

The Ethical Use of Machine Learning in Cybersecurity

By Frank Gearhart – ISSA Senior Member – Colorado Springs Chapter



The ongoing arms race between cyber defense and attackers is entering a new phase as the use of machine learning-based tools increases on both sides. In this article the author explores some of the ethical issues involved in using machine learning-based tools and what this means for us as cybersecurity researchers and practitioners.

*“I just think we're living in a time of massive, amazing change, like the Industrial Revolution on acid.”
~ Kelly Lynch, actor*

Abstract

The ongoing arms race between cyber defense and attackers is entering a new phase as the use of machine learning-based tools increases on both sides. One challenge of using these tools for cyber defense is that we often do not understand how they arrive at their conclusions. Current research into machine learning indicates these tools can have hidden biases. As cybersecurity professionals we are obligated to act ethically. Using machine learning-based tools that we don't fully understand challenges our ability to act responsibly. We should understand the ethical risks involved and the work being done to make machine learning more understandable and transparent. In this article I will look at some of the ethical issues involved in using machine learning-based tools and what this means for us as cybersecurity researchers and practitioners.

Much of humanity is now digitally connected. Facebook is used by over one-third of the human population [18]. Farmers in remote areas can check

crop prices, send and receive payments, and monitor their solar-powered irrigation pumps using simple text messaging [13]. Banking and finance, medicine, government business at all levels, and even military operations either take place or are dependent on digital information and global data networks [7]. Cyber defenses and adversaries continue to be engaged in an arms race—there are indications that adversaries are beginning to use artificial intelligence and machine learning (AI/ML), and cyber defenses will have to keep up [8]. Machine learning-based cyber defense tools bring opportunities and challenges we have not faced before—tools that may be effective, but that we may not fully understand. How we use (and trust) these tools will have a major impact on how successful we are with them.

Use of machine learning in cybersecurity

ML-based cybersecurity is a growing market and a hot area of research [10][17]. Since IBM Watson's highly publicized win over the human champions on the show *Jeopardy* in 2011, AI has exploded in the public's consciousness. This year (2020) CB Insights lists one hundred startups in thirteen countries

that are working to deliver AI-based capabilities in fifteen industries: health care, finance and insurance, energy, agriculture, and supply chain management to name a few [2]. In 2019, CB Insights listed eight cybersecurity startups using AI, eleven in 2018, and six in 2017. This does not include industry leaders such as IBM, Palo Alto, and FireEye that use AI/ML in many of their products.

Unintended biases in machine learning

The issue of unintended biases in machine learning tools gained visibility after two public incidents: the mislabeling of a photo of two African Americans by Google's ML-based photo tagging tool and the removal, after less than sixteen hours online, of Microsoft's automated Twitter chatbot, Tay, due to its unintended derogatory and offensive postings [11]. Machine learning algorithms are particularly vulnerable to unintended bias when they are learning, often because the datasets used in ML training are unintentionally biased [9]. In domains such as predictive policing, both the victim's and the suspect's age, race, and gender are part of the predictive attributes used in the training datasets [3]. There is substantial evidence that human social bias can enter these training datasets, as there is normally a great deal of manual work required to turn raw data into useful datasets.

Secondly, most machine learning algorithms perform better when they have large amounts of data to work with. The advent of big data gives researchers and practitioners the opportunity to develop and use new models. However, these very large datasets can be unintentionally biased. For example, the word2vec natural language algorithm developed by Google and used by many natural language processing systems often demonstrates gender stereotyping, since the training dataset used by word2vec is the huge corpus of human-written documents Google has collected [9].

This sort of unintended bias can be reduced, but it requires deliberate work by a diverse group of subject experts. Cyber defense ML-based tools that only look at technical data such as network traffic patterns may be immune to these types of bias. However, more sophisticated cyber defense tools that combine user behavior analysis with automated access to individual HR information such as promotion denials or lower than average performance reviews could make insider risk determinations that are unintentionally biased [12].

Ethics in machine learning

Machine learning algorithms such as deep learning, convolutional neural networks, and recurrent neural networks are capable of amazing results [7] but come with new and unprecedented challenges. Unlike previous human tools, machine learning is often a black box where we do not understand how those answers were arrived at. If the tool is a shopping search engine, this may be no more than amusing. When the tool may have significant impact on a person's job, career, or life, such ignorance should not be accepted.

In a March 2020 interview, Murat Sönmez, the director of the World Economic Forum, discussed an approach to preventing social harm from AI using what he called an ethics switch: "We have a concept of an ethic switch where countries define their ethics rules. We download these rules to the smart devices. When they're asked to do something that's harmful, the switch says no" [15].

Developing an ethical switch is obviously easier said than done. While there is work being done to design ethical behavior into computer systems, there are many challenges. Some of them are technical, while many are social. A culture's ethics is an expression of that culture's values—what they deem acceptable and unacceptable behavior. For example, a culture



Members Join ISSA to:

- Earn CPEs through Conferences and Education
- Network with Industry Leaders
- Advance their Careers
- Attend Chapter Events to Meet Local Colleagues
- Become part of Special Interest Groups (SIGs) that focus on particular topics

Join Today: www.issa.org/join

Regular Membership \$95*
(+Chapter Dues: \$0-\$35*)

CISO Executive Membership \$995
(Includes Quarterly Forums)

**US Dollars/Year*

that values individual freedoms over group expectations may have a different set of values and priorities than a culture that values group cohesion over individual desires. It is this common ethical framework that defines acceptable and unacceptable behavior for a culture.

The importance and influence of cultural standards on individual ethics and behavior (when applied to machines) was presented in a March 2020 article in *Communications of the ACM*. In this article Awad, et al. discussed the use of crowd-sourced machine morality through a process they called “society in the loop,” noting that social scientists and computational social scientists have a role to play in articulating social ethics into artificially intelligent systems [1]. They recognized several limitations of crowd sourcing in this context, pointing out that language is often imprecise, and the same word can have subtle but important differences in meaning even in a generally homogeneous culture. Further, they noted that cultures change over time, so even a successful attempt at programmatically codifying social ethics will not be static.

It is unlikely that fully autonomous cyber defensive systems will be available soon. However, ML-based cyber defensive tools are in use now. What these systems lack is a way to explain how they arrived at the answers they present to their human users. Explainable artificial intelligence (XAI) is a way for ML-based systems to do this and is another strong area of research [17]. Getting an understandable answer becomes more critical as ML-based tools are used in systems that have greater impact on people—finance, medicine, law, and security.

Machine learning and the law

There are legal impacts to using AI/ML-based tools. The European Union’s General Data Protection Regulation (GDPR), which took effect in May of 2018, includes requirements regarding the use of ML-based tools and the right of a person to receive an explanation regarding decisions made that affect him or her [6]. Article 22 of the GDPR notes that (with certain exceptions) “The data subject shall have the right not to be subject to a decision based solely on automated processing...” [16]. Article 15 also requires that the data subject (the person) receives “meaningful information about the logic involved” regarding any decisions impacting them [16]. If and how these would apply to AI/ML cyber defense systems has not been determined, but the concepts are worth considering.

While GDPR is likely the most well-known law that addresses the impact of ML-based tools on people, it is not the only one. Miller reports several ways in which ML-based tools are having consequential impacts on people’s lives [11]. In the United States, machine learning-based tools are finding use in some courts to assist judges in deciding which defendants should be held in jail before trial, which defendants should be released on bail, and how much bail should be required. Often these tools are being used by well-meaning organizations and people who do not understand the tool’s sometimes subtle and critical biases and limitations [4]. The resulting damage may not be understood until well after the fact.

The Partnership on AI’s report on the use of AI in US courts of law includes three requirements that are applicable to ML-based cyber defense tools [14]:


1. Predictions and how they are made must be easily interpretable
2. Tools should produce confidence estimates for their predictions
3. Users must attend trainings on the nature and limitations of the tools

These requirements support the concept of transparency—a characteristic necessary for ML-based tools to be accepted.

Ethical cybersecurity

Cyber attacks occur at the speed of computers—far faster than humans can detect and respond to. Sophisticated attacks can find and exploit vulnerabilities that are either unknown to vendors or for which there is no effective protection, and these challenges will not subside in the future. Effective cyber defense will have to become faster and smarter than the attacks, and that will require smarter, more automated defensive tools that can learn and adapt at wire speed. This will eventually require a high level of trust in the tools—eventually to the point where we will trust the tools to actively respond to attacks.

As machine ethics are researched, argued over, and eventually implemented, and as explainable AI makes its way into commercially available tools, cybersecurity practitioners


ISSA International Series:



Breaking Down Zero Trust: What Does it Actually Mean?

120-minute Live Event: Tuesday, April 28, 2020
9 a.m. US-Pacific/ 12 p.m. US-Eastern/ 5 p.m. London

Igenerously supported by



For more information on this or other webinars:
[ISSA.org => Events => Web Conferences](https://www.issa.org/Events/Web-Conferences)

must learn to use ML-based cyber defensive tools in a manner consistent with our professional and social obligations. I propose three general guidelines that may be of benefit to us as we move into a world of human-machine cybersecurity:

1. Where you can, campaign for machine learning-based cyber defense tools that utilize explainable AI (XAI).
2. Move carefully when taking action based on the output of an AI/ML-based tool. Consider if you would take the same or similar action if a human came to you with the same recommendation or observation.
3. Understand that AI/ML-based tools—particularly those using deep learning and big data—are heavily dependent on probability, may have subtle biases, and do not provide certainty. Act appropriately.

As cybersecurity practitioners, we have an obligation to act ethically. We often have access to privileged and sensitive information as well as the trust of leadership. We help protect the crown jewels of organizations: intellectual property, financial information, personal information, etc. As the amount and detail of the personal data collected by governments and corporations grows every day, while the attacks against that data become more sophisticated, our defenses must become faster, better, and more capable. Our professional obligations do not change even as our tools do. The following two scenarios may help to present the ideas discussed earlier in a more concrete way:

Scenario 1: ML-based insider threat tool

Let's assume that your cybersecurity department uses a machine learning tool to detect malicious network activity. The tool looks at user activity, network traffic (including attempts to access various databases), and other metadata about the network. Based on the activity it sees, the tool sends an alert to the security team that the behavior of one of the network administrators is showing significant indicators of insider threat activity. The tool lists all the traffic it has flagged as suspicious, but there is no explanation of why it considers this activity suspicious.

The tool has previously detected several previously unknown attacks that were later verified as coming from technically sophisticated attackers associated with foreign governments.

The cybersecurity investigator trusts the tool (it comes from a well-respected vendor and was expensive), so she begins an investigation. As part of the investigative process, the network administrator's supervisor is notified, and the supervisor decides to temporarily reassign the network administrator until the investigation is complete (without letting the administrator know the real reason for being reassigned).

Question 1: Should the cybersecurity team begin an investigation based solely on the report from the ML-based tool?

Question 2: Assuming the company decides to terminate the network administrator, should it do so based on the report from the ML-based tool (including the details of the

behavior flagged as indicative of insider threat activity), or should the decision to terminate or not be based on independent evidence not collected by the tool?

Question 3: Would an investigation be less necessary if the tool provided human-understandable reasons for its determination that the administrator is likely an insider threat?

Question 4: What if the tool were sophisticated enough to have access to the individual's HR records—when hired, work history, performance evaluations, etc.? Would that change the confidence the company has in the tool's assessment of threat from the administrator?

Scenario 2: Automated offensive cybersecurity operations

Let's move forward several years and consider a fully automated cybersecurity tool designed to protect against external threats—perhaps a cybersecurity system based on the winner of a future DARPA Cyber Grand Challenge [5]. This automated system has detected a moderate level cyber attack and attributes it to a specific group located in a foreign country with a high level of assurance. Using what the system knows regarding the nation-state and the specific unit that sent the attack, it autonomously generates a cyber-based response designed to disable (but not destroy) the power system in that area for a limited time. A hospital near the location is also affected by the power outage. While no one dies or suffers injury thanks to the hospital's backup generators, some operations and treatments are postponed, so some patients experience increased pain and possibly cost due to the delays.

Question 1: Is this cyber response legal with respect to international law?

Question 2: Would this cyber response be legal (or at least more legally defensible) if the origin of the attack was in the same country as the target?

Question 3: If the damage from the original attack were greater and had resulted in physical injury but no loss of life, what level of response would be appropriate?

Question 4: Assume the system that was attacked was part of national defense and that no individuals were injured. Would a fully autonomous cyber response be an act of war? If so, who would be considered responsible for it?

Conclusion

Developing a programmatic ethical framework for machine learning and artificial intelligence is needed now. We cannot wait until human-level AI arrives—tools that use machine learning are being used now, and often in ways we are not aware of. In current machine learning techniques, large datasets are used for training. These datasets can have inherent biases, and the manual normalization and preprocessing needed before the training data can be used can also add unintentional biases. Removing these, or at least minimizing them and taking their impacts into account when using

ML-based tools, is necessary before ML-based tools can be trusted.

At some point fully automated cyber defense systems will operate next to human experts—there are indications that adversaries are beginning to use AI, and cyber defenses will have to keep up [9]. The trust and responsibility inherent in our positions as defenders of personal, corporate, and government data does not change as AI/ML-based tools become part of our cyber defense toolkit, and that will not change if (perhaps when) fully autonomous AI/ML-based cyber defenses are implemented.

References

1. Awad, Edmond, Sohan Dsouza, Jean-Francois Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. "Crowdsourcing Moral Machines." *Communications of the ACM (ACM)* 63 (3): 48-55 – <https://dl.acm.org/doi/10.1145/3339904>.
2. CB Insights. 2020. "AI 100: The Artificial Intelligence Startups Redefining Industries," – <https://www.cbinsights.com/research/artificial-intelligence-top-startups/>.
3. Elluri, L., Mandalapu, V. and Roy, N., 2019. "Developing Machine Learning Based Predictive Models for Smart Policing," In 2019 IEEE International Conference on Smart Computing, pp.198-204 – <https://ieeexplore.ieee.org/document/8784006>.
4. Eubanks, Virginia. *Automating Inequality*. New York, NY: St. Martin's Press, 2017.
5. Frazee, Dustin. 2020. "Cyber Grand Challenge, DARPA - <https://www.darpa.mil/program/cyber-grand-challenge>.
6. ICO. 2019. "Project ExplAIIn - Interim Report," Information Commissioner's Office – <https://ico.org.uk/media/2615039/project-explain-20190603.pdf>.
7. Jackson, Phillip C. Jr. 2019. *Introduction to Artificial Intelligence*. 3rd. Mineola, NY: Dover publications, Inc.
8. Kaloudi, Nektaria, and Jingyue Li. 2020. "The AI-Based Cyber Threat Landscape: A Survey." *ACM Computing Surveys (ACM)* 53 (1): 1-34.
9. Kearns, Michael, and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. 1st. New York, NY: Oxford University Press.
10. Miller, Seumas. 2019. "Machine Learning, Ethics and Law," *Australasian Journal of Information Systems* 23: 1-13 – <https://journal.acs.org.au/index.php/ajis/article/view/1893/851>.
11. Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. New York, NY: Farrar, Strauss, and Giroux, 2019.
12. O'Neil, Cathy. *Weapons of Math Destruction*. New York, NY: Broadway Books, 2017.
13. Opiyo, Nicholas. 2019. "How Mobile Money Platforms and Other Innovative Technologies Have Stimulated Energy Revolution in Rural Sub-Saharan Africa." 36th European Photovoltaic Solar Energy Conference. Marseille. 2013-2018. doi:10.4229/EUPVSEC20192019-7DV.2.29.
14. Partnership on AI. 2019. "Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System," Partnership on IA – <https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/>.
15. Patterson, Dan. 2020. "Forecasting the Future of Artificial General Intelligence," *Tech Republic* (Mar 6, 2020) – <https://www.techrepublic.com/article/forecasting-the-future-of-artificial-general-intelligence/>.
16. Proton Technologies AG. 2020. "GDPR: Chapter 3: Rights of the Data Subject," Intersoft Consulting – <https://gdpr-info.eu/chapter-3/>.
17. Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines," *Fordham Law Review* 87 (3): 1085-1140 – <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5569&context=flr>.
18. Zephoris. 2020. "The Top 20 Valuable Facebook Statistics—Updated January 2020," Zephoris – <https://zephoris.com/top-15-valuable-facebook-statistics/>.

Defend Forward

Continued from [page 5](#)

presence of private industry in all areas of the Internet (including the IoT) necessitate (a) greater public/private interaction, (b) specific attention to developing a defend forward posture and appropriate active cyber response policies, (c) increased information sharing, and (d) an institution run by the private sector (but with continual government oversight) for sharing cyber threat indicators and other relevant threat information at network speed. With these components and activities in place, both the government and private industry will benefit, and the resulting increase in resiliency should reduce (possibly significantly reduce) the effects of malicious cyber actors and nation-state adversaries.

About the Author

Randy V. Sabett, J.D., CISSP, is an attorney with Cooley (www.cooley.com/rsabett), a member of the advisory board of the Georgetown Cybersecurity Law Institute and the RSA Selection Committee, a member of the Cyber Leadership Council in the US Chamber of Commerce, and is the former Senior VP of ISSA NOVA. He has completed FBI Citizen Academy training, was a member of the Commission on Cybersecurity for the 44th Presidency, and was named ISSA Professional of the Year. He can be reached at rsabett@cooley.com.

About the Author

Frank Gearhart is a lead cybersecurity engineer (contractor) at DoD Missile Defense Agency, Colorado Springs, CO. He is a past president of Colorado Springs Chapter and ISSA Volunteer of the Year – 2017. He has ten years of cybersecurity experience, served in the US Coast Guard, is a PhD candidate at North Central University, and may be reached at frank.gearhart@outlook.com.

