Public Comments Ahead of March 14, 2019 Defense Innovation Board Listening Session

Table of Contents

1) Mark Gubrud, University of North Carolina

We stand today at the beginning of a revolution: the rapid advance and wide use of AI. Because this technology replaces human intelligence & judgment, it has the potential to cause catastrophic errors whose specific causes might be identifiable but often are not and in general are not foreseeable.

In the military and global security context, the most critical danger arises from the unforeseeable and untestable interactions of complex networks of competing and adversarial systems, which could erupt into crisis or open combat which humans might not understand or be able to interrupt. Nations may be driven to undertake the risk of reliance on such systems by competitive pressures as automated systems acquire capabilities to aggregate and integrate more information than any human decision maker and to react to unambiguous signals of hostile action faster than any human could.

For these and other reasons, it is imperative that the global community undertake ambitious arms control initiatives, including a mandate of real-time accountable human control over all lethal weapons systems.

Questions about AI and autonomous weapons are too often framed only in terms of ethical questions - is it right to use such [powerful, useful] weapons? - and not enough in terms of the need to avoid an accelerating arms race toward the loss of human control and the occurrence of war by accident or by misconceived design. Is it ethical to lead a global race to oblivion, instead of leading toward a strong regime of binding, verified arms control, global governance and global security?

Second Public Comment

We stand today at the start of a revolution: the rapid advance and wide use of AI. Because this technology replaces human intelligence and judgment, it has the potential to cause catastrophic errors with consequences in proportion to the responsibilities being delegated to machines.

The specific causes of error in AI may be foreseen but in general will not be, and may not even be identifiable. In complex systems it becomes in principle impossible to foresee all exceptional situations that may and will arise. Computer algorithms are particularly brittle, but also complex networks of "analog" and living systems exhibit unpredictable collective behavior and sudden crises. As AI advances toward human level capabilities, its potential for instability is clear.

The most severe danger arises from the unforeseeable, untestable (except in real-life confrontation) interactions of networked, complex, competing and adversarial systems. Experience with such networks, such as the instability of high-speed trading which has produced several hugely expensive stock market "flash crashes," has demonstrated the likelihood that confronting & interacting adversarial networks will erupt into crisis or open combat, spontaneously or as triggered by unforeseen circumstances, and in any case once ignited into a condition of ongoing violence may execute and escalate that violence so rapidly and in such a complicated and opaque way as to resist or frustrate any human attempt to intervene. The ongoing confrontation, competition and adversarial nature of such systems, and the adversarial relationships of their creators, contradict and will frustrate any effort to coordinate between them so as to mitigate the risk of unauthorized or uncontrollable conflict.

Why would nations undertake to construct and rely on such obviously dangerous systems? For the same reasons as in the Cold War, as in the arms race today, and under the competitive pressure of an advancing technology that is already able to aggregate and correlate more data than any human could, and is increasingly able to integrate that information and make high-level decisions, particularly where signals are unambiguous, more rapidly than any human.

We must avoid taking that road. But in fact it is the road we are already on, so we must avoid going further. The global community must undertake ambitious arms control initiatives, including a mandate of real-time accountable human control over all weapons systems. The DoD and the US cannot do this alone, but America's preference should be for arms control, we should say so and everything we do should be consistent with that. Unilateral disarmament or renunciation of strategically decisive technology would not be effective or possible, but the opposite extreme of trying to win the arms race should be equally strongly rejected.

Questions about AI and autonomous weapons are too often framed only in terms of ethics: is it right to use such [useful, powerful] weapons? We need to consider the ways in which these weapons are eroding our control and creating a threat to ourselves. We must avoid an accelerating arms race toward the loss of human control and the occurrence of war by accident or misconceived design. Is it ethical to lead a global race to oblivion, instead of leading toward a strong regime of binding, verified arms control, global governance and human security?

2) Geoffrey Odlum, Private Sector

While it is a welcome and timely request from DOD to ask the DIB to draft a set of ethical principles in the use of artificial intelligence in warfare, it is equally imperative that the State Department then take these principles and try to secure consensus from other states developing military AI (China, Russia, Israel, ROK, etc) to commit to them as well. If the USG binds itself to certain ethics with regard to AI in warfare and our adversaries do not, adversarial military AI weapon systems will almost certainly defeat ours.

3) Matthew Reyburn, DCMA

Here is an article of an AI performing a "moral hazard" (from the principal-agent problem in economic theory) –

https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/

Here is a good series of articles detailing current research and work –

https://intelligence.org/embedded-agency/

4) Alexis Prest-Simpson, The Software Engineering Institute of Carnegie Mellon University

How do you see the FFRDC’s role in software development for Artificial Intelligence for Defense?

5) Link Parikh, AI Vendor

AI can become another black hole of spending. The best way to move forward is to work with small prototyping labs with formal large vendor IBM Watson, for example. These firms have expertise in designing and delivering increasingly deployable capability at 15% the cost of large integrator RFPs and internal DIY projects. Note that AI is not only about the tool, but the art of data science in terms of model development and model training, for example. Contract vehicles such as IMPAX at Navair is the right way to acquire these 6 to 12 month iterations to a pilot-ready 12 month effort to full operational capability. This industry-style engineering approach will drive AI modernization in 24 months vs. 5 or 6 years to stay ahead of the well-funded threat.

6) Julian Kline, Kline Studios LLC

Dear DIB, here are my suggestions: 1. While AI is very powerful for information gathering, data analysis and reactive cyber-defenses, any AI dealing with humans should have a margin of error for a living creature's mistakes, confusions and improvisations. No human nor animal should be held to mathematical expectations. 2. Humans cannot be "backed up", "redownloaded", nor "rebooted". Any physical-world AI should do its best to preserve human life in a humane way, despite any further coded tasks. 3. Any powerful AI should come with equally-powerful AI (minimum of 3) which can create a "checks and balances" scenario. If one AI begins acting strange, the other 2 can fix it with permission override keys/regulations. One cannot override the other 2. In the event all 3 are corrupted, the Owner or Developer should have a kill-switch and back door code to delete all three. 4. We need to clean the internet with AI. This sort of cyber-regulation will rely on a communicative tech society. Segregating tech experts and what they know may seem safer but It's detrimental to our minimum-bar of tech education. A collaborative cyber-community will have the creative answers to defense issues. 5. Using AI to effect a peoples' way of life, sway a peoples' opinions, or cause chaos or confusion is different than a media campaign based of AI-collected and configured data. The former is a systematic, intentional and cultural intrusion. The latter is an educated broadcast. Thank you for your time and work.

7) Joshua Darrow, Department of Navy – NAVAIR

Innovation begins and ends with technical capability. Over decades the government has outsourced its most technical work, eroding the technical skill of the government workforce. To have an innovative government workforce, dealing with the distribution of technical work (ie designing, building, testing, redesigning, testing) is essential.

8) Thomas Creely, US Naval War College

The U.S. Naval War College has recently established a special graduate certificate program in Ethics and Emerging Military Technology. As its director, I work with a dozen competitively- selected students to engage in ethics- and technology-related coursework as well as conduct research on the ethical implications of emerging technologies. Each student produces a lengthy professional paper analyzing some aspects of the ethics-technology nexus, many of them dealing with various forms of AI. We have developed connections with DARPA, ONR, Boston Global Forum, the Director for Defense Intelligence, and a number of academic institutions exploring the ethics of AI.

9) David Simon, NNData

What is the difference between DoD's JAIC and the newly formed DoD AI Commission?

With the proliferation of so many new Other Transaction Authorities (OTA); how and where is DoD tracking the listings of all the OTA's that exists? How can procurement officers, operational elements with requirements and funding, and industry able to find a listing of these OTA's?

Where and how are all the white papers, proposals, and awards for projects being listed and made publicly available?

10) Michelle Kinsy, Boston University

The Department of Defense (DoD), in its push to expand the intelligence, autonomy, and mobility of systems supporting the dismounted soldier's real-time tactical decision-making capabilities, will have to address a number of design challenges related to the safe deployment of artificial intelligence learning models and techniques. Machine learning (ML) models are often trained using private datasets that are very expensive to collect, or highly sensitive, using large amounts of computing power. The models are commonly exposed either through online APIs, or used in hardware devices deployed in the field or given to the end users. This gives incentives to adversaries to attempt to steal these ML models as a proxy for gathering datasets. While API-based model exfiltration has been studied before, the theft and protection of machine learning models on hardware devices have not been explored as of now. In this work, we examine this important aspect of the design and deployment of ML models. We illustrate how an attacker may acquire either the model or the model architecture through memory probing, side-channels, or crafted input attacks.

11) Matthew Dodd, Kinospher/NIH

I will make a very brief submission by email for your consideration. I would like to thank the Board for their Leadership and sage counsel to the Department, Industry, Academia, and indeed, the Nation. It is a profoundly overwhelming undertaking of responsibility - I thank you all and I am humbled by the purity of competence demonstrated by the Board, this stands in stark contrast to the “anti-expert”-virus that has infected every corner of society, one participation trophy at a time. While demonstrating that, primarily, one has a duty to serve their country and one has unique experiences and tools with which to innovate the notion of “service”. One has the moral imperative, obligation to find out how to manifest it, “innovate”, so as to be of service.

12) **Eric Van Hoose, FlowVU**

The Federal Risk and Authorization Management Program (FedRAMP) is a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services which provides a path for the small to midsize companies in our supply chain at the lower tiers to become cybersecure and protect our IP from thief through high level cloud service providers. This needs to be explored as a complimentary alternative to the DOD NIST 800-171 approach currently in place.

13) **Michael Tsai, Milpitas Unified School District**

I would like to know more about educational programs and initiatives that are in motion, or can be put into motion, to prepare the next generation for the changes discussed today.

14) **Seth Kazzim, Airsoft Group**

Can AI assist right down to the platoon level?

15) **Aaron Johnson, Carnegie Mellon University**

I urge the DoD to carefully consider the use of AI in situations where moral judgement is required, most critically in lethality decisions. While computer systems can learn how to mimic moral actors, they cannot grant themselves the moral authority to properly make such decisions. I draw an important distinction between "Acting" moral, where system learns what actions a moral human would be likely to take, and "Being" moral, where a human will make moral decisions based on their individual morality. For a full exposition please see:

Aaron M. Johnson; and Sidney Axinn. "Acting vs. Being Moral: The Limits of Technological Moral Actors." In Proceedings of the IEEE Intl. Symposium on Ethics in Engineering, Science, and Technology, Chicago, IL, May 2014.

Aaron M. Johnson; and Sidney Axinn. "The Morality of Autonomous Robots." Journal of Military Ethics, 12(2): 129–141. 2013

16) **William Powers, MIT**

OPINION | NICK OBRADOVICH, WILLIAM POWERS, MANUEL CEBRIAN, AND IYAD RAHWAN

Beware corporate ‘machinewashing’ of AI

MARK LENNIHAN/AP PHOTO

It was revealed in March of last year that the political data-mining firm swept up the personal information of millions of Facebook users for the purpose of manipulating national elections.

By Nick Obradovich, William Powers, Manuel Cebrian and Iyad Rahwan

JANUARY 07, 2019

Back in the late 1960s and early ’70s, when the fossil fuel industry and other corporate polluters came under fire for harming the environment, the polluters launched massive ad campaigns portraying themselves as friends of the earth. This cynical practice was later dubbed “greenwashing.”

Today we may be witnessing a new kind of greenwashing in the technology sector. Addressing widespread concerns about the pernicious downsides of artificial intelligence (AI) — robots taking jobs, fatal autonomous-vehicle crashes, racial bias in criminal sentencing, the ugly polarization of the 2018 election — tech giants are working hard to assure us of their good intentions surrounding AI. But some of their public relations campaigns are creating the surface illusion of positive change without the verifiable reality. Call it "machinewashing."

Last year, Google posted a list of seven AI principles, beginning with "Be Socially Beneficial." Microsoft published "The Future Computed," a book calling for "a human-centered approach to AI that reflects timeless values," and launched a program to support developers working to meet humanitarian needs. Germany-based SAP, one of the world's largest software companies, now has an AI ethics advisory panel that includes a theologian, a political scientist, and a bioethicist.

On seeing these initiatives, the natural response is to applaud. If the most powerful tech companies are on the case, surely these problems will soon be solved.

Or will they? Facebook's response to the intense public scrutiny it has received since the election has been to treat it as a public-relations challenge. After a sell-off of its stock in the wake of the Cambridge Analytica scandal in early 2017, Facebook spent $1.7 million on an ad campaign in subway stations and trains in the Boston area. Its slogan was "The best part of Facebook isn't on Facebook," and the accompanying images showed people engaged in healthy, fun offline activities such as hiking and dancing. The message: Facebook is all about making our world a better, more harmonious place. Yet, as The New York Times recently reported, the company had also hired lobbyists and opposition-research firms "to combat Facebook's critics, shift public anger toward rival companies, and ward off damaging regulation."

As experts on the societal effects and ethics of AI — a term that broadly refers to all technologies that use decision-making algorithms — we are keenly aware of how much work remains to be done in understanding how this new form of intelligence works once it's released in the real world.

The tech industry has a long history of humanistic intentions and pronouncements — and in fact is responsible for all kinds of progress. Yet somehow we've gotten into the most serious AI crisis since the dawn of these technologies. As with climate change and environmental degradation, if we leave oversight of intelligent machines solely to the companies that build and sell the technologies, we'll see many more crises in the coming decades.

Why? In a word, capitalism. The tech economy is driven by massive complex enterprises that exist to maximize short-term profits. High-minded rhetoric notwithstanding, serving the best interests of society is not the industry's primary objective.

To compound the problem, the baleful effects of AI are often rooted in the very algorithms that drive many tech companies' profit streams. Economists call such societal costs "negative externalities." A key component of a negative externality is that the selling or buying of the product itself doesn't price in the costs borne by others in society as a result of this transaction.

For instance, if people are clicking like crazy on ideologically divisive content served up by personalized algorithms designed to manipulate emotions, it may make both the social media company and the individual user happy in the moment. But it's inarguably a bad thing for the world. However, those clicks equal money, and when you're answering to impatient shareholders, greed has an edge over principle.

MODERN INDUSTRY IShighly skilled at concealing its true agenda with happy-talk. Greenwashing began when long-simmering concerns about pollution and other environmental threats finally gained prominence following the publication of "Silent Spring," the landmark book by Rachel Carson. It's too soon to say if most of the humanistic rhetoric and initiatives flowing from the tech industry really deserves to be called machinewashing. But as long as the profits keep flowing, the tech giants have little incentive to change the values and practices that drove their success.

Environmentalists learned long ago that if you want to know where big corporations are headed, don't follow the rhetoric — follow the money. Public grandstanding is much cheaper for corporations than implementing costly but socially beneficial solutions. Executives who make idealistic public pronouncements often act very differently when they’re behind closed doors choosing between profits and the public good.

Imagine if the titans of 1980s Wall Street had brought in ethicists to advise them on humane investment practices. Imparting moral wisdom on the Gordon Gekko generation might have sparked some great after-hours conversations over martinis. But do we seriously believe it would have prevented the systemic risk that Wall Street actually created and the price we all wound up paying for it?

AI, which is still emerging, is simply too powerful a force to entrust entirely to self-interested businesses. So beware of machinewashing. The only way to ensure that today’s technologies evolve in a healthy direction is through thoughtful, truly independent oversight. Some government regulation seems inevitable, as Apple’s Tim Cook and others now concede. But since aggressive regulation tends to stifle innovation, there should also be a role for nongovernmental oversight, perhaps through independent, transparent standards created for this purpose.

> “
>
> AI, which is still emerging, is simply too powerful a force to entrust entirely to self-interested businesses.

There will inevitably be some who view such oversight as a threat to the world’s most vibrant industry. But the environmental oversight that emerged in the last half-century — with both governmental and third-party elements — didn’t bankrupt the affected industries. It changed how they did business (though regrettably, not enough), helped clean up our air and water, and gave birth to a slew of new, genuinely green industries.

AI is the new framework of our lives. We need to ensure it’s a safe, human-positive framework, from top to bottom. If we leave it solely to the corporations, we’ll never get there.

The authors work in the Scalable Cooperation research group at the MIT Media Lab, where Nick Obradovich is a research scientist, William Powers is head of strategic partnerships, Manuel Cebrian is a research scientist, and Iyad Rahwan is an associate professor.