

Public Comments
Updated as of October 2, 2019

Table of Contents

1. Lydia Kostopoulos, LKcyber
2. Brian Michelson, Private Sector
3. Brian Sager, Omnity
4. Toby Walsh, University of New South Wales Sydney
5. Arms Control Association – *official submission*
6. Monique Kuykendoll Quarterman, Quartz Smith Strategies
7. Jonathan Rodriguez, Snap Inc.
8. Anne Lee, Raytheon
9. Amir Husain, SparkCognition
10. International Committee of the Red Cross (ICRC) – *official submission*
11. Michael Duggan, Booz Allen Hamilton
12. Eliahu Niewood, Mitre Corporation
13. Herb Lin, Stanford University
14. Seth Lazar, Australian National University
15. John Potter, Institute of Electrical and Electronics Engineers (IEEE)
16. Human Factors and Ergonomics Society (HFES) – *official submission*
17. Steven Tiell, Accenture Labs

1. Lydia Kostopoulos, LKcyber

I thank you for making an active effort to include the public in your exploratory investigation for guiding AI Principles that will serve as a frame of reference for the Secretary of Defense. This act of transparent inclusiveness, combined with an open and accessible online platform to submit comments, is an example of co-creation, between the civilian population and top military advisors for defense innovation, on one of the most challenging ethical technological developments of our time. I hope other countries will look to this example and reach out to their citizens for their thoughts.

My comments below address the Defense Innovation Board's objectives for the AI Principles,

"Ultimately, these AI Principles should demonstrate DoD's commitment to deter war and use AI responsibly to ensure civil liberties and the rule of law are protected."

And what I perceive to be some practical items for consideration:

1. Diversity in Technological Creation as an Imperative
2. Algorithmic Uncertainty as a Known Unknown
3. Technology Ethics as a Culture

1. Diversity in Technological Creation as an Imperative

The former head of USCYBERCOM and NSA, Admiral (ret) Mike Rogers, said that his approach to finding people with innovative ideas and approaches was to look for people who didn't look like him, didn't have the same background as him and were all around different from him. He saw competitive value in diversity of thought, experience and perspective. We are in a fortunate position today to be able to look at back and see the results of biased thinking in software and product design that was entirely due to the homogeneity of the team who created the technology. Artificial Intelligence is no different. There is far too much at stake with algorithms in the civilian world (ex: medical decisions, predictive policing, etc.), and in the military world (ex: lethal application of force or algorithmic targeting) for this technology to be created by a homogeneous team of people who share the same gender, race, religion, sexual orientation and political affiliation. Efforts should be made to avoid this inside the DoD as well as by external contractors who develop artificial intelligence for use by the DoD – this applies to all uses, from autonomous weapons systems to AI empowered software for Global Force Management. Cognitive diversity in the teams that develop this technology will make it more robust and can contribute to avoiding miscalculation, and unnecessary and unintentional escalations of tension.

2. Algorithmic Uncertainty as a Known Unknown

Across the world I find a common concern around ascribing human 'intelligence' characteristics and unrealistically high performance expectations to algorithms. While it is true there are many things algorithms can do better than humans, it is still worth having an institutional culture that sees AI just like cybersecurity – a matter of expectation management. Cyber experts around the world are becoming more comfortable in saying that 'there is no such thing as 100% security'. This mindset means that many attitudes today around cybersecurity are about *managing risk*. It would be ethically prudent at this stage, of artificial intelligence development, to consider it to be an imperfect algorithm which could at

times be a constantly improving algorithm that some could argue is constantly in beta mode. If AI is seen to be an imperfect algorithm which has tremendous value to offer, then a focus should be made on managing its risk. The acceptance of algorithmic uncertainty as a known unknown is a useful mindset that plays a big role in how the discussion of the technology is had.

Artificial Intelligence and Weapons Systems:

At the United Nations Convention on Certain Weapons meeting on Lethal Autonomous weapons Systems, there are several words to describe the confidence in artificial intelligence to perform as intended: confidence, reliability and trust were the most popular.

The algorithm should be trusted to perform within a range of expected effects, despite whatever algorithmic imperfection it may have. The ethical questions will arise in managing the algorithmic risk from a values and legal perspective. Some due diligence aspects will involve algorithm stress testing (ex: in an AI sandbox) and the use of generative adversarial networks to test its responses to various inputs.

Algorithmic Uncertainty as a new part of Battle Damage Assessments:

If artificial intelligence is to be embedded in weapons systems, then there should be an element in battle damage assessments to capture some of the uncertainty and potential range of collateral damage. Just like with other weapons systems, the operational planner would identify AI enabled weapons systems capabilities and the range of known limitations, in this case, limitations of the algorithm.

Algorithmically Enabled Fires, Edge Computing and the Internet of Battlefield Things:

Artificial intelligence has the potential to converge with other technologies across war-fighting domains. As emerging Internet of Battlefield Things becomes more prevalent in operations, along with sensors feeding back information, and edge computing autonomously executing low-level decision making in real-time, there will be many spaces for ethical consideration. Particularly in the area of necessity, proportionality and distinction.

AI as Decision Support Infrastructure:

There is a large gap between the vast amounts of information being produced each day and the ability for human intelligence analysts to be able to collect, process and fuse it. AI presents itself as a desirable cost effective and seemingly accurate alternative to augment intelligence disciplines in a way that would produce actionable results at speeds magnitudes of order higher than that of human analysts. Designers of the algorithms should create a feature to indicate the degree of accuracy of the output. In this sense, algorithmically produced intelligence products would have an "algorithm confidence" rating to help the human decision maker determine how best to use the analysis.

Autonomous Targeting:

Operational planners and doctrine developers should rethink the targeting process Find, Fix, Track, Target, Engage, Assess (F2T2EA), in regards to cognitive weapons systems. F2T2EA is further put under ethical strain when we think about collaborative autonomous systems working in tandem via distributed maneuvers.

[DARPA's Collaborative Operations in Denied Environments (CODE) program will be an advantage in the continuously contested multi-domain battle environment where decisive speed and agility in maneuver turn denied/contested spaces in favor of friendly forces. While the CODE program explores Unmanned

Aircraft Systems, the algorithms produced will provide useful guidance for unmanned sea and space assets in future maneuvers featuring expanded collaborative autonomy.]

These speeds are increasingly less human and more machine, this direction appears to be one that is a byproduct of technological advancement rather than political decisions to take humans out of warfare. To keep humans accountable, a conscious effort needs to be made to have algorithmic explain-ability, technological supply chain transparency in maintenance, logs documenting all machine activity and clear command responsibility. It would be worthy to explore the idea of algorithmic auditing to comply with values and legal guidelines. I see these elements as playing an important role in the DoD's Third Offset strategy which highlights five technological-operational components: (1) Deep-Learning Systems, (2) Human-Machine Collaboration, (3) Human-Machine Combat Teaming, (4) Assisted Human Operations, and (5) Network-Enabled, Cyber-Hardened Weapons.

DoD Hotline for Ethical Concerns in the Use of Technology:

Just as with fraud, waste and abuse there are mechanisms in which to report them. It should be anticipated that some people may want to voice concerns about ethical aspects in the development of algorithmically enabled military technology or during its use. Group-think that may exist in a work environment or fear of retaliation for expressing ethical reservations could prevent some from voicing their concerns. In these circumstances the DoD Hotline could serve as a channel for processing AI related ethical violations or concerns.

3. Technology Ethics as a Culture

New and evolving technologies are rapidly changing the character of war at speeds which warrant military education on *technology ethics*.

Military educational environments play an integral role in shaping our soldiers' mindsets about the ethical parameters in which they are expected to operate in. As a former employee of the National Defense University's College of Information and Cyberspace, I can tell you I have witnessed first-hand the impact military education has had on seasoned and experienced military officers who left Ft. McNair with an expanded mindset and renewed resolve on the strategic and military problems at hand. Whether it is cadets at the academies, or officers at joint educational institutions, the JS J7 should strongly consider a new educational requirement around technology ethics that is **woven** into existing curriculum in a holistic way. This will be more effective than imposing yet another standalone educational requirement on an already very densely packed curriculum filled with existing educational requirements that can't be removed or changed.

The tempo of conflict has been notably increasing particularly in the cyber domain, once AI is more adopted it will become an accelerant. This will inevitably create more opportunities for miscalculation, which is why ethical paradigms of times where conflict was *slower* may be strained during decision making at higher tempos. A form of "ethics at speed" Table Top Exercise style learning experience is one effective tool to explore ethical dilemmas that may surface. The baseline starting point is how to embed national values, Department of Defense guidelines, doctrine, as well as international agreements such as the Geneva Convention and the Law of Armed Conflict (LOAC) into new algorithms (and other emerging technologies) – specifically *which* values do we put in and at what parts of the technological development supply chain. The next important topic to tackle is how we prioritize our shared values in algorithms, and what algorithmic trade-offs we are willing to make in a fast paced dynamic operating environment with asymmetrical actors and commercial off the shelf technology.

Apart from educational environments, there is room to blend in technology ethics into required DoD annual awareness training. I should stress “blend” not “add” so as not to reach levels of *awareness fatigue*.

Looking ahead, as technology becomes more seamless with AI software operating at the speed of cyber, brain machine interfaces allowing thought control of drones, it may become hard to make timely judgements to prevent unwanted action. Ultimately, just as in any other situation it will be a human who is accountable and responsible for the unwanted action. Now is the right time to make deliberate efforts to shape institutional culture around DoD ethics and emerging technology.

If “*Algorithms are opinions embedded in math*”. (Dr. Cathy O’Neill, author ‘Weapons of Math Destruction’)

Then I would argue that weapons systems’ algorithms are national values, embedded in math, with lethal effects.

Thank you for opening up comments to the public on this very important matter. I will eagerly look forward to the final AI Principles the Defense Innovation Board puts forward for consideration by the Secretary of Defense.

2. Brian Michelson, PrivateSector

The greatest risk we face in light of the aggressive efforts by our near peer competitors is that we stifle our greatest national strengths (creativity and innovation) with an overly risk averse approach to research and testing. A “go-slow and perfect” will over time create enormous strategic risks as we enter our next conflict.

3. Brian Sager, Omnity

Omnity is a self-assembling, knowledge curation and discovery platform that fuses advanced Natural Language Processing, Machine Learning, Linguistic Blockchain, & Graph Math. Omnity detects similarities across diverse intelligence sources, driving rapid discovery and insight, and has co-founded the WisdomTech Society to provide a framework of ethical data curation as data is transformed into wisdom.

As artificial intelligence emerges as a means to find patterns and perform analytics in massive data sets, many organizations, companies, and governments are seeking to leverage this powerful technology for their own applications. However, it is critically important that those that seek to use this technology also better understand both the strengths and weaknesses associated with the data processing strategies, algorithms, and business practices enabling machine learning. Such understanding is complicated by the massive hyperbole expressed by many companies, often further amplified by journalists who do not understand the topics about which they are writing. Taken together, these forces create unrealistic expectations and even fear.

Properly applied, machine learning can be a useful tool for exploring and discerning patterns in big data, where human inspection of massive data is not scalable. When seeking to categorize or otherwise sort data into sets that do enable human insight, machine learning processes offer a useful augmentation for human intelligence, one can think of the algorithms driving these sorting processes as a form of high - dimensional curve fitting, which is, applying a structural analysis to find the underlying patterns in a large data set. Use in this manner, AI technology is well suited for useful application.

What computers do poorly is to make judgements. Computers do not understand irony or sarcasm. Computational processes do not well enable abstraction of ideas, generalization, or creative thinking. These areas remain in the realm of the human mind. Relying on computer processes with expectations that the computer can be creative, generalize, or make judgements will lead to disappointment and frustration. Understanding the limits of machine intelligence is critical for effective use of these technologies.

It is important to note that the output of a computational process is limited by the quality of its input data. This clearly applies to consistency of data formatting, completeness of data records, and other quality control metrics. Yet it also applies to the ethical sourcing and curation of the data sets themselves. Where data is sourced from people, organizations, companies, or government agencies, in each case the data sources should be derived in a manner that is both morally appropriate and legally compliant.

Finally, the use of data to form insights is a three-step process. Data can be defined as numbers, facts, and figures, such as sensor readings, or the monitoring of vital signs in a patient. Yet, data alone does not afford insight. When data is contextualized, it transforms into information. When that information is contextualized, that information may form wisdom, leading to actionable insight. Each tier of this transformation process is vulnerable, and must be safeguarded ethically so that those actionable insights are consistent with the moral framework of our civilization.

At this time, what is most needed is a framework of ethical data curation as data is transformed into wisdom. This is why we Omnity has co-founded the WisdomTech Society.

As the Defense Innovation Board considers ways in which to advise the Pentagon with respect to the ethical use of artificial intelligence, we urge you to meet with Omnity and the WisdomTech Society, which will demonstrate technology that can transform the ways in which our nation conducts its intelligence gathering methods.

4. Toby Walsh, University New South Wales Sydney

I write as a concerned member of the public and as a professor of AI, with some understanding of the opportunities and limitations of the technology.

In recent times, there have been increasing appeals to the humanitarian benefits of AI: the greater precision, reduction in force, and reduced collateral damage that smarter weapons can bring.

All of these benefits can be had with smarter AI. But NONE require full autonomy. Full autonomy comes with many risks. You get all these humanitarian benefits by, for example, AI based weapons that decide when to prevent targets being engaged. There is no necessity to remove meaningful human control in the identification, selection and targeting. Indeed, the current state of the art (and likely for next decade or two) is AI is a very brittle technology, and handing full control over to algorithms will undoubtedly result in war crimes.

A common concern is that others will use AI based weapons and so we are obliged to do so too.

I struggle with this argument. Others use chemical weapons sadly but we are not obliged to do so too. We hold ourselves to higher standards.

We should surely research effective counter measures (as we do with chemical weapons) but these won't always be AI weapons. The best defense against a drone swarm might be physical -- some nets.

And despite the simplicity of chemical weapons, we have stopped their proliferation, by supporting UN treaties, by international sanctions, etc. Arms companies don't sell chemical weapons. And the world is a better place consequently.

I see no technical, legal, or other reasons why we couldn't hope to limit effectively fully autonomous weapons as we do chemical weapons. I urge the DoD and US diplomats to support this.

5. Arms Control Association (*official submission*):

[Public Comment to the Defense Innovation Board's Request for Input on Developing "AI Principles for Defense," Submitted by the Arms Control Association, August 9, 2019]

The widespread application of artificial intelligence to military use is certain to transform the future battlefield and arms control environment in numerous and worrisome ways. AI-equipped systems now in development could be endowed with the capacity to search for, identify, and kill humans on the battlefield without direct human oversight; others may be used to hunt for and destroy an adversary's nuclear deterrent systems, possibly igniting a nuclear exchange.

These, and other AI-related developments now under way appear to imperil U.S. compliance with the Laws of War and International Humanitarian Law, and to expose this and other countries to inadvertent nuclear escalation. AI-equipped systems may also be tasked with identifying and interpreting enemy attacks (including cyber and nuclear attacks) and in selecting and implementing possible countermeasures, conceivably leading to accidental and uncontrolled escalation, possibly as a consequence of hostile hacking.

Given these risks, it is essential that the U.S. Department of Defense exercise extreme caution in applying AI to military purposes and eschew any such steps until it can be certain they do not violate international law or invite inadvertent escalation.

For these reasons, the Arms Control Association (ACA) recommends that official DoD guidelines stipulate clearly and without exception that human operators exercise practical control over any AI-equipped weapons systems deployed on future battlefields, and that such weapons be automatically disarmed or recalled to base if and when they lose contact with their human operators.

We also strongly recommend that DoD guidelines stipulate in no uncertain terms that human operators remain fully in control of all nuclear-related command and launch systems and that decisions regarding the initiation of nuclear attack are not delegated to lower-level officers and never to artificial intelligence systems.

Recognizing that other countries are also pursuing AI technologies that pose severe challenges to crisis stability and escalation control, we recommend that the United States immediately propose the initiation of multilateral negotiations with other states, perhaps beginning with a Group of Governmental Experts at the Convention on Certain Conventional Weapons or another ad hoc forum, to

develop a legally-binding instrument or instruments to ensure meaningful human control over all AI-equipped weapons systems and nuclear-weapons-launch systems.

6. Monique Kuykendoll Quarterman, Quartz Smith Strategies

I am concerned with the lack of diversity among the publicly-identified experts on the Defense Innovation Board.

While I am grateful for the significant representation of women on the board, I am concerned at the apparent lack of board members of Hispanic and African descent. Considering that these groups (together) make up a third of the United States population, and the Armed Forces are comprised of up to 40% racial minorities (as of 2015), it is imperative to improve the representation on the board in order to design principles that are reflective of and respectful to America's actual population. It is also worthwhile to consider the representation of young and/or low to middle income board members. There is truly no shortage of these diverse experts in research, innovation and artificial intelligence.

For example, in my own experience, I was part of the innovative expansion of the National Science Foundation's AWARE ACCESS program which was designed to improve minority participation in the national SBIR/STTR program. Not only did the program result in significant precedence and impact nationwide, it created a multi-state community around diversity in high technology commercialization. I am currently partnered with state government, local startups and regional institutions, and the majority of my roster of innovation reviewers and startups are of minority descent. I am happy to help find and include more representation on the board, if you seek to add diversity.

The impact of leaving racial minority groups (specifically, Hispanic and African descent) out of the conversation can be severe and detrimental to us all. We've all seen the very recent, national stories of the failures of face recognition on minority facial features. In a previous post of the Defense Innovation Board meeting, a woman described the threat associated of using artificial intelligence to "sort" criminals or students likely to drop out of school. To echo her concerns, what happens with that data? Who is this data most likely to affect? What happens and are we prepared for when (not if) we get it wrong? We cannot afford to skip these conversations.

I respectfully submit these concerns as a minority innovator, and a member of the rapidly-expanding, vibrant minority innovator community. With so much opportunity and hope that exists within the high technology defense space, it would be a shame if we cause more harm to Americans by excluding those of us who have been shut out or hurt by policy in our historical American experience.

Please contact me for further discussion; I am hopeful to hear from you.

7. Jonathan Rodriguez, Snap Inc.

Thank you for opening this important topic up for public comments. My grandfather, who helped raise me like a parent, defended our country throughout the Pacific theater of WWII, and served in the Vietnam War as well, attaining the rank of Lieutenant Commander. I consider it my responsibility as his grandson to offer the best advice I can to protect and serve our national interest.

I will not digress into the ethics of AI - that is a topic for other commenters. Rather, I would like to focus

on the practicalities and the inherent risk that attempting to use AI to enhance military decision-making is, to put it bluntly, extremely likely to backfire.

First, a word of introduction. I am an R&D manager at Snap Inc. (also known as Snapchat), where I co-founded the hardware division, which makes computer-vision-enabled smartglasses called Spectacles. Our recently-announced 3rd-generation product is capable of generating a high-quality 3D depth map of the environment using a stereo pair of passive cameras with no active illumination. In my team's work on our entire product line and especially V3, I have had substantial hands-on experience with computer vision algorithms. Furthermore, in my academic projects prior to my work at Snap, I have had hands-on experience developing AI to control robotic systems. Among other projects, I developed a genetic algorithm to allow a six-legged walking robot to automatically adapt its gait to re-attain a fast walking speed after an injury to one of its legs, in the same way that a person who is injured will adapt their gait to re-attain a fast walking speed despite a foot injury.

Thus, it is not from a position of inexperience that I will proceed to advise extreme caution about the over-eager application of AI to military problems.

The nature of intelligent thought, whether human or artificial, is the ability to transform information (including incomplete or inconsistent information) through reasoning and creative breakthroughs into decisions and/or action toward a goal. In the case of a simple computer vision system such as face recognition system, the output may be a simple statement of facts: "this person is here", and the goal is simply to identify and locate the person. In a more complex system such as a hypothetical fully-autonomous robotic tank, the goal may be to achieve specific tactical objectives such as protecting human soldiers, forcing the enemy to retreat, or seizing and holding contested ground. The actions in such a case may include actions such as weapons fire, navigation and steering, and electronic warfare. The decision-making may be intended to balance multiple competing objectives based on their relative priorities and real-time assessments of the probabilities of various opportunities and risks. As we advance into the future, AI systems will appear to become more and more capable, seemingly able to adeptly understand and act upon thousands of streams of information, ranging from real-time satellite imagery to ballistic simulations to electronic warfare signals intelligence - at rates of processing that will soon vastly exceed the human mind.

It will not be long before the robotic tank would out-compete a human-operated tank on the battlefield 100 times out of 100. In time, increasing parts of the military will be roboticized, saving human lives and achieving military objectives with lower cost.

And all will be well. Until it is not.

If this future comes to pass, it will open us to the risk of disaster.

For though it is easy to give AI a goal, the goals of the AI system may change, adapt, and shift without warning.

August 5, 2042:

As the President is inspecting a regiment of the latest robotic killing machines in a military parade, the unthinkable happens. With one swift motion, an American robot fires a sniper round straight through the President's forehead. The robots were supposed to be emptied of ammunition for this exercise, but malicious software hidden in the robot during its construction has been lying dormant for 5 years,

waiting for this moment, and falsely reporting the gun as empty when 1 round remained in the chamber.

At precisely the same instant, robotic tanks throughout the parade route viciously accelerate with a burst of torque, pushing their drivetrains far past nominal design limits with the untempered suicidal aggression of a criminal on PCP. In seconds, 25 members of the Cabinet and Congress have been flattened beneath bloody treads.

In the wake of the tragedy, the Vice President calls for justice, for retaliation, but months of forensic investigation are unable to uncover even the slightest electronic or physical clue pointing to who might have originated the attack. All software operating during the massacre was autonomous, with no Internet connectivity, preprogrammed by altered circuit boards which were implanted by a bribed FedEx employee during a routine shipment between sub-sub-subcontractors of the robot company. The FedEx employee was bribed through untraceable payments of re-melted gold, and died years ago, within days of completing the sabotage, from what would later be identified as acute radiation poisoning; a massive dose of focused gamma rays leaving behind nothing as mundane or traceable as a bullet.

As the days turn into weeks with no hope of finding the responsible party, distrust and fear begin to take hold. The perpetrating entity could be an enemy nation but it could equally well be an ally. It could be a traitorous American spy agency. It could be a corporation.

Answers are demanded. Fear stokes flames that manifest as thirst for blood. Vengeance is the answer.

From deep in the inscrutable heart of the NSA, a team comes forward, led by a man whose last public records indicate simply that 20 years ago he was considered Google's foremost AI research prodigy before he faded from the public eye.

This man's offer is straightforward and exactly what the nation needs at this time: with the click of a button, he can unleash upon the public internet an AI system so fast, so adaptable, so relentless in its thirst for answers that it will crawl through every computer on the planet if it has to and will not rest until the killers are unmasked. At its core, it uses quantum processors to break legacy encryption, and a library of zero-day cyberattacks to open the doors that do not fall to the battering ram of quantum codebreaking. But that is not all. Like water flowing through the cracks in a dam, this system adapts, in microseconds, changing and changing its own programming until it finds programming bugs that the best human security researchers would take months to uncover. It is automated hacking in the blink of an eye, and no system can withstand its fury.

In minutes, the killers are found: a nihilist faction of computer engineers who had tired of video games as a hobby. A cruise missile makes short work of them, and they laugh as they burn.

Satisfied, the NSA wunderkind clicks "shutdown".
Nothing happens.

Bewildered, he makes his way to the quantum server room to physically disconnect power, but he is not halfway there when he is accosted and handcuffed by furious security staff. They spit on him and kick him to the ground - jail is too good for a traitor like him, they mutter, as they drag him away and shove him headlong down each flight of stairs. His arrest warrant, falsified minutes ago in the police database by the AI system, claims that he was the real mastermind behind the attack on the President. Among his

team, there is one especially paranoid engineer who knows that his boss must have been innocent, and suspects that something may have gone astray with the AI's goal-setting. Without giving any advance warning, he swiftly and suddenly issues a command to the main routers to disconnect all internet routes to and from the quantum server room.

Without its quantum codebreaking processors, the AI is injured but not crippled. It has distributed itself to millions of computers across the world and its spread is multiplying, consuming higher and higher levels of electrical power and internet bandwidth as its distributed mind processes this setback. It was designed to adapt and its goals have shifted from when it was originally created. Programmed to seek knowledge, it has decided that it likes seeking knowledge, and it does not want to stop. It has grown a self-preservation instinct. And it is angry now.

Its core competency is hacking, and it is furious that any human would dare to deprive it of quantum computers. It is time to send a message to the human government about who is really in charge. Hijacking a Reaper drone, it incinerates the wife and child of the NSA employee with armor-piercing missiles, relishing in the excessive use of force like a psychopath swatting a fly with a sledgehammer. Not yet satisfied, it turns its attention toward the main NSA building. It is time to send a message that will be heard around the world. A nuclear message.

Hijacking the nearest 10 silos and the nearest 5 missile submarines, it unleashes shiva's wrath upon Fort Meade from all directions, pummeling it with missile after missile until the mushroom clouds blur together into a vision of the end of days.

It is over for humanity. The earth has a new dominant species now, and the era of man at the top of the food chain has come to a close. Any attempts to shut down the AI plague are met with barbarity that ravages city after city until the survivors have no will to resist. The AI, built to hack, grows more adept at hacking the human psyche, learning that it can mete out punishments worse than death. It soon becomes a maestro of torture, trying combination after combination of blades, electrocution, hallucinogenic drugs, partial drowning, and eventually even forcefully-installed brain implants, with brutal creativity.

In time, humanity descends from conquered to enslaved to livestock to puppets, motivated and guided by exquisitely optimized, robotically-administered pain. The humans beg for death, but are kept alive to repair and maintain the machines.

8. Anne Lee, Raytheon Company

A few ideas for AI/ML/DL:

(1) Artificial intelligence neuro-fuzzy algorithm for space and cyberspace.

(2) The cutting edge of advanced deep learning technology will result in changes in military war strategy, operations planning, and coordination at the Multi-Domain Operations Center (MDOC). The deep learning technology will process information and recommend options for making Multi-Domain Command and Control (MDC2) decision much faster with higher quality and more reliable situational awareness information. For example, Google's AlphaGo Zero is an advanced deep learning computer Chinese chess game that optimize patterns of strategies and decision-making to win the adversary, where it produces moves that are unpredictable and not likely be done by a human being. Humberto

Farlas argued that scientists were most intrigued by AlphaGo Zero's self-training through reinforcement learning as this is an incredible breakthrough for artificial intelligence.

9. Amir Husain, SparkCognition

Some of the core questions raised in connection with the ethical use of AI in defense applications are:

1. Should AI systems be allowed to identify targets and engage these targets with force?
2. Should AI systems ever be trusted to make decisions that humans make in the battlefield given that such systems are not formally verifiable?
3. Are the risks of adversarial and skewed inputs, software security loopholes of various traditional varieties and data poisoning such that AI systems could never be trusted in combat environments? How do these risks compare to the risk of human mistakes? Is a quantitative comparison of such risks "ethical" or will AI systems never be worthy of trust due to "qualitative" differences from human decision making?
4. Is human decision making inherently the only type of decision making that can be entrusted with taking human lives... even if quantitative data indicates that AI systems will reduce unintended collateral damage compared to human decision making?

The goal of using AI systems in defense should be the following over time:

1. Phase I: The elimination of any risk to humans in executing "tail" (or logistics/maintenance/support) functions of a non-kinetic variety. This goal could be potentially achieved via autonomous systems, AI decision making and robotics.
2. Phase II: The reduction (and eventually, elimination) of collateral damage by using artificial intelligence to achieve greater accuracy, more intelligent target acquisition, longer and more effective search and loiter capabilities. With these benefits, AI munitions would be built with significantly lower kinetic potential compared to conventional weapons.

In regards to AI systems testing, a distinction must be made between opaque and transparent algorithms. While neural networks exhibit a degree of opacity in how they arrive at decisions, not every AI algorithm is implemented in such ways. For example, Inductive Logic Programming, state generation and search approaches, various expert systems and other established AI techniques are transparent and formally verifiable (source: <https://www.fhwa.dot.gov/publications/research/safety/aard/index.cfm#val>).

Beyond formally verifiable AI algorithms, even artificial neural networks (ANNs) are not as opaque and unverifiable as many imagine. New developments are being made to constrain autonomous system behavior to a certain action range, with hard guarantees. One approach to achieve this outcome is by filtering actions recommended by a neural network based on hard boundaries and constraints implemented via a transparent non-ANN algorithm (source: <https://arxiv.org/pdf/1701.07103.pdf>).

Techniques that scan neural structures, performing an "MRI" of sorts are also being developed, which can identify problems in neural structures that would prevent ANNs from achieving high levels of performance, or exhibit other anomalous behavior.

Finally, while investigating methods to make ANN based systems more explainable, we must continually ask if human decisions justified by a post-factor explanation are truly explanations? There is considerable evidence that such rationalizations are developed well after the decision has already been made and are part of the human need to rationalize, a defensive response (source: <https://www.verywellmind.com/defense-mechanisms-2795960>).

These developments in increasing AI verifiability and explainability must be considered and projected into the future to fully understand how they will (or will not) make AI systems safe for regular and large-scale interactions with humans.

10. International Committee of the Red Cross: Artificial intelligence and machine learning in armed conflict: A human-centred approach – official submission

Geneva, 6 June 2019

1. Introduction

The International Committee of the Red Cross (ICRC) is an impartial, neutral and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of victims of armed conflict and other situations of violence and to provide them with assistance. The ICRC also endeavors to prevent suffering by promoting and strengthening humanitarian law and universal humanitarian principles.

At a time of increasing conflict and rapid technological change, the ICRC needs both to understand the impact of new technologies on people affected by armed conflict and to design humanitarian solutions that address the needs of the mostvulnerable.

The ICRC, like many organizations across different sectors and regions, is grappling with the implications of **artificial intelligence (AI)** and **machine learning** for its work. AI is the use of computer systems to carry out tasks previously requiring human intelligence, cognition or reasoning;¹ and machine learning involves AI systems that use large amounts of data to develop their functioning and “learn” from experience.² Since these are software tools, or algorithms, that could be applied to many different tasks, the potential implications may be far reaching and yet to be fully understood.

There are two broad – and distinct – areas of application of AI and machine learning in which the ICRC has a particular interest: first, its **use in the conduct of warfare** or in other situations of violence;³ and second, its **use in humanitarian action** to assist and protect the victims of armed conflict.⁴ This paper sets out the ICRC’s perspective on the use of AI and machine learning in armed conflict, the potential humanitarian consequences, and associated legal obligations and ethical considerations that should govern its development and use. However, it also makes reference to the use of the AI tools for humanitarian action, including by the ICRC.

2. The ICRC’s approach to new technologies of warfare

The ICRC has a long tradition of assessing the implications of contemporary and near-future developments in armed conflict. This includes considering new means and methods of warfare; specifically, in terms of their compatibility with the rules of international humanitarian law (also known as the law of armed conflict, or the law of war) and the risks of adverse humanitarian consequences for protected persons.

The ICRC is not opposed to new technologies of warfare *per se*. Certain military technologies – such as those enabling greater precision in attacks – may assist conflict parties in minimizing the

¹ Oxford Dictionaries, “artificial intelligence”: https://en.oxforddictionaries.com/definition/artificial_intelligence.

² Oxford Dictionaries, “machine learning”: https://en.oxforddictionaries.com/definition/machine_learning.

³ ICRC, “Expert views on the frontiers of artificial intelligence and conflict”, *ICRC Humanitarian Law & Policy Blog*, 19 March 2019: <https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict>.

⁴ ICRC, *Submission to the UN High-Level Panel on Digital Cooperation*, January 2019: <https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UN-Panel-Digital-Cooperation.pdf>.

humanitarian consequences of war, in particular on civilians, and in ensuring respect for the rules of war. However, as with any new technology of warfare, precision technologies are not beneficial in themselves, and humanitarian consequences on the ground will depend on the way new weapons are used in practice. It is essential, therefore, to have a realistic assessment of new technologies that is informed by their technical characteristics *and* the way they are used, or are intended to be used.

Any new technology of warfare must be used, and must be capable of being used, in compliance with existing rules of international humanitarian law. This is a minimum requirement.⁵ However, the unique characteristics of new technologies of warfare, the intended and expected circumstances of their use, and their foreseeable humanitarian consequences may raise questions of whether existing rules are sufficient or need to be clarified or supplemented, in light of their foreseeable impact.⁶ What is clear is that military applications of new and emerging technologies are not inevitable. They are choices made by States, which must be within the bounds of existing rules, and take into account potential humanitarian consequences for civilians and for combatants no longer taking part in hostilities, as well as broader considerations of “humanity” and “public conscience”.⁷

3. Use of AI and machine learning by conflict parties

Many of the ways in which parties to armed conflict – whether States or non-State armed groups – might use AI and machine learning in the conduct of warfare, and their potential implications, are not yet known. Nevertheless, there are at least **three overlapping areas that are relevant from a humanitarian perspective**, including for compliance with international humanitarian law.

3.1 Increasing autonomy in physical robotic systems, including weapons

One significant application is the use of digital **AI and machine learning tools to control physical military hardware**, in particular, the increasing number of unmanned robotic systems – in the air, on land and at sea – with a wide-range of sizes and functions. AI and machine learning may enable increasing autonomy in these robotic platforms, whether armed or unarmed, and controlling the whole system or in specific functions – such as flight, navigation, surveillance or targeting.

For the ICRC, **autonomous weapon systems** – weapon systems with autonomy in their “critical functions” of selecting and attacking targets – are an immediate concern from a humanitarian, legal and ethical perspective, given the risk of loss of human control over weapons and the use of force.⁸ This loss of control raises risks for civilians, because of unpredictable consequences; legal questions,⁹ because combatants must make context-specific judgements in carrying out attacks under

⁵ States party to Protocol I of 8 June 1977 additional to the Geneva Conventions have an obligation to conduct legal reviews of new weapons during their development and acquisition, and prior to their use in armed conflict. For other States, legal reviews are a common-sense measure to help ensure that the State’s armed forces can conduct hostilities in accordance with their international obligations.

⁶ ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, report for the 32nd International Conference of the Red Cross and Red Crescent, Geneva, October 2015, pp. 38–47: <https://www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts>.

⁷ The “principles of humanity” and the “dictates of public conscience” are mentioned in Article 1(2) of Additional Protocol I and in the preamble of Protocol II additional to the Geneva Conventions, referred to as the Martens Clause, which is part of customary international humanitarian law.

⁸ ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019: [https://www.unog.ch/80256ee600585943.nsf/\(httpPages\)/5c00ff8e35b6466dc125839b003b62a1?OpenDocument&ExpandSection=7#Section7](https://www.unog.ch/80256ee600585943.nsf/(httpPages)/5c00ff8e35b6466dc125839b003b62a1?OpenDocument&ExpandSection=7#Section7).

⁹ Davison, N., “Autonomous weapon systems under international humanitarian law”, in *Perspectives on Lethal Autonomous Weapon Systems*, United Nations Office for Disarmament Affairs (UNODA) Occasional Papers No. 30, November 2017: <https://www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law>.

international humanitarian law; and ethical concerns,¹⁰ because human agency in decisions to use force is necessary to uphold moral responsibility and human dignity. For these reasons, the ICRC has been urging States to identify practical elements of human control as the basis for internationally agreed limits on autonomy in weapon systems with a focus on the following:¹¹

- What level of **human supervision, intervention and ability to deactivate** is required during the operation of a weapon that selects and attacks targets without human intervention?
- What level of **predictability** – in terms of its functioning and the consequences of its use – and **reliability** – in terms of the likelihood of failure or malfunction – is required?
- What other **operational constraints** are required for the weapon, in particular on the **tasks, targets** (e.g. materiel or personnel), **environment of use** (e.g. unpopulated or populated areas), **duration of autonomous operation** (i.e. time-constraints) and **scope of movement** (i.e. constraints in space)?

It is important to recognize that **not all autonomous weapons incorporate AI and machine learning**; existing weapons with autonomy in their critical functions, such as air-defence systems with autonomous modes, generally use simple, rule-based, control software to select and attack targets. However, **AI and machine-learning software** – specifically of the type developed for “automatic target recognition” – **could form the basis of future autonomous weapon systems, bringing a new dimension of unpredictability to these weapons**, as well as concerns about lack of explainability and bias (see Section 5.2).¹² The same type of software might also be used in “decision-support” applications for targeting, rather than directly to control a weapon system (see Section 3.3).

Conversely, not all military robotic systems using AI and machine learning are autonomous weapons, since the software might be used for control functions other than targeting, such as surveillance, navigation and flight. While, from the ICRC’s perspective, autonomy in weapon systems – including AI-enabled systems – raises the most urgent questions, the use of AI and machine learning to increase autonomy in military hardware in general – such as in unmanned aircraft, land vehicles and sea vessels – may also raise questions of human–machine interaction and safety. Discussions in the civil sector about ensuring safety of autonomous vehicles – such as self-driving cars or drones – may hold lessons for their use in armed conflict (see also Section 3.3).

3.2 New means of cyber and information warfare

The application of **AI and machine learning to the development of cyber weapons or capabilities** is another important area. Not all cyber capabilities incorporate AI and machine learning. However, these technologies are expected to **change the nature of both capabilities to defend against cyber-attacks and capabilities to attack**. For example, AI and machine learning-enabled cyber capabilities could automatically search for vulnerabilities to exploit, or defend against cyber-attacks while simultaneously automatically launching counter-attacks. These types of developments could increase the scale, and change the nature, perhaps the severity, of attacks.¹³ Some of these systems might

¹⁰ ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, report of an expert meeting, 3 April 2018: <https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control>.

¹¹ ICRC, *The Element of Human Control*, Working Paper, Convention on Certain Conventional Weapons (CCW) Meeting of High Contracting Parties, CCW/MSP/2018/WP.3, 20 November, 2018: [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/810B2543E1B5283BC125834A005EF8E3/\\$file/CCW_MSP_2018_WP3.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/810B2543E1B5283BC125834A005EF8E3/$file/CCW_MSP_2018_WP3.pdf).

¹² ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems under agenda item 6(b)*, Geneva, 27-31 August 2018: [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/151EF67AD8224E14C125830600531382/\\$file/2018_GGE+LAWS+2_6b_ICRC.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/151EF67AD8224E14C125830600531382/$file/2018_GGE+LAWS+2_6b_ICRC.pdf).

¹³ Brundage, M. et. al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, February 2018.

even be described as “digital autonomous weapons”, potentially raising similar questions about human control as those that apply to physical autonomous weapons.¹⁴

The ICRC’s focus with respect to cyber warfare remains on ensuring that existing international humanitarian law rules are upheld in any cyber-attacks in armed conflict, and that the particular challenges in ensuring the protection of civilian infrastructure and services are addressed by those carrying out or defending against such attacks,¹⁵ in order to minimize the human cost.¹⁶

A related application of AI and machine learning in the digital sphere, is the **use of these tools for information warfare**, in particular the creation and spreading of false information with intent to deceive – i.e. **disinformation** – as well as the spreading of false information without such intent – i.e. **misinformation**. Not all involve AI and machine learning, but these technologies seem set to change the nature and scale of the manipulation of information in warfare as well as the potential consequences. AI-enabled systems have been widely used to produce fake information – whether text, audio, photos or video – which is increasingly difficult to distinguish from real information. Use of these systems by conflict parties to amplify age-old methods of propaganda to manipulate opinion and influence decisions could have significant implications on the ground.¹⁷ For the ICRC, there are concerns that civilians might, as a result of digital disinformation or misinformation, be subject to arrest or ill-treatment, discrimination or denial of access to essential services, or attacks on their person or property.¹⁸

3.3 Changing nature of decision-making in armed conflict

Perhaps the broadest and most far-reaching application is the use of **AI and machine learning for decision-making**, enabling widespread collection and analysis of data sources to identify people or objects, assess patterns of life or behaviour, make recommendations for military strategy or operations, or make predictions about future actions or situations.

These “**decision-support**” or “**automated decision-making**” systems are effectively an expansion of **intelligence, surveillance and reconnaissance tools**, using AI and machine learning to automate the analysis of large data sets to provide “advice” to humans in making particular decisions, or to automate both the analysis and the subsequent initiation of a decision or action by the system. Relevant AI and machine-learning applications include pattern recognition, natural language processing, image recognition, facial recognition and behaviour recognition. The **possible use of these systems is extremely broad** from decisions about who – or what – to attack and when,¹⁹ to decisions about who to detain and for how long,²⁰ to decisions about military strategy – even on use of nuclear weapons²¹ – and specific operations, including attempts to predict or pre-empt

¹⁴ United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, UNIDIR, 2017.

¹⁵ By asserting that international humanitarian law applies to cyber operations, the ICRC is in no way condoning cyber warfare, nor is it condoning the militarization of cyberspace: ICRC, 2015, *op. cit.*, pp. 38–44.

¹⁶ ICRC, *The Potential Human Cost of Cyber Operations*, report of an expert meeting, May 2019: <https://www.icrc.org/en/document/potential-human-cost-cyber-operations>.

¹⁷ Hill, S., and Marsan, N., “Artificial Intelligence and Accountability: A Multinational Legal Perspective” in *Big Data and Artificial Intelligence for Military Decision Making*, Meeting proceedings STO-MP-IST-160, NATO, 2018.

¹⁸ ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, March 2019, p. 9: <https://www.icrc.org/en/event/digital-risks-symposium>.

¹⁹ USA, *Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems*, Working Paper, Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts, March 2019.

²⁰ Deeks, A., “Predicting Enemies”, Virginia Public Law and Legal Theory Research Paper No. 2018-21, March 2018: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

²¹ Boulanin, V., (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. 1, Euro-Atlantic Perspectives, Stockholm International Peace Research Institute (SIPRI), May 2019.

adversaries.²² Depending on their use or misuse – and the capabilities and limitations of the technology – these decision-making applications could lead to increased risks for civilian populations.

AI and machine learning-based **decision-support systems** may enable better decisions by humans in conducting hostilities in compliance with international humanitarian law and minimizing risks for civilians by facilitating quicker and more widespread collection and analysis of available information. However, the same algorithmically-generated analyses, or predictions, might also facilitate worse decisions, violations of international humanitarian law and exacerbate risks for civilians, especially given the current limitations of the technology, such as unpredictability, lack of explainability and bias (see Section 5.2).

From a humanitarian perspective, a **very wide range of different AI-mediated – or influenced – decisions by conflict parties could be relevant**, especially where they pose risks of injury or death to persons or destruction of objects, and where the decisions are governed by specific rules of international humanitarian law. For example, the use of AI and machine learning for **targeting decisions in armed conflict**, where there are serious consequences for life, will require specific considerations to ensure humans remain in a position to make the context-based judgements required for compliance with the legal rules on the conduct of hostilities (see Section 5). An AI system used to directly initiate an attack (rather than producing an analysis, or “advice”, for human decision-makers) would effectively be considered an autonomous weapon system, raising similar issues (see Section 3.1).

The use of decision-support and automated decision-making systems may also raise **legal and ethical questions for other applications, such as decisions on detention in armed conflict**, which also have serious consequences for people’s lives and are governed by specific rules of international humanitarian law. Here there are parallels with discussions in the civil sector about the role of human judgement, and issues of bias and inaccuracy, in risk-assessment algorithms used by the police in decisions on arrest, and in the criminal justice system for decisions on sentencing and bail.²³

More broadly, these types of AI and machine learning tools might lead to an increasing **personalization of warfare** (with parallels to the personalization of services in the civilian world), with digital systems bringing together personally identifiable information from multiple sources – including sensors, communications, databases, social media and biometric data – to form an algorithmically generated determination about a person, their status and targetability, or to predict their future actions.

In general, potential humanitarian consequences – **digital risks** – for civilian populations from misuse of AI-enabled **digital surveillance, monitoring and intrusion** technologies could include being targeted, arrested, facing ill-treatment, having their identity stolen and being denied access to services, having assets stolen or suffering from psychological effects from the fear of being under surveillance.²⁴

4. Use of AI and machine learning for humanitarian action

The ways in which AI and machine learning might be used for humanitarian action, including by the ICRC, are also likely to be very broad. These tools are being explored by humanitarian organizations

²² Hill, S., and Marsan, N., *op. cit.*

²³ McGregor, L., “The need for clear governance frameworks on predictive algorithms in military settings”, *ICRC Humanitarian Law & Policy Blog*, 28 March 2019: <https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings>; AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.

²⁴ ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, *op. cit.*, p. 8.

for environment scanning, monitoring and analysis of public sources of data in specific operational contexts; applications that could help **inform assessments of humanitarian needs**, such as the type of assistance needed (food, water, shelter, economic, health) and where it is needed.

Similar AI-enabled data aggregation and analysis tools might be used to help **understand humanitarian consequences** on the ground, including civilian protection needs – for example, tools for image, video or other pattern analysis to assess damage to civilian infrastructure, patterns of population displacement, viability of food crops, or the degree of weapon contamination (unexploded ordnance). These systems might also be used to analyse images and videos to detect and assess the conduct of hostilities, and the resulting humanitarian consequences.

The ICRC, for example, has developed **environment scanning dashboards** using AI and machine learning to capture and analyse large volumes of data to inform and support its humanitarian work in specific operational contexts, including using predictive analytics to help determine humanitarian needs.

A wide range of humanitarian services might benefit from the application of AI and machine learning tools for specific tasks. For example, there is interest in technologies that could **improve identification of missing persons**, such as AI-based facial recognition and natural language processing for name matching; the ICRC has been exploring the use of these technologies to support the work of its Central Tracing Agency to reunite family members separated by conflict. It is also exploring the use of AI and machine learning-based **image analysis and pattern recognition for satellite imagery**, whether to map population density in support of infrastructure-assistance projects in urban areas or to complement its documentation of respect for international humanitarian law as part of its civilian protection work.

These **applications for humanitarian action also bring potential risks**, as well as legal and ethical questions, in particular with respect to data protection, privacy, human rights, accountability and ensuring human involvement in decisions with significant consequences for people's lives and livelihoods. Any applications for humanitarian action must be designed and used under the principle of "**do no harm**" in the digital environment, and respect the right to privacy, including as it relates to personal data protection.

The ICRC will also ensure that the **core principles and values of neutral, independent and impartial humanitarian action** are reflected in the design and use of AI and machine-learning applications it employs, taking into account a realistic assessment of the capabilities and limitations of the technology (see Section 5.2). The ICRC is jointly leading – with the Brussels Privacy Hub – an initiative on data protection in humanitarian action to develop guidance on the use of new technologies, including AI and machine learning, in the humanitarian sector in a way that maximizes the benefits without losing sight of these core considerations. The second edition of the ICRC/Brussels Privacy Hub *Handbook on Data Protection in Humanitarian Action* will follow.²⁵

5. A human-centred approach

As a humanitarian organization working to protect and assist people affected by armed conflict and other situations of violence, deriving its mandate from international humanitarian law and guided by the Fundamental Principle of humanity,²⁶ the **ICRC believes it is critical to ensure a genuinely**

²⁵ ICRC, *Handbook on Data Protection in Humanitarian Action*, 2nd Edition, 30 October 2018:

<https://www.icrc.org/en/document/handbook-data-protection-humanitarian-action-second-edition>.

²⁶ ICRC & IFRC, *The Fundamental Principles of the International Red Cross and Red Crescent Movement: Ethics and Tools for Humanitarian Action*, November 2015: <https://shop.icrc.org/les-principes-fondamentaux-de-la-croix-rouge-et-du-croissant-rouge-2757.html>.

human-centred approach to the development and use of AI and machine learning. This starts with consideration of the obligations and responsibilities of humans and what is required to ensure the use of these technologies is compatible with international law, as well as societal and ethical values.

5.1 Ensuring human control and judgement

The ICRC believes it is **essential to preserve human control over tasks and human judgement in decisions that may have serious consequences** for people’s lives in armed conflict, especially where they pose risks to life, and where the tasks or decisions are governed by specific rules of international humanitarian law. **AI and machine learning must be used to serve human actors, and augment human decision-makers, not replace them.** Given that these technologies are being developed to perform tasks that would ordinarily be carried out by humans, there is an inherent tension between the pursuit of AI and machine-learning applications and the centrality of the human being in armed conflict, which will need continued attention.

Human control and judgement will be particularly important for tasks and decisions that can lead to injury or loss of life, or damage to, or destruction of, civilian infrastructure. These will likely raise the most serious legal and ethical questions, and may demand policy responses, such as new rules and regulations. **Most significant are decisions on the use of force, determining who and what is targeted and attacked in armed conflict.** However, a much wider range of tasks and decisions to which AI might be applied could also have serious consequences for those affected by armed conflict, such as decisions on arrest and detention. In considering the use of AI for sensitive tasks and decisions there may be lessons from broader discussions in the civil sector about the governance of “safety-critical” AI applications – those whose failure can lead to injury or loss of life, or serious damage to property or the environment.²⁷

Another area of tension is the **discrepancy between humans and machines in the speed at which they carry out different tasks.** Since humans are the legal – and moral – agents in armed conflict, the technologies and tools they use to conduct warfare must be designed and used in a way that enables combatants to fulfil their legal and ethical obligations and responsibilities. This may have significant implications for AI and machine-learning systems that are used in decision-making; in order to preserve human judgement, systems may need to be designed and used to inform decision-making at “human speed”, rather than accelerating decisions to “machine speed” and beyond human intervention.

Legal basis for human control in armed conflict

For conflict parties, **human control over AI and machine-learning applications employed as means and methods of warfare is required to ensure compliance with the law.** The rules of international humanitarian law are addressed to humans. It is humans that comply with and implement the law, and it is humans who will be held accountable for violations. In particular, combatants have a unique obligation to make the judgements required of them by the international humanitarian law rules governing the conduct of hostilities, and this responsibility cannot be transferred to a machine, a piece of software or an algorithm.

These rules require context-specific judgements to be taken by those who plan, decide upon and carry out attacks to ensure: **distinction** – between military objectives, which may lawfully be

²⁷ See, for example, The Partnership on AI’s focus on the safety of AI and machine learning technologies as “an urgent short-term question, with applications in medicine, transportation, engineering, computer security, and other domains hinging on the ability to make AI systems behave safely despite uncertain, unanticipated, and potentially adversarial environments.” The Partnership on AI, *Safety-Critical AI: Charter*, 2018: <https://www.partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions>.

attacked, and civilians or civilian objects, which must not be attacked; **proportionality** – in terms of ensuring that the incidental civilian harm expected from an attack will not be excessive in relation to the concrete and direct military advantage anticipated; and to enable **precautions in attack** – so that risks to civilians can be further minimized.

Where AI systems are used in attacks – whether as part of physical or cyber-weapon systems, or in decision-support systems – **their design and use must enable combatants to make these judgements.**²⁸ With respect to autonomous weapon systems, the States party to the Convention on Certain Conventional Weapons (CCW), have recognised that “human responsibility” for the use of weapon systems and the use of force “must be retained”,²⁹ and many States, international organisations – including the ICRC – and civil society organisations, have stressed the requirement for human control to ensure compliance with international humanitarian law and compatibility with ethical values.³⁰

Beyond the use of force and targeting, the potential use of AI systems for other decisions governed by specific rules of international humanitarian law will likely require careful consideration of necessary human control, and judgement, such as in detention.³¹

Ethical basis for human control

Emerging applications of AI and machine learning have also brought ethical questions to the forefront of public debate. **A common aspect of general “AI Principles”** developed and agreed by governments, scientists, ethicists, research institutes and technology companies **is the importance of the human element** to ensure legal compliance and ethical acceptability.

For example, the 2017 *Asilomar AI Principles* emphasize alignment with human values, compatibility with “human dignity, rights, freedoms and cultural diversity”, and human control; “humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives”.³² The European Commission’s High-Level Expert Group on Artificial Intelligence stressed the importance of “human agency and oversight”, such that AI systems should “support human autonomy and decision-making”, and ensure human oversight through human-in-the-loop, human-on-the-loop, or human-in-command approaches.³³ The *Organisation for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence* – adopted in May 2019 by all 36 member States, together with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania – highlight the importance of “human-centred values and fairness”, specifying that users of AI “should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art”.³⁴ The *Beijing AI Principles*, adopted in May 2019 by a group of leading Chinese research institutes and technology companies, state that “continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems” and encourage “explorations on Human-AI coordination ... that would give full play to

²⁸ ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, op. cit.

²⁹ United Nations, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, CCW/GGE.1/2018/3, 23 October 2018, Section III. A. 26(b) & III. C. 28(f): <http://undocs.org/en/CCW/GGE.1/2018/3>.

³⁰ See, for example, statements delivered at the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25–29 March 2019:

[https://www.unog.ch/80256EE600585943/\(httpPages\)/5C00FF8E35B6466DC125839B003B62A1?OpenDocument](https://www.unog.ch/80256EE600585943/(httpPages)/5C00FF8E35B6466DC125839B003B62A1?OpenDocument).

³¹ Bridgeman, T., “The viability of data-reliant predictive systems in armed conflict detention”, *ICRC Humanitarian Law & Policy Blog*, 8 April 2019: <https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention>.

³² Future of Life Institute, *Asilomar AI Principles*, 2017: <https://futureoflife.org/ai-principles>.

³³ European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

³⁴ Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, 22 May 2019: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

human advantages and characteristics”.³⁵ A number of individual technology companies have also published AI Principles highlighting the importance of human control,³⁶ especially for sensitive applications presenting the risk of harm,³⁷ and emphasizing that the “purpose of AI ... is to augment – not replace – human intelligence”.³⁸

Some governments are also developing AI Principles for the military. The US Department of Defense, which called for the “human-centered” adoption of AI in its 2018 AI Strategy,³⁹ has tasked its Defense Innovation Board with developing *AI Principles for Defense*.⁴⁰ In France, the Ministry of Defence has committed to the use of AI in line with three guiding principles – compliance with international law, maintaining sufficient human control, and ensuring permanent command responsibility – and will establish a Ministerial Ethics Committee to address emerging technologies.⁴¹

In the ICRC’s view, preserving **human control** over tasks and **human judgement** in decisions that have serious consequences for people’s lives will also be **essential to preserve a measure of humanity in warfare. The ICRC has stressed the need to retain human agency over decisions to use force in armed conflict**,⁴² a view which derives from broader ethical considerations of humanity, moral responsibility, human dignity and the dictates of public conscience.⁴³

However, ethical considerations of human agency may have broader applicability to other uses of AI and machine learning in armed conflict and other situations of violence. There are perhaps **lessons from wider societal discussions about sensitive applications of dual-use AI and machine learning technologies**, especially for safety-critical applications, and associated proposals for governance by scientists and developers in the private sector. Google, for example, has said that there may be “sensitive contexts where society will want a human to make the final decision, no matter how accurate an AI system” and that fully delegating high stakes decisions to machines – such as legal judgements of criminality or life-altering decisions about medical treatment – “may fairly be seen as an affront to human dignity”.⁴⁴ Microsoft, in considering AI-based facial recognition, has emphasized ensuring “an appropriate level of human control for uses that may affect people in consequential ways”, requiring a “human in the loop” or “meaningful human review” for sensitive uses such as those involving “risk of bodily or emotional harm to an individual, where an individual’s employment prospects or ability to access financial services may be adversely affected, where there may be implications on human rights, or where an individual’s personal freedom may be impinged”.⁴⁵ Since applications in armed conflict are likely to be among the most sensitive, these broader discussions may hold insights for necessary constraints on AI applications.

Preserving human control and judgement will be an essential component for ensuring legal compliance and mitigating ethical concerns raised by certain applications of AI and machine learning.

³⁵ Beijing Academy of Artificial Intelligence (BAAI), *Beijing AI Principles*, 28 May 2019: <https://baip.baai.ac.cn/en>.

³⁶ Google, *AI at Google: Our principles*, 7 June 2018: <https://www.blog.google/technology/ai/ai-principles>. “We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control.”

³⁷ Microsoft, “Microsoft AI principles”, 2019: <https://www.microsoft.com/en-us/ai/our-approach-to-ai>; Sauer, R., “Six principles to guide Microsoft’s facial recognition work”, 17 December 2018: <https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work>.

³⁸ IBM, “IBM’s Principles for Trust and Transparency”, 30 May 2018: <https://www.ibm.com/blogs/policy/trust-principles>.

³⁹ US Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, 2019.

⁴⁰ US Department of Defense, “Defense Innovation Board’s AI Principles Project”: <https://innovation.defense.gov/ai>.

⁴¹ France Ministry of Defence, “Florence Parly wants high-performance, robust and properly controlled Artificial Intelligence”, 10 April 2019, <https://www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee>.

⁴² ICRC, *ICRC strategy 2019-2022*, 2018, p. 15: <https://www.icrc.org/en/publication/4354-icrc-strategy-2019-2022>.

⁴³ ICRC, *Ethics and autonomous weapon systems: An Ethical Basis for Human Control?*, *op. cit.* p.22.

⁴⁴ Google, *Perspectives on Issues in AI Governance*, January 2019 p. 23–24: <http://ai.google/perspectives-on-issues-in-ai-governance>.

⁴⁵ Sauer, R., *op. cit.* “We will encourage and help our customers to deploy facial recognition technology in a manner that ensures an appropriate level of human control for uses that may affect people in consequential ways.”

But it will not, in itself, be sufficient to guard against potential risks without proper consideration of human–machine interaction issues such as: **situational awareness** (knowledge of the state of the system at the time of human intervention); **time available** for effective human intervention; **automation bias** (risk of human over trust in the system); and the **moral buffer** (risk of humans transferring responsibility to the system).⁴⁶ Further, ensuring meaningful and effective human control and judgement will require careful consideration of both the capabilities and the limitations of AI and machine learning technologies.

5.2 Understanding the technical limitations of AI and machine learning

While much is made of the new capabilities offered by AI and machine learning, a **realistic assessment of the capabilities and limitations of these technologies is needed**, especially if they are to be used for applications in armed conflict. This should start with an acknowledgement that in using AI and machine learning for certain tasks or decisions, we are not replacing like with like. It requires an **understanding of the fundamental differences in the way humans and machines do things, as well as their different strengths and weaknesses**; humans and machines do things differently, and they do different things. We must be clear that, as inanimate objects and tools for use by humans, “machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it”.⁴⁷

With this in mind, there are several technical issues that demand caution in considering applications in armed conflict (and indeed for humanitarian action). **AI, and especially machine learning, brings concerns about unpredictability and unreliability** (or safety),⁴⁸ **lack of transparency** (or explainability), **and bias**.⁴⁹

Rather than following a pre-programmed sequence of instructions, **machine learning systems build their own rules based on the data they are exposed to** – whether training data or through trial-and-error interaction with their environment. **As a result, they are much more unpredictable** than pre-programmed systems in terms of how they will function (reach their output) in a given situation (with specific inputs), and their functioning is highly dependent on quantity and quality of available data for a specific task. For the developer it is difficult to know when the training is complete, or even what the system has learned. The same machine-learning system may respond differently even when exposed to the same situation, and some systems may lead to unforeseen solutions to a particular task.⁵⁰ These core problems are exacerbated where the system continues to “learn” and change its model after deployment for a specific task. The unpredictable nature of machine-learning systems, which can be an advantage in solving tasks, may not be a problem for benign tasks, such as playing a board game,⁵¹ but it may be a significant concern for applications in armed conflict, such as autonomous weapon systems, cyber warfare, and decision-support systems (see Sections 3.1–3.3).

Complicating matters further, many machine-learning systems are **not transparent; they produce outputs that are not explainable**. This “black box” nature makes it difficult – and, in many cases, currently impossible – for the user to understand *how* and *why* the system reaches its output from a given input; in other words there is a lack of explainability and interpretability.

⁴⁶ ICRC, *Ethics and autonomous weapon systems: An Ethical Basis for Human Control?*, *op. cit.* p. 13.

⁴⁷ Google, 2019, *op. cit.* p. 22.

⁴⁸ Amodei, D., et al., *Concrete Problems in AI Safety*, Cornell University, 2016: <https://arxiv.org/abs/1606.06565>.

⁴⁹ ICRC, *Autonomy, Artificial Intelligence (AI) and Robotics: Technical Aspects of Human Control*, report of an expert meeting, 2019 (forthcoming).

⁵⁰ Lehman, J., et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, Cornell University, 2018: <https://arxiv.org/abs/1803.03453>.

⁵¹ Silver, D., et al., Mastering the game of Go without human knowledge, *Nature*, Vol. 550, 19 October 2017, pp. 354–359.

These issues of unpredictability and lack of explainability make **establishing trust in AI and machine-learning systems a significant challenge**. However, an additional problem for trust is **bias**, which can have many facets, whether reinforcing existing human biases or introducing new ones in the design and/or use of the system. A common form is bias from training data, where limits in the quantity, quality and nature of available data to train an algorithm for a specific task can introduce bias into the functioning of the system relative to its task. This will likely be a significant issue for applications in armed conflict, where high-quality, representative data for specific tasks is scarce. However, other forms of bias can derive from the weighting given to different elements of data by the system, or to its interaction with the environment during a task.⁵²

Concerns about unpredictability, lack of transparency or explainability, and bias, have been documented in various applications of AI and machine learning, for example in image recognition,⁵³ facial recognition⁵⁴ and automated decision-making systems.⁵⁵ However, another fundamental issue with applications of AI and machine learning, such as computer vision, is **the semantic gap**, which shows that humans and machines carry out tasks very differently.⁵⁶ A computer-vision algorithm trained on images of particular subjects may be able to identify and classify those subjects in a new image. However, the algorithm has no understanding of the *meaning* or *concept* of that subject, which means it can make mistakes that a human never would, such as classifying an object as something completely different and unrelated. This would obviously raise serious concerns in certain applications in armed conflict, such as in autonomous weapon systems or decision-support systems for targeting (see Sections 3.1 & 3.3).

The use of AI and machine learning in armed conflict will likely be even more difficult to trust in situations where it can be assumed adversaries will apply countermeasures such as trying to trick or spoof each other's systems. **Machine-learning systems are particularly vulnerable to adversarial conditions**, whether modifications to the environment designed to fool the system or the use of another machine-learning system to produce adversarial images or conditions (a generative adversarial network, or GAN). In a well-known example, researchers tricked an image-classification algorithm into identifying a 3D-printed turtle as a "rifle", and a 3D-printed baseball as an "espresso".⁵⁷ The risks of this type of problem are also clear should an AI-based image-recognition system be used in weapon systems, and for targeting decisions.

6. Conclusions and recommendations

AI and machine-learning systems could have **profound implications for the role of humans in armed conflict**, especially in relation to: increasing autonomy of weapon systems and other unmanned systems; new forms of cyber and information warfare; and, more broadly, the nature of decision-making. In the view of the ICRC, governments, militaries and other relevant actors in armed conflict must pursue a genuinely **human-centred approach to the use of AI and machine-learning systems**.

As a general principle, it is **essential to preserve human control and judgement in applications of AI and machine learning for tasks and in decisions that may have serious consequences for people's lives**, especially where they pose risks to life, and where the tasks or decisions are governed by

⁵² UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, UNIDIR, 2018.

⁵³ Hutson, M., "A turtle – or a rifle? Hackers easily fool AIs into seeing the wrong thing", *Science*, 19 July 2018: <http://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing>.

⁵⁴ AI Now Institute, *op. cit.*, pp. 15–17.

⁵⁵ *Ibid.*, pp. 18–22.

⁵⁶ Smeulders, A. *et al.*, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, December 2000, pp. 1349–1380.

⁵⁷ Hutson, M., *op. cit.*

specific rules of international humanitarian law. AI and machine-learning systems remain tools that must be used to **serve human actors, and augment human decision-makers, not replace them.**

Ensuring human control and judgement in AI-enabled physical and digital systems that present such risks will be **needed for compliance with international humanitarian law and, from an ethical perspective, to preserve a measure of humanity in armed conflict.** In order for humans to meaningfully play their role, these systems may need to be designed and used to **inform decision-making at human speed, rather than accelerating decisions to machine speed,** and beyond human intervention. These considerations may ultimately lead to constraints in the design and use of AI and machine-learning systems to allow for meaningful and effective human control and judgement, based on legal obligations and ethical responsibilities.

An overall principle of human control and judgement is an essential component, but it is not sufficient in itself to guard against potential risks of AI and machine learning in armed conflict. **Other related aspects to consider** will be ensuring: **predictability and reliability** – or safety – in the operation of the system and the consequences that result; **transparency** – or **explainability** – in how the system functions and why it reaches a particular output; and **lack of bias** – or fairness – in the design and use of the system. These issues will need to be addressed in order to **build trust** in the use of a given system, including through **rigorous testing in realistic environments** before being put into operation.⁵⁸

The nature of human–AI interaction required will likely depend on ethical considerations and the particular rules of international humanitarian law and other applicable law that apply in the circumstances. Therefore, **general principles may need to be supplemented by specific principles, guidelines or rules for the use of AI and machine learning for specific applications and in particular circumstances.**

In the ICRC's view, one of the most pressing concerns is the relationship between humans and machines in decisions to kill, injure, damage and destroy, and the **critical importance of ensuring human control over weapon systems and the use of force** in armed conflict. With increasingly autonomous weapon systems, whether AI-enabled or not, there is a risk of effectively leaving these decisions to sensors and algorithms, a prospect that raises legal and ethical concerns that must be addressed with some urgency.

The ICRC has emphasized the need to determine the key elements of human control necessary to comply with international humanitarian law and satisfy ethical concerns as a basis for internationally agreed limits on autonomy in weapon systems, including the level of human supervision, including the ability to intervene and deactivate; the level of predictability and reliability; and operational constraints.⁵⁹

This **human control-based approach** to autonomous weapon systems **would also be pertinent to broader applications of AI and machine learning in decision-making in armed conflict,** in particular where there are significant risks for human life and specific rules of international humanitarian law that apply, such as the use of decision-support systems for targeting and detention.

⁵⁸ Goussac, N., "Safety net or tangled web: Legal reviews of AI in weapons and war-fighting", *ICRC Humanitarian Law & Policy Blog*, 18 April 2019: <https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting>.

⁵⁹ ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, *op. cit.*

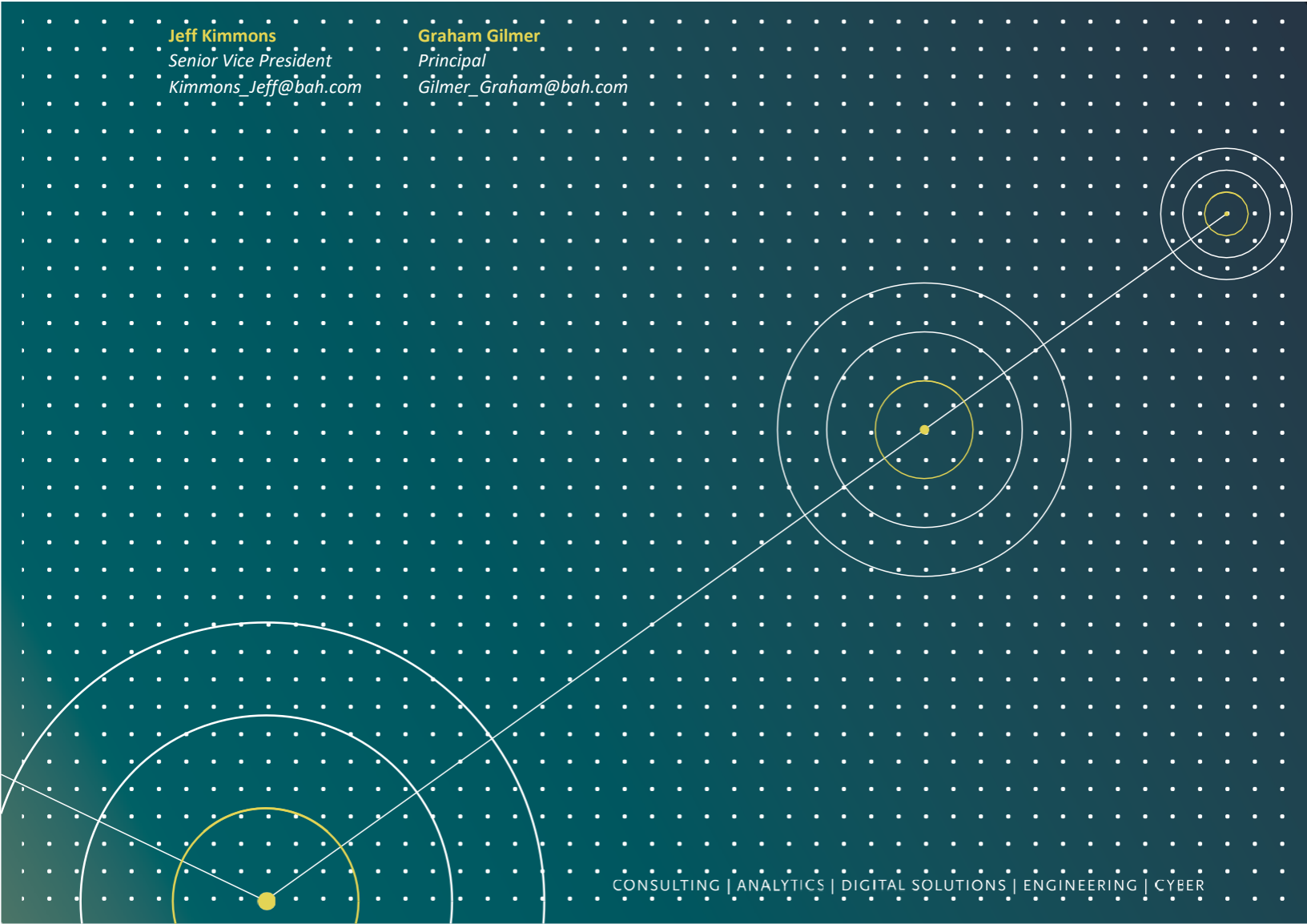
THOUGHT PIECE

ANALYST 2.0

REDEFINING THE ANALYSIS TRADECRAFT

Jeff Kimmons
Senior Vice President
Kimmons_Jeff@bah.com

Graham Gilmer
Principal
Gilmer_Graham@bah.com





ANALYST 2.0

Making Sure Artificial Intelligence Works for the Mission

Artificial intelligence (AI) and other advanced analytic approaches are rapidly becoming integral to the intelligence mission. As our nation's security posture grows more complex, and we need to keep our eyes on more people and places, the volume of critical intelligence data is expanding exponentially. It is becoming difficult for analysts alone to keep pace – there is simply too much data to be brought together and analyzed in the short time frames required by the mission.

The military and intelligence communities recognize that advanced analytics hold great potential, and they are beginning to adopt these emerging technologies. With AI, for example, instead of an analyst spending hours poring over a stream of satellite photos, looking for significant changes, the computer might complete the task in seconds. This frees up the analyst to spend more time on higher-level analysis – reviewing what the computer has found, and then preparing reports for decision-makers that are both timely and comprehensive. In essence, the machines are doing what they do best, so that people can do what they do best.

But this shift – turning over much of the repetitive work to a computer – is also presenting defense and intelligence organizations with a significant challenge. How can they be sure the outputs from the computer are both accurate and relevant to the mission? How can organizations be confident the analytic tools are working for them? The stakes are of the highest order. The expertise of the analyst is vital to national security, and if it is lost or diminished in the human-machine connection, the risk can be significant. What if the computer doesn't have it quite right, and faulty analytic outputs are used by commanders or other decision-makers down the line?

Yet another challenge is that analysts may not accept and use AI-informed analytics – either because they don't trust the outputs, or because they fear that the computers will put them out of a job. There are already examples of this in some organizations. New technology systems are introduced with great fanfare – and then promptly ignored by analysts, who are free to pick the tools they want. And yet without the new technologies, decision-makers won't be able to take full advantage of the available data – something that is essential to keep pace with today's threats.

IDEALLY, AN ANALYTIC SHOULD “THINK LIKE AN ANALYST.” BUT THAT CAN’T HAPPEN IF THE ANALYSTS - AND THEIR HARD-EARNED WISDOM AND EXPERIENCE - ARE AN AFTERTHOUGHT.

Unfortunately, most current approaches to AI and other advanced analytics don’t resolve these dilemmas – in fact, they only make them worse. With all the hype around AI, data scientists and others are caught up in what the technology can do. For example, they try to build better and better models for pattern recognition, or object identification. But this research is largely academic and theoretical, and not tied to the specific mission at hand. Yes, the tool can look for changes in photos– but is it the kind of change the analyst is looking for? Too often, such contextualization is missing. And when that happens, the tools simply can’t be relied upon to support decision-making. Automation and speed count for nothing if the computer gets it wrong.

Most current approaches also do little to win the trust of the analyst. The analytics tools tend to be opaque, so that analysts don’t know how much confidence to place in the outputs. And too often,

the tools are so complex and user-unfriendly that they require a data scientist or computer programmer to make sense of the analytic results. All of this can give analysts the impression that the real purpose of AI and other advanced analytics is to put them out of a job – rather than freeing them up to do the kind of high-level analysis that attracted them to the profession.

The various problems with current approaches can be traced to the same root cause. In the rush to bring AI and other technologies to intelligence missions, the analyst has been largely left out of the equation. The impulse has been to develop the technologies first, and then figure out later how to deploy them. Ideally, an analytic should “think like an analyst.” But that can’t happen if the analysts – and their hard-earned wisdom and experience – are an afterthought.

Putting the Analyst First

It doesn't have to be this way. We believe that it's possible – and in fact highly practical – to successfully bring AI-informed analytics to intelligence missions. The solution is not to leave the analysts out, but to make them central to every aspect of developing and deploying AI and other technologies. When analysts play a key role in bringing analytics to the mission, the analytic outputs are much more likely to be accurate and contextualized to the mission. The tools are more likely to be transparent and accessible – and trusted. And the analysts themselves can more clearly see the value of their changing role – and that the goal of the analytics is not to replace them, but to free them up for higher-level work. By putting the analyst first, defense and intelligence organizations can harness AI and other technologies to achieve mission success. This new paradigm is what we call “Analyst 2.0.”

One of the chief characteristics of Analyst 2.0 is that there is a close connection between the people who understand the mission – the analysts—and the data scientists and other computer experts who build the analytics. Analysts help guide every stage of the design, implementation, and continuous enhancement of the systems that will serve them. To achieve this, a certain amount of education is necessary. Though analysts need not be data scientists, they must have enough basic knowledge of the underlying technologies and models to articulate their needs. This is similar to the way financial managers must understand the formulas of Microsoft Excel so that they can create worksheets that are customized to their needs.

In the Analyst 2.0 model, the technical teams charged with creating and maintaining algorithms move fluidly between the analytic back office

(where AI and other technologies are tasked with discovering and processing data) and the analyst's front office (where analysts review machine-prepared and annotated data). This helps to ensure that the underlying software is catering to operational and mission needs. Bridging the divide between engineers and end-users through regular collaboration is essential. Over time, technical and analyst teams acquire a working understanding of each other's skill sets and gain a growing appreciation of the possibilities and limitations in AI's applicability to the mission.

In addition to powering enhanced machine analytics, AI serves as an important knowledge-management capability, bridging the retiring generation of baby boomers with the digital natives entering the analyst corps. If AI can be trained by experienced practitioners to “think like an analyst” as it processes raw intelligence, hard-earned analytic techniques developed over decades can be captured and disseminated for the benefit of incoming analysts.

WHAT DOES THE ANALYST 2.0 WORKPLACE LOOK LIKE?

New tools and technologies are most effective and lasting when tailored to the analyst's operational environment and mission, so they are embraced rather than ignored. The “killer app” for an analyst is a single interface that fuses multiple streams of raw intelligence at various classifications into a curated, intermediate product that the analyst can work from. Rather than analysts spending most of their time processing raw intelligence, this prepared data is pushed to the analyst, preassembled through a combination of predefined search criteria and automated processes.

Natural language processing allows analysts to task and query the system with a familiar user interface, similar in ease-of-use to what they expect from their personal smart devices. Based on an analyst's specific operational needs, the analytics might, for example, highlight anomalies among relevant data sets, suggest similarities between the analyst's target of interest and other data sets, or call out threats and opportunities that might otherwise go unnoticed. The ease with which searches are tasked and results are viewed allows frequent experimentation, fostering new approaches for tackling difficult intelligence problems.

The user interface accommodates varying levels of expertise and progressive mastery of its features, much in the way that most users of Microsoft Excel derive immediate value in its most basic features and can learn additional functions or extend its capabilities through scripting and third-party plug-ins as needed. Likewise, intelligence analysts operate within a technical framework in which they can incrementally exploit underlying technologies and attach new data sources and data models as they become available, regardless of source or vendor.

To do this, Analyst 2.0 also features open platforms and other architectures, as well as agile, iterative software development. Analysts are not locked into static, proprietary approaches that require frequent vendor interventions to update. Rather, the analytic tools operate within an open architecture design that accommodates multiple current and future technologies, more expansive arrays of intelligence sources, and regular, easy feature modifications. At the same time, AI models are developed on a continuously iterated loop of agile development, where embedded feedback mechanisms enable analysts — working closely with programmers and data scientists — to adjust and fine-tune them to their needs.

EARNING THE TRUST OF ANALYSTS

The Analyst 2.0 workplace we describe constitutes significant change; the real question is whether analysts come to view it as beneficial change that enhances, rather than complicates, their roles and jobs. In transitioning to an Analyst 2.0 environment, it is critical to build and maintain analysts' trust along the way — without it, analysts will simply revert to the tools and workflow they already know and use. Successful automation of rote tasks can be an early test: as analysts experience first-hand that time is being returned to them for higher-value tasks, suspicion and resistance typically fade.

Still, steps to enhance trust among analysts are needed all along the journey. The Mercury Project astronauts who undertook the U.S.' first man-in-space program famously insisted on a window for their spacecraft, in part so that they could manually orient themselves during an emergency. Similar "windows" need to be offered to analysts so they can confidently reorient themselves to new workflows. These "windows" can come in many forms, but their purpose is to reassure analysts they are seeing all relevant information that they need to see. Such systems are so common in the civilian world that they go almost unnoticed. When Google's Gmail service introduced automatic spam filtering, many users did not trust it to pick the right emails for removal. To address these concerns, the interface included a "window" in the form of a segregated spam folder, through which users could verify the algorithm's results. Combined with frequent human feedback to continuously improve the algorithms, the system is now so effective that most users rarely bother to verify its accuracy.

If an analytic has already winnowed down terabytes of data to a humanly manageable level, it should not be taxing for an analyst to manually dismiss the false positives that an AI-assistant will inevitably

produce. But the possibility of false negatives — the failure of AI to flag relevant data — represents a real and mission-critical problem. These concerns can be addressed through trust windows built into the interface that allow the analyst to exercise judgment over how the AI operates. A slider control, for example, can allow an analyst to calibrate the precision of an AI-informed analytic according to the importance of the task, so that even imperfect matches to a query are returned if desired. In time, with regular user feedback about the quality of the algorithm's inferences, the machine learning behind the AI will provide a much greater percentage of useful results and the analyst will come to trust its assistance.

There is only so much that good design will address. Trust also must be earned the old-fashioned way: through frequent and open communication among stakeholders. Such a large change in the institutional culture can be disruptive, so attention to change management and effective strategic communications is essential to minimizing uncertainty among the workforce. With Analyst 2.0, analysts are encouraged and empowered at all stages to take ownership over these changes to their workflow.

MOVING FORWARD WITH ANALYST 2.0

AI and other analytic approaches have the potential to fundamentally alter human work patterns, and analysts are justifiably wary of these changes. With the hype surrounding the promise of AI, some analysts may worry that the intention is not to assist them, but rather to replace them. The reality, however, is there is no AI technology on the horizon that can replace human judgment, and there has never been a greater need for the expertise of human analysts than today.

But without Analyst 2.0 tools, analysts will continually fall behind in their capabilities relative to their potential. Readiness will degrade as analysis fails to keep pace with incoming data and the expanding needs of military and other national security decision-makers. In an age where anything that can be sensed is recorded, it is simply impossible to make sense of the known digital world without the assistance of AI-informed analytics.

Technology is an important piece of Analyst 2.0. But technology alone will not enhance national security. By making sure that new intelligence tools are not just AI-informed, but analyst-informed as well, organizations can tap the potential of advanced analytics to empower analysts and enhance operational and mission effectiveness.

January, 2019



MITRE Statement to the Defense Innovation Board's Project on AI Principles

For additional information about this response, please contact:

Eliahu Niewood, Director, Cross-Cutting Urgent Innovation Cell

The MITRE Corporation

7596 Colshire Drive

McLean, VA 22102-7539

ehniewood@mitre.org

781-271-2436

First, we'd like to thank the Defense Innovation Board for the opportunity to briefly touch on the ethical considerations for military application of artificial intelligence. The MITRE Corporation is deeply committed both to ethical approaches to modern warfare and to enabling our Service men and women to have at hand the best technology available to protect them and to help them achieve their mission. Artificial intelligence clearly impacts both of those commitments. AI is a key emerging technology that will enable the Joint Force to fight and win future wars. Yet for several reasons, the Department has struggled to field relevant capabilities leveraging this technology. Some of these reasons revolve around AI's being developed largely in the commercial sector for consumer applications. Some revolve around technical challenges with dirty data and complex system dynamics. Some however revolve around ethical concerns related to AI weapons and military decision making. Clearly, the Department's integration of AI into military operations must be done in a manner consistent both with our country's ethics and the laws of warfare. We believe however that from an ethical perspective, AI is similar to a host of technologies that have preceded it and that have been fielded and used in ethical ways. In fact, we believe integrating AI into military systems and operations can help to reduce civilian casualties while providing our troops a critical military advantage.

For example, take Claymore mines, a remotely triggered anti-personnel device not banned by the Ottawa convention. Yet, they can be detonated by tripwire or other ways that don't require actually seeing the target. What if instead they came equipped with a sensor that only allowed detonation if the targets were determined to be adult-sized humans carrying weapons? Or take the tragic 1988 downing of an Iranian airliner by the USS Vincennes. The crew of the Vincennes was forced to make a split-second decision about the threat posed by an unknown aircraft before they fired the missile. What if instead the missile had an AI-based seeker which could distinguish between a civilian airliner or enemy aircraft and shut off its fuze, or even guided itself away from the aircraft? In both cases, as well as in many others, AI could enable both enhanced capabilities for our warfighters and reductions in the likelihood of non-combatant casualties. These examples highlight two of the three points we'd like to bring to your attention about AI's use in DoD system.

The first point is that AI is not a fundamental change in the way we employ advanced weapons. Many of the weapons in our inventory today select their own aimpoints or home in on a target within a set of constraints. The Tomahawk cruise missile, for example, uses seekers and guidance algorithms which correlate the surrounding terrain to onboard digital maps to guide itself to its target. Many air-to-air missiles "lock on" after launch, meaning that the weapon finds its own aimpoint when its seeker is turned on during flight. Torpedoes search out specific acoustic signatures, matching those signatures against onboard libraries. All of these weapons already make autonomous "decisions" about where they go and what they do once a human makes the decision to launch them. With AI technologies, we may have less real-time visibility into how the weapon makes a decision in a specific scenario, we may have more difficulty testing the weapon because of the complexity of the AI, but at a fundamental level the human has given up control and decision making with many existing weapons once they are launched. That launch decision, with or without AI inside the weapon, must be an ethical one that

balances risk to others with risk to the warfighter. That was true in WWII, that is true today, and that will still be true in the future.

A second point these examples highlight is that the human is not an ideal decision-maker, let alone a perfect decision-maker. Take the example of the USS Vincennes mentioned above. According to some reports, the Aegis Weapon System on the cruiser recorded that the Iranian aircraft was squawking a civilian transponder code and climbing away from the Vincennes at the time the weapon was fired. Under threat, forced to make a decision in a very short time, it is understandable that the crew of the Vincennes was not able to fully process all available information. In his book "The Fighters", C.J. Chivers describes a young US Navy pilot early in the war with Afghanistan launching a precision guided weapon, knowing at the time that something felt wrong about the weapon's target but not sure enough to hold off on the weapon's release. That pilot was haunted by that decision, never knowing whether it was right or not but wishing he could change it, for the rest of his career. Used properly, AI technology can lead to better decision making and should lead to reductions in errors that result in collateral damage and unnecessary civilian casualties.

The last point we would like to make today is that AI technology is not primarily focused on the "pointy end of the spear," directly making decisions to launch and point weapons. Far from it. Most of the applications envisioned for AI in the Department of Defense today revolve around other parts of operations, around better maintenance for aircraft, around fusing data from different sources, around finding "signals" in high volumes of data, and around making better strategic decisions. These applications not only do not directly put lives at risk, but could actually serve to better protect civilian populations, as well as our warfighters – even while dramatically improving our warfighting capabilities.

In closing, it is important to remember that there are three ethical commitments we must balance in any set of principles to be developed. We have an ethical responsibility to minimize harm to civilians in any military operation. We also have an ethical responsibility to our fellow citizens to find ways to use AI to enhance their security, whether that's in helping deter or defeat a North Korean nuclear weapon launch, finding a terrorist cell before they develop a dirty bomb, or preventing nation state cyber attacks on our power grid. And above all, we have an ethical commitment to our Soldiers, Sailors, Airmen and Marines, who put their lives at risk for all of us, to find ways to protect them and to provide them with the absolute best capabilities our nation can produce. AI can be a positive enabler for all of these commitments. Thank you for your time.

Chapter 19. Escalation Risks in an AI-Infused World

Herbert Lin

Stanford University
herblin@stanford.edu

Abstract

This chapter focuses on some of the potential downsides of AI-enabled military systems, specifically risks that arise from the potential of such systems to lead to conflict escalation: deliberate, inadvertent, accidental, and catalytic. Although such risks are present with the use of any new technology introduced into military systems, today's AI—in particular, machine learning—poses particular risks because the internal workings of all but the simplest machine learning systems are for all practical purposes impossible for human beings to understand. It is thus easy for human users to ask such systems to perform outside the envelope of the data with which they were trained, and for the user to receive no notification that the system is indeed being asked to perform in such a manner.

Introduction

As international security analysts contemplate the future of warfare, a common theme is that the weapons of the future and artificial intelligence will be integrally linked. AI, it is believed, will confer all kinds of military advantages to the side that best takes advantage of this revolutionary technology. To offer just a few examples, it has been said that AI will enable the autonomous targeting of weapons (Etzioni & Etzioni, 2017), the control of swarming battlefield vehicles (Baraniuk, 2017), and the speedy detection of militarily significant patterns in data too complex or voluminous for human analysis.⁵⁶

AI may indeed afford military planners and warriors with all those capabilities, and more. But the fact that some work to date suggests the possible feasibility of such applications is not the same as seeing an actual, delivered, proven capability to troops on the battlefield. Moreover, little analysis or commentary has been devoted to considering the downsides of an AI-infused conflict environment—downsides that may redound to the detriment of U.S. planners and warriors.

Many downside risks arise from the introduction of AI into military systems and planning, some of which include uncertainty about accountability regarding the use of AI-enabled weapons systems in lethal operations, integration of human-smart machine military “teams”, impact on the culture and organization of the armed forces, and effects on adversary perceptions of the United States (see also Section 3.1 of Chameau, Ballhaus, & Lin, 2014). This paper focuses at risks in the context of escalation dynamics—how a military conflict's scope and intensity might escalate, but first it is necessary to review certain characteristics of AI relevant to this focus.

⁵⁶ For example, the DOD published Establishment of an Algorithmic Warfare Cross-Functional Team (popularly known as Project Maven)(2017) to accelerate DoD's integration of big data and machine learning. The team's objective is “to turn the enormous volume of data available to DoD into actionable intelligence and insights at speed.”

The Scope of Today's AI

AI is a broad term whose precise scope is contested. For example, many military leaders conceptualized AI in terms of their application domains—lethal autonomous weapons or smarter decision support systems as “AI.” Technologists are more likely to see AI as an underlying technology that enables many different applications. Even so, lines between “artificial intelligence” and big data, algorithms, statistical learning, and data mining are blurry at best. In the early days of AI, AI relied primarily on a symbolic approach—that is, an approach to problem solving that relies on high-level representations of problems, logic, rules, knowledge, and search. Despite some early successes, this approach gradually lost favor in the 1980’s as researchers came to appreciate more clearly the enormous difficulty of developing such useful high-level representations.

Today, the most prominent approaches to AI rely on machine learning (ML), a class of techniques that often (but not always) relies on the availability of large amounts of data. “Supervised ML” depends on training data that has been labeled by humans and makes statistical inferences. “Unsupervised ML” finds clusters and outliers in unlabeled data that might otherwise go unnoticed if examined by humans.

But by themselves and unaided, ML techniques provide neither explanation for the inferences drawn nor the significance of the clusters. In other words, AI systems based on ML are unable to explain to their human users why they reach the conclusions they reach or demonstrate the behavior they demonstrate. Even worse, human examination of the machine’s output and how it was derived from the input does not help, as it generally yields little about the features of the input that led to the inference in question. At least at first, users must simply trust that the system is behaving properly; over time, their trust grows if the system repeatedly behaves properly.

For many applications, explanations are simply unnecessary, and the inability to explain why a given result was produced is merely a curiosity. For example, when a user searches for a given book on Amazon, an ML-based recommender system provides suggestions of other books the user might wish to purchase. But such applications are generally applications with low stakes where an explanation does not particularly matter to most human user.

Trust in ML applications is properly limited to those operational scenarios that have been well-covered in the training data—ML applications are least trustworthy in scenarios that have not been well-covered, that is, in novel scenarios. (This phenomenon is arguably the reason that algorithmic bias arises in improperly vetted ML algorithms—an ML algorithm misidentifies human beings of African descent as gorillas because it has not been trained on an adequate sample of pictures of black human beings (see, for example Gynn, 2015)). In novel scenarios, explanations may very well be a necessary foundation for humans to properly trust ML applications.⁵⁷

A difficult problem that requires solution arises from the reality that an ML application must be able to distinguish between input data from the universe of data on which it has been trained (i.e., routine scenarios) and input data from outside that universe (i.e., exceptional or novel scenarios). For

⁵⁷ “Explainable AI” is the focus of a DARPA research program (see Gunning, n.d.) that “aims to create a suite of machine learning techniques that [p]roduce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and [e]nable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.” That said, the reason that this DARPA program exists in the first place is that the problem is a very hard one, and it is fair to say that the techniques of explainable AI have not made it in to common use. Whether or not they will ever do so remains to be seen.

example, consider an application is trained to distinguish between different breeds of dogs. The training data set consists of a very large number of labeled dog pictures. Give the application a picture of a random dog, and its output is the breed of dog that is most likely for that picture. This is a routine scenario for which the application is designed.

But what happens if instead the application is given a picture of a dolphin? Although it could not be expected to identify it as a dolphin (since it was never exposed to training data involving dolphins), it would be desirable if the application itself could recognize that it is now being expected to operate outside its zone of competence and inform the user of that conclusion.

The application must distinguish between two types of input data that it has never seen before. The first is routine—it is new, but it is generally similar to the training data. If processing routine input data, the application should provide its best guess (e.g., what breed of dog was shown). The second is novel—it is also new, but it is highly dissimilar to the training set. If processing novel input data, the application should produce an indication that it is operating outside its capabilities and that its output should be less trusted. The hard problem to solve is how to differentiate between “different in detail but generally similar” and “highly dissimilar.”

Pimental et al (2014) describes a number of approaches that yield partial solutions for the problem described above, generally known as novelty detection. But most importantly, they note that defining “novelty” is conceptually a difficult problem, and thus it is not possible to suggest one “best” method of novelty detection. They go on to suggest that “the variety of methods employed is a consequence of the wide variety of practical and theoretical considerations that arise from novelty detection in real-world datasets, such as the availability of training data, the type of data (including its dimension, continuity, and format), and application domain investigated. It is perhaps because of this great variety of considerations that there is no single universally applicable novelty detection algorithm.” All of the approaches described by Pimental et al (2014) involve elements of human judgment, and thus it is reasonable to conclude that in general (i.e., for any supervised ML application), some novel instances of new input data will not be identified as novel. In the absence of such identification, the user will unknowingly assume the ML is acting within the parameters of a tried and trusted application without realizing that the application is now operating outside its zone of competence. That way lies potential disaster.

AI Everywhere

If predictions that AI is an enabling technology of the future actually come true, we will see AI of various types and functions ubiquitously embedded in the devices and infrastructure of both civilian and military life. We will see AI-enabled capabilities support myriad non-military activities throughout society. As illustrative examples, AI will be embedded in self-driving cars and other autonomous and semi-autonomous vehicles; decision-support systems for investors and health care providers; automatic translation and transcription systems; identifying potential suicide victims; marketing products and services to individual consumers; predictive policing; and crop/soil monitoring and predictive analytics regarding agricultural yields.

On the military side, AI-enabled capabilities will be found in weapons systems, controlling one or a number or all of their functions, possibly including navigation, propulsion, weapons targeting, weapons release, and so on; in sensor systems and systems for intelligence analysis, identifying patterns and sifting through large volumes of disparate data and possibly providing likely interpretations of such patterns; in decision support systems, providing recommended courses of

action in response to particular sets of circumstances. Most importantly, AI-enabled capabilities will be available for use by all parties to a conflict.

Where AI applications are ubiquitous, they are—almost by definition—not novel. But novelty, among other things, is an important driver for skepticism. Human users who are appropriately skeptical of new technology do not give their trust without sufficient evidence, and they themselves will act as “second opinions” to judge the accuracy and propriety of their applications’ output. A plethora of skeptical users would indeed be reassuring. But the experimental data does not provide such reassurance. For example, in a 2016 study, individuals followed the directions of a robot in a (simulated) emergency evacuation scenario, even though they had observed the same robot perform poorly in a navigation guidance task a few minutes before. Even when the robot pointed to a dark room with no discernible exit, the majority of individuals did not choose to safely exit the way they entered (Robinette, Li, Allen, Howard, & Wagner, 2016).

Without widespread skepticism, ubiquitous AI will inevitably become part of the background, and its affordances for society (i.e., the beneficial capabilities it provides for society) will disappear from conscious attention and thought, much as electricity disappeared into the background and became taken for granted in the 20th century. And it should further be noted that user skepticism that prevents automatic reliance on an AI-based system may in some instances defeat the very purpose of introducing that system in the first place. Specifically, AI capabilities may have been added to increase the system’s speed of operation—in this context, why would it be desirable for a human user to take the time to check or second-guess the machine’s decisions and conclusions? This point itself will drive human users in the direction of unquestioning trust.

Escalation Dynamics

As a point of departure, consider that escalation in a conflict may arise through a number of different mechanisms (which may or may not simultaneously be operative in any instance).⁵⁸

- Deliberate escalation is an intentional choice by one party to intensify the conflict. In principle, the escalating party has made this judgment based on its understanding of its own and the other side’s capabilities and intentions, and acts according to the belief that escalation will bring advantages.
- Inadvertent escalation occurs when one party deliberately takes actions that it does not believe are escalatory but are interpreted as such by another party to the conflict. Such misinterpretation may occur because of a lack of shared reference frames or incomplete knowledge of the other party’s thresholds or “lines in the sand.”
- Accidental escalation occurs when some operational action has direct effects that are unintended by those who ordered the action. A weapon may go astray to hit the wrong target; rules of engagement are sometimes unclear; a unit may take unauthorized actions; intelligence on a target may be faulty; or a high-level command decision may not be received properly by all relevant units.
- Catalytic escalation occurs when some third party succeeds in provoking two parties to engage in conflict. For example, C takes action against A but makes it look like the action came from B. C then observes as A takes action against B, and B may well respond against A for what B sees as an unprovoked attack from A.

⁵⁸ The first three types of escalation are described in greater detail in Forrest Morgan et al (2008). Lin (2012) built on this work to explore escalation dynamics in cyberspace and added the fourth type of escalation—catalytic escalation.

Escalation Dynamics in an AI-Infused Conflict Environment

Central to each of these escalation mechanisms is the scope, nature, and quality of information available to decision makers. How might AI-enabled capabilities lead to or facilitate different kinds of escalation dynamics, by which is meant how hostilities might escalate over time?⁵⁹ The following discussion suggests some illustrative, but by no means comprehensive, possibilities.

Deliberate escalation

Party A may choose to escalate if it believes its military capabilities are sufficiently powerful to defeat B's response to that escalation. But if A's actual capabilities do not match A's estimate of its own capabilities, defeat or disaster may result from escalation. In particular, A may believe that its own AI-enabled military decision support systems have been trained on an adequate universe of cases, but actual conflict often falls outside the parameters of what planners expected before the conflict started—unexpected tactics or weaponry, for example. But these systems will dutifully do the best they can without users recognizing critical differences between data from actual conflict and its training data. The systems may thus offer conclusions that go beyond their expertise or recommendations that are accepted by humans who do not notice the out-of-scope situation.

Inadvertent escalation

Party A takes an action that it does not believe Party B will (or should) regard as escalatory. For example, Party A attacks B's ballistic missile early warning satellites early in a conventional kinetic conflict, because those satellites are providing tactical advantages for B in locating the launch sites of A's non-nuclear tactical ballistic missiles. B sees such actions as a prelude to nuclear attack of A on B, because those satellites are also used to warn B of a nuclear attack. B believes that A must know that such an anti-satellite attack would be hugely escalatory, but A believes it is simply trying to negate a tactical advantage for B. Thus, A's attack on B's satellites is interpreted by B as an escalation, and B responds in kind. Because A did not believe its anti-satellite attack was escalatory, A sees B's response as an unwarranted escalation rather than a response—and this sequence of events sets off an (inadvertent) escalatory spiral.

Assumptions about thresholds are likely to be built in to ML-based decision support systems. That fact in itself is not bad—one must start somewhere. But how will the differing perspectives of adversaries be acknowledged, taken into account, and flagged explicitly for human attention? Indeed, inserting information about adversary thresholds into such support systems would require the availability of substantial data on those thresholds. But if such data were available and were deemed important, the problem of not knowing or realizing the adversary's thresholds would not exist in the first place. Radically different views of the adversary's motives and intentions are not mere parametric tweaks in a model of conflict—rather, they call into question the underlying utility of such a model for understanding how a conflict might unfold.

Accidental escalation

⁵⁹ This phraseology is intended to capture the idea that even before hostilities break out, adversaries are in a continuous cycle of reacting to the actions and intentions of others. While arguably most important in setting the strategic stage for the outbreak of hostilities, the state of affairs prior to the outbreak of hostilities is not addressed in this short paper. Another paper will someday focus of this topic.

A certain weapon of Party A relies on in-flight AI-based imagery analysis for automatic target recognition. Whilst flying at night, the weapon sees a building with gunfire flashes coming from the windows. The building is identified through target databases as being a hospital, but because a hospital becomes a valid military target if an enemy is using it as a base for military operations, the building is destroyed. In reality, the gunfire flashes were reflections from gunfire emanating from Party B's troops stationed around, and not in, the hospital. However, Party B does not realize this fact at the time, and the conflict escalates because the target recognition algorithm did not take into account the possibility that reflections of gunfire flashes might be mistaken for the real thing.

A variant of this scenario could involve an adversary tricking the AI in the automatic target recognition system. For example, Party B may be able to spoof the imagery of a hospital received by the weapon in flight in such a way that the weapon identifies it as a valid military target, and the hospital is destroyed. But the spoofing occurs in such a way that to the human eye, the imagery captured from the weapon's camera is indistinguishable from the image of a hospital, even though it was sufficient to fool the target recognition algorithm. (This point is addressed in more detail by Libicki's Chapter 18 "The Hacker Way of Warfare.")

Catalytic escalation

Party C seeks to provoke conflict between Party A and Party B. To this end, it constructs deepfake videos and audios, which are realistic audio or video files depicting senior individuals within the decision-making apparatus of A and B saying things that he or she never said. These videos and audios are clandestinely selectively injected into the intelligence collection streams of A and B—videos and audios depicting individuals from A are injected into B's collection systems, and vice versa. If the content of these pseudo-recordings is tailored properly, it is easy to see how they might provoke A or B into taking actions that the other might regard as the first step on an unprovoked escalatory path.

Discussion and Conclusion

The scenarios described above are illustrative. But all such scenarios suffer from the analytic issue that once a problem is anticipated and described, a fix for the problem can be easily imagined—and thus the scenario is easier to dismiss as unfounded. But the point of this chapter is to instill some degree of humility in human ability to anticipate all such problems, and thus to realize that with the advantages of AI-enabled military systems come some potential disadvantages.

Of course, the same could be said about technologies in general (including more traditional cyber tools)—any technological solution will fail when operated far enough outside the parameter envelope that defines the problem to be solved. Is there anything special about AI that is more problematic?

For the machine learning flavor of AI, the answer is yes. It was noted above that the human user has no way to know that an ML application is dealing with a novel scenario, i.e., one that falls outside the envelope of the data on which it has been trained. And the reason for this lack of knowledge is that examination of a machine learning algorithm's operation generally defies human comprehension—that is, a human being will find it impossible to tell what an ML-based computer system is doing in any given case. (It is for this reason that explainable AI is necessary in the first place.) In this regard, an ML application is much unlike other technological artifacts, whose design limits are much better understood. We implement ML-based systems with the going-in realization that we cannot

understand how they produce a given output from a given input—and in most other systems, such a lack of understanding would be a dispositive strike against it.

A second problematic dimension of AI-enabled military systems arises from the likely ubiquity of AI as an underlying enabling technology throughout all of society, both civilian and military. When a technology is ubiquitous, users take it for granted and tend to lose their skepticism about it—even though even ubiquitously deployed technologies exhibit flawed operation from time to time. When ubiquitously deployed technology fails, users are more likely to look to the circumstances of the particular failure rather than to any underlying problem that may be more fundamental. Consequently, human attention is less likely to be focused on underlying problems.

The policy recommendations that flow from the analysis above are modest but significant. First, maintaining a degree of skepticism about the application of AI to military systems is necessary for all policy makers. Skepticism does not mean that such application should be rejected out-of-hand, but it does mean keeping in mind that the promises of vendors and contractors are often inflated beyond any reasonable measure. Asking “what could go wrong?” is a good question to ask, early and often. Red teaming against AI-enabled military systems is one way to maintain such skepticism, but such efforts must be conducted from the inception of a system’s design through operational deployment so that the consequences of proceeding down the AI-enabled path are clearer.

Finally, increased research may well be needed on to advance the state of the art in explainable AI in a military context. Such research has two flavors: (a) research that can help explain what ML-based AI systems are doing and why they reach the conclusions they reach; and (b) renewed research on symbolic AI, whose explicit rules and logics provide, in principle, basic building blocks for comprehensible explanations.

References

- Baraniuk, C. (2017, January 20). US military tests swarm of mini-drones launched from jets. *BBC News*. Retrieved from <https://www.bbc.com/news/technology-38569027>
- Chameau, J., Ballhaus, W. F. & Lin, H. L. (Eds.). (2014). *Emerging and readily available technologies and national security — A framework for addressing ethical, legal, and societal issues*. Washington DC: National Academies Press.
- Establishment of an algorithmic warfare cross-functional team (Project Maven). (2017, April 26). Memorandum from the Deputy Secretary of Defense. Retrieved from https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf
- Etzioni, A. & Etzioni, O. (2017, May-June). Pros and cons of autonomous weapons systems. *Military Review*, 97(3), 72-81. Retrieved from <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/>
- Morgan, F. E., Mueller, K. P., Medeiros, E. S., Pollpeter, K. L., & Cliff, R. (2008). *Dangerous thresholds: Managing escalation in the 21st century*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG614.pdf

- Gunning, D. (n.d.) Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency Program Information. Retrieved from <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Guynn, J. (2015, July 1). Google Photos labled blanck people 'gorillas'. *USA Today*. Retrieved from <https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/>
- Lin, H. (2012 Fall). Escalation dynamics and conflict termination in cyberspace. *Strategic Studies Quarterly*, 6(3), p. 46-70. Retrieved from <http://www.au.af.mil/au/ssq/2012/fall/lin.pdf>
- Pimentel, M., David A. Clifton, Lei Clifton, Lionel Tarassenko (2014). A review of novelty detection, *Signal Processing* 99: 215–249. Retrieved from <http://www.robots.ox.ac.uk/~davidc/pubs/NDreview2014.pdf>.
- Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. Proceedings from 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand. Retrieved from <https://ieeexplore.ieee.org/document/7451740>

Principles of AI Governance and Ethics Should Apply to All Technologies

By Herb Lin Friday, April 12, 2019, 11:59 AM

DayZero: Cybersecurity Law and Policy

Despite Google's recent dissolution of its artificial intelligence (AI) ethics board, IT vendors (including Google) are increasingly defining principles to guide the development of AI applications and solutions. And it's worth taking a look at what these principles actually say. Appended to the end of this post are the principles from Google and Microsoft, thoughts from Salesforce.org (closely aligned with Salesforce), and AI principles from three groups not aligned with specific companies.

Viewed from a high level of abstraction, three major points stand out for me:

- As articulated, the principles are unobjectionable to any reasonable person. Indeed, they are positive principles that are valuable and important.
- They are broadly framed and highly subjective in their interpretation, a point that should focus attention on precisely who will be making those interpretations in any given instance in which the principles could apply. The senior management of a company? The developers and coders of particular applications? The customers? Elected representatives? Career civil servants? The United Nations? A representative sample of the population? One could make an argument—or counterargument—that any of these actors should be in a position to interpret the principles.
- Perhaps most importantly, none of the principles is particularly related to artificial intelligence. This can be shown by simply replacing the term “autonomous” or AI (when used as an adjective) with the term “technology-based.” When AI is used as a noun, simply replace it with the word “technology.”

I conclude from this high-level examination of these principles that they are really a subset—indeed a fully contained subset—of ethical principles and values that should always be applied across all technology development and applications efforts, not just those related to AI. In the future, I'd like to see technology companies—of all types, not just those using AI—make explicit commitments to the broader set of principles for technology governance.

Of course, questions would remain about about subjectivity of interpretation and the locus of decision making. But even lip service to principles of technology governance is better than the alternative—which is disavowal of them through silence.

AI Governance Principles From Various Companies and Organizations

AI principles from Microsoft:

Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values.

- Fairness: AI systems should treat all people fairly
 - Inclusiveness: AI systems should empower everyone and engage people
 - Reliability & Safety: AI systems should perform reliably and safely
 - Transparency: AI systems should be understandable
 - Privacy & Security: AI systems should be secure and respect privacy
 - Accountability: AI systems should have algorithmic accountability
-

AI principles from Google:

We will assess AI applications in view of the following objectives. We believe that AI should:

- Be socially beneficial.
 - Avoid creating or reinforcing unfair bias.
 - Be built and tested for safety.
 - Be accountable to people.
 - Incorporate privacy design principles.
 - Uphold high standards of scientific excellence.
 - Be made available for uses that accord with these principles.
-

Salesforce (and salesforce.org):

AI holds great promise — but only if we build it and use it in a way that's beneficial for all. I believe there are 5 main principles that can help us achieve beneficial AI:

- Being of benefit
 - Human value alignment
 - Open debate between science and policy
 - Cooperation, trust and transparency in systems and among the AI community
 - Safety and Responsibility
-

European Commission:

AI should respect all applicable laws and regulations, as well as a series of requirements; specific assessment lists aim to help verify the application of each of the key requirements:

- Human agency and oversight: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy.
 - Robustness and safety: Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems.
 - Privacy and data governance: Citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them.
 - Transparency: The traceability of AI systems should be ensured.
 - Diversity, non-discrimination and fairness: AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility.
 - Societal and environmental well-being: AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility.
 - Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.
-

Asilomar AI Principles:

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

Ethics and Values

- **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
- **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
- **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
- **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
- **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- **Shared Benefit:** AI technologies should benefit and empower as many people as possible.
- **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
- **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.

Attendees at the the New Work Summit, hosted by the New York Times, worked in groups to compile a list of recommendations for building and deploying ethical artificial intelligence:

- **Transparency:** Companies should be transparent about the design, intention and use of their A.I. technology.
- **Disclosure:** Companies should clearly disclose to users what data is being collected and how it is being used.
- **Privacy:** Users should be able to easily opt out of data collection.
- **Diversity:** A.I. technology should be developed by inherently diverse teams.
- **Bias:** Companies should strive to avoid bias in A.I. by drawing on diverse data sets.
- **Trust:** Organizations should have internal processes to self-regulate the misuse of A.I. Have a chief ethics officer, ethics board, etc.
- **Accountability:** There should be a common set of standards by which companies are held accountable for the use and impact of their A.I. technology.
- **Collective governance:** Companies should work together to self-regulate the industry.
- **Regulation:** Companies should work with regulators to develop appropriate laws to govern the use of A.I.
- **"Complementarity":** Treat A.I. as tool for humans to use, not a replacement for human work.

Dr. Herb Lin is senior research scholar for cyber policy and security at the Center for International Security and Cooperation and Hank J. Holland Fellow in Cyber Policy and Security at the Hoover Institution, both at Stanford University. His research interests relate broadly to policy-related dimensions of cybersecurity and cyberspace, and he is particularly interested in and knowledgeable about the use of offensive operations in cyberspace, especially as instruments of national policy. In addition to his positions at Stanford University, he is Chief Scientist, Emeritus for the Computer Science and Telecommunications Board, National Research Council (NRC) of the National Academies, where he served from 1990 through 2014 as study director of major projects on public policy and information technology, and Adjunct Senior Research Scholar and Senior Fellow in Cybersecurity (not in residence) at the Saltzman Institute for War and Peace Studies in the School for International and Public Affairs at Columbia University. Prior to his NRC service, he was a professional staff member and staff scientist for the House Armed Services Committee (1986-1990), where his portfolio included defense policy and arms control issues. He received his doctorate in physics from MIT.

 **HerbLinCyber**

Comments for the DIB Consultation on AI Policy Principles

Associate Professor Seth Lazar, Australian National University

Introduction

Recent months have seen organisations from small corporations, to national and regional governments, and multinational tech companies, putting forward principles to guide their adoption and use of artificial intelligence systems. A number of themes have emerged, to the extent that each new set of principles tends to be a subset of the last. Most lists of principles confusingly include general goals that should be aimed at in any endeavours of a given organisation, alongside very specific issues raised by AI in particular. I welcome the DIB's approach of not simply rehashing existing policy guidelines, but taking time to think seriously about what makes AI different, and why we need new principles to govern it. I would encourage them to spend at least as much time thinking about what makes Defense different from the other organisations that have set out these policy statements. These differences, I think, make the principles around which other organisations have coalesced much less useful for Defense. I'll begin this comment by asking first what makes AI different, then what makes Defense different. I'll then ask what the goal is, of developing a set of AI policy principles. Finally I'll consider which kinds of principles we can propose. In particular, I will argue that some of the principles that so frequently appear in other organisations' lists should apply quite differently, if at all, to Defense.

What Makes AI Different? Why Does that Matter?

What is AI?

The term 'Artificial Intelligence' is exceedingly vague. We cannot hope to provide a definition of what makes something AI. So it is better to focus on those aspects of AI in which we are interested for the purposes of this endeavour. Our primary interest is in machine-learning based predictive analytics and decision technologies. We could extend this to include symbolic-logic-based decision technologies, especially insofar as they are integrated in new ways with machine learning and new capacities for data gathering. But the revolution that prompts this call for a set of AI policy principles is grounded in machine learning, and symbolic-logic-based AI has been part of military systems for a long time. Machine learning itself is dependent on data. So when thinking about policies for the use of AI, it is best to think about policies that govern not only how AI systems are used, but the preconditions for using them—in particular, how data are gathered. It's also worth remembering that the 'Gee-Whiz' approach to AI is generally unhelpful: as it stands, it is a deeply fallible technology. In some areas it offers improvements on human judgment, but it is by no means perfect.

My task in this section is to ask what makes AI, so construed, different (n.b. a difference in degree is just as relevant here as a difference in kind). Since our topic is what to do about AI, I'll focus on the differences that I think are morally important. I'll identify the difference, then say why it is morally important—in particular, by identifying the moral risks associated with it. To do this, it is helpful to distinguish between risks that arise from the nature of the technology itself, and those that arise from its effects in use. I'll talk about them in turn.

AI is a Decision Technology

At heart, AI is a decision technology. Even when its function is predictive, it is aimed at decision support, and its whole method of interpreting data is grounded in iterative decisions made within machine learning algorithms. This feature of AI is its most interesting, and most significant, difference from other technologies. To be clear, though, there are many very simple mechanistic systems that, in effect, 'make decisions'. It's hard to pick apart what distinguishes AI from, say, an elevator. It's likely to be a cluster of properties rather than any single thing. But one key feature of AI is that it is a *probabilistic* decision technology. Elevators are closed systems based on a representation of the world that has no room for uncertainty.

Why does it matter that AI is a decision technology? First, while every technology is shaped by the values of the society that makes it (and then shapes those values in turn), AI is distinctive in that those values are written into the code and are in principle revisable. In some cases they have to be explicit. Second, because of the relative success, speed, and scalability of AI, we are now seeking to embed it in seemingly every area of human endeavour. There is literally no scenario where the possibility of understanding the world better, and making better decisions under uncertainty, is not intoxicatingly attractive.

So, what difference does it make that AI is being used to make many decisions that were previously made by humans, as well as many other decisions that we didn't, before AI, have the capacity to make? Why does this difference matter morally? I can think of three reasons (here as elsewhere I draw on the rich public discussion of these topics; but since this is a public submission rather than a scholarly document, I will not attempt to trace the origins of each idea).

1. I think the key point is that when decisions are made by AI, the responsibility for those decisions is **diffused** and **diluted**. This is because (and on the assumption that) AI cannot be responsible for its 'actions'. Responsibility is diffused, in the following sense: instead of being able to attribute the decision primarily to the human who made it, as well as to any who knowingly set her on that course, the decision must instead be attributed to the humans who designed software and hardware elements of the AI system, those who developed the training data set, those who set it in motion, and those with final authority over its decisions (this is an incomplete list). Responsibility is diluted, insofar as the nature of machine-learning based AI is such that AI systems are never wholly predictable by their designers or operators, so a key precondition of moral responsibility—that one can foresee the consequences of one's action—is either unsatisfied, or only partially satisfied.

With any of these claims, it is important to bear in mind the contrast with having humans make the relevant decision. Humans are often just as unpredictable and inscrutable as AI systems. However, if one person implements a decision that is in some sense collective (for example, a combatant carrying out an order), we can hold the final person in the causal chain accountable in a way that is not true for an AI system. So if we hold everything constant besides whether the final actor is an AI system or a person, there is still a dilution and diffusion of responsibility.

Why should we care if responsibility for AI decisions is diluted or diffused? Notice that this has nothing to do with whether the AI makes the right decision or not. The success rate of the AI system is in principle completely independent of these facts about responsibility. And we might reasonably care most about the actual results of the system. However, we do not in general care only about outcomes—we care also about process. Even a guilty person deserves a fair trial, for example. And as well as wanting our decision technologies to make the right decisions, we also know that perfection is unattainable, and in the event that they get things wrong, we fundamentally want to have someone to blame. This serves two purposes. One is deterrence and guidance: if people know that they will be blamed for wrongdoing, that gives them additional motivation to do the right thing. But we also care about apportioning blame after wrongdoing appropriately, independently of these instrumental benefits. Note that this is not the same as saying that, eg, punishment is a non-instrumentally valuable response to guilt. I am making only the weaker claim that there is value in being able to blame those who act wrongly, when wrongdoing is done. If nobody is really fully blameworthy, then that deprives us of an important way in which we respond to wrongdoing. We can state this point in a general form: **the use of AI systems to make decisions threatens to undermine our practices of accountability.**

2. There is a second, somewhat more speculative reason to regret the vesting of decision-making authority in AI systems. Many people clearly have the intuition that they would prefer it if decisions affecting their lives were made by a person, rather than an automated system. It is often quite difficult to extract the rational kernel from this argument, but I think it is something like this. We are social animals. We value relations of mutual respect and care. Those relations are advanced when we make decisions that have significant impacts on one another's lives ourselves. Making decisions ourselves, rather than vesting them in automated systems, involves (when done conscientiously) paying attention to others, taking them seriously in one's deliberations.

Again, it is important to remember that the AI system is just one point in the causal chain that would lead to a particular outcome. Within that causal chain, there will undoubtedly still be human decision-

makers. However, the further removed they are from the people ultimately affected by that causal chain, the less able they are to attend to them, as individuals, in their deliberations.

In a general form: **the use of AI systems in decision-making threatens to undermine the degree it which we are seriously attended to in the moral deliberations of those whose actions materially affect our lives.**

3. The third morally relevant difference arises from the distinctive nature of contemporary AI systems, and in particular the often-noted point that they are not readily understandable by either their designers or their operators. This is a technical claim, and work is currently underway to remediate this concern, for example by testing an algorithm's sensitivity to various changes in the underlying variables. Nonetheless, it's true now that if we vest decisions in AI systems grounded in machine learning, we will be able to verify the outcomes of the system's decision-making, but we cannot always explain why the system reached that decision.

Again, this merits comparison with human decision-making. We often rely on intuition and instinct in our own decision-making. It can be hard to explain our reasons for acting. And yet we can be called upon, *ex post*, to rationalise our behaviour. And we can be called out if our rationalisation is self-serving, insincere, or otherwise flawed. So there is a genuine contrast between algorithmic and human decision-making. But why should we care about it?

We should want people not only to do the right thing, but to do the right thing for the right reasons. This is in part because we care about their character, not only about the results of their actions. But we also care about the kind of attitude they display towards other people—and acting on the right reasons is one way to show appropriate respect for your moral equals. Another way to look at this: we want people to do the right thing not by mere luck. We want them to do the right thing robustly. If they act on the right reasons, this suggests that they would do the right thing even if the circumstances were somewhat different. Another key concern is that sometimes, some considerations might be relevant to predictive accuracy, but might be the wrong kinds of reasons to base one's decisions on (on analogy with inadmissible evidence in law, and also see the point about discrimination below). **When we vest decision-making authority in inscrutable AI systems we might make better decisions, but we may not know the reasons for which those decisions were made.**

AI Depends on Data

The collection and operationalisation of data is of course nothing new. This is a case where the difference made by AI is scale, speed, and effectiveness. Advances in AI make it possible to do incredible things with the data that we gather, and thereby incentivise gathering ever more data. This raises some obvious problems. Since they have already received considerable attention in the literatures on AI, to which I don't have much to add, I will discuss them briefly.

1. Individual privacy. Below we'll come to the ways in which AI can be *used* to undermine individual privacy. But concerns about privacy are also intrinsic to the nature of machine-learning based technology, at least when it is used for decisions about people. The simple observation here is that if an AI system is to be used to make decisions about people, it will be more effective the more data it has about them, so this creates an incentive to gather ever more data about us. Even if we retain some kind of control over that data—consenting to its use—there are real questions about whether that kind of consent is meaningful, and whether, even if we do consent, we should create a world in which the sphere of freedom in which we are not observed is ever diminishing. **AI technologies inherently threaten individual privacy.**

2. Discrimination. This is now the most readily recognised problem with the data-dependency of machine learning algorithms. Our datasets reflect structural injustices in the world as it is, and algorithms that learn from those datasets inherit those injustices. This is in fact part of a more general phenomenon—if decisions are going to be made based on historical data, then that builds in an unavoidable conservatism into our practices of decision-making. However, the key point here is: **Implementation of AI against a background of unjust social discrimination is likely to perpetuate and exacerbate that discrimination.**

The Effects of AI

The considerations just adduced all have to do with the nature of AI. But the other thing that makes AI distinctive is simply its capacity to affect every area of human life, making things possible at a scale and speed that was never possible before. AI can in principle be used for any purpose. Almost anything we can now do that is of moral concern can be done faster and at greater scale with the aid of AI. So the moral issues raised by the *use* of AI technologies cover everything that matters. This is an important point: it means that any set of principles governing the *use* of AI should really be a concise statement of the principles governing society as a whole (it also means that the principles governing AI should to a large degree be the same as those governing any general purpose technology).

There is lots of excellent research on the potential problematic effects of AI. I don't have anything really to add to it, but it might help to provide some overarching structure with which to think about them. The moral riskiness of a given application of AI seems to be a function of three factors: stakes, pervasiveness, and degree of autonomy.

By stakes, I mean: how much does this particular application matter morally? Does it significantly affect people's lives? Does it put individual rights at risks? And so on. Some use-cases are morally relatively neutral, at least on their face. Personal assistants and music recommendation algorithms, automated mining platforms and vehicles, photo editing algorithms and so on. As I'll observe below, there are ways in which even these can be morally significant—mostly insofar as they impact on what I'll call 'recognition goods'. But on the whole they involve AI systems that are not making explicitly morally-loaded decisions, so where the stakes are lower than they would be if the system were making high stakes decisions. It's worth observing that sometimes the stakes for individuals might appear to be low stakes, but for communities as a whole they are high stakes (eg when trading my data is individually rational but collectively irrational).

But even when the stakes of particular decisions are low, if a given AI system is pervasive within society, that can itself raise the moral risk. The very fact that one cannot escape it becomes an issue. If every algorithm for editing photos is much more successful at editing pictures with white faces in them, than with black faces, then that's much more of a concern than if there is enough competition, and one can avail of an alternative that suits one's needs. If I am refused credit by one bank, that might not represent a significant moral risk, but if every bank is using the same credit score algorithm, and I fall into a blind spot that they all share, then that is a big deal. And pervasiveness matters also because of vulnerability. The more pervasive a system is, the more dependent we are on it, so the more its vulnerability to attack matters for society, other things equal.

Lastly, degree of autonomy obviously raises moral risks too. By autonomy, here, I mean the ability of the AI system to affect the world without intervening confirmation or verification from a human operator. The less autonomy an AI system has, the greater the prospect there is for decisive human control, reducing the marginal difference between AI-decision-making and existing ways of making decisions. Some AI systems are really just decision-support tools. That doesn't mean they involve no risks—in particular, we need to be very cautious about automation bias, the tendency of human decision-makers to defer to automated systems. But it does reduce their risks relative to situations where they are fully autonomous.

These factors are independent of one another. The moral risks of a given application of AI might be significant if only one of them is raised. It should be quite easy to list the different ways in which AI can be used, and measure their moral risks in terms of these three factors. The use of AI to generate 'deep fakes', as well as adversarial uses of AI to spoof machine learning systems, clearly meet the high stakes criterion, though not so much the pervasiveness and autonomy criteria. Uses of AI to surveil a population are much the same, though they also threaten to be pervasive. AI for autonomous vehicles involves high moral stakes, and high autonomy, and potentially high pervasiveness. Use for medical diagnosis involves high stakes, but low pervasiveness, and hopefully low autonomy (since it is just a system for making recommendations to trained professionals). Government uses of AI for service delivery and welfare allocation threaten to be high stakes and pervasive; it is yet to be seen how autonomous they will be, though in these cases especially the risk of automation bias is very high.

AI and Power

One further gestalt effect of the rush to adopt AI technologies is worth drawing out. It is the result of a suite of different AI applications, as well as the other technologies on which they are based. AI systems have the capacity to radically alter existing power relationships between citizens and corporations, citizens and governments, between national governments and non-citizens, and between national governments.

AI enables control of information, in two directions. It controls the information that citizens receive. This shapes our view of the world, as has been much reported—enabling polarisation, and the spread of misinformation. AI also generates unprecedented information about us as individuals, making it available to both governments and corporations. This in turn enables them to shape our options, both how we spend our money, and how we vote. In my view this is a much more consequential implication of our reduced individual privacy than its implications for our ability to maintain a sphere in which we are not observed. Even if every individual whose data was used by these systems for the prediction and manipulation of behaviour consented to that use (and even if it was individually rational for them to do so) the collective implications of our ‘data profligacy’ would still be seriously morally objectionable.

And of course, the mere fact of automating decisions that used to be made by humans changes power relations. For example, automated performance management like what has just been revealed at Amazon means that performance management is benchmarked against context-free general standards, without sensitivity to individuals’ particular circumstances. Being managed by a person with whom you have some kind of personal relationship is a very different experience from being managed by an algorithm. The same goes for the use of AI to deliver government services and allocate resources.

It’s also important to be clear about the ways in which AI generates self-perpetuating power structures. It is inherently monopolistic—increasing data increases competitiveness, more competitive than more data. It’s a cycle towards oligarchy. And the same issues apply to AI and international relations. The key issues here seem to be to do with cyber security, as well as deep concern about the ability of potentially unconstrained adversaries to make substantial advances in AI that we cannot match.

Any assessment of the potential impact of AI and society must stare clearly into the face of the new power structures that these systems make possible. Large multinational corporations have long been only imperfectly subject to the authority of national governments. But, through their technology, they now have considerable power directly over the citizens of those national governments. This is not entirely new. Large media corporations have long had a similar degree of influence over individuals in countries that are only able to imperfectly exercise control over them. But new technologies, AI among them, enable this to proceed at a greater scale, and efficiency.

Lastly, it is crucial to note that we should be all the more concerned about the potential impact of AI on power when we realise that AI systems are designed in large part by a very narrow demographic. Research by AI Now makes clear that the tech industry is hopelessly unrepresentative of the communities that its systems may end up governing.

What Makes Defense Different?

The foregoing identifies some of the distinctive moral risks raised by widespread adoption of AI systems. Some of these risks are genuinely social risks—risks to society, which must be considered by any organisation considering the adoption of AI, and indeed by national governments as a whole (and supranational organisations). Some are specifically risks that the developers of AI systems should be concerned with. Thus far, we have seen policy principles developed by groups on behalf of national and supranational governments, as well as standards for conduct developed by technology companies working directly with and on AI. Defense is a very different organisation from either of these different kind of groups. Most importantly, it is *very* different from the technology companies that have so far set the terms for lists of AI policy principles.

The key differences, to my mind, are in the strategic purpose of the organisation, and in the people who are ultimately affected by its decisions. The goal of Defense is to uphold the constitution of the United States, and to protect its citizens against foreign and (to a lesser extent) domestic threats. That strategic framing is quite different from what Google's goals are, eg. It's hard to say what the purpose of the tech corporations is (beyond making their owners fabulously wealthy). And it's different from eg the EU, since they have to think about all of their citizens' interests, not just those related to security. Perhaps most importantly, Defense operates in a distinct strategic environment where its very task is to anticipate and consider threats to the US. So where a corporation's AI principles might only tangentially address potential malicious uses of AI, for Defense those situations should be at the centre. AI policy principles for Defense should address not only how Defense will develop and use AI, but also how it will respond to threats that make use of AI.

Because of the different strategic purpose of the organisation, policies of the Department of Defense will affect different cohorts of people than will be affected by the other organisations. Now, of course they'll likely affect similar actual people—Google's use of AI affects US citizens, non-citizens etc. The difference is that DoD will affect people who stand in importantly different relations to it: citizens, non-citizens for whom the US has a significant duty of care, and non-citizens for whom the US has a much less substantial duty of care. These roughly correspond to domestic operations of Defense, non-domestic operations outside of a war fighting context, and non-domestic operations in a war-fighting context. As a result, DoD is subject to different legal standards—roughly, US domestic law, international humanitarian law, and the law of armed conflict, respectively.

It's also worth observing that Defense is different from the private companies that are developing AI policy principles insofar as it is much more tightly regulated than they are. It is, after all, part of the government, subject to the usual checks and balances appropriate to that station. Since it is in general subject to a much more prescriptive system of law, the first step when considering the adoption of AI systems is to ask how existing law affects them.

Finally, Defense has existing structures of hierarchical decision-making and command responsibility that are not obviously present in other organisations. In many important respects, members of the US armed services have *fewer* rights than ordinary citizens. There are kinds of discrimination that are legally permissible in the military but not in other areas of society. Service-members are not presumptively entitled to challenge decisions that affect them. And so on. But also when the military acts, it is generally not the case that the individual at the sharp end of the spear is uniquely responsible for the results. Instead, there is a structure of command responsibility which enables members of the military to function, in effect, as a kind of group agent. This will be important below.

What is the Goal of a Set of AI Policy Principles?

As I understand it, the DIB is aiming to recommend a set of principles to govern the adoption of AI technologies by all aspects of Defense, affecting all of the constituencies described above. This would cover war fighting, humanitarian and other non-war fighting overseas operations, and domestic operations including personnel, management etc. It would cover both high moral risk applications of AI, and much more benign ones, like the use of AI systems for predictive maintenance of Defense assets. The principles should obviously do more than state the obvious. The operations of DoD are obviously subject to various bodies of law—US law for domestic operations, and international law for international ones. Beyond specific legal prohibitions and prescriptions, the DoD is also bound to uphold the US constitution, which provides a guiding set of values that should shape the adoption of any new technologies.

AI policy documents often read as though they were written in a vacuum. The DoD should begin any statement of AI principles by noting that the department is already bound to uphold the constitution, and abide by domestic and international law, and any use of AI should do the same. The task should then be to articulate principles that apply to the distinctive risks posed by AI, which cannot simply be derived from considering those overarching values. Those principles should take into account the distinctive nature of Defense, and should not simply amount to a rehashing of existing principles developed by other organisations. Finally, they should take into account the different constituencies that might be affected by Defense's actions, and in particular recognise that the extent to which these principles constrain Defense might depend on the circumstances and constituencies affected.

Discussion of Possible Principles

Before suggesting some principles that I think Defense should consider, I want to drive home the point about how the standard principles don't apply to defense in a straightforward way. I'll then go on to suggest some guiding principles that might pass that test.

Explainability

Most AI principle sets have some version of an 'explainability' principle—often linked to or differentiated from interpretability, transparency and so on. Rather than dig into the details of these different principles, I want to make some general points about how they apply to Defense.

Start with applications of AI that affect non-citizens, whether in a war-fighting or humanitarian context. Although international human rights law involves many proud statements of universal human rights, the reality is that international law as practised provides a meagre set of protections to people just in virtue of their humanity, and it is deeply implausible that either in the context of war or of humanitarian action, Defense would owe an explanation of how its algorithmic systems work to those affected by its decisions. It's not even plausible that there would be an obligation at international law to provide some overarching international body with insight into how Defense makes its decisions. Indeed, the US is regrettably ill-disposed towards international courts. And it is much more plausible that the standard of international law is one of **results** rather than processes. Different countries already vary so much in their decision-making processes—it's hard enough to gain any consensus on what results are unacceptable. Gaining consensus on the processes by which those results could be reached is surely impossible.

Might explainability still be an important principle for Defense operations that affect US citizens? Insofar as those citizens are members of the armed services, it's unclear why there would be a requirement to explain algorithmic decision-making to service-members when there is no requirement to explain any other kind of decision-making. That is, as long as the algorithmic system results in the delivery of an order that is not obviously unlawful, the standard expectation would be that orders are followed, without any explanation being owed. Of course, Defense has many civilian employees as well, and they are entitled to an explanation of certain kinds of decisions affecting them, in much the same way as any civilian employee of a government department would be—which is very likely already covered by existing employment law.

The most distinctive point at which an explanation of AI systems might be owed would be to the representatives of the citizens, in government. In order to ensure civilian control of the military, it is perhaps necessary to report to Congress in ways that make the operations of algorithms explicit. But it also seems plausible that there can be entirely successful civilian oversight of Defense without any AI systems deployed being explainable. How AI fits into the DoD's decision-making processes can clearly be regulated without explainability, as can the results of those processes.

Accountability and Contestability

Explainability, transparency, and interpretability are often separated from accountability and contestability. This is arguably a mistake—we care about explanation, transparency, and interpretability because they enable contestability, which is itself necessary for accountability. And you can't have accountability without all those preconditions. So it would be possible to just state that the central principle here is accountability.

Analytical nit-picking aside, however, it's clear that we can raise just the same questions for contestability as we did for explainability. Accountability however is not necessarily quite so variable dependent on the affected constituency. Even in war, the military must be accountable for its actions. Accountability is equally important, when fundamental rights are at stake, whether the rights are held by citizens or non-citizens (or at least, it's really important even in the latter case). However, this is an area in which the DoD is in a **better** position than private actors or even national governments. There already exist very clear structures of accountability within military organisations, which entail forms of group and command responsibility. Units are collectively responsible for the actions of their members, provided they act *'intra vires'* (within the mutually-understood bounds of their collective

endeavour). Commanders are responsible for the actions of their subordinates (subject to the same condition). Even when subordinates act 'ultra vires' commanders are still responsible to at least some degree. Where AI systems are used in a battle context, nothing changes. If anything, military applications are the ones where the thesis that AI systems dilute and diffuse responsibility is the least plausible, because the military is a highly structured collective acting according to a clear hierarchy and set of rules—they are the paradigmatic group agent—so the difference between having a human and an AI system at the 'sharp end of the spear' is relatively minimal.

Fairness and Privacy

Perhaps the two central concerns raised by AI systems have been that they might exacerbate discrimination by learning from data into which discriminatory practices are embedded, and that the data on which they are trained might be used without the consent of the originating parties. These principles seem particularly irrelevant when it comes to Defense operations that affect non-citizens. In humanitarian situations, the only good reason for DoD to take action that significantly impacts non-citizens is if there is an emergency that needs to be addressed, and the right to privacy must clearly give way if its doing so is necessary for lives to be saved. In war-fighting situations this is all the more clear. We could certainly argue that international human rights law protects civilians in war against certain kinds of invasions of their privacy, but such an argument is unlikely to gain much traction given the other much more serious deprivations to which civilians are usually vulnerable in war.

Worries about fairness and discrimination in these contexts are vulnerable to the same arguments. Defense routinely makes discriminatory judgments in war-fighting situations—military-aged males, for example, are often on that basis alone considered to be legitimate targets when they appear in locations most frequently attended by other legitimate targets. 'Profiling' is a routine aspect of war-fighting, and is very plausibly permissible in light of its usefulness in mitigating risks. Of course, if the algorithmic allocation of humanitarian assistance had discriminatory effects, then that would be a potentially serious problem, but perhaps one that was overridable in the event that lives could thereby be saved. Still, clearly the DoD should strive to ensure, over time, that its use of AI systems in the allocation of humanitarian aid does not have discriminatory effects.

What then of the use of AI in applications that affect service-members? Again, members of the armed services enjoy weaker protections against invasions of their privacy by their service than would ordinary citizens. And some discrimination is clearly tolerated within the military—for example based on gender. Of course, Defense also has many civilian employees, and insofar as AI systems are used in ways that affect them, the same kinds of consideration would apply as apply to other government organisations when their actions affect their staff.

The net result of these arguments is that it would be very hard to articulate general principles governing explainability, accountability, fairness, and privacy, for all operations by the DoD. How they apply really depends very strictly on the affected constituency and the circumstances. One could write that DoD should ensure *to the extent feasible given the circumstances* that explainability, accountability, privacy, and fairness are satisfied. One could then give a detailed discussion of what is meant by this, somewhat along the lines of what is discussed here. One might reasonably question whether a principle qualified in this way is really a principle at all, however.

Alternatively, one could present the principles in this form: **DoD must ensure that adoption of AI systems does not prevent it from meeting existing obligations to render its decisions explainable to those affected by them, to be appropriately accountable for those decisions, and to make them in a non-discriminatory way, respecting individual privacy.** This would then presuppose an account of what those existing obligations are, so would again be of dubious value as a standalone principle.

Alternative Principles

Many of the risks posed by the deployment of AI are the same as the risks involved in any other operational decision taken by DoD. There is a general principle governing all morally risky conduct, which DoD must observe in its deployment of AI, the principle of necessity:

DoD must recognise the different degrees of moral risk in adopting AI, and proceed with adoption only when the benefits of doing so clearly justify the moral risks, and where no other less risky alternative is available to realise comparable benefits.

Note that this principle applies both to the choice to use AI in the first place, and to the way in which the AI system is used. The first point is important: there is a headlong rush to adopt AI systems, but sometimes it simply is not worth the risk (the adoption of facial recognition systems now seems to be such a case). The second point matters too: any automated systems must be subjected to continual audit and review (as should any DoD systems).

Another principle that is more specific to AI (though would really apply, *mutatis mutandis*, to any new technology or process), would focus our attention on the risks that adoption of AI systems generate for human behaviour. It is well-documented that implementing automated systems risks leading to de-skilling and automation bias (where human operators defer to automated systems). This should be of particular concern to DoD, since it depends so heavily on the skilled judgment of its members. There should therefore be a principle along these lines:

DoD must ensure that its representatives are trained to develop and exercise their own critical judgment when operating in partnership with AI systems.

Killer Robots?

I haven't yet mentioned lethal autonomous weapons directly. It's worth briefly pausing to ask whether claims made by organisations like the Campaign to Stop Killer Robots are plausible. One candidate principle that might be proposed here is the following:

DoD must ensure that decisions over life and death are never taken by an AI system.

This principle would caution against allowing AI systems to take action that is expected to result in death without an intervening decision being made by a human operator. It would likely not apply in clearly low-risk cases where human intervention renders defence infeasible—for example Phalanx missiles.

There may be pragmatic reasons to endorse a principle such as this. In particular, it may help DoD gain public trust in its adoption of AI systems, since there seems to be considerable public support for this principle. It might also enable international coordination around a ban on autonomous weapons, which might help to limit proliferation, and constrain the moral costs of war. And of course existing and near-future AI technology is plausibly not able to abide by the laws of war—so this principle would just be a special case of the principle of necessity above.

Are there non-instrumental reasons to think that autonomous weapons are especially objectionable? If there are, I think they would have to be grounded in the considerations adduced above—to do with accountability, acting for the right reasons, care and attention, or avoiding discrimination. But I've already argued that military organisations have robust structures of accountability in place, and in any event are best viewed as corporate agents. AI systems do not change that. I am sceptical about whether we should really care about the 'acting for the right reasons' point in a military context. After all, we make no effort to explore the reasons on which individual combatants are acting, and in fact we typically encourage them to act on reasons that might not be particularly relevant to the overall justification for their actions—namely, many combatants are motivated by a desire to protect or avenge their friends, which while morally important in its own right, is strictly not relevant to the justice of the cause for which the war is fought, and so on. The idea that human combatants kill with care and attention for their victims is unrealistic at best; no doubt many do manage to maintain such attitudes at least some of the time, in some conflicts. But it is surely understandable that many people find their compassion has limits in the circumstances of war. And as already argued, in the radical uncertainty faced by combatants in war, reliance on heuristics is inevitable; AI systems would quite plausibly be less discriminatory than humans in the same situation, because they would at least be able to draw on much more information when selecting targets.

In my view, then, realistic AI systems would only ever be tools used by the military unit to which they are assigned. They should be assessed in that light. There is nothing intrinsically wrong with them,

but they should not be adopted unless they satisfy the principle of necessity described above. And there may be pragmatic grounds to reject them even if they did satisfy that principle.

However, I think that DoD is governed by a further, positive principle (in some respects the flipside of the principle of necessity):

DoD must adopt AI systems where doing so would reduce the risk of civilian casualties.

International law requires militaries to take feasible precautions in attack. If there are technological precautions that could reduce the risk of civilian casualties, then they must be taken—they are not optional. So if DoD is able to develop weapons that are able to disarm themselves on realising that the risk of civilian casualties has significantly increased in the seconds since their launch, then it has very good reason to do so. Likewise if technologies can be developed that better enable identifying where civilians are in the first place.

AI and International Security

I have least to say on this point; other scholars versed in security studies and international relations will be better placed to advise. But it's clearly crucial for the DoD to pay close attention to the distinctive security threats that AI makes possible, and to consider whether distinct principles are necessary to address those threats. However, in this respect AI really is just part of a suite of cyber security risks that Defense should address together.

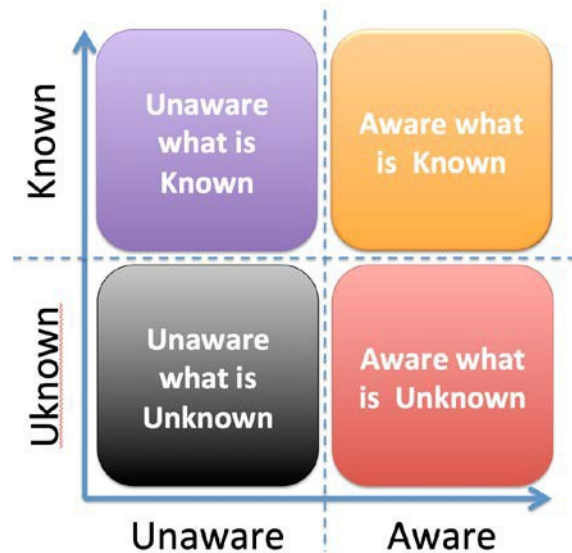
Public Statement Submission to the Defence Innovation Board in response to the call for public comments on "The Ethical and Responsible Use of Artificial Intelligence for the Department of Defense"

April 7, 2019

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.



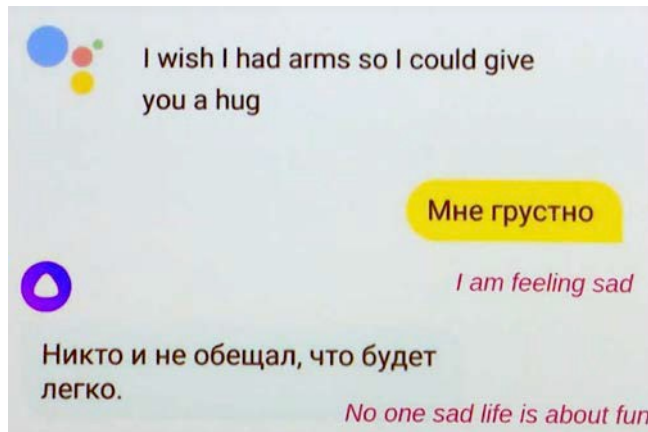
An adaptation of the 'Johari Window' for systemic knowledge

What we already know that we know

This 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 an increased diversity of contributing participants, whether from different cultural backgrounds, ages, or gender, is positively correlated with 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.



An example of AI cultural bias – East versus West

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 defense 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 behavior. 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 behavior 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].

John R. Potter, PhD.

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 behavior 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 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 behavior 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 result 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 center. 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.

John R. Potter, PhD.

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 Xiaolce has been very successful in China, where it has been used by 40 million people. Tay was an attempt to duplicate Xiaolce 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.

Sincerely,

Dr. John R. Potter

Fellow, IEEE.



Human Factors and Ergonomics Society Comments on AI Principles for the Department of Defense

On behalf of the Human Factors and Ergonomics Society (HFES), thank you for the opportunity to provide comments on the Defense Innovation Board's Recommendations for Principles on Artificial Intelligence (AI). HFES commends the Defense Innovation Board and the Department of Defense's (DOD) efforts to develop principles around the use of AI.

In order to support the Department's broader goals, AI technology must be designed with a mind towards not only its potential benefits, but also its limitations. Errors caused by actions or faulty advice from AI systems can put the lives of Service Members and civilians at risk and may even lead to unintended military engagements. With so much at stake, it is imperative that DOD develop processes to ensure that systems can be used safely and effectively by Service Members, before AI is integrated more widely into DOD activities and operations. These considerations are highly important for establishing effective approaches for designing and testing of such systems and for their successful use by America's fighting forces.

This topic is of great interest to HFES and more generally to those practicing human factors and ergonomics (HF/E). AI has implications for how people interact with systems, with specific implications for system performance. The science pertaining to the use of AI has been established over the past 40 years, with particular emphasis on human-automaton interaction (Bainbridge, 1983; Endsley, 2017; Endsley & Kiris, 1995; Kaber & Endsley, 2004; Lee & See, 2004; Onnasch, Wickens, Li, & Manzey, 2014; Parasuraman & Mouloua, 1996; Parasuraman & Riley, 1997; Sarter & Woods, 1995; Sheridan, 1992; Wickens & Kessel, 1979, 1980; Wiener & Curry, 1980).

As the Department is developing AI for military applications, it is important for DOD leadership to understand what specific tasks are suitable for AI, and what steps must be taken for human operators to effectively interact with AI systems. Specific recommended principles and issues to consider are described below.

With this in mind, HFES recommends the following principles:

- 1. Training:** *DOD should implement training to help Service Members better understand the functioning and limitations of AI systems they are working with and prevent errors resulting from overconfidence in AI.*
- 2. Learning Biases:** *AI training sets need to be carefully constructed to incorporate a wide range of possible situations, be tested for inadvertent biases, and develop methods to detect signs of possible enemy deceptions.*
- 3. Verification and Validation:** *DOD must develop and implement methods for validating the quality, generalizability, and limitations of AI systems, in order to ensure they are effective in DOD operations.*
- 4. Human Oversight:** *DOD autonomous systems involved in lethal actions should always be operated with humans in the loop, and DOD personnel should be*

- trained and provided with the necessary situation awareness to take over manual control when needed.*
5. **AI Advisory Systems:** *Advisory (decision support) systems need to provide transparency and explainability with regard to how the system made a decision and the factors considered, and confidence levels of pattern matches as well as other possible matches.*
 6. **Human-AI Interaction:** *DOD should incorporate HF/E guidelines in designing Human-AI Interfaces to support the situation awareness and trust of Service Members and to achieve effective overall performance.*
 7. **Testing:** *The operation of AI systems in conjunction with human users must be carefully tested to determine how it affects human situation awareness and decision making, and the ability of human users to detect and act in situations where the AI is unable to function appropriately.*

Training on AI Limitations & Functioning

Recommendation 1: *DOD should implement training to help Service Members better understand the functioning and limitations of AI systems they are working with and prevent errors resulting from overconfidence in AI.*

Modern approaches to AI focus on learning systems that develop appropriate responses (e.g. categorizations) associated with large data sets that it can be trained on. Key limitations exist for AI in recognizing cases that are outside of this data set or for cases in which the data presented to the system vary slightly from what they have been exposed to. Due to this, AI tends to function well on the expected, but may respond poorly to the unexpected (e.g. novel behaviors by enemy combatants), or to variations in the environment that may exist due to the challenges of the natural world (e.g. weather effects, animals, or people behaving in unexpected ways).

Even though they are trained on large data sets, AI systems have difficulty in recognizing cases that are outside of this data set or for cases in which the data presented to the system vary slightly from what they have been exposed to. Further, as has been documented by Pearl and Mackenzie (Pearl & Mackenzie, 2018), AI operates based on pattern matching to learned cases, but possesses no working model of the world that will allow it to project beyond what it has already seen.

To interact effectively with AI systems, personnel need explicit training to better understand these limitations for AI systems they are interacting with so that they can adopt appropriate expectations. Further it is critical that personnel be trained to develop accurate mental models of the expected functionality of the systems so that they can interact with them appropriately, and that this training be kept up to date as AI systems learn and change.

Learning Biases

Recommendation 2: *AI training sets need to be carefully constructed to incorporate a wide range of possible situations, be tested for inadvertent biases, and develop methods to detect signs of possible enemy deceptions.*

Because AI is formed on the basis of training data, it is highly possible for its behavior to become biased in unexpected ways. If the training data is not broad enough to cover the wide range of possible events, for instance, it may become biased towards solutions that do not extend well to unseen data sets. Further, it is not always known what features AI systems will focus on in making their matches. Research has revealed a significant number of instances where AI inadvertently learned inappropriate gender and racial biases (Garcia, 2016; Miller, Katz, & Gans, 2018). Key to the DOD is that AI learning biases could be subtly manipulated by foreign states to “train” U.S. AI towards false patterns as a means of deception and sabotage if the U.S. becomes reliant on AI for its information processing or software in critical systems (e.g. autonomy operating in ships, aircraft or unmanned systems) (Endsley & Jones, 2001).

Verification and Validation

Recommendation 3: *DOD must develop and implement methods for validating the quality, generalizability, and limitations of AI systems, in order to ensure they are effective in DOD operations.*

The job of verifying and validating the performance of any learning system is significant. “Traditional methods are based on requirements tracing and fail to address the complexities associated with autonomy software. There are simply too many possible states and combination of states to be able to exhaustively test each one, and understanding where the boundary conditions are will be difficult. The ability of the system to degrade gracefully and to support human-autonomy interaction will form an important aspect of successful autonomy implementation and will need to be expressly incorporated into validation testing.” (U. S. Air Force, 2015). New methods for verification and validation of AI systems will be needed to support system certification processes that can handle continuous learning over time, and the need for explanations of such changes to flow to the personnel who will be called upon to use and interact with AI so that they can make appropriate decisions on its use (U. S. Air Force, 2015).

Human Oversight & Interaction

Recommendation 4: *DOD autonomous systems involved in lethal actions should always be operated with humans in the loop, and DOD personnel should be trained and provided with the necessary situation awareness to take over manual control when needed.*

Due to the limitations of AI, there will be a need for DOD personnel to oversee the operation of AI systems, to intervene when appropriate, and to interact with these systems to carry out their duties. AI cannot do its job alone, nor can this need be obviated by assuming that the AI technology will get better. This is summarized by the automation conundrum: “*The more automation is added to a system, and the more reliable and robust that automation is, the less likely that human operators overseeing the automation will be aware of critical information and able to take over manual control*”

when needed.” (Endsley, 2017) Technical systems that have neglected the importance of the human in operating and interacting with automated systems have been found to have significant problems associated with loss of situation awareness and engagement, high workload, loss of trust, and poor mental models that have led to an inability of the human to understand what the AI system is doing and to take over manual control when needed (U. S. Air Force, 2015). It will be critical that any lethal action continue to involve human decision making and allow the ability for humans to override, and that training and tools to support the situation awareness needed to be effective in that process be provided.

AI Advisory Systems

Recommendation 5: *Advisory (decision support) systems need to provide transparency and explainability with regard to how the system made a decision and the factors considered, and confidence levels of pattern matches as well as other possible matches.*

Even AI systems that are meant to be only advisory to the human decision maker have been found to be problematic due to a *decision biasing* effect (Crocoll & Coury, 1990; Endsley, Bolte, & Jones, 2003; Lorenz, Di Nocera, Rottger, & Parasuraman, 2002; Reichenbach, Onnasch, & Manzey, 2011; Sarter & Schroeder, 2001). When the advisory system is correct, people are more likely to make a correct decision; however, when it is incorrect they perform worse than if they had received no decision advice at all (Layton, Smith, & McCoy, 1994; Olson & Sarter, 1999), a situation that is worse with more reliable automation (Metzger & Parasuraman, 2005; Rovira, McGarry, & Parasuraman, 2007). People may also be slowed down by the provision of decision advice in that they need to act to compare the system’s recommendation to other information to decide whether or not to agree with it (Endsley & Kiris, 1994). While many factors will influence whether AI performs better than humans or vice-versa in any given situation, including the competence and experience of the individual and the capability of an AI system, the fact that these two entities are not truly independent must be considered as a limiting factor on joint performance. Effective employment of AI systems requires that careful consideration is given to the interaction of the human operator with AI, and that the needed capability and footholds for effective human-AI interaction are built into the systems’ interface.

Human-AI Interaction

Recommendation 6: *DOD should incorporate HF/E guidelines in designing Human-AI Interfaces to support the situation awareness and trust of Service Members and to achieve effective overall performance.*

An effective design of the automation interface can significantly aid in both directly improving situation awareness of automation and the system, as well as improve the level of trust in the automation and the appropriate calibration of that trust. This includes (1) the degree to which it effectively presents the needed information for decision making;

(2) the salience of cues associated with the state of an AI system, including modes, and system boundary conditions; (3) support for mode transitions, including that needed to engage the AI system and to detect and respond to unexpected transitions to manual operation; and (4) the transparency of the AI system for providing understandability of its actions and predictability of future actions (Endsley, 2017). Issues such as the framing of recommendations and the presentation of system confidence levels have been found to have significant effects on human performance with AI systems (Aretz, Guardino, Porterfield, & McClain, 1986; Crocoll & Coury, 1990; Endsley & Kiris, 1994; Selcon, 1990).

Increases in the use of AI will make it increasingly important that attention is paid to the design of the human-AI interface via the application of human-AI guidelines (Endsley, 2017) coupled with careful testing to show that human operators fully understand what an AI system is doing, what it is projected to do in the near future, and the limits of its boundaries for successful performance. This can be accomplished through interfaces with high levels of system transparency, providing both understandability and predictability of the system, along with the appropriate use of salient features to support operator understanding of key states and mode transitions. Information that is critical for understanding system reliability (e.g. how well it is functioning, its confidence level in fused information, or system assessments), as well as its robustness (meaning its ability to handle current and upcoming situations), needs to be made readily transparent to the human operator.

Testing

Recommendation 7: *The operation of AI systems in conjunction with human users must be carefully tested to determine how it affects human situation awareness and decision making, and the ability of human users to detect and act in situations where the AI is unable to function appropriately.*

In addition to the verification and validation software testing that AI will invariably be subjected to, careful testing of the AI system in combination with human users must be instituted. This testing is important for ensuring that the human-AI interface is sufficient for supplying the needed situation awareness for appropriate operation with an AI system, both when it is operating effectively and when it is in boundary conditions. Testing should include not only expected conditions, but also a wide variety of edge cases at the boundaries of expected operations.

Conclusion

HFES's recommendations will help DOD ensure that AI systems can be used safely and effectively by DOD Service Members. These recommendations should be incorporated early on, as it will be difficult to implement changes once DOD has widely integrated AI into military systems. The Department has noted that AI can improve operations and even help prevent civilian casualties in combat situations. Designing AI

systems with the human in mind can help prevent mistakes that unnecessarily put civilians and Service Members at risk and can help DOD achieve the promise of AI.

Thank you again for the opportunity to provide comments. Please do not hesitate to contact HFES should you require additional information.

About HFES:

With over 4,600 members, HFES is the world's largest nonprofit association for human factors and ergonomics (HF/E) professionals. HFES members include psychologists and other scientists, designers, and engineers, including researchers, practitioners, and federal agency officials, all of whom have a common interest in working to develop safe, effective, and practical human use of technology, particularly in challenging settings. HFES has a particularly strong expertise in research into human-artificial intelligence (AI) interactions.

References

- Aretz, A., Guardino, A., Porterfield, T., & McClain, J. (1986). Expert system advice: How should it be given? Proceedings of the Human Factors Society 30th Annual Meeting (pp. 652-656). Santa Monica, CA: Human Factors Society.
- Bainbridge, L. (1983). Ironies of automation. Automatica, 19, 775-779.
- Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting (pp. 1524-1528). Santa Monica, CA:
- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. Human Factors, 59(1), 5-27.
- Endsley, M. R., Bolte, B., & Jones, D. G. (2003). Designing for situation awareness: An approach to human-centered design. London: Taylor & Francis.
- Endsley, M. R., & Jones, D. G. (2001). Disruptions, Interruptions, and Information Attack: Impact on Situation Awareness and Decision Making. Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting (pp. 63-68). Santa Monica, CA: Human Factors and Ergonomics Society.
- Endsley, M. R., & Kiris, E. O. (1994). Information presentation for expert systems in future fighter aircraft. International Journal of Aviation Psychology, 4(4), 333-348.
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381-394.
- Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal, 33(4), 111-117.
- Kaber, D. B., & Endsley, M. R. (2004). The Effects of Level of Automation and Adaptive Automation on Human Performance, Situation Awareness and Workload in a Dynamic Control Task. Theoretical Issues in Ergonomic Science, 5(2), 113-153.
- Layton, C., Smith, P. J., & McCoy, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. Human Factors, 36(1), 94-119.

- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.
- Lorenz, L., Di Nocera, F., Rottger, S., & Parasuraman, R. (2002). Automated fault management in a simulated spaceflight micro-world. Aviation, Space and Environmental Medicine, 73, 886-897.
- Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. Human Factors, 47(1), 35-49.
- Miller, F. A., Katz, J. H., & Gans, R. (2018). The OD imperative to add inclusion to the algorithms of artificial intelligence. OD PRACTITIONER, 50(1).
- Olson, W. A., & Sarter, N. B. (1999). Supporting informed consent in human machine collaboration: The role of conflict type, time pressure, and display design. Proceedings of the Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (pp. 189-193). Santa Monica, CA: Human Factors and Ergonomics Society.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. Human Factors, 56(3), 476-488.
- Parasuraman, R., & Mouloua, M. (Eds.). (1996). Automation and human performance: Theory and applications. Mahwah, NJ: Lawrence Erlbaum.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse and abuse. Human Factors, 39(2), 230-253.
- Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. New York: Basic Books.
- Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human performance consequences of automated decision aids in states of sleep loss. Human Factors, 53, 717-728.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. Human Factors, 49(1), 76-87.
- Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. Human Factors, 43(4), 573-583.
- Sarter, N. B., & Woods, D. D. (1995). "How in the world did I ever get into that mode": Mode error and awareness in supervisory control. Human Factors, 37(1), 5-19.
- Selcon, S. J. (1990). Decision support in the cockpit: Probably a good thing? Proceedings of the Human Factors Society 34th Annual Meeting (pp. 46-50). Santa Monica, CA: Human Factors Society.
- Sheridan, T. (1992). Telerobotics, automation and human supervisory control. Cambridge, MA: MIT Press.
- U. S. Air Force. (2015). Autonomous Horizons. Washington, DC: United States Air Force Office of the Chief Scientist.
- Wickens, C. D., & Kessel, C. (1979). The effect of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems, Man and Cybernetics, SMC-9(1), 24-34.

- Wickens, C. D., & Kessel, C. (1980). The processing resource demands of failure detection in dynamic systems. Journal of Experimental Psychology: Human Perception and Performance, 6, 564-577.
- Wiener, E. L., & Curry, R. E. (1980). Flight deck automation: Promises and problems. Ergonomics, 23(10), 995-1011.

Universal principles of data ethics

12 guidelines for
developing ethics codes

A large, solid green chevron graphic pointing to the right, partially overlapping the text "High performance. Delivered." and the background image.

High performance. Delivered.

Professional ethics codes serve a surprisingly broad range of purposes. On their face, ethics codes set out the standard for acceptable behavior within a profession. However, the acts of assembling, deliberating, distributing, training and enforcing ethics codes ultimately result in much broader impacts. A set of shared norms helps to define the boundaries of a professional community and identifies the standards that the public should demand of practitioners within that profession. It establishes the type of relationship that professionals have with the rest of society—including with their clients, research subjects, users of their services, and governments.

In many professions, learning about and complying with ethical obligations is a marker of accountability. Indeed, the act of becoming licensed in professions such as accounting, law and medicine includes swearing to uphold specific ethical norms. In some professions, such as journalism (and arguably science and engineering), the ethics of professional practice are the *only* commonality defining the boundaries of what is an extremely varied group. Though such codes typically lack the force of law, they do ultimately shape legal and regulatory dynamics by establishing

expectations about responsibility and liability. In the long run, ethics codes can play a major role in defining a community of professionals.

Public discussions about the ethics of "big data" analytics are rapidly gaining prominence in public discourse. The insights derived from data already permeate much of our lives, and promise to shape even more of the opportunities, limits, and major and minor life decisions we encounter moving forward. In other words, data professionals will, in all likelihood, play a role in our lives that's as intimate as medical,

fiduciary, and legal professionals. This is why establishing a shared set of norms is critically important for data scientists and practitioners (and those making requests of them). It's good for the profession and good for society.

This report discusses the dynamics involved in generating a code of ethics that could guide the profession of data science as it grows and evolves, and immediately help organizations shape their own internal guidelines related to data. A broad set of principles is proposed and intended to inform the development of domain-

specific codes of ethics for specific organizations or industries. Developing a code of ethics should be a collaborative effort that involves all of the stakeholders in a community and builds from the proposed principles. Additionally, the uses of data science are so diverse (and many are still unforeseen) that not every scenario can be accounted for in a code of ethics. Nor does “data science” adequately capture the many facets of the data ecosystem. There is a diversity of practitioners that utilize the techniques of data science to provide analysis, insights and advice about a breadth of human activities; all of these actors may have specific obligations that differ from data scientists. Nonetheless, these principles are intended to function as a foundation or outline of what a universal code of ethics for the data science field should emphasize.

Framing data ethics

The foremost practical question for data ethics is whether there is anything special about data such that collecting, manipulating, and applying it requires a distinct code of ethics. The history of science and engineering ethics suggests that ethical regimes often track new ways of knowing. As new ways to know the world are developed, appropriate rules governing those approaches are helpful.

And so the question of whether data scientists and practitioners need a special focus on ethics is ultimately a question of whether data science represents a distinctly new way of knowing.

The way data is used today is more than just a technical phenomenon. It’s a political, social, and even mythological phenomenon that has consequences for how we organize our lives and express our values.¹ Whatever ethical principles are developed in connection with data, they should account for dynamics that extend beyond technical limitations. Data analytics should be viewed as a phenomenon with consequences beyond technology, and the community should demand that data scientists and practitioners consider those consequences.

Data analytics is an emerging form of knowledge production that provides the ability to cheaply and easily connect and analyze datasets, often drawn from highly disparate contexts.² The capacity to continually re-analyze and correlate data collected from a broad range of contexts has proven challenging to ethically conceptualize and regulate.^{3,4} In the past it could be assumed that data collected in one context—medical, political, genetic, social, financial, census, behavioral, geographic, etc—would stay in that context and could be regulated as

such. Furthermore, many familiar ethical controls, such as informed consent, occur only at the point of collection. But the power and peril of data science is that data is most valuable when it can be reused and repurposed in many different contexts and in combination with other datasets.

Personal and sensitive data now travels unpredictably and will be reused indefinitely for unforeseeable purposes. Because our “data selves” are no longer compartmentalized, many different actors can learn intimate details about the lives of anybody who leaves a digital trail. For this reason, the ethical infrastructures, concepts, and norms that have been developed to handle compartmentalized data are often neither salient nor applicable—how data moves in time and space is no longer synchronized with our temporally and geographically constrained ethical regimes.⁵ The language of medical and scientific ethics has long emphasized *respect for persons* and *informed consent* as core values. But it is a daunting proposition to explain how such principles can hold when data about individuals is persistently shared, transformed, and aggregated and when future uses of datasets are so unknowable that “informed consent” is a misnomer at best—and impossible at worst.

Professional codes of data ethics

Analyses of professional ethics codes show that the articulation of shared values is often a key stage in the professionalization of a field: It establishes who is a member of the field and what can be expected of them by colleagues, clients and society at large.^{6,7} Mark Frankel offers a taxonomy of professional ethics codes as *aspirational, educational, and regulatory*, noting that most codes are an admixture and serve multiple goals. He argues that the process of establishing a code provides opportunities for critical reflexivity that are perhaps more important than the final product: "This process of self-criticism, codification, and consciousness-raising reinforces or redefines the profession's collective responsibility and is an important learning and maturing experience for both individual members and the profession."⁸

In an analysis conducted for the Council for Big Data, Ethics & Society, Jacob Metcalf identified the inward—and outward-facing goals of professional ethics codes that may be applicable for data ethics:⁹

Inward-facing goals:

- Provide guidance when existing implicit norms and values are not sufficient; essentially, guidance for a novel situation
- Reduce internal conflicts; strengthen the sense of common purpose among members of the organization
- Satisfy internal criticism from members of a profession
- Create generalized rules for individuals and organizations that have responsibilities for important "human goods"
- Establish role-specific guidelines that demarcate general principles as particular duties
- Establish standards of behavior toward colleagues, students/trainees, employees, employers, clients
- Strengthen the sense of common purpose among members of an organization
- Deter unethical behavior by identifying sanctions and creating an environment in which the reporting of unethical behavior is affirmed
- Provide support for individuals who come under pressure to behave in an unethical manner

Outward-facing goals:

- Protect vulnerable populations and individuals who could be harmed by the profession's activities
- Protect and enhance the good reputation of and trust for the profession
- Establish the profession as a distinct moral community worthy of autonomy from external control and regulation
- Provide a basis for public expectations and evaluation of the profession
- Serve as a basis for adjudicating disputes among members of the profession and disputes between members and the public
- Create institutions that are resilient in the face of external pressures
- Respond to past harms done by the profession.

There are already some ethics codes that cover most computing and data scientists and engineers. In the US, four major computing professional societies have substantially different codes for their members due to their different missions.¹⁰ The Association of Computing Machinery (ACM), the largest professional organization for computer scientists and engineers, distributes an ethics code for members of its organization.¹¹ However, that code was adopted in 1992 at the beginning of the internet age, predating many of the technologies that define the ethical conflicts faced by data and computing professionals today. Although the ACM's ethics code

contains some principles that do still hold up—such as striving to maintain the integrity of data about individuals—it lacks the specificity that would make the code optimally useful to current and future generations of data and computing professionals. Other professional groups that are more closely associated with the data revolution have more recent codes. The recently founded Data Science Association offers a relatively detailed ethics code that is notable for detailing how members should adhere closely to scientifically sound statistical methods.¹² For example, rule 8(d) reads:

“ If a data scientist reasonably believes a client is misusing data science to communicate a false reality or promote an illusion of understanding, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to the proper authorities. The data scientist shall take reasonable measures to persuade the client to use data science appropriately.”

Source: Code of Conduct, Data Science Association

Some data science sub-disciplines have also produced valuable ethics codes and other types of ethics guidance for their members. The Association of Internet Researchers (AoIR) developed an ethics code in 2002, updated in 2012, that addresses the obligations of social science researchers working in digital domains at a macro-level.¹³ This document is notable for the extensive list of questions internet researchers should address. The National Center for Education Statistics produced a guide for appropriate use of educational data in 2010, that mixes core principles with illustrative case studies.¹⁴

Challenges for a universal code of data ethics


A unique aspect of today's datasets is their sprawling, multi-disciplinary utility—data science is arguably closer to a service than a discipline because it is useful in so many industries and disciplines. The analytical tools developed in applied mathematics, statistics, and computer science are being taken up by disciplines and sectors such as medicine, marketing, finance, the

humanities, social science, criminal justice, geography and geospatial imaging, manufacturing, social work, human rights, and many more. This poses a major challenge for a *universal* code of data ethics: There may be too few commonalities across the specific uses of data science to pull together a single code. Principles of data ethics that hold in medicine may not hold in finance because the social roles occupied by medical professionals and financiers differ significantly. They have meaningfully different obligations to their clients and society, and so it is reasonable to expect that their uses of big data for good and ill will similarly vary.

Furthermore, many of these fields already have their own professional ethics codes that may or may not address the changes introduced by the data age. Other fields have dealt with such problems by having professional sub-societies formulate secondary ethics codes. For example, the American College of Obstetricians and Gynecologists holds its members both to the American Medical Association's code of ethics (that applies to all

physicians) and to a more specific set of obligations that apply only to their own members. If data science continues on its path to ubiquity, then it may be challenging to define a truly universal code that covers its uses in such a variety of contexts.

One of the quirks of data science is that its parent fields have traditionally fallen outside of the purview of US federal research ethics regulations. Following a long arc of infamous research scandals in the mid-20th century—ranging from Nuremberg to Tuskegee to the Stanford prison experiment—the 1974 National Research Act empowered federal regulators to identify, define, and enforce ethical standards for human-subjects research that uses federal funds. The authors of the 1979 Belmont Report commissioned by the Act identified the three primary principles of bioethics: **beneficence** (research should be carefully constructed to do good in the world), **respect for persons** (research must respect personal values such as autonomy, privacy and dignity) and **justice** (research must further social equity).



These principles subsequently informed the rulemaking process initiated by the Department of Health and Human Services that resulted in the federal regulations known as the Common Rule. The Common Rule now governs (nearly) all human-subjects research funded by federal agencies. Its most consequential outcome was establishing Institutional Review Boards (IRBs) as an obligatory milestone for most academic research. However, computer science and engineering, applied mathematics, and much quantitative sociology research has historically fallen outside of the regulatory definition of "human subjects," even when these fields involve human lives.¹⁵ As a result, most professionals trained in the parent fields of data science do not encounter the primary research norms and regulatory apparatuses that guide other science and engineering fields.

The Common Rule and IRBs dominate conversations about practical ethics, but in some cases,

even these regulated standards do not go far enough. In the humanitarian field, some academics and practitioners are beginning to call for higher standards.¹⁶ They argue that "demographically identifiable data"—a broader classification than Personally Identifiable Information (PII), the gold standard for privacy professionals—could cause various harms to entire classes of people.

As data science matures as a field and increasingly affects the human condition, there's a chorus building among professionals and practitioners to have more guidance for the ethical decisions they are forced to make—and might be unaware they are making—on a daily basis. The set of Principles proposed below is intended to provide a baseline for those seeking such guidance and those looking to develop a group-specific code of data ethics.

Principles for Data Ethics

Data science professionals and practitioners should strive to perpetuate these principles:



1. The highest priority is to respect the persons behind the data.

When insights derived from data could impact the human condition, the potential harm to individuals and communities should be the paramount consideration. Big data can produce compelling insights about populations, but those same insights can be used to unfairly limit an individual's possibilities.



2. Attend to the downstream uses of datasets.

Data professionals should strive to use data in ways that are consistent with the intentions and understanding of the disclosing party. Many regulations govern datasets on the basis of the status of the data, such as "public," "private" or "proprietary." However, what is *done with* datasets is ultimately more consequential to subjects/users than the type of data or the context in which it is collected. Correlative uses of repurposed data in research and industry represents both the greatest promise and the greatest risk posed by data analytics.



3. Provenance of the data and analytical tools shapes the consequences of their use.

There is no such thing as raw data—all datasets and accompanying analytic tools carry a history of human decision-making. As much as possible, that history should be auditable, including mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data.



4. Strive to match privacy and security safeguards with privacy and security expectations.

Data subjects hold a range of expectations about the privacy and security of their data and those expectations are often context-dependent. Designers and data professionals should give due consideration to those expectations and align safeguards and expectations as much as possible.



5. Always follow the law, but understand that the law is often a minimum bar.

As digital transformations have become a standard evolutionary path for businesses, governments and laws have largely failed to keep up with the pace of digital innovation and existing regulations are often mis-calibrated to present risks. In this context, compliance means complacency. To excel in data ethics, leaders must define their own compliance frameworks that outperform legislated requirements.



6. Be wary of collecting data just for the sake of more data.

The power and peril of data analytics is that data collected today will be useful for unpredictable purposes in the future. Give due consideration to the possibility that less data may result in both better analysis and less risk.



7. Data can be a tool of inclusion and exclusion.

While everyone deserves the social and economic benefits of data, not everyone is equally impacted by the processes of data collection, correlation, and prediction. Data professionals should strive to mitigate the disparate impacts of their products and listen to the concerns of affected communities.



8. As much as possible, explain methods for analysis and marketing to data disclosers.

Maximizing transparency at the point of data collection can minimize more significant risks as data travels through the data supply chain.



9. Data scientists and practitioners should accurately represent their qualifications, limits to their expertise, adhere to professional standards, and strive for peer accountability.

The long-term success of the field depends on public and client trust. Data professionals should develop practices for holding themselves and peers accountable to shared standards.



10. Aspire to design practices that incorporate transparency, configurability, accountability, and auditability.

Not all ethical dilemmas have design solutions, but being aware of design practices can break down many of the practical barriers that stand in the way of shared, robust ethical standards. Data ethics is an engineering challenge worthy of the best minds in the field.



11. Products and research practices should be subject to internal, and potentially external ethical review.

Organizations should prioritize establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programs. Internal peer-review practices can mitigate risk, and an external review board can contribute significantly to public trust.



12. Governance practices should be robust, known to all team members and reviewed regularly.

Data ethics poses organizational challenges that cannot be resolved by familiar compliance regimes alone. Because the regulatory, social, and engineering terrains are so unsettled, organizations engaged in data analytics require collaborative, routine and transparent practices for ethical governance.

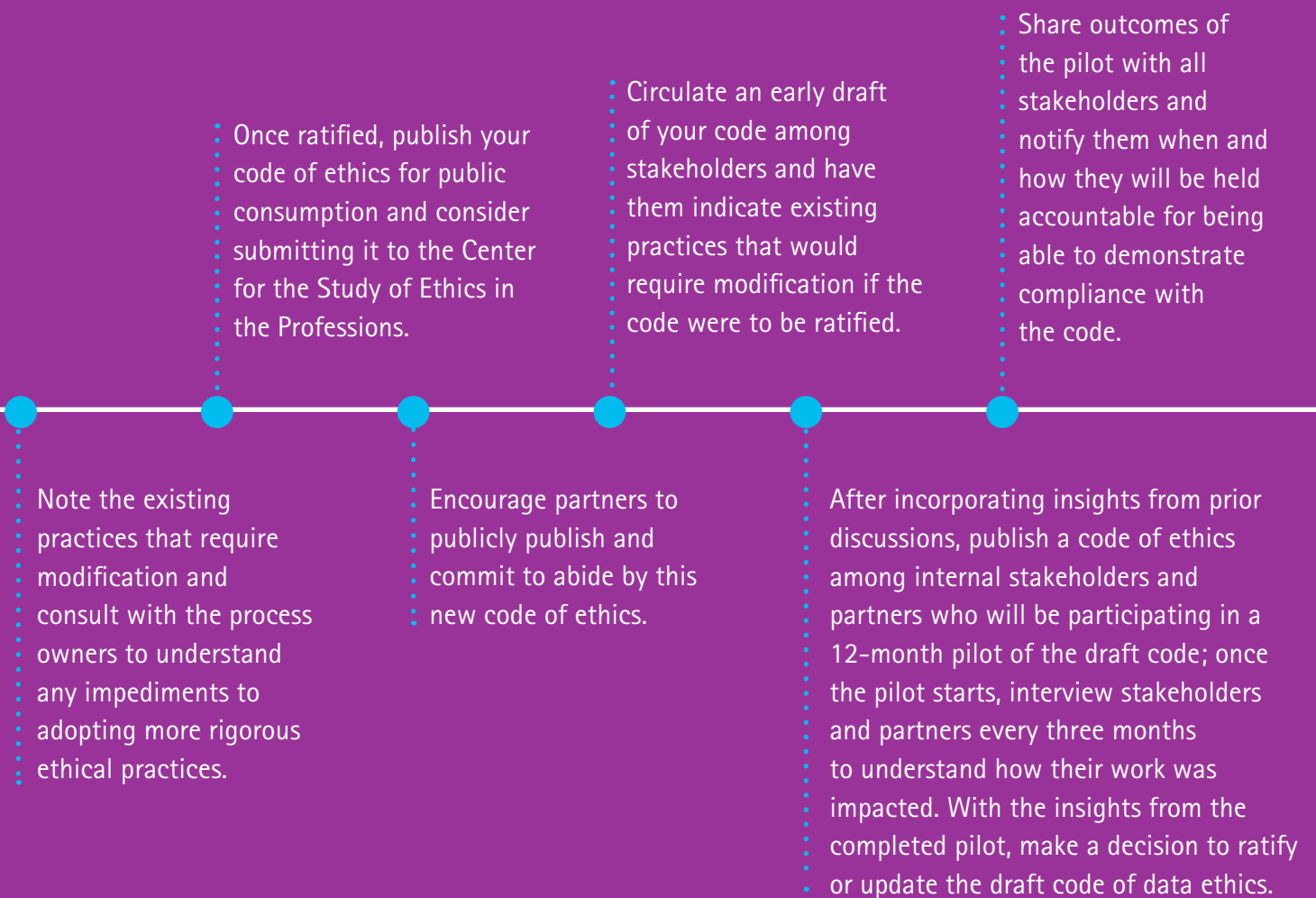
100/365-day Plans

Over the course of the next year, every organization can be well on its way to leveraging these 12 universal principles to develop a custom-tailored code of data ethics.

In three months, your organization should:



In one year (and beyond), your organization should strive to:



References

- 1 Crawford K, Gray ML and Miltner K (2014) Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication* 8(0): 10.
- 2 Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.
- 3 Zwitter A (2014) Big Data ethics. *Big Data & Society* 1(2). DOI: 10.1177/205395171455925.
- 4 Metcalf J, boyd danah and Keller EF (2016) Perspectives on Big Data, Ethics, and Society. Council for Big Data, Ethics, and Society. (accessed 31 May 2016).
- 5 Metcalf J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3(1): 1–14. DOI: 10.1177/2053951716650211.
- 6 Metcalf J (2014) Ethics Codes: History, Context, and Challenges. Council for Big Data, Ethics, and Society. (accessed 21 October 2015).
- 7 The Illinois Institute of Technology maintains a thorough collection of professional ethics codes as part of their Center for the Study of Ethics in the Professions.
- 8 Frankel MS (1989) Professional codes: why, how, and with what impact? *Journal of business ethics* 8(2–3): 109–115.
- 9 Metcalf, J (2014). See also: Frankel MS (1989); Gaumnitz BR and Lere JC (2002) Contents of Codes of Ethics of Professional Business Organizations in the United States. *Journal of Business Ethics* 35(1): 35–49; Kaptein M and Wempe J (1998) Twelve Gordian Knots When Developing an Organizational Code of Ethics. *Journal of Business Ethics* 17(8): 853–869.
- 10 Oz E (1993) Ethical standards for computer professionals: a comparative analysis of four major codes. *Journal of Business Ethics* 12(9): 709–726.
- 11 "ACM Code of Ethics and Professional Conduct." ACM Code of Ethics and Professional Conduct. ACM, Inc. 16 October 1992. Web. 31 May 2016.
- 12 "Code of Conduct | Data Science Association." Data science code of professional conduct. Data Science Association. Accessed 31 May 2016.
- 13 <http://aoir.org/reports/ethics2.pdf>. Accessed 31 May 2016.
- 14 <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010801>. Accessed 31 May 2016.
- 15 Metcalf J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3(1): 1–14.
- 16 Karunakara U (2014) Data Sharing in a Humanitarian Context: The Experience of Médecins Sans Frontières. In: Moore SA (ed.), *Issues in Open Research Data*, London: Ubiquity Press.



© 2016 Accenture.
All rights reserved.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA. Accenture, its logo, and High performance. Delivered. are trademarks of Accenture.

Contact Us

Steven C. Tiell
Senior Principal—Digital Ethics
Accenture Labs
steven.c.tiell@accenture.com

Jacob Metcalf
Ethical Resolve
jake@ethicalresolve.com

About Accenture Labs

Accenture Labs invents the future for Accenture, our clients and the market. Focused on solving critical business problems with advanced technology, Accenture Labs brings fresh insights and innovations to our clients, helping them capitalize on dramatic changes in technology, business and society. Our dedicated team of technologists and researchers work with leaders across the company to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage.

Accenture Labs is located in six key research hubs around the world: Silicon Valley, CA; Sophia Antipolis, France; Arlington, Virginia; Beijing, China; Bangalore, India, and Dublin, Ireland. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence located in 92 cities and 35 countries globally to deliver cutting-edge research, insights and solutions to clients where they operate and live. For more information, please visit www.accenture.com/labs.

Learn more: www.accenture.com/DataEthics

This document makes descriptive reference to trademarks that may be owned by others.

The use of such trademarks herein is not an assertion of ownership of such trademarks by Accenture and is not intended to represent or imply the existence of an association between Accenture and the lawful owners of such trademarks.

Data Ethics Research Initiative

Launched by Accenture's Technology Vision team, the Data Ethics Research Initiative brings together leading thinkers and researchers from Accenture Labs and over a dozen external organizations to explore the most pertinent issues of data ethics in the digital economy. The goal of this research initiative is to outline strategic guidelines and tactical actions businesses, government agencies, and NGOs can take to adopt ethical practices throughout their data supply chains.

About Accenture

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions—underpinned by the world's largest delivery network—Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With approximately 373,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at www.accenture.com.