Enhancing Cybersecurity via Artificial Intelligence: Risks, Rewards, and Frameworks

Joshua A. Kroll, Naval Postgraduate School
James Bret Michael, Naval Postgraduate School
David B. Thaw, University of Pittsburgh

Recent advances in artificial intelligence challenge classical models of productivity by increasing the scale, complexity, and range of tasks that can be meaningfully automated, including those associated with cybersecurity.

for cybersecurity. Although it will radically change practices, AI will not change fundamentals, such as an attack retaining asymmetric advantage over defense and human operators continuing to be the primary locus of control. We provide a taxonomy of issues for assessing the effects of AI on cybersecurity. AI deployments should be viewed as rich sociotechnical systems rather than mere technical tools. When assessing their behaviors, it is critical to include people, policies, and interactions that comprise a broader system and to situate specific tools within that context. This systemic framing provides the best view to understand how to control new failure modes. We conclude with three principles for practitioners as they learn to operate in the evolving world of AI.

Artificial intelligence (AI) refers to technologies for automating tasks. Increased automation drives a race for new capabilities, presenting new opportunities and challenges

AI affords attackers opportunities to further automate their tasks, such as discovery, prioritization, and exploitation of vulnerabilities. For example, contemporary data science enables new inference capabilities, exposing

Digital Object Identifier 10.1109/MC.2021.3055703
Date of current version: 4 June 2021

EDITOR **JAMES BRET MICHAEL**
IEEE Senior Member; bmichael@nps.edu

sensitive and protected information from seemingly nonsensitive data with less need for input from human analysts. Likewise, defenders can leverage automation for monitoring, surveillance, and other defensive activities. For example, management tasks, the maintenance of situational awareness, threat-indicator triage, and the allocation of human resources can be automated and improved.

As AI-enhanced defenses improve, there will be reciprocal improvements in AI-enabled offensive capabilities—a cat-and-mouse game that already defines the domain. As automation increases, human operators balance against each other using ever-more-sophisticated technology, with attendant changes in operating paradigms, such as the shift from information superiority to decision superiority.[1] Higher-tempo and more complex operations must be balanced by more nimble governance, faster decision making, and new policies that protect security and resilience of the information infrastructure.

At the nation-state level, AI is an instrument of great power competition, as evidenced by investments by the United States, Russia, and China in research, development, and the deployment of military and civilian applications of AI. However, nation-states need to balance what is technologically possible with what is permissible (that is, law) and preferable (that is, policy).[2] Will the advantage in the future of cyberspace be driven by technological superiority; the integration of technology into existing conventional and cyberoperations; or the integration of technology, policy, and doctrine?

What will the practice of cybersecurity look like in a more automated world? What benefits will cybersecurity practitioners accrue from adopting AI technology? What new threats arise in highly automated systems, and what countermeasures are appropriate to manage them? How much does automation increase the level of control that human operators have over their domain? Does automation give a human-in-the-loop greater situational awareness and a stronger ability to project desires into actions, or does it reduce humans to components in a system, who struggle to keep up with the rapid pace and increased complexity of machine-driven activities? Will human decision makers understand and rely on recommendations from AI systems or will inscrutability in machine-generated insights limit their application and utility? Can automation improve decision making, or will it amplify failings—biases, preconceptions, and foibles of existing systems—as much as it enhances capabilities?

Answering these questions depends on the context for use of automation in cybersecurity. Although AI has transformative effects, many dynamics will remain as they are today: core threats and vulnerabilities endure; cyberconflict remains primarily asymmetric. This article provides a framework to explore this space of new risks and opportunities.

TRANSFORMATION AND RISK

To leverage new capabilities successfully, we need to understand what AI can achieve and where it will solve existing problems as well as where its limitations will limit applicability or require alternative solutions. It is also necessary to understand why and how technological advances augment the existing capacity of both attackers and defenders so that there is a clear link between the new technology and its purported effects.

AI technologies have increased the breadth of tasks that can be automated by technological systems. These advances create excitement about the ways that novel software-driven automation will change the world. And while these tools indeed hold great opportunity, that potential leads too often to failed promises, technological underperformance, unsubstantiated claims, or illusory visions of an imagined future where new technology wholly solves existing problems. These misconceptions and overambitious visions often depend, ironically, upon solving problems for which the technology is unsuited or in some cases provably inadequate. AI has gone through many iterations of the hype cycle.[3]

AI is best conceptualized as a collection of technological advances that broaden the set of tasks that can be automated usefully—the scaling-up of data storage and computational resources coupled with long-known tools such as machine learning, automated entity extraction, and the use of structured knowledge bases and inference rules to compute beyond what a programmer can write into an explicit program. Automation embodies tasks in tools and rules. AI systems discover these rules during their training or operation instead of relying on subject matter experts to develop the rules and on engineers and programmers to encode the rules into machines. Automation enables a spectrum of *autonomy*, the property that a system takes actions or influences the world on its own, without additional input from human operators. Automated systems may be highly autonomous or tightly teamed with a human: some automated systems exist purely to support human decision making (for example, a system that ranks detected alerts for response by a security operations center); others operate primarily unattended and reify their decisions without human intervention (for example, a firewall or spam filter).

Automation can enable humans to replace repetitive or high-vigilance tasks with more time for strategic

thinking. However, automation comes with risks, taking humans out of the immediate operation of systems and thus potentially lowering their situational awareness. Additionally, automation may relocate or outright remove less apparent embedded discretionary tacit decision making (for example, a police officer not issuing a ticket for someone who broke the speed limit by executing an emergency maneuver to avoid a crash). Whether the resulting human–machine partnering is more resilient and capable than the human or the machine alone depends heavily on the context of the AI application, the design of the entire system and of the automated tools, and the way these tools are situated in surrounding systems that contain and govern them. Establishing that AI is dependable—trusted to be safe, reliable, effective, free of bias, and sufficiently understandable to the humans affected by it for the purpose for which it is put into use—requires thinking about both the technology and its context, which in turn shapes system behavior. However, users often apply systems in ways the systems' designers did not intend or envision, which in turn can necessitate changes in law, policy, doctrine, strategy, and tactics.

By viewing the entire system-in-context as the locus of function and the technical tools as components, we can come to a clearer understanding of the capabilities, limitations, and transformative effects of AI technologies and how they will change cybersecurity practice and address cybersecurity risk. Systems have not only AI components but also human components, explicitly programmed software components, and nonsoftware components (that is, hardware). Interactions between components drive the most change in both capacity and risk from AI—the broadening of what can be automated.

For cybersecurity automation, this means tasks will increasingly be situated in increasingly capable tools, leaving operators more time to consider their actions while quickening the pace of decision making and execution of plans. Individuals and teams of operators could become more efficient and effective, performing existing tasks in less time with higher fidelity and gaining the ability to perform new tasks. However, as tasks are automated, operators become less familiar with details, less aware of the specifics of their actions, and less able to view or understand the internal workings of the automation, raising issues of dependability and trust. Poor design of automation can also lead to task overload for the operator, introducing risk and potentially negating some of the benefits of automation.

For attackers, automation can help synthesize new vulnerabilities;[4] scan for existing vulnerabilities; identify pivot opportunities during an attack; or suggest and prioritize targets to an attacker based on their value, ease of engagement, or risk of exposure. Furthermore, the increased volume of data and sophistication of data analysis approaches means that attackers and analysts can often make inferences about data they do not yet have access to reveal its contents with high probability or even certainty. While inference attacks are not new,[5] the capacity of modern methods to automate them is concerning. For example, a Massachusetts Institute of Technology graduate student was able to take anonymized medical records and news stories and identify which record belonged to the governor of Massachusetts despite the fact that the records contained no explicit information that singled out an individual.[6] Reidentification has been automated to the point that it can be performed on entire data sets, such as the Netflix Prize data set, which contained the movie ratings of 500,000 subscribers,[7] yielding significant amounts of protected information about individuals. Moreover, automated tools can help attackers conduct shaping operations at scale, such as influencing social media with bots.[8] The expansion of automation exacerbates the asymmetry of cyberattacks and online influence operations.

Automation also helps defenders. New tools and analytics make it easier to separate outlying or anomalous activity from normal background; automated tools can triage system alerts to raise to human operators only the items that are most critical or actionable.[9] Furthermore, tools can make decisions about system management, identifying which systems have patchable vulnerabilities and which of those patches are unlikely to be disruptive.[10] Together, such tools refocus human work, eliminating repetitive incident management or system-state tracking tasks, and facilitating more time for better tool development and other tasks.

The same tools that automate the work of attackers and defenders, along with AI tools used outside of cybersecurity, also present new risks. The most examined of these is the problem of adversarial AI in which inputs to AI systems are modified by an adversary to trigger incorrect behavior.[11] But AI systems also suffer from all of the supply-chain risks of other software. Additionally, sophisticated automation drives novel risks driven by its complexity, which can limit the efficacy of traditional approaches to ensuring that a system meets its mission-effectiveness and dependability goals.

In addition, automated systems often suffer the risk of "poisoning," where changes to the underlying data lead to changes in the decision rules, possibly in targeted ways. Such poisoning could cause the system to misrepresent the world to decision makers or fail to detect or properly interpret actions of an adversary. For example, an adversary who feeds large numbers of legitimate emails into an email system may become trusted and thus better able to insert spam or malware links at a later time. Alternatively, poisoning could simply destroy a system's effectiveness or perceived legitimacy: rather than exfiltrating data, an attacker might instead modify financial records in an attempt to affect

corporate reporting and decision making or to make profitable trades on the company's stock.

As in other domains of adversarial competition, advances in defensive technology lead to advances in attack sophistication, which, in turn, require more sophisticated defensive technology. We do not see the introduction of AI into cybersecurity changing this fundamental security dilemma either for traditional software or for AI systems.

ANALYZING AI IN CYBERSECURITY

We offer a framework for understanding the way that AI systems will affect cybersecurity. Our framework sets out a number of functional tradeoffs that break down AI application risks versus benefits to make tradeoffs legible to decision makers.

Doctrine versus generalization: Following rules is not the same as knowing things

All automation operates according to a fixed set of rules, determined in advance of when the automation is deployed. When such rules are situated in software, they can be difficult for those outside the software development process to understand, leading to opacity in the AI system; this is not the same as arguing that the system operates other than according to a fixed rule.[12] In traditional software applications, the set of rules is specified as a set of requirements and a design by human programmers and product managers and then explicitly encoded into software. AI systems, by contrast, use traditional software to knit together components that specify their decision rules implicitly (often as the set of rules that maximize some goal). Whether we are talking about early AI systems that used knowledge bases and the rules of logic to combine assertions about the world into new claims or the newest deep learning natural language models that digest huge volumes of data into the equivalent of a programmed function, the end result is always a component that maps inputs to outputs in a well-defined way—a rule.

The idea of operating a process according to a fixed rule is powerful—it enables operation at a speed and scale that could not be achieved by manual-only means, broadening the reach of sophisticated functionality, such as by scouring threat intelligence feeds from thousands of sources to extract patterns that distinguish a group of threat actors from the background din of Internet and host-system behavior.

Rules allow greater speed and scale but also create blindness around behavior not considered when they were formulated. For this reason, automated systems can fail when taken out of the context for which they were designed. For example, a computer vision system that identifies dog breeds in photos is unlikely to perform well at cat breed identification without substantial adaptation. Similarly, rules that work well on laboratory data may only do so because of quirks in that data. Adversaries who can respond to the rules can learn to fool the target system, as in the ongoing battle between better spam filtering rules and the resulting changes in the behavior of spammers. Similar dynamics exist in malware development, identity theft, and other adversarial situations.

Thus, the system's stakeholders need to know whether the rules their system embodies are suitable to the task for which the system is put to use. For traditional software-driven automation, there are a number of approaches to verification and validation (V&V), but V&V itself is challenging to do well.[13]

For AI systems, rules are generated implicitly from information built into the system. Hence, although it can be hard to state the rules by which an AI system operates, it cannot be said that such systems are built or put to use without requirements and specifications. Although V&V techniques for AI systems are not as well developed as for traditional software systems, it is eminently possible to assess the fitness of AI systems for specified tasks and goals as they are designed and built.[12]

We may prefer more flexible standards that can be applied ex post rather than needing to define an ex ante rule.[14] Think of building an automated speed enforcement camera. If the threshold for enforcement is driving at a given speed, a safe driver traveling just over the limit will receive a citation, while an unsafe driver traveling well under the limit will not. Should the enforcement threshold be the same as the legal speed limit or slightly higher to accommodate this disparity? Or should a speed limit be determined according to *standards*, flexible decision criteria a decision maker must apply and that bind that decision maker to specific desiderata (for example, "no faster than is safe for the conditions") but that do not bind that decision maker to specific outcomes? Standards bind the structure of decisions but do not mechanically translate fact patterns into decisions, and so they do not translate well to automated systems. And what of enforcement discretion, where legitimate grounds for exceeding the limit may justify noncitation? This enforcement discretion provides a low-cost, low-friction mechanism for the system to make space for interpretation of the rules, while ex post adjudication can be extraordinarily costly.

Automation versus autonomy: Human–machine teams versus direct-effect systems

Freeing up humans makes them more capable if systems are designed properly but also can lead operators astray and introduce new failure modes and safety hazards, as Tesla and Uber learned from mishaps with automated driving. The same problems with human–machine teaming were at play in the crash of Air France Flight 447 and the at-sea collision between the USS *John S. McCain* and the chemical tanker *MV ALNIC*

MC. Operator confusion about what the automated system was doing played a role in each of these mishaps. Rigorous V&V and human factors evaluation might have led to finding each system's flaws or identified scenarios that require other system interventions, such as additional operator training. The recent global compromise of Microsoft Exchange servers provides an interesting parallel: even organizations that patched their systems immediately when Microsoft announced the vulnerability were at high risk of compromise (a gap between the standard operating playbook of the system—patch critical vulnerabilities as soon as possible—and system safety—not being compromised). Furthermore, the capacity for attackers to locate and breach tens of thousands of targets indiscriminately demonstrates the power of automation for attack, while similar detection tools available to defenders provide lists of affected organizations. Although the attack seems straightforwardly detectable, incident response requires substantial human adjudication, and even automation-enhanced response teams cannot operate fast enough to meet the scale of the breach. This both shows the continued asymmetry between attack and defense in a more automated world and highlights the continued need for operator intervention to maintain system integrity. A similar problem exists with more "traditional" security tools such as malware scanning and perimeter defenses. Each requires configuration and exception management. For example, a law firm investigating a security breach may need to "loosen" the "malware" restrictions in its perimeter defenses to facilitate legitimate investigative work. Automating this process can be difficult and overreliance on automation can result in greater exposure, as distinguishing between legitimate investigative content related to the breach and potentially risky content is difficult at best without individualized human adjudication. The privilege separation requirements (within the firm) of a sensitive investigation of this nature further complicates this problem, and large-scale organizations would face a similar problem in other business contexts.

While automated systems introduce new failure modes and can conflate delegation of authority with delegation of responsibility, they also deal well with speed of decision making and execution, which is often necessary in cyberdefense.

Measurement versus reflection: Data are not the truth, and neither are models

All automated systems are models of the real encompassing assumptions about the world. Collapsing these models' abstractions onto the real world can lead to loss of functionality or even disastrous failure. Rather, we need to understand the nature of each model's assumptions. Systems based on machine learning and data science take their assumptions in the form of data gathered from the world. Even non-data-science-driven systems, such as traditional software, make assumptions about the real world when establishing their decision rules and use data to validate these assumptions. The quality and fidelity-to-reality of those data matter. Data are not objective truth. Someone decides what data to collect; how collection happens; and what to do about missing data items, outlying examples, and other data imperfections. These decisions present opportunities for data gathered to be skewed or biased representations of reality. Automated systems rely heavily on the theories of measurement and metrology, which describe how to establish whether the world models implied by gathered data are faithful to reality, valid, and reliable in both qualitative and quantitative ways.[15]

Systematic deviation from reality can lead to systematic error. However, accurate reflection of the world can also capture existing social structures and reinforce them through the operation of automated systems. In either case, this is the problem of bias in AI. Automating activities using poorly defined models of the world leads to poor results. Thus, while it is possible to recognize whether an image contains a face and identify the person identified by that face, predicting the emotion displayed by the face or whether the features of the face are predictive of attributes such as "criminality" or "employee performance" is not possible. These latter applications overlay a veneer of data-driven objectivity onto a world model that is not, and cannot be, well established or validated against reality. Automating the discovery of exploitable vulnerabilities corresponds to a well-defined, valid construct. Establishing whether a social network post represents a threat or a good-natured joke is challenging for human moderators; it is difficult or even impossible for automated systems limited by their rule-driven nature and limited linguistic capacity. It is also easy to ignore basic tenets of experimental design and conflate correlation with causation—the fact that a model finds relationships among objects or co-occurrences of phenomena does not imbue those interactions with meaning. In cybersecurity, the fact that Internet Protocol (IP) generated traffic at the time of a denial-of-service attack does not mean the IP was involved in the attack.

Permissible versus possible: Technology can challenge ethics

What is technologically possible may not correspond to what is preferable policy-wise or legally permissible. What is legally permissible is controlled by what laws and policies are operative and applicable. What is ethical depends on the context of human use and on the operating organization's chosen values and normative commitments in that context. Balancing these requirements is the domain of requirements engineering: establishing how systems can uphold legal demands or behave in an ethical manner by design. To view AI applications as systems, we must incorporate both technical components

and their context, including human interactions. However, that is not to say that the technical components of a system are mere neutral tools—rather, they afford humans a set of choices. It may be that none of these choices is ethically unproblematic. Is it then reasonable to hold a human responsible for making an ethically problematic choice? To be employed ethically, systems must afford humans ethical choices. Furthermore, the structure of the human-machine assemblage may obscure accountability.

It is possible to delegate authority but not responsibility. Automated tools enhance the capabilities of humans in sociotechnical systems, but the humans remain responsible for how the system as a whole behaves. Even if the automation robs responsible humans of detailed situational awareness, those humans always retain responsibility in hindsight. For critical applications, rigorous systems engineering, human factors evaluation, and careful design can limit the likelihood of failure while trimming the range of failure scenarios available or address other dependability concerns such as safety and security.

It is important to provide avenues for recourse outside the rule set provided by automated tools.[16] These channels for challenging a system's outputs enhance the autonomy of those affected by automation, giving them a way to respond when rules cannot or do not apply appropriately. In cybersecurity, this might mean giving discretion to human operators for how to respond to threats, maintaining records of operational decisions and actions for review by competent authorities, or providing for escalation paths when outcomes might be contested. Helping leaders and operators avoid becoming war criminals was one of the impetuses for formulating the black letter rules and commentary that comprise the *Tallinn Manual*,[17] which has influenced cyberoperations doctrine and suggests similar channels for resources within that context.

Possible versus practical: Technology versus operational practice

Just as technological possibility may not map to lawfulness or policy preference, likewise it may not correspond with operational practicality. Just because something is theoretically *possible* (that is, its existence/possibility has not been disproven) does not mean that it is a *plausible* (or even likely) occurrence. Conflating possibility and plausibility was a common error made by data protection advocates working at the intersection of privacy and cybersecurity during early efforts to define data protection rules. Contemporary cybersecurity research is gradually refocusing on risk-oriented analytical frameworks, leaving aside absolutist views of how protections and controls should apply. Cybersecurity practitioners and researchers must resist recommitting this error when considering AI.

The information revolution of the late 1990s and early 2000s saw an explosion of concern with the misuse of information. Voluminous scholarship decried threats to privacy as a result of information availability driven by new technology, notwithstanding that much of this information was already public record.[18] Pieces of information commonly used as authentication tokens in security protocols, such as an individual's mother's maiden name or place of birth, were designated as sensitive. Curiously, this information is legally required to be a matter of public record in nearly every U.S. jurisdiction—making the use of that information for security seemingly absurd. Yet, as a *practical* matter in the 1970s or 1980s, identifying where an arbitrary individual was born let alone obtaining a copy of birth records (or at least parents' names) was a time- and resource-intensive task. Today, it is as easy as cross-referencing social media profiles with some basic online searches. What changed in this time period was not the secrecy of this information but rather the *practicality* of accessing that information at scale.

In a similar vein, AI changes the operational practicality of certain types of attacks. As discussed earlier, for example, automation increases the practicality of inference attacks. Similarly, identity theft operations and associated techniques of large-scale human impersonation, noise-generating activities in verification channels, phishing, and other operations premised on access to large volumes of reasonable-quality identity-relevant information are all now much more practical than before, complicating the traditional problems of authentication. Likewise, automation of the deployment of this information—for example, submitting large volumes of credential-reset requests for employees at a targeted company—could render practical a type of denial-of-service attack that always has been theoretically *possible*, but never would have been considered practical, let alone likely.

It will always remain critical for cybersecurity to be able to distinguish between the *possibility* of a particular threat and its *practicality*. AI does not change the basic risk calculus for what attacks are feasible, which has been a core tenet of security analysis for centuries. But it might change the relative practicality of certain attacks, changing the inputs to that risk assessment.

DIMENSIONS OF CYBERSECURITY ISSUES FOR AI SYSTEMS

Similarly, we provide dimensions that cybersecurity decision makers can use to analyze problems managing the cybersecurity of AI systems used for applications other than enhancing the practice of cybersecurity.

Integrity risks: Poisoning, evasion, and trickery

If we accept that data are a model of the world, necessarily imperfect and not a true and objective reflection of reality, then we must also accept

that maintaining the integrity of that model is of paramount importance. If an adversary can modify data, they can redefine truth for the purpose of an automated system. In doing so, the adversary can trick an automated system into giving outputs undesired by the system's controller but desired by the adversary, for example, allowing access to protected data and facilities or tricking predictive maintenance systems into leaving certain vehicles unmaintained (and thus unready when needed). Although primitives for establishing the integrity of data stored at rest and data in transit are well established, defenses against data poisoning and assurance methods, which establish that data are traceable to their source and an accurate reflection of their generating measurement process, are not widely deployed. This is an area for future attention.

Confidentiality risks: Privacy and data governance

As noted earlier, AI leads to a world where the problem of data aggregation can be robustly automated by adversaries. Suddenly, every small action of an individual or an organization can become an indicator of identity or reveal sensitive activity. The retailer Target famously used the history of items purchased by individual customers to determine which of those customers were pregnant, enabling targeted marketing and advertising to customers likely to engage in drastic shifts in purchasing patterns. Thus, data privacy presents new and critical risks in an age of AI systems. Similarly, existing notions of what data are sensitive or not are likely outdated. These inference issues can be managed through the use of cover noise (for example, via differential privacy, which can guarantee that no analyst can infer certain facts with certainty greater than a given probability[19]) and more careful data governance (for example, establishing when data should be retained, for what they can be used, and when they should be destroyed).

Human factors, automation, and coordinating inauthentic behavior

AI enables the more robust adversarial manipulation of digital systems, increasing the risk of external and internal abuses. For example, adversaries can shape the nature of discourse on an online forum or social media platform by using bots to inject irrelevant spam or to respond to legitimate messages with replies that drive propaganda messages and stifle legitimate discussion. Establishing whether such activity is authentic is a difficult problem in isolation, based solely on the text of the reply. However, by drawing patterns of activity across time, space, and accounts, automation can help address the problem at the systemwide level (for example, if thousands of accounts post the same message at the same time or related times, post the same message across multiple platforms, or connect from the same network-level address).

PRINCIPLES FOR CYBERSECURITY IN AN AI-ENABLED WORLD

We have already entered the world where AI enhances the range of tasks that can be successfully automated. The changes driven by new technology are important and large but leave the fundamental outlines of cybersecurity intact. To properly understand how AI interacts with cybersecurity, it is necessary to understand the particulars of the application as well as the capabilities and limitations of the technology being introduced. Furthermore, we need to view that technology in the context of an entire system, which includes human operators, policies, organizational ethics, and many concerns beyond technology. To navigate these difficult waters, we offer an analytic framework for relating a new technology to its application. Furthermore, we conclude with three broad principles for any decision maker considering how AI will affect their cybersecurity by applying our framework.

First and foremost: automation takes tasks done by humans and embodies it in technology. Because the same work happens, but in a different way and within a different process workflow, this represents a delegation of authority but must not reflect a delegation of responsibility. Humans must remain accountable for the operation of the system and its outcomes. To establish sufficient oversight, the actions of any system must be sufficiently traceable that an oversight entity can determine what led to them and whether they might have been manipulated by insiders or outside adversaries. Designing systems to support this level of traceability is a key challenge.

Second, systems should act as their controllers intend them to, fulfilling the requirements set forward for them and capturing the needs of their controllers in those requirements. To establish this, systems must be subjected to rigorous test and evaluation processes, including robust whole-system V&V. Again, test and evaluation for AI systems is a topic of active research, though many sophisticated assessments are already feasible.

Finally, to take a complete system-level perspective, it is necessary to consider the human factors, including educating AI system developers, human operators (for example, system administrators), and other stakeholders (for example, policy makers). Identifying stakeholders and their relationship to the AI system does a great deal to establish and clarify the system's goals and requirements and aids its smooth operation and adoption. Otherwise, automation can suffer failures as fewer humans are responsible for more output but less aware of how that output is generated. Enhanced by automation, humans are both more critical to the points where the technology hands off control and less able to take on that control either in nominal or degraded modes of system operation.

In short, AI is transformative not because it magically solves previously unsolvable problems but because it

augments the capabilities of existing organizations, enabling new structures and solutions that reorganize existing problems and the work needed to address them.

REFERENCES

1. W. Williamson III, "From battleship to chess," U.S. Naval Inst., Annapolis, MD, 2020. [Online]. Available: https://www.usni.org/magazines/proceedings/2020/july/battleship-chess
2. T. C. Wingfield and E. Tikk, "Frameworks for international cyber security: The cube, the pyramid, and the screen," in *Int. Cyber Security Legal Pol. Proc.*, Tallinn, 2010, pp. 16–22.
3. J. Petro. "Beyond the AI hype cycle: Trust and the future of AI." MIT Technology Review, July 6, 2020. https://www.technologyreview.com/2020/07/06/1004823/beyond-the-ai-hype-cycle-trust-and-the-future-of-ai/ (accessed Jan. 25, 2021).
4. S.K.Cha, T.Avgerinos, A. Rebert, and D.Brumley, "Unleashing mayhem on binary code," in *Proc. 2012 IEEE Symp. Security Privacy*, pp. 380–394. doi: 10.1109/SP.2012.31.
5. C. Farkas and S. Jajodia, "The inference problem: A survey," *ACM SIGKDD Explorations Newslett.*, vol. 4, no. 2, pp. 6–11, 2002. doi: 10.1145/772862.772864.
6. L. Sweeney, "Only you, your doctor, and many others may know," *Technol. Sci.*, no. 9, p. 29, Sept. 28, 2015. [Online]. Available: https://techscience.org/a/2015092903/
7. A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. 2008 IEEE Symp. Security Privacy*, pp. 111–125. doi: 10.1109/SP.2008.33.
8. B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, "Evidence of complex contagion of information in social media: An experiment using Twitter bots," *PLOS ONE*, vol. 12, no. 9, p. e0184148, 2017. doi: 10.1371/journal.pone.0184148.
9. R. S. S. Kumar. "Reducing security alert fatigue using machine learning in Azure Sentinel." Microsoft Azure, Mar. 19, 2019. https://azure.microsoft.com/en-us/blog/reducing-security-alert-fatigue-using-machine-learning-in-azure-sentinel/ (accessed Aug. 14, 2020).
10. A. Ramesh and M. Stephenson. "Using machine learning to improve the Windows 10 update experience." Windows IT Pro Blog, Sept. 26, 2019. https://techcommunity.microsoft.com/t5/windows-it-pro-blog/using-machine-learning-to-improve-the-windows-10-update/ba-p/877860 (accessed on Aug. 14, 2020).
11. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Security Privacy*, 2018, pp. 399–414. doi: 10.1109/EuroSP.2018.00035.
12. J. A. Kroll, "The fallacy of inscrutability," *Philosoph. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 376, no. 2133, p. 20,180,084, 2018. doi: 10.1098/rsta.2018.0084.
13. J. B. Michael, G. W. Dinolt, and D. Drusinsky, "Open questions in formal methods," *Computer*, vol. 53, no. 5, pp. 81–84, 2020. doi: 10.1109/MC.2020.2978567.
14. J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," *Univ. Pennsylvania Law Rev.*, vol. 165, no. 3, pp. 633–705, 2017. [Online]. Available: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3
15. A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proc. Conf. Fairness, Accountability, Transparency*, 2021, pp. 375–385. doi: 10.1145/3442188.3445901.
16. A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Fordham Law Rev.*, vol. 87, no. 3, pp. 1085–1139, 2018. [Online]. Available: https://ir.lawnet.fordham.edu/flr/vol87/iss3/11
17. M. N. Schmitt, Ed. *Tallinn Manual on the International Law Applicable to Cyber Warfare*. New York: Cambridge Univ. Press, 2013.
18. S. Garfinkel, *Database Nation: The Death of Privacy in the 21st Century*. Sebastopol, CA: O'Reilly Media, 2000.
19. C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *J. Privacy Confidentiality*, vol. 7, no. 3, pp. 17–51, 2017. doi: 10.29012/jpc.v7i3.405.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of their employers. The U.S. Government is authorized to reproduce and distribute reprints for government purposes, notwithstanding any copyright annotations thereon.

JOSHUA A. KROLL is an assistant professor of computer science at the Naval Postgraduate School, Monterey, California, 93943, USA. Contact him at jkroll@nps.edu.

JAMES BRET MICHAEL is a professor of computer science and electrical and computer engineering at the Naval Postgraduate School, Monterey, California, 93943, USA. Contact him at bmichael@nps.edu.

DAVID B. THAW is an associate research professor of law and computing and information at the University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, USA. Contact him at dbthaw@pitt.edu.