Presentation to NSPE PECon18

Beachhead of the Coming AI Tsunami

By Anthony Patch

Beachhead: a secure initial position that has been gained and can be used for further advancement www.dictionary.com

Introduction

Today's Artificial Intelligence has secured its worldwide beachhead. Immediately, it is advancing. Predominant evidence is the proliferation of blockchain architectures driven forward by quantum computing invading and encompassing every facet of human existence.

Hyperbole? Hardly. For example, cryptocurrency coverage dominates our media. However, it is but a singular sampling of the ground gained by a manmade system threatening its creators. We ourselves are giving aid and comfort to this clear and present danger.

Targeting threats to employment, the multiplicity of levels from Average General Intelligence (AGI) to the tip-of-the-spear implements of Deep Machine Learning (DML), AI's processing power has exponentially exploded within man's domain.

Worldwide deployment of quantum computing has established for AI a firm beachhead. Entrenching and advancing its positions are the ever accelerating forking and branching of its blockchain systems. ***It is especially important that organizations with strong ethics and social applications today enter this battlefield.*** Quantum computing is powerful and can be employed in the breaking of encrypted systems such as its own blockchain and attendant cryptocurrencies; as well, to the steering of financial markets, and facilitating secret communications among terror groups and criminal organizations.

AI Threat Identification and Mitigation Strategies

Less attention has historically been paid to the ways in which artificial intelligence can be used maliciously. It is necessary to survey the terrain of potential security threats from malicious uses of artificial intelligence technologies, and strategically develop tactics to better forecast, prevent, and mitigate these threats. Herein are some of the attacks likely to be seen soon if adequate defenses are not developed:

1. Policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate potential malicious uses of AI.

2. Researchers and engineers in artificial intelligence should take the dual-use nature of their work seriously, allowing misuse considerations to influence research priorities and norms, and proactively reaching out to relevant actors when harmful applications are foreseeable.

3. Best practices should be identified in research areas with more mature methods for addressing dual-use concerns, such as computer security, and imported where applicable to the case of AI.

4. Actively seek to expand the range of stakeholders and domain experts involved in discussions of these challenges.

Rapidly Evolving Threats Posed by AI

As AI capabilities become more powerful and widespread, the expected growing use of AI systems lead to the following changes in the landscape of threats:

• Expansion of existing threats. The costs of attacks may be lowered by the scalable use of AI systems to complete tasks that would ordinarily require human labor, intelligence and expertise. A natural effect would be to expand the set of actors who can carry out particular attacks, the rate at which they can carry out these attacks, and the set of potential targets.

• Introduction of new threats. New attacks may arise through the use of AI systems to complete tasks that would be otherwise impractical for humans. In addition, malicious actors may exploit the vulnerabilities of AI systems deployed by defenders.

• Change to the typical character of threats. We believe there is reason to expect attacks enabled by the growing use of AI to be especially effective, finely targeted, difficult to attribute, and likely to exploit vulnerabilities in AI systems.

AI Threat Analysis

Threat analysis is conducted by separately considering three security domains, and illustrate possible changes to threats within these domains through representative examples:

• Digital security. The use of AI to automate tasks involved in carrying out cyberattacks will alleviate the existing tradeoff between the scale and efficacy of attacks. This may expand the threat associated with labor-intensive cyberattacks (such as spear phishing). We also expect novel attacks that exploit human vulnerabilities (e.g. through the use of speech synthesis for impersonation), existing software vulnerabilities (e.g. through automated hacking), or the vulnerabilities of AI systems (e.g. through adversarial examples and data poisoning).

• Physical security. The use of AI to automate tasks involved in carrying out attacks with drones and other physical systems (e.g. through the deployment of autonomous weapons systems) may expand the threats associated with these attacks. We also expect novel attacks that subvert cyberphysical systems (e.g. causing autonomous vehicles to crash) or involve physical systems that it would be infeasible to direct remotely (e.g. a swarm of thousands of micro-drones).

• Political security. The use of AI to automate tasks involved in surveillance (e.g. analyzing mass-collected data), persuasion (e.g. creating targeted propaganda), and deception (e.g. manipulating videos) may expand threats associated with privacy invasion and social manipulation. Also expected are novel attacks that take advantage of an improved capacity to analyze human behaviors, moods, and beliefs on the basis of available data. These concerns are most

significant in the context of authoritarian states, but may also undermine the ability of democracies to sustain truthful public debates.

In addition to the high-level recommendations listed above, exploration of several open questions and potential interventions within four priority research areas requires rapid deployment:

• Learning from and with the cybersecurity community. At the intersection of cybersecurity and AI attacks, highlight the need to explore and potentially implement red teaming, formal verification, responsible disclosure of AI vulnerabilities, security tools, and secure hardware.

• Exploring different openness models. As the dual-use nature of AI and ML becomes apparent, highlight the need to reimagine norms and institutions around the openness of research, starting with pre-publication risk assessment in technical areas of special concern, central access licensing models, sharing regimes that favor safety and security, and other lessons from other dual-use technologies.

• Promoting a culture of responsibility. AI researchers and the organizations that employ them are in a unique position to shape the security landscape of the AI-enabled world. Highlighting here the importance of education, ethical statements and standards, framings, norms, and expectations.

• Developing technological and policy solutions. In addition to the above, surveying a range of promising technologies, as well as policy interventions, that could help build a safer future with AI. High-level areas for further research include privacy protection, coordinated use of AI for public-good security, monitoring of AI-relevant resources, and other legislative and regulatory responses. These interventions require attention and action not just from AI researchers and companies but also from legislators, civil servants, regulators, security researchers and educators. The challenge is daunting and the stakes are as high as the AI tsunami already breaking upon the beachhead.

AI Threat Intervention, Regulation and Control

Much of the above focuses on interventions that can be carried out by researchers and practitioners within the AI development community. However, there is a broader space of possible interventions, including legal ones that should

be considered. Note that ill-considered government interventions could be counterproductive, and that it is important that the implications of any specific policy interventions in this area should be carefully analyzed. A number of questions concerning the proper scope for government intervention in AI security arise; some initial examples:

- Is there a clear chain of responsibility for preventing AI security related problems?

- Which government departments, marketplace actors or other institutions would ideally have what responsibilities, and what would the interactions with the academic and industry communities be?

- How suitable would existing institutions be at playing this role, and how much will it require the establishment of new institutions founded on novel principles or innovative structures in order to effectively operate in such an evolving and technical field?

- Are relevant actors speaking to each other, and coordinating sufficiently, especially across political, legal, cultural and linguistic barriers?

- Are liability regimes adequate? Do they provide the right incentives for various actors to take competent defensive measures?

- How prepared does e.g. the US government feel, and how much appetite would there be for focused offices/channels designed to increase awareness and expertise?

- Should governments hold developers, corporations, or others liable for the malicious use of AI technologies? What other approaches might be considered for pricing AI security-related externalities ?

- What are the pros and cons of government policies requiring the use of privacy-preserving machine learning systems or defenses against adversarial examples and other forms of malicious use?

- Are data poisoning and adversarial example attacks aimed at disrupting AI systems subject to the same legal penalties as traditional forms of hacking? If not, should they be?

- Should international agreements be considered as tools to incentivize collaboration on AI security?

• What should the AI security community's "public policy model" be - that is, how should we aim to affect government policy, what should the scope of that policy be, and how should responsibility be distributed across individuals, organizations, and governments?

• Should there be a requirement for non-human systems operating online or otherwise interacting with humans (for example, over the telephone) to identify themselves as such to increase political security?

• What kind of process can be used when developing policies and laws to govern a dynamically evolving and unpredictable research and development environment?

• How desirable is it that community norms, ethical standards, public policies and laws all say the same thing and how much is to be gained from different levels of governance to respond to different kinds of risk (e.g. near term/long term, technical safety / bad actor and high uncertainty / low uncertainty risks)?

Conclusions

It seems unlikely that interventions within the AI development community and those within other institutions, including policy and legal institutions, will work well over the long term unless there is some degree of strategic and tactical coordination between these groups. Ideally discussions about AI safety and security from within the AI community should be informing legal and policy interventions, and there should also be a willingness amongst legal and policy institutions to devolve some responsibility for AI safety to the AI community, as well as seeking to intervene on its own behalf. Achieving this is likely to require both a high degree of trust between the different groups involved in the governance of AI and a suitable channel to facilitate proactive collaboration in developing norms, ethics education and standards, policies and laws; in contrast, different sectors responding reactively to the different kinds of pressures that they each face at different times seems likely to result in clumsy, ineffective responses from the policy and technical communities alike.

Therefore, the sense of immediacy not only in recognizing the tsunami of AI has broken upon and established its beachhead, but in understanding the outright

clear and present dangers it imposes advancing upon society as a whole and to the individual on this brave new worldwide quantum battlefield of human existence.