service Laboratories and Academia.

4) **Frederic Filbert, Department of Defense**

While I am all for ethical and moral use of AI, consideration has to be made that imposing a "black and white" approach to AI ethics cannot occur without a somewhat "shades of gray" aspect because AI in DOD will be supporting leader decisions that run counter to normal moral and ethical standards (i.e., war and conflicts designed to kill adversaries). Example: intelligence gathering requires breaking laws and that is acceptable to humans but may not be acceptable to AI in their absolute "ones and zeros" approach to things; particularly as "stretched ethics" from a human standpoint is already incorporated into warfare. Further, as DOD is focused on defensive and offensive operations that will result in deaths of humans (specifically for an adversary and often inadvertently with non-combatants) AI code and sub-routines written for ethical approaches could limit their effectiveness potentially resulting in defeat for our forces because the AI's ethics don't allow support. We have a human example of "stretched ethics": Google employees protesting that AI support to Project MAVEN is unethical as it could result in "death by AI" while support to Google's project to develop a censorship capable surveillance search engine for China will result in an expanded security state, potentially resulting in increased humanitarian problems, further loss of individual rights, and even imprisonment and death is apparently ok.

5) **John Potter, Jocara**

Dear Defense Innovation Board,

First of all, I thank you for creating an opportunity for the public to contribute to forming guiding principles for the development and deployment of AI for the Secretary of Defense. I believe that the coming decade will prove pivotal in terms of the development of AI and its impact on our society and regard the active involvement of the public as a critical ingredient to mapping out our collective future. This is a strong example of value co-creation for our society, addressing one of the most challenging ethical issues of our time. I hope that this initiative will inspire other countries to follow suit, and for an international consensus to develop on the guiding principles that we, as a global society, want to implement. In an adaptation of the Johari Window, I would like to frame my comments in the context of the four quadrants of knowledge, as shown in the figure, and address issues by quadrant. What we already know that we know is the domain of conscious competency, what we already know how to do and the problems we already know that we have. We are at a point in the development of AI where we are experiencing an exponential increase in this conscious competency, which is being rapidly employed to enable self-driving cars, ships, aircraft, spacecraft, automated face recognition, target identification, and much more. What ethical issues have we already seen arise in these applications? The imperative for cognitive and cultural diversity in technology creation It is well recognized that increased diversity of contributing participants, whether from different cultural backgrounds, ages, or gender, is positively correlated with the improved performance of teams. The converse is also true. While great care is generally taken to reduce bias in the databases on which AI learns, several recent cases have highlighted ethical bias problems resulting from a lack of broad inclusion, most recently in how FaceBook (FB) targets advertisements, for which FB is now under threat of being taken to court for discrimination, and in facial recognition software developed by Amazon, which is better at recognizing white males than women or people of colour. Even before an AI algorithm is exposed to data, biases inherent in the coding and project management team will inevitably become hard-wired into the algorithmic approach.

For a wonderful example of inherent cultural imprinting on AI algorithm performance, one need look no further than the screenshot that went viral in Russia in 2017 (as reported by Polina Aronson at the Digital Society Conference 2018 Discussion Panel) with the answers from Google Assistant and Alisa (the Russian counterpart) to the statement ‘I feel sad”. Google Assistant replied with “I wish I had arms so I could give you a hug“while Alisa responded with “No one said life is about fun”. Alisa, apparently, is designed and/or trained to dispense dark humour and irony more than comfort. When asked if it was OK to hit your wife, Alisa answered: “Of course, if a wife is beaten by her husband she still needs to be patient, love him, feed him, and never let him go”. A product of emotional socialism, Alisa dispenses hard truths and tough love. Alisa is more likely to view suffering as unavoidable, and thus better taken with a clenched jaw rather than with a soft embrace. Anchored in the 19th-century Russian literary tradition, emotional socialism doesn’t rate individual happiness very highly, but prizes one’s ability to live with atrocity. So cultural bias in AI may be inevitable, but in which case, we must have checks and tests to be sure AI is aligned with what we want, ethically, in our society. We already know that we have a strong gender bias in science and technology, which will surely also show its hand in AI. I would therefore implore the DoD to take all possible measures to reduce, or at least quantify, all types of bias in AI development teams and in the data employed to train AI algorithms, establishing metrics to test and measure degrees of bias. With AI likely to be applied to almost every aspect of our technological lives in the coming decade, there is far too much at stake for this technology to be created by a homogeneous team of people who share the same gender, race, religion, sexual orientation and/or political affiliation. The DoD needs to make inclusiveness a primary objective and also to extend its ethical requirements to sub-contractors. Achieving greater diversity and balance in the teams and data used to create AI will yield benefits in robustness and precision in AI performance, which in the context of security and defence applications translates to fewer miscalculations and associated human cost.

Known Unknowns In this second quadrant, we deal with the things that we already know that we do not yet know. In the case of AI, these known unknowns arise both from unexpected pathological behaviours of algorithms and from uncertain outcomes of the AI learning feedback process. Quality and safety performance risks in adaptive learning AI The first class of issues arise from possible unexpected behaviours of algorithms, even ones which are in principle deterministic and unchanging. If an algorithm is not extensively tested with the best practices of software design, exploring every possible outcome in the (often very large) parameter space, including anticipating the outcome of faulty input, it may produce unanticipated pathological behaviour. No more dramatic, and tragic, example is available than the current furore over the anti-stall flight control software that Boeing installed in the 737 Max 8 aircraft, with tragic results. In this case, it appears that Boeing not only failed to find and deal with a potentially pathological behaviour of the software in the case of degraded sensor input, but it likely also failed to detect flaws in its standard pilot procedures for disabling the software to allow pilots to regain control of the aircraft. The first class of issues arise from possible unexpected behaviours of algorithms, even ones which are in principle deterministic and unchanging. If an algorithm is not extensively tested with the best practices of software design, exploring every possible outcome in the (often very large) parameter space, including anticipating the outcome of faulty input, it may produce unanticipated pathological behaviour. No more dramatic, and tragic, example is available than the current furore over the anti-stall flight control software that Boeing installed in the 737 Max 8 aircraft, with tragic results. In this case, it appears that Boeing not only failed to find and deal with a potentially pathological behaviour of the software in the case of degraded sensor input, but it likely also failed to detect flaws in its standard pilot procedures for disabling the software to allow pilots to regain control of the aircraft.

Three-time US presidential candidate Ralph Nader is taking Boeing to court over the Ethiopia Airlines crash, saying “This Boeing 737 Max 8 disaster is a harbinger, for all technologies that are going to be controlled by AI, where the robotics, the arrogance of the algorithms, will take control, and the Boeing experience where the software took control of the plane, in a wrong way, away from its own pilots” [Ralph Nader, interview April 2019]. If a widely-respected and very large, aero-space company such as Boeing, with extensive internal quality procedures, can fall victim, so can any organization. While there can never be any guarantee that all eventualities have been uncovered (the parameter space may be so large as to be uncountable, even in deterministic cases) we need to develop a rigorous methodology to minimize this risk. But AI goes far beyond the risks of unexpected behaviour from deterministic algorithms, it takes our known unknowns into a new and much bigger domain, the products of the AI learning process and what that produces. As we move on from older paradigms developed for deterministic algorithms, that ultimately could only produce results lying within a set of outcomes based on the programming and direct data inputs, to adaptive learning algorithms that evolve their non-linear decision-making processes in the light of experience we must embrace the fact that the outcomes are no longer deterministic (even if uncountable) and cannot be guaranteed to lie within any given performance envelope, even if there are no programming errors and all foreseen error conditions have been explored and proven safe. This uncertainty must be managed as a dynamic risk, with estimates of the probability of occurrence and severity of outcome considered in the risk management and contingency planning. It would thus be ethically prudent to consider any and all AI algorithms to be imperfect and in continuous development, subject to continuous risk analysis and management. The acceptance of AI behavioural uncertainty as a known unknown is a useful perspective that will provide a valuable framing for how such technology is developed and effectively controlled. From this perspective, we must accept that the AI behaviour itself cannot be uniquely determined or predicted, and we must seek instead a more ‘fuzzy’ set of constraints based on confidence limits around the possible behavioural outcomes, so that we may develop a level of trust in the reliability of AI performance to

achieve the objectives we desire. Every deployed AI algorithm must be trustworthy to perform within a limited range of expected outcomes, whatever algorithmic or sensory input imperfection it may have.

That AI algorithms will be handling critical systems, at rates far surpassing human capacity to track or understand the evolution of the situation, makes it imperative that there be a sophisticated framework in place to implement code auditing, decision traces and transparent accountability. At each point, there must ultimately be an identified human individual who is responsible for each action taken. The ethical questions arise in managing the algorithmic risk from a values and legal liability perspective. Certainly we will need specifications for due diligence and algorithm stress testing, which must include extensive simulation and the use of generative adversarial networks to test its responses over the largest range of inputs. Unknown knowns: The third quadrant is very poorly understood and even more rarely considered. This is the domain of the unknown knowns, that is, the things that we know subconsciously, intuitively, but which we do not consciously recognize. This is where the wisdom of 'sleeping on it' before making a big decision lies. If we do not have sufficient objective conscious information to be able to make an informed choice between options, each with its associated risks and benefits, then 'sleeping on it' allows our subconscious to weigh in with additional competencies obtained from subconscious sensory inputs and evaluations, below our conscious horizon, that often results in clarity come morning. Algorithms are (mostly) written in full consciousness, and thus do not encode this unconscious wisdom.

To minimize the risk of AI taking erroneous action, we need high-level control processes that include people in the control loop, providing a 'sanity check' before any decision is taken that has the likely outcome of significant collateral damage. A powerful example comes to mind from the depths of the cold war, on September 26 1983, when lieutenant colonel Stanislav Yevgrafovich Petrov was in charge of a Soviet nuclear early warning centre. On this night, his satellite sensing network reported five American nuclear missile launches. Rather than immediately retaliate, as protocol demanded, Stanislav followed his gut feeling and went against protocol, delaying his report to seniors and eventually convincing the armed forces that it was a false alarm. With his decision to ignore algorithms and instead follow his gut instinct, Stanislav prevented an all-out US-Russia nuclear war. Going partly on gut instinct and believing the United States was unlikely to fire only five missiles, he told his commanders that it was a false alarm before he knew that to be true. Unknown Unknowns: And finally, we come to the darkest quadrant, that which we do not even know we do not know. These are the things that will come from the outfield to surprise us. Developers of artificial intelligence diligently work to control for errors in the data, human bias and changes in the context of which the AI is used. The algorithms are researched and tested for accuracy and reliability. However, despite all this, there are unexpected unknown unknowns that will arise, particularly when under deliberate attack by counter-AI forces, something that the DoD can reasonably expect to occur in the battlespace. A prime example is the Microsoft chatbot 'Tay' that was designed to be a friendly teenager to entertain on twitter. Tay was not Microsoft's first online AI application, a chatbot called XiaoIce has been very successful in China, where it has been used by 40 million people. Tay was an attempt to duplicate XiaoIce but for a very different culture. Tay was given a Twitter account and autonomously tweeted and interacted with others. Despite extensive prior user studies with diverse user groups, Microsoft failed to identify a vulnerability that was exploited in a coordinated attack. In the process, absorbing and learning from data provided by tweets addressed to Tay, the chat bot rapidly diverged from the intended character role, becoming a racist fascist within hours of launch. Microsoft had to shut down Tay's account only 16 hours after it was released. Following Tay's breakdown, Microsoft Research Corporate VP, Peter Lee, said "Looking ahead, we face some difficult – and yet exciting – research challenges in AI design. AI systems feed off of both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical. We will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes. To do AI right, one needs to

iterate with many people and often in public forums. We must enter each one with great caution and ultimately learn and improve, step by step, and to do this without offending people in the process. We will remain steadfast in our efforts to learn from this and other experiences as we work toward contributing to an Internet that represents the best, not the worst, of humanity. I hope that DoD will take a similar approach. I look forward to seeing the AI Principles the Defense Innovation Board puts forward for consideration by the Secretary of Defense.

6) Zac Taschdijan, H2O.ai, Georgia Institute of Technology

AI, machine learning and related technologies must be human interpretable, explainable, fair and accurate. The reasons for this are fairly self-evident and well-documented. To do this, I believe we need "human-centered AI", based on (and extending) the principles of human-computer interaction.

7) Future of Life Institute – *official submission*

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 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 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 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 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).

8) Stephen Rapp, DIU and USA CCDC

I'm looking forward to spreading innovation into our tank - automotive RD&E at the USA CCDC - Ground Vehicle Systems Center. AI is a crucial current and future technology we need to innovate and integrate, for the US to maintain its technology overmatch.