TECHNOLOGY FOR GLOBAL SECURITY

Roundtable Discussion: AI and Human Decision-Making

Technology for Global Security and Center for Global Security Research

November 28, 2018

Recommended Citation

T4GS, "Roundtable Discussion: AI and Human Decision-Making", T4GS Reports, November 28, 2018, http://www.tech4gs.org/ai-and-human-decision-making.html

Roundtable Discussion: AI and Human Decision-Making

Workshop Summary

On June 29, 2018, Technology for Global Security and the Center for Global Security Research hosted a roundtable discussion to explore how artificial intelligence related techniques and tools will impact international security policymaking. The discussion specifically investigated the potential security implications of these technologies as they are considered for use in military capacities. The discussion was attended by a mix of academics, research scientists, venture capitalists, civil society, and industry. This discussion was the first in a series of workshops to better understand the potential role AI will play in international stability and deterrence. This summary is an attempt at capturing the essence and critical takeaways from the discussion.

As the 21st century geopolitical balance shifts in uncertain ways, there is an increasing eagerness to deploy AI technologies into the both the physical and digital battlefields, to gain both tactical and strategic advantage over adversaries. However, the nature of increasingly powerful and unpredictable AI demands a cautious approach to releasing it before the limitations, risks, and vulnerabilities are fully understood and addressed. The consensus among the discussants was that these technologies are not currently "ready for primetime", on a number of levels. First, assumptions regarding the ability for AI technologies to "predict" are over-hyped. Second, an increase in power in a specific task does not translate to unrelated tasks: the current generation of AI remains limited to constrained environments - which warzones are not - making the deployment of current AI technologies in a military context highly unpredictable. As one participant explained, "machine learning is still very reactive - its just not sustainable." The human-machine interface remains an extremely important element of the development of these technologies, which was highlighted in the discussion of the 'black-box' problem, which makes it difficult for the user to understand the exact process by which the tech comes to it's conclusions/decisions. The speed with which decisions must be made - and increasingly so, including in wartime - means there can be limited human interference/participation, necessitating a predictability of the system well beyond current capabilities.

The concern of black-box systems was raised multiple times during the roundtable discussion, as many participants noted the importance of the user experience in the development of AI. In imagining a military commander relying upon an AI system for information on enemy whereabouts and logistical planning for military positioning, if there is not a transparent understanding of the machine's decision-making process, the veracity of the information being provided for real-time human decision making results in almost impossible decision points - not to mention that the after action review of the situation is rendered unbelievably complex not only due to the opacity of machine-learning decision making, but also due to the dynamic between the human and machine coming to a decision at all. Therein, as these systems are increasingly relied on, humans will likely no longer be able to participate in the decision making process, as attacks become either too complicated and/or fast, and the human cannot decide quickly enough which course of action to choose. This will present a situation where it is easier to attack with AI rather than defend an AI attack, which results in a bias towards conflict and first strike - while they still can. Machine learning is advancing the ability to quickly reconstruct different scenarios as part of the planning phase. In terms of deployment, the current limitations of AI relate back to the system design of these technologies. It is within the system design that determines if the developments in AI can become suitable for warfare or not.

The limitations of AI for military purposes is also rooted in the sources of data used within the training process of the systems. More specifically, incorrectly identifying the input and output of the system, as well as the context of its application(s). Currently the data on policy outputs is limited due to the small sample set that the data could be drawn from. Models that involve policy data will be limited in problem solving because there are fundamental limitations within the processes. The models lack fidelity as a result of systems viewing the world through a specific lens, and their bias within the decision making process. Additional limitations for data sets include AI's lack of ability to capture and comprehend complex geo-political concepts. Certain concepts are unable to be learned by systems of AI because they cannot be quantified. Within a military context, much of the high-level decision processes has to do with abstract ideas that cannot be broken down into correlated pairs. What is needed is not just data - of which we have an abundance - but correlated pairs. Thus, another main concern with the deployment of such AI is data validation, because models of complex real-world situations require enormous amounts of human-tagged correlated pairs of data points, which direct the computer to its goal through human involvement and supervision. This remains a significant barrier in AI. For AI to learn something, it has to be "representable" quantitatively. If it cannot be represented, the system cannot learn it. In looking at the development of AI towards human-level performance, there is a narrow context for AI systems in their specific problem domains, meaning that machine performance may degrade dramatically if the original task is modified even slightly. This means it would take a very different set of AI systems to perform a set of diverse tasks that would require only the intelligence of a single human. While there is great potential in AGI research, the exponential gains in "easy problems" with narrow scope are not being correlated to moderate gains in the "hard problems" that necessitate a holistic view of a highly complex, open-ended problem space such that a battlefield will demand (digital or physical).

Within this same discussion, it was noted that if the data set has been polluted, the question remains as to how it can be evaluated or degraded based off of the sources of data used for training systems. With massive data sets it can be difficult to converge sets of different types and values. Within a military setting it can be difficult to establish correlated pairs between the input and output data sets, and due to the different ways data is used, if there is not data on the specific goal there is a high risk with a lot of unintended consequences (i.e. proxy data).

Additionally, the uptick in spending in AI by commercial applications, and increasingly by the U.S. Department of Defense and other government globally, risks the likelihood of a race to the bottom - both in industry but also more dangerously between nation states to leverage technological developments to deter, gain, and hold advantage. The discussants agreed that overpromising was problem within the community, which is a risk at multiple levels for the development and deployment of the technology. Technology is a field that is driven by profit. As a result, AI tends to be over emphasized without a clear understanding of what is possible and what is lacking with these technologies ("we do not have the capabilities, but people think we do"). Current cheap consumer level AI can be dangerously repurposed. To determine a resolution to this problem there should be a focus on placing constraints on systems to avoid repurposing. Cyber attacks and other forms of disinformation have the potential to destabilize and undermine entire governments and their societies, and such technology is increasingly feasible without large military capabilities/expenditure. The potential threats of election manipulation and digital terrorism are very real, and the speed with which such technology is being developed/deployed is worrisome to those concerned with establishing ethical guidelines

to correlate to certain human based principles and norms. The issue of autonomous weapons engagement was vocalized by the discussion members, as such decisions made by an autonomous weapons system would need to be calculated in such a way that took into account ethics such as shooting a child vs an adult, etc. The incentives of the systems manifest different behaviors, and if the functions of a system are not clearly understood, there is a risk of a lack of control and direct responsibility of the technology that is created as decision-making process is delegated to machines.

In reality, the incentives for deploying powerful AI technology onto the battlefield and beyond will likely outweigh any ethical apprehensions, as both money and political power are at stake when competing for technological dominance. However, the present gap between technology specifications and policy makers is a real challenge. In order to determine the exact risks and policy solutions to malpurposing AI for unintended contexts, there needs to be a focus on bridging the gap between AI researchers and developers, and policy makers who make the decision to deploy such technology. The present gap between technology specifications and policy makers is a real challenge. Without the Department of Defense (DoD) understanding how to correctly deploy technology, there is a concern that if AI is applied in a military context that it will not be used in the way it was intended to. By hosting collaborative functions to bridge this gap, blind spots and technical understanding of AI systems can be promoted through discussions of both the needs of the government on a shifting, violent geopolitical landscape, as well as the current capabilities to serve those needs from a technical perspective. The policy community does not have the familiarity of using AI advancements as a strategic lever. The binary functions and lopsided capabilities of the learn patterns of AI are therefore misunderstood within this community. Thus allowing the policy community to become misguided in recognizing opportunities and misunderstanding limits. This gap represents the need to educate policy makers on the limitations of AI and how to develop alternative frameworks that are consistent and sustained. The foundation of these functions need to serve as catalysts for conversations that will lead to broader and diversified understandings of AI developments. Through repeated engagements with these groups, an understanding of the value of geo-political concepts within the context of AI as well as the reality of current AI would be further defined. Finally, as these technologies continue to evolve, it was noted that while AGI may yet be far in the future, there is a need to plan for if and when AI becomes truly powerful, instead of waiting until it actually is.