Human-Machine Teaming and Cyberspace

Fernando J. Maymí[1] and Robert Thomson[2(✉)]

[1] Soar Technology, Ann Arbor, MI 48105, USA
fernando.maymi@soartech.com
[2] Army Cyber Institute, West Point, NY 10996, USA
robert.thomson@usma.edu

Abstract. Artificial Intelligence is becoming the key enabler of solutions to a variety of problems including those associated with cyberspace operations. Based on our analysis of cyber threats and opportunities in the coming years, we assess it as very likely that teams consisting of humans and synthetic agents will routinely work together in many if not most organizations. To fully leverage the potential of these teams, we must continue to develop new paradigms in human-machine teaming. Specifically, we must address three areas that are currently in their infancy. Firstly, we need interfaces that allow all teammates to communicate effectively with each other and seamlessly transfer tasks among them. This must be true regardless of whether the endpoints are human or not. Secondly, we will need cybersecurity operators with broad knowledge and skills. They must know how their synthetic teammates "think," when to task them and when to question their reports. Thirdly, our AI systems must be able to explain their decision-making processes to their human teammates. This paper provides an overview of cyberspace threats and opportunities in the next ten years and how these will impact human-machine teaming. We then apply the key lessons we have learned while working a multitude of advanced research projects at the intersection of human and AI agents to cyberspace operations. Finally, we propose areas of research that will allow humans and machines to better collaborate in the future.

Keywords: Human-machine teaming · Artificial intelligence · Cyberspace

1 Introduction

The United States Department of Defense (DoD) defines cyberspace as a global domain consisting many different and often overlapping networks (Joint Pub 3-12 2013). Though many people equate cyberspace with the Internet, the latter is simply a subset of the former. Cyberspace, after all, includes many networks (e.g., classified intelligence networks) and systems that are not directly reachable from the Internet. Though it is difficult to characterize the nature of these other networks and systems that comprise cyberspace, we know a fair amount about the Internet. We know, for instance that it is the largest, most complex system ever built by humans. By some estimates, it consists of over 8 billion devices (Tung 2017) exchanging over 4 billion bytes of data every second ("Internet Live Stats" 2018). Cyberspace, by definition, is even bigger.

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10915, pp. 299–315, 2018.
https://doi.org/10.1007/978-3-319-91470-1_25

The size and speed of the Internet, coupled with its growth rate, prompted the development and application of artificial intelligence (AI) techniques for performing tasks that humans alone could no longer effectively do at this scope. It also spurred the creation of novel capabilities that take advantage of, and indeed require, the very large data sets that are available in cyberspace. The development of these techniques has been organic and, while enabling localized capabilities, has sometimes hindered other ones. In particular, we are concerned that some trends in both human and synthetic (i.e., AI-enabled) operator development are not supportive of effective human-machine teaming. In Sect. 2 of this paper, we provide a brief introduction to AI in general and to some of the specific concepts we'll discuss later in the paper. On this foundation, we describe in Sect. 3 future threats that motivate the need for better human-machine teaming. We then describe advances in AI that allow the creation of synthetic cyberspace actors in Sect. 4. In Sect. 5, we address the human members of future cyberspace operations teams. Section 6 presents the need for AI that is explainable to humans as the foundation of trust in these teams. These human-machine teams of the future are described in Sect. 7. Finally, we offer our conclusions and recommended future work in Sect. 8.

2 A Brief Introduction to AI

AI is fundamentally concerned with machines that solve problems and make decisions or appear to think analogously to a human at some level of approximation. While there is no single definition for the term, there exist different classes of AI that allows us to formulate a tentative ontology, which we show in Fig. 1. A high-level bifurcation is possible by differentiating the approach used to represent information or knowledge. Symbolic approaches, as the name implies, use symbols (e.g., words) to represent the atomic components of thought and generally rely on some kind of semantic rules to process information. Alternately, non-symbolic approaches use numerical and often distributed representations (reflected by patterns of activity across numerous processing units).

In symbolic approaches to AI, system developers model real-world concepts, their relationships and how they interact to solve a set of problems. This effort requires considerable knowledge of both the problem and solution domains, which makes it fairly labor-intensive. However, it yields results that are inherently explainable to humans since they are derived from human knowledge models in the first place. Symbolic AI systems include the expert systems that became prolific in the 1970s and 80s. These relied on extensive interviewing of subject matter experts and time-consuming encoding of their expertise in a series of conditional structures. An example of this approach is MYCIN, one of the first practical rule-based systems that was developed to help physicians select antimicrobial therapies (Shortliffe 2012). These early expert systems suffered from a fundamental inability to adapt or learn absent human intervention in updating the knowledge base.

Non-symbolic AI gained momentum after many in the AI community, disappointed with the limitations of symbolic approaches, looked to animal brains for inspiration. In Artificial Neural Networks (ANN) each node receives multiple inputs from other

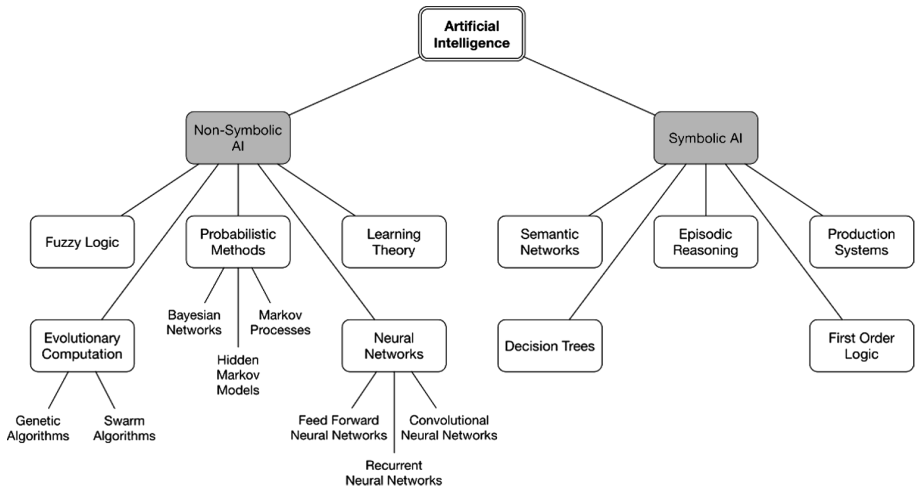

Fig. 1. A general ontology of AI techniques

nodes, typically in the form of a real number, and produces one or more outputs that are the result of applying some function to those inputs. By applying weights to each connection and allowing those weights to be modified through a feedback loop, the ANN can be trained. There are many other non-symbolic approaches, such as probabilistic ones, that have been successfully applied to problem sets in which the knowledge engineering required in symbolic AI is not a feasible option.

Many modern AI systems are able to learn from experiences. Machine learning (ML) refers to techniques that allow AI systems to adapt to changing inputs and ideally improve their performance as a result. Though ML is equated with non-symbolic approaches, it is also possible for some symbolic AI systems to learn. The Soar cognitive architecture, for instance, is a symbolic production system capable of episodic learning. A Soar agent could achieve a goal through a circuitous series of intermediate steps, some of which will be successful. Over multiple experiences or episodes, the agent condenses these steps into a shorter, more effective and efficient chain. This process, called "chunking," is one of the main ways in which Soar agents learn.

ML can take place with or without human help. In supervised ML, the system is presented with inputs and must then produce an output, typically in terms of a classification (e.g., an email message is or is not spam). If the output is correctly classified, the system receives positive reinforcement; otherwise, it may receive negative reinforcement. The learning or, more accurately, training process can be automated by using labeled training data sets. If you remove the human from the process and don't use labeled data, a system can still learn through reinforcement learning. In this form of ML, the system interacts with its environment in a sequence of observation-action pairs where a reward is presented after each action. Using this approach, a system could learn how to efficiently route network packets using as rewards the inverse of the number of hops required. The key requirement for these ML approaches is a feedback process that allows the agent to determine when its decisions are correct. This process can be

artificial (e.g., tagged data sets) or natural (e.g., observing the behavior of routed packets).

3 Future Threats

The ability of AI to make increasingly complex decisions much faster than humans could, all the while learning from its experiences has already delivered many benefits in the service of humanity. The same capabilities, however, can cause unexpected and undesirable effects as Microsoft learned when it developed chatbot that learned to compose racially and sexually offensive tweets (Metz 2016) from its interactions with thousands of people. Perhaps more concerning are scenarios in which AI systems are intentionally developed and deployed to cause harm. It is, after all, logical to assume that malicious cyberspace actors will leverage emerging technologies for their own purposes.

If an attacker is using AI to operate at machine speed, the defender must be able to work at least as quickly in order to be effective. This idea of synthetic agents attacking and defending information systems with no humans in the decision-making loop inspired the Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge (CGC), which brought together seven finalists to Las Vegas, Nevada in August of 2016. The goal was for these cyber reasoning systems (CRS) to perform automated vulnerability detection, exploit generation, software patching, and to determine when it would be most advantageous to patch a vulnerability or exploit it on a competing team's CRS without human intervention (Brooks 2017). The message is clear: in the future of cyberspace both attackers and defenders will, at least partially, be autonomous agents. In fact, the leader of the winning team, David Brumley, founded the company For All Secure to take autonomous vulnerability detection (and potentially patching) to market.

It is not only machines who will be threatened by autonomous agents. Many security experts anticipate a new breed of phishing emails generated by ML algorithms that will be much more targeted, compelling and effective than human-generated ones (Emmanuel 2017). One of the reasons why these messages will be more threatening is that they will leverage the ability of data analytics and ML to scour vast data sources for information with which to precisely target individuals at scale. The U.S. Army already identified this micro-targeting trend as a feature of future wars (Kott et al. 2015) for which our current counter-measures are ineffective.

The Social Network Automated Phishing with Reconnaissance (SNAP_R) system (Seymour and Tully 2016) demonstrated a recurrent neural network (RNN) that is able to tweet phishing messages that target specific users. During a limited experiment, SNAP_R was four times faster than humans at sending out targeted attacks while achieving an order of magnitude improvement in the target click rate. One year later, DARPA announced its Active Social Engineering Defense (ASED) program, aimed at autonomously identifying, disrupting, and investigating social engineering attacks. The very existence of ASED underscores the difficulty and long-term significance that DARPA attributes to this threat.

Finally, as AI in general and ML in particular become increasingly important in our lives, adversaries will develop attacks aimed directly at the ML mechanisms that are designed to improve and defend our lives. Adversarial ML (AML) is an emerging field of study concerned with attacks against online ML algorithms (Huang et al. 2011). Early research has shown that ML classifiers are susceptible to three types of attacks (Papernot et al. 2016). Confidentiality attacks entail gaining information on data used to train the ML system (Shokri et al. 2017), its internal model (i.e. weights), or architectural (i.e. learning rate) parameters. Integrity attacks attempt to modify input to the ML classifier in order to induce a particular output or behavior, such as causing an image recognition system to misclassify a 2 as a 9 by modifying a few image pixels (Carlini and Wagner 2017). Availability attacks attempt to deny access to the ML classifier such as by generating numerous false positives. An ML-based IDS/IPS, for example, is vulnerable to such attacks. As AML techniques mature, malicious actors will employ them to manipulate the outputs of intelligent systems.

4 Synthetic Actors

Against this backdrop of technological opportunities and threats, research and development of autonomous synthetic actors proceeds apace. Though much work to date has focused on applications of ML to the detection and mitigation of cybersecurity incidents, research is also taking place towards the development of more robust defensive agents that can hunt for and neutralize threats on their networks. As we mentioned in our threat discussion, we are seeing similar moves in the development of attack capabilities. In fact, one of the noteworthy aspects of DARPA's CGC was that it demonstrated the feasibility (and, one might argue, inevitability) of autonomous offensive and defensive agents fighting against each other with humans out of the loop at least in some cases. While we have not yet seen documented cases of autonomous synthetic attackers conducting real operations, many think that these incidents are not too far in the future (Dvorsky 2017). Indeed, SoarTech has already demonstrated cognitive agents that can perform defensive and offensive (e.g., penetration testing) activities in virtualized environments.

SoarTech's Simulated Cognitive Cyber Red-team Attacker Model (SC2RAM) is a synthetic, offensive, cognitive agent that emulates real attackers by modeling the complex thoughts, decision-making, and contextual understanding of a human interactive operator. Its goal-seeking behavior results in a virtually unlimited range of realistic attacks. The current attacker agent, built on the Soar Cognitive Architecture (a symbolic AI platform) can conduct multiple attacks including phishing with malicious documents, remote exploitation, and SQL injection. A custom remote access toolkit developed for this project provides additional persistent on-target capabilities such as lateral movement and file exfiltration, providing a realistic experience for training network defenders. The premise of red teaming and penetration testing, exemplified by SC2RAM, is that it is better to test one's own defenses against realistic but benign attackers than it is to wait until the real adversaries do so. Since human penetration testers are rare and expensive experts, it is logical to leverage synthetic agents in this manner.

It also makes sense to employ such agents when the scale of a problem requires a very large number of interactions. Much of the research at the intersection of cybersecurity and AI uses non-symbolic approaches. Some of the first successful applications of ML to cybersecurity were in classification of spam email messages (Cohen 1996). Over the last two decades, these approaches have become remarkably accurate. Today's ubiquitous spam filters improve their performance through interaction with the humans whose inboxes they protect. When the agent misclassifies a message, the human has an opportunity to correct the error thus allowing the ML system to learn to improve itself.

Given the role of these agents as first lines of defense for end points, much research is needed in identifying vulnerabilities to AML in systems such as these spam filters or the newer breeds of antimalware products that use ML to detect malicious software. Here one could utilize machine learning techniques to make inferences on the training set of another machine learning classifier in order to manipulate inputs to generate desired outputs. For example, given an ML system that classifies software as benign or malicious (e.g., an anti-malware application), one could imagine another system that generates multiple variants of malware, each with small perturbations that don't affect its functionality. These variants could be sent to the classifier until it incorrectly decides that the malware sample is benign. Given enough such misclassified samples, the AML system can make inferences about what it takes to fool the defender. This AML versus ML assessment could serve to harden network security applications by evaluating the robustness of an already trained model, particularly when the internal classifier parameters are unknown. Since this sort of assessments require many thousands or millions of attempts to characterize the system under test, synthetic agents would be well-suited to perform them.

Despite their ability to analyze vast amounts of information, non-symbolic approaches like those in use for spam, malware and intrusion detection are less effective at reasoning over the context and meaning of cyberspace activities. They are ideally suited to answer the questions of *what* and even the *how*, but not the *why*. Symbolic approaches, such as rule-based systems, on the other hand, are oftentimes better for this purpose because they model higher-level cognitive processes and human expertise. A promising area of research for more effective synthetic cyberspace actors is the integration of symbolic and non-symbolic approaches to help us identify not just the threats, but also their possible implications to our organizations and systems. Such hybrid systems would be more capable in a wider variety of situations. It will be at that point that synthetic actors could become real teammates to their human counterparts, significantly enhancing the performance of our workforce.

5 Human Actors

One of the challenges in reviewing the current state of the cyber workforce is that there is a paucity of quantitative assessment regarding the cognitive aptitudes, work roles, or team organization required by cyber professionals to be successful. We argue that the people who operate within the cyber domain need a combination of technical skills,

domain specific knowledge, and social intelligence to be successful. They, like the networks they operate, must also be secure, trustworthy, and resilient.

A concern in writing about human actors is that cyber professionals are generally seen as a homogeneous, holistic classification. That said, due to the complexity and rapid evolution of the tasks involved in cyber defense, it is important to note that there is substantial heterogeneity between work roles and individual skillsets. By virtue of this complexity in the task environment, cyber professionals need to work in teams. While in the military context cyber teams tend to be teams of diverse talents, in the private sector it is much more likely for smaller teams to be composed of similarly-talented individuals rather than a group with diverse work roles and backgrounds (Champion et al. 2012). Recent research has identified that cybersecurity teams are better able to solve complex tasks than individual analysts, potentially due to the distribution of expertise across analysts (Rajivan 2014; Rajivan et al. 2013; Rajivan and Cooke, in press). For instance, performance on incident triage was highest with a diverse group of heterogeneous talents as opposed to a team with members of similar background and skills. (Rajivan 2013). A limitation of research into cyber teamwork is that they have not examined different organizations of teams or combinations of teams. This future research is essential to determine the correct make-up of the future cyber workforce.

Champion et al. (2014) investigated the contribution of informal education to developing cyber security expertise and found that 69 of 82 professionals reported that informal education supplementation was a prerequisite for career success. Furthermore, 40% of professionals felt that job experience was the highest factor in positive performance over degree of knowledge/education (12%). Many professionals anecdotally reported that those receiving supplemental on-the-job training and mentoring exhibited the highest performance benefits as measured by future career success. Similarly, Asgharpour et al. (2007) found that operators who subjectively rated themselves with higher levels of expertise tended to have both more and more diverse competencies than those with less self-professed expertise.

Cognitive task analyses have identified that cyber professionals need to exhibit strong situational awareness (Jajodia et al. 2010), including juggling concurrent sources of information regarding the health of the network, historical and current network activity, and performing a continual assessment of risk. For recent meta-analyses see Franke and Brynielsson (2014), and Onwubiko and Owens (2011). Similarly, through the use of structured interviews, Goodall et al. (2009) interviewed twelve cyber professionals and identified that the requirement for situated knowledge (i.e., knowledge of the local environment) made intrusion detection a relatively unique task and challenging to transfer expertise to other tasks in the cyber domain. This required triage teams to interface with local workers to understand the topology and peculiarities of the local network to determine whether an intrusion had occurred and what remedies were available.

There are numerous tools to process this incoming information (e.g., Bro and Snort for intrusion detection), however, there is just too much information for a human actor to successfully process, and critical misses are inevitable. A human teamed with a machine, however, has the potential to cover a much wider set of attack vectors

because the machine does not have the same attentional limitations and can do a more thorough assessment of making sense of large swaths of incoming data.

Before proceeding to discuss the importance of AI systems that can interact with human actors, it is important to understand how we are training our cyber workforce and to identify any gaps in training. The Department of Homeland Security's National Initiative for Cybersecurity Careers and Studies (NICCS) developed a Cybersecurity Workforce Framework (Newhouse et al. 2016) to provide a base set of work roles for the cyber workforce. While this ontology was not empirically justified, it represents the most well-documented rostering of work roles in the cyber domain. This collection includes nine work-role categories, 31 specialty areas, and over 1000 types of knowledge, skills, and abilities. Major categories are described in Table 1.

Table 1. Cybersecurity Workforce Framework. Reproduced from *(Newhouse et al. 2016, p. 14)*.

Work-role category	Description
Securely provision	Conceptualizes, designs, and builds secure information technology (IT) systems, with responsibility for aspects of systems and/or networks development
Operate and maintain	Provides the support, administration, and maintenance necessary to ensure effective and efficient information technology (IT) system performance and security
Oversee and govern	Provides leadership, management, direction, or development and advocacy so the organization may effectively conduct cybersecurity work
Protect and defend	Identifies, analyzes, and mitigates threats to internal information technology (IT) systems and/or networks
Analyze	Performs highly specialized review and evaluation of incoming cybersecurity information to determine its usefulness for intelligence
Collect and operate	Provides specialized denial and deception operations and collection of cybersecurity information that may be used to develop intelligence
Investigate	Investigates cybersecurity events or crimes related to information technology (IT) systems, networks, and digital evidence

Securely Provision roles revolve around the more traditional information technology field including software developers, computer programmers, and network architects. The Operate and Maintain roles include System Administrators, Knowledge Management, and Security Analysts. The Oversee and Govern roles include managerial roles, Cyber Law, Policy Development, and Education. The Protect and Defend roles include Cyber Analysts (Operators) and Network Defenders. The Analyze, Collect and Operate, and Investigate roles all encompass the broad field of Digital Forensics and will tend to be government or law enforcement positions (Caulkins et al. 2016).

In general, cyber professionals in the Securely Provision, Operate and Maintain, and Protect and Defend work roles must have good mental flexibility and pattern matching abilities (Baker 2016; Ben-Asher and Gonzalez 2015; Champion et al. 2014).

They will have to possess significant skill and knowledge about computer operating systems and using analytical tools for such things as network scanning, network mapping, and vulnerability analysis. This task environment involves scanning large numbers of network events and (generally false) alerts across multiple computer screens with the goal of identifying threats while minimizing false alerts (D'Amico and Whitley 2008).

A limitation of the NICCS Workforce Framework is that, of the 1060 types of knowledge, skills, and aptitudes, fewer than ten describe teamwork or working with AI. This implies that the Framework paints an incomplete picture of workforce proficiency (Cook 2014). Furthermore, the development of any cyber workforce that neglects the social aspect of human behavior on the network neglects a critical component of the cyber domain. For instance, cyber defense would be aided by an understanding of human behavior and how it introduces risk to the network (Asgharpour et al. 2007; Pfleeger and Caputo 2012). We should leverage AI and humans' capabilities to maximize information exchange so each level processes the right 'kinds' of information to be most effective. Under this view AIs should process the large swaths of incoming poorly-structured data and distill this data into a format that can be readily presented to a human operator. The human operator can then perform high-level strategic inference over this well-structured information from the AI. We now know that human operators, though, will not use this data unless they can understand why the AI makes its recommendations.

6 Explainable AI

Most AI systems today are not designed to (nor can they usually) explain to their human users the manner in which they arrived at their conclusions. The reason is that most AI developed to date for cybersecurity applications is non-symbolic. As we explained in our introduction to AI earlier, these approaches, unlike symbolic ones, are not inherently explainable. System designers would have to deliberately develop explanation mechanisms, which is something seldom seen in the field. Faced with such opacity, many users choose to blindly trust the computer, which is a phenomenon that has been called the "in screen we trust" effect (Aiken 2017). The option is to distrust the computer and ignore its decisions if they seem unreasonable. Some systems, however, might not allow this option if their AI mechanisms are part of closed decision loops that don't allow real-time human interference.

In order to develop and maintain the trust that is inherent in teaming, AI systems must be able to explain their conclusions to human teammates. In this regard, symbolic AI approaches such as expert systems and cognitive architectures are better suited because they model human knowledge and thought processes respectively. Their very nature is similar to higher level human thought constructs, which in most cases makes it simpler for them to present their causal chains to humans. Conversely, this nature also makes it easier for humans to point out errors or omissions in their synthetic teammates. The Soar cognitive architecture, for instance, uses goal graphs to simulate human cognitive processes, which naturally lends itself being explainable to people by representing the synthetic decision-making processes as goal trees.

Visualizing non-symbolic systems like ML processes, on the other hand, has traditionally been more difficult. The reason for this is that they mostly rely on mathematical models and processes. To this end, the Defense Advanced Research Projects Agency (DARPA) is pursuing its eXplainable AI (XAI) program, to which the authors of this paper are both contributing. One of projects in this program is XAI for the Veterans' Transition Assistance Program (XAI-VTAP), which is geared towards matching the resumes of veterans to open job postings. Some of the work being done in this project uses novel techniques to provide an unprecedented level of visibility into how ML algorithms arrive at conclusions. Figure 2 shows how and why the system matched a specific resume to multiple occupational categories. The top part of the figure shows how good of a fit a candidate is against each category and provides examples from that person's resume. The bottom part illustrates how various indicators were ultimately mapped to various categories. One could imagine job seekers using this feedback as a training aid to create better resumes in general, as well as resumes that improve their odds of getting specific jobs.

Occupational Category	Maximum Activation	Occurrences in Resume	Example from Resume
general and operations managers	32.3	29	"employment history ai cyber rd director 2016 present us"
software developers, applications	60.8	13	"arduino pic developer tools eclipse vi ms visual studio svn git uml avr studio matlab octave artificial intelligence systems soar clips jess machine learning robotics player stage ros gazebo"
chief executives	18.3	9	"accomplished public speaker speaking in both technology focused and senior executive level sessions"
software developers, systems software	12.1	9	"automated army mobilization and deployment processes by architecting and overseeing development of several multi million dollar distributed cloud based applications"
information security analysts	16.2	7	"technical experience in cyber security autonomous systems cognitive systems and computer science and engineering research development and education"

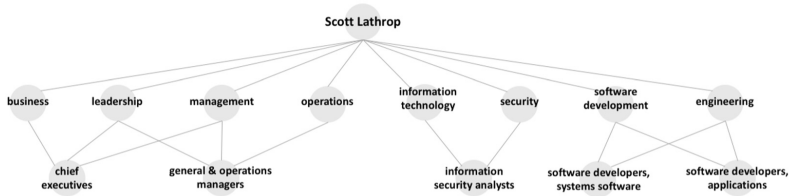

Fig. 2. User interface prototype for Explainable AI to support Veterans Transition Programs (XAI-VTAP)

While such explainability could lead to better employment opportunities and possibly improve resume-writing skills for veterans and other job seekers, it also enables the threats to AI systems posed by AML. While there are many types of AML, the one that is most relevant to our discussion is the deliberate manipulation of the data inputs to an ML mechanism so that it fails to function as intended. This could happen if an adversary determines how an ML-based span filter works and then crafts span messages that are not identified as such and thus are delivered to a victim's inbox. It could also happen if the adversary pollutes the training data set for an ML-based product so that it is trained to correctly identify spam messages except those that have a particular set of characteristics that only the adversary knows. This would then allow

only that adversary to bypass the detection mechanism. The knowledge that can be gained through explainable AI facilitates AML techniques.

Still, explainability is crucial to our human-machine teaming efforts for three reasons. Firstly, it allows trust to be developed between humans and their synthetic teammates. Autonomous AI agents are likely to reach some seemingly far-fetched conclusions that may stretch their credulity of their human counterparts. In those situations, it is necessary to be able to walk the human through the thought process. Secondly, the AI system's conclusions will only be as good as the models they have and the learning they have been able to do on their own. It is entirely possible that some misfits may occur, in which case the human will be able to detect the error, point it out to the AI system, and allow it to learn from the experience. Thirdly, synthetic teammates have tremendous potential as training tools, which can only be realized if it is able to explain itself to those who are learning from or with that system.

7 Human-Machine Teaming

The notion of shared mental models between humans and machines is a common thread when examining human-centered big data research. Mental models provide a representation of situation, various entities, capabilities, and past decisions/actions. These models are dynamic, with analyst and model engaged in a continuous production loop. In addition, from a purely human level there is research on teamwork (Baumann and Bonner 2013) and the degree to which teammates from different backgrounds have overlapping shared mental models (Bearman et al. 2010). There is also research on the degree to which multiple agents can recognize a common plan from reading large corpora (Paletz 2014).

Teams of security analysts are in many instances, a loose association of individuals, rather than a functioning team (Champion et al. 2012). A functioning professional team is a "purposive social system" (Hackman and Katz 2010), in which members of the team have diverse backgrounds, identified by role and work together in an interdependent manner towards common objectives (Salas et al. 1992). Team effectiveness largely depends upon appropriate leadership, team structure, communication, collaboration and distribution of tasks. Communication is the key medium by which human teams form relationships, collaborate and share information (Cooke et al. 2013). Communication is the conduit to transform individual expertise and situational awareness to team level knowledge and situational awareness.

Field studies with security analysts found that communication and collaboration between security analysts was an integral aspect of effective defense particularly during a widespread security crisis (Goodall et al. 2009; Jariwala et al. 2012). Lab experiments on collaboration during the threat detection have also found evidence that cooperation between security analysts during triage analysis augments signal detection performance, particularly in novel and complex situations (Rajivan et al. 2013). However, during collaborative analyses, analysts may fail to contribute requisite expert knowledge and demonstrate biases in the way information is pooled from each other, leading to communication losses affecting threat detection performance (Rajivan 2014). Communication across the hierarchy of security analysts have also been observed to be

inefficient and largely one-directional (bottom-up). Tools for collaborative threat detection developed using human systems engineering principles would help in mitigating such losses in communication between security analysts (Rajivan 2014).

Leadership is also crucial to security defense team development and performance (Buchler et al. 2017). Typically, an individual in a leadership role is expected to: develop team capabilities, facilitate problem solving, provide performance expectations, synchronize and integrate team member contributions, clarify team member roles and engage in meetings and feedback (Salas et al. 1992; Simsarian 2002). Field studies on security leadership showed that leadership is a significant predictor of defense performance. In one such study, two security teams, otherwise equivalent in skills, experience and knowledge, was observed to demonstrate widely different defense performance primarily due to differences in leadership approach and amount of collaboration (Jariwala et al. 2012). In a subsequent study, it was found that functional specialization and adaptive leadership strategies are important predictors of security defense performance (Buchler et al. 2017). Except for these handful of studies, the determinants of effective teamwork and leadership among security analysts is still an emerging area.

Collaboration, communication and knowledge integration is necessary for accurate and expeditious correlation analysis. From past team research, it is evident that teams often don't realize their full potential and could fail for a multitude of reasons. Loss in team processes such as communication would lead to sub-optimal decision making. For example, collaborative threat detection requires the exchange of expert information between security analysts. Previous research has demonstrated that teams may not be effective in exchanging novel information. Particularly, uneven information distribution biases people to share, more often, information that are known to majority in the team and prevents them from sharing and associating unique information available with them (Stasser and Titus 1985). The effect of such team-level biases on security team collaborations are largely unknown.

Experiments on team interactions need to be conducted ideally in context (through field studies) or using simulation environments. Due to restricted access to real world cyber protection teams and due to lack of importance currently given to team process metrics in cyber defense exercises (Granåsen and Andersson 2016), experiments on team interactions in cyber defense can instead be conducted in the lab using simulation systems that recreate realistic team interactions and work flows between study participants which would in turn require the participants to exercise some of the same cognitive process involved while conducting cyber defense in the real world (Cooke and Shope 2004).

We argue that in order to incorporate machines into human teams effectively, they must be natural to use, seamlessly integrate into the task environment, and provide a subjective improvement in effectiveness. Ideally, a single human operator (or small team of operators) would be able to supervise multiple AIs (Chen and Barnes 2014; Pellerin 2015; Trexler 2017). The goal of the AI is to process the massive amount of incoming information, present it efficiently to the human operator, make low-level decisions, and help the human operator make high-level strategic decisions. This AI will be able to make decisions at the speed of cyberspace and adapt to new attack vectors in near real-time, which is orders of magnitude faster than a human operator.

We foresee that within the next decade, the war for cyberspace will be fought between nations' AIs, and the skill of the operators and effectiveness of the AI's algorithms will be the deciding factor.

As such, it is essential for human operators to trust their AIs. Petraki et al. (2015) argue that it is important to have mutual predictability and adaptability in order engender trust. As previously discussed, that is one of the main goals of DARPA's eXplainable AI Program. The ability of the AI to be able to adapt to a human operator's goals, and for the operator to query the underlying question as to 'why' a decision was made is key to trusting in the AI's automation. One such technique is to supplement traditional AI techniques with models that approximate human behavior, such as in the Soar cognitive architecture and the ACT-R cognitive architecture.

In summary, by leveraging AIs to do much of the complex sensemaking required in many cyber operations tasks, we argue that it is possible to maximize a human operator's ability to conduct strategic operators effectively, even in the face of an overwhelming amount of incoming data. We argue that AIs need to seamlessly integrate with humans, and that they need to be explainable in order for human teammates to trust their output.

8 Conclusions

From the foregoing, we posit that there are three key elements of effective human-machine teaming in cyberspace: effective intra-team communications mechanisms, a sophisticated and diverse cyber workforce, and AI systems that can readily explain the rationales for their decisions to their human teammates.

We have already established that communication is the key medium by which teams form relationships, collaborate and share information. It is a logical extension of this premise to assert that whatever the team composition (e.g., human, synthetic), as long as there is at least one human in the mix, effective communications will be required to build and maintain the team's effectiveness. Even if there are no humans in a team of cyberspace actors, communications will be key, albeit in a somewhat different form.

It will also be important to ensure that the human actors that are teaming with AI systems are knowledgeable of the capabilities and limitations of the underlying technologies. In other words, to fully leverage the potential of our synthetic teammates, we will need cybersecurity operators with broad knowledge and skills, and who know when to task agents and when to question their reports. There is a dearth of research in this area, so much work needs to be completed before we can quantify the requirements for humans in an effective human-machine cybersecurity team.

Finally, the skills of the human actors will be excessively tasked unless their synthetic teammates are able to explain to them the manner in which they reached a specific decision. This requirement for explainable AI addresses two critical aspects of effective teaming: trust and correctness. An important element of teamwork is trust, which can be eroded by unexpected behaviors, particularly those that could seem to undermine or threaten mission accomplishment. If a synthetic agent is incapable of explaining to its teammates how it arrived at a particular conclusion, it will not

engender (and may erode) trust. Furthermore, since it may likely be infeasible to develop a perfectly correct AI system, the ability to explain itself will allow its human teammate to identify logical or syntactical errors.

Given that it is likely that AI will play an increasingly important role in the future of cybersecurity, it is imperative that we develop better constructs for human-machine teaming. These should be focused on effective communications, human workforce development, and explainable AI. Though much research is needed in all three areas, we can't afford to take the risk of not getting this right. Our cybersecurity depends on it.

References

Abbas, H., Petraki, E., Kasmarik, K., Harvey, J.: Trusted autonomy and cognitive cyber symbiois: open challenges. Cogn. Comput. **8**(3), 1–24 (2015)

Aiken, M.: The Cyber Effect: A Pioneering Cyberpsychologist Explains How Human Behavior Changes Online. Spiegel & Grau, New York (2017)

Asgharpour, F., Liu, D., Camp, L.J.: Mental models of computer security risks. In: Dietrich, S., Dhamija, R. (eds.) International Conference on Financial Cryptography and Data Security. Lecture Notes in Computer Science, vol. 47886, pp. 367–377. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77366-5_34

Baumann, M.R., Bonner, B.L.: Member awareness of expertise, information sharing, information weighting, and group decision making. Small Group Res. **44**, 532–562 (2013)

Baker, M.: Striving for Effective Cyber Workforce Development. Software Engineering Institute, Carnegie Mellon University, Pittsburgh (2016)

Ben-Asher, N., Gonzalez, C.: Effects of cyber security knowledge on attack detection. Comput. Hum. Behav. **48**, 51–61 (2015)

Bearman, C.R., Paletz, S.B.F., Orasanu, J., Thomas, M.J.W.: The breakdown of coordinated decision making in distributed systems. Hum. Factors **52**, 173–188 (2010)

Brooks, T.N.: Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems (2017). arXiv preprint arXiv:1702.06162

Buchanan, A., Goodall, L., Walczak, D'Amico, P.: Mission impact of cyber events: Scenarios and ontology to express the relationship between cyber assets (2009). http://www.dtic.mil/cgibin/GetTRDoc?AD=ADA517410

Buchler, N., Rajivan, P., Marusich, L., Lightner, L., Gonzalez, C.: Sociometrics and observational assessment of teaming and leadership in a cyber security defense competition. J. Comput. Secur. **73**, 114–136 (2017)

Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE, May 2017

Caulkins, B.D., Badillo-Urquiola, K., Bockelman, P., Leis, R.: Cyber workforce development using a behavioral cybersecurity paradigm. In: Connelly, C., Brantly, A., Thomson, R., Vanatta, N., Maxwell, P., Thomson, D. (eds.) International Conference for Cyber Conflict US. Army Cyber Institute, West Point (2016)

Champion, M.A., Rajivan, P., Cooke, N.J., Jariwala, S.: Team-based cyber defense analysis. In: 2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), pp. 218–221 (2012)

Champion, M., Jariwala, S., Ward, P., Cooke, N.J.: Using cognitive task analysis to investigate the contribution of informational education to developing cyber security expertise. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **58**(1), 310–314 (2014)

Chen, J.Y., Barnes, M.J.: Human-agent teaming for multirobot control: a review of human factors issues. IEEE Trans. Hum. Mach. Inter. **44**(1), 13–29 (2014)

Cohen, W.W.: Learning rules that classify e-mail. In: AAAI Spring Symposium on Machine Learning in Information Access, vol. 18, p. 25, March 1996

Cook, M.: Cyber Acquisition Professionals Need Expertise (But They Don't Necessarily Need to be Experts). Defense Acquisition University, Fort Belvoir (2014)

Cooke, N.J., Gorman, J.C., Myers, C.W., Duran, J.L.: Interactive team cognition. Cogn. Sci. **37**, 255–285 (2013). https://doi.org/10.1111/cogs.12009

Cooke, N.J., Shope, S.M.: Designing a synthetic task environment. In: Scaled Worlds: Development, Validation, and Application, pp. 263–278 (2004)

Dvorsky, G.: Hackers Have Already Started to Weaponize Artificial Intelligence 11 September 2017. https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425. Accessed 22 Feb 2018

Emmanuel, Z.: Security experts air concerns over hackers using AI and machine learning for phishing attacks, 5 October 2017. http://www.computerweekly.com/news/450427653/Security-experts-air-concerns-over-hackers-using-AI-and-machine-learning-for-phishing-atttacks. Accessed 22 Feb 2018

Franke, U., Brynielsson, J.: Cyber situational awareness: a systematic review of the literature. Comput. Secur. **46**, 18–31 (2014)

Gonzalez, C., Ben-Asher, N., Oltramari, A., Lebiere, C.: Cognitive models of cyber situation awareness and decision making. In: Wang, C., Kott, A., Erbacher, R. (eds.) Cyber Defense and Situation Awareness. Springer (in press)

Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cogn. Technol. Work **18**(1), 121–143 (2016)

Hackman, J.R., Katz, N.: Group Behavior and Performance, pp. 1208–1251. Wiley, New York (2010)

Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 43–58. ACM, October 2011

Internet Live Stats - Internet Usage & Social Media Statistics. Accessed 19 Feb 2018. http://www.internetlivestats.com/

Jajodia, S., Liu, P., Swarup, V., Wang, C.: Cyber Situational Awareness. Springer Publishing, New York (2010)

Jariwala, S., Champion, M., Rajivan, P., Cooke, N.J.: Influence of team communication and coordination on the performance of teams at the iCTF competition. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56(1), pp. 458–462. SAGE Publications, September 2012

Joint Publication 3–12: Cyberspace Operations, Washington, DC: Joint Chiefs of Staff, U.S. Department of Defense (2013)

Knowles, W., Prince, D., Hutchison, D., Disso, J.F., Jones, K.: A survey of cyber security management in industrial control systems. Int. J. Crit. Infrastruct. Protections **9**, 52–80 (2015)

Kott, A., Alberts, D., Zalman, A., Shakarian, P., Maymi, F., Wang, C., Qu, G.: Visualizing the tactical ground battlefield in the year 2050: Workshop report (No. ARL-SR-0327). Army Research Lab Adelphi Maryland (2015)

Metz, R.: Why Microsofts teen chatbot, Tay, said lots of awful things online, 24 March 2016. https://www.technologyreview.com/s/601111/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/. Accessed 23 Feb 2018

Newhouse, B., Keith, S.S., Witte, G.: NICE Cybersecurity Workforce Framework. National Institute of Standards and Technology, Gathersburg (2016)

Onwubiko, C., Owens, T.J.: Situational Awareness in Computer Network Defense: Principles, Methods and Applications. Information Science Reference, Hershey (2011)
Paletz, S.B.F.: Multidisciplinary teamwork and big data. In: Human-Centered Big Data Workshop, At Raleigh, NC (2014). https://doi.org/10.1145/2609876.2609884
Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning (2016). arXiv preprint arXiv:1611.03814
Pellerin, C.: Work: Human-Machine Teaming Represents Defense Technology Future (2015). https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/. Accessed 1 Feb 2018
Pfleeger, S.L., Caputo, D.D.: Leveraging behavioral science to mitigate cyber security risk. Comput. Secur. **31**(4), 597–611 (2012)
Proctor, R.W., Chen, J.: The role of human factors/ergonomics in the science of security decision making and action selection in cyberspace. Hum. Factors J. Hum. Factors Ergon. Soc. (2015). https://doi.org/10.1177/0018720815585906
Rajivan, P., Champion, M., Cooke, Nancy J., Jariwala, S., Dube, G., Buchanan, V.: Effects of teamwork versus group work on signal detection in cyber defense teams. In: Schmorrow, Dylan D., Fidopiastis, Cali M. (eds.) AC 2013. LNCS (LNAI), vol. 8027, pp. 172–180. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39454-6_18
Rajivan, P.: Information Pooling Bias in Collaborative Cyber Forensics (Doctoral dissertation, Arizona State University) (2014)
Rajivan, P., Cooke, N.: Information Pooling Bias in Collaborative Security Incident Analysis. Human Factors (in press)
Rajivan, P., Moriano, P., Kelley, T., Camp, J.: Factors in an end user security expertise instrument. Inf. Comput. Secur. **25**(2), 190–205 (2017)
Salas, E., Dickinson, T.L., Converse, S.A., Tannenbaum, S.I.: Toward an understanding of team performance and training. In: Teams their Training and Performance, pp. 3–29 (1992)
Seymour, J., Tully, P.: Weaponizing data science for social engineering: Automated E2E Spear Phishing on Twitter. Black Hat USA, 37 (2016)
Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE, May 2017
Shortliffe, E.: Computer-Based Medical Consultations: MYCIN. Elsevier, New York (2012)
Simsarian Webber, S.: Leadership and trust facilitating cross-functional team success. J. Manage. Dev. **21**(3), 201–214 (2002)
Spiro, R.J.: Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains. Technical Report No. 441 (1988). http://eric.ed.gov/?id=ED302821. Accessed 5 Oct 2017
Srinidhi, B., Yan, J., Tayi, G.K.: Allocation of resources to cyber-security: the effect of misalignment of interest between managers and investors. Decis. Support Syst. **75**(1), 49–62 (2015). http://doi.org/10.1016/j.dss.2015.04.011
Stasser, G., Titus, W.: Pooling of unshared information in group decision making: biased information sampling during discussion. J. Pers. Soc. Psychol. **48**(6), 1467 (1985)
Trexler, E.: Why Human-Machine teaming is the future of cybersecurity (2017). https://federalnewsradio.com/commentary/2017/11/why-human-machine-teaming-is-the-future-of-cybersecurity/. Accessed 1 Feb 2018
Tung, L.: IoT devices will outnumber the world's population this year for the first time 13 February 2017. http://www.zdnet.com/article/iot-devices-will-outnumber-the-worlds-population-this-year-for-the-first-time/. Accessed 19 Feb 2018

Veksler, B.Z.: Visual search strategies and the layout of the display. In: Salvucci, D.D., Gunzelmann, G. (eds.) Proceedings of the 10th International Conference on Cognitive Modeling, pp. 323–324. Drexel University, Philadelphia (2010)

Vieane, A., Funke, G., Mancuso, V., Greenlee, E., Dye, G., Borghetti, B., Brown, R.: Coordinated displays to assist cyber defenders. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60(1), pp. 344–348. SAGE Publications, September 2016

Whitley, J., D'Amico, K.: The real work of computer network defense analysts. In: Goodall, J.R., Conti, G., Ma, K.L. (eds.) Workshop on Visualization for Computer Security, pp. 19–37 (2008)