Journal of Cybersecurity Education, Research and Practice

Volume 2024 | Number 1 Article 10

11-20-2023

Integrity, Confidentiality, and Equity: Using Inquiry-Based Labs to help students understand AI and Cybersecurity

Richard C. Alexander
Texas Christian University, curby.alexander@tcu.edu

Liran Ma
Texas Christian University, l.ma@tcu.edu

Ze-Li Dou
Texas Christian University, z.dou@tcu.edu

Zhipeng Cai
Georgia State University, zcai@gsu.edu

Yan Huang
yhuang24@kennesaw.edu

Recommended Citation
Alexander, Richard C.; Ma, Liran; Dou, Ze-Li; Cai, Zhipeng; and Huang, Yan (2023) "Integrity, Confidentiality, and Equity: Using Inquiry-Based Labs to help students understand AI and Cybersecurity," *Journal of Cybersecurity Education, Research and Practice*: Vol. 2024: No. 1, Article 10.
DOI: 10.32727/8.2023.34
Available at: https://digitalcommons.kennesaw.edu/jcerp/vol2024/iss1/10

Integrity, Confidentiality, and Equity: Using Inquiry-Based Labs to help students understand AI and Cybersecurity

Abstract

Recent advances in Artificial Intelligence (AI) have brought society closer to the long-held dream of creating machines to help with both common and complex tasks and functions. From recommending movies to detecting disease in its earliest stages, AI has become an aspect of daily life many people accept without scrutiny. Despite its functionality and promise, AI has inherent security risks that users should understand and programmers must be trained to address. The ICE (integrity, confidentiality, and equity) cybersecurity labs developed by a team of cybersecurity researchers addresses these vulnerabilities to AI models through a series of hands-on, inquiry-based labs. Through experimenting with and manipulating data models, students can experience firsthand how adversarial samples and bias can degrade the integrity, confidentiality, and equity of deep learning neural networks, as well as implement security measures to mitigate these vulnerabilities. This article addresses the pedagogical approach underpinning the ICE labs, and discusses both sample activities and technological considerations for teachers who want to implement these labs with their students.

Keywords

Artificial Intelligence (AI), Security, Cybersecurity Labs, Inquiry-Based Labs

Cover Page Footnote

This study and the development of Eureka Labs was funded by the US National Science Foundation grants 2244220, 2244219, 2315596, 2244221, and 2315595.

Integrity, Confidentiality, and Equity: Using Inquiry-Based Labs to help students understand AI and Cybersecurity

1st Curby Alexander
Department of Counseling, Societal Change, and Inquiry
Texas Christian University
Fort Worth, TX USA
ORCID: 0000-0002-3888-6930

2nd Liran Ma
Department of Computer Science
Texas Christian University
Fort Worth, TX USA
ORCID: 0000-0002-1003-1770

3rd Ze-Li Dou
Department of Mathematics
Texas Christian University
Fort Worth, TX USA
ORCID: 0009-0008-9241-0619

4th Zhipeng Cai
Department of Computer Science
Georgia State University
Atlanta, GA USA
ORCID: 0000-0001-6017-975X

5th Yan Huang
Department of Software Engineering & Game Development
Kennesaw State University
Kennesaw, GA USA
ORCID: 0000-0001-7775-4597

***Abstract*—Recent advances in Artificial Intelligence (AI) have brought society closer to the long-held dream of creating machines to help with both common and complex tasks and functions. From recommending movies to detecting disease in its earliest stages, AI has become an aspect of daily life many people accept without scrutiny. Despite its functionality and promise, AI has inherent security risks that users should understand and programmers must be trained to address. The ICE (integrity, confidentiality, and equity) cybersecurity labs developed by a team of cybersecurity researchers addresses these vulnerabilities to AI models through a series of hands-on, inquiry-based labs. Through experimenting with and manipulating data models, students can experience firsthand how adversarial samples and bias can degrade the integrity, confidentiality, and equity of deep learning neural networks, as well as implement security measures to mitigate these vulnerabilities. This article addresses the pedagogical approach underpinning the ICE labs, and discusses both sample activities and technological considerations for teachers who want to implement these labs with their students..**

***Index Terms*—Artificial Intelligence (AI), Security, Cybersecurity Labs, Inquiry-Based Learning**

I. Introduction

Artificial intelligence (AI) can be found in just about every aspect of daily life. Netflix suggests movies and shows based on past viewing choices. Smartphones recognize a person's face or fingerprint and simplify the login process. Chess and Go models consistently beat human world champions. Maps calculate the fastest route without taking a toll road, and cars can drive themselves in city traffic. Google seems to know what you are searching for before you finish typing, and ChatGPT can answer your queries in a structured essay that follows the basic conventions of writing. In each of these examples, AI scientists have developed models that enable computers to perform tasks previously only humans were capable of. Whether Spotify is building a playlist based on listening history or molecular imaging models are detecting cancer cells in their earliest stages [1], the underlying truth is the same: We have given over common, and sometimes not so common, aspects of our cognition and decision-making to machines.

The rapid development of AI capabilities brings about many possibilities for society. AI can efficiently perform many of the same functions as humans without fear of errors due to fatigue, input errors, and distraction. If Netflix can predict a movie a person may enjoy, or if ChatGPT can provide a quick summary of Romeo and Juliet, why not let it perform those functions? In many instances, AI frees us of mundane tasks and allows us to focus our attention on other tasks, and the negative consequences are trivial even if AI fails. We may spend 90 minutes watching a movie we did not enjoy or we missed an important plot element Shakespeare was hoping we would notice. As AI is applied to more complex and serious processes and systems, however, the aftermath may yield more devastating consequences.

Deep learning AI models are increasingly utilized in privacy-sensitive and safety-critical applications ranging from

This study and the development of Eureka Labs was funded by the US National Science Foundation. Curby Alexander, Ze-Li Dou and Liran Ma are supported in part by the US National Science Foundation under grant 2244220. Zhipeng Cai is supported in part by the US National Science Foundation under grant 2244219 and 2315596. Yan Huang is supported in part by the US National Science Foundation under grant 2244221 and 2315595. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

biometric user authentication to autonomous driving. This trend is bound to continue: many open-source frameworks and tools from online code repositories (e.g., GitHub) are embedded with deep learning modules. It is well known to experts in the field that many deep learning models contain weaknesses that could be exploited by attackers, which may pose significant risks to user privacy and safety. Because of their complexity, however, such vulnerabilities often appear hidden. Prospective data engineers must develop security awareness and become equipped with knowledge and strategies for designing trustworthy deep learning-based applications.

Finding effective strategies for preparing students on the secure use of deep learning models is critical to supplying the workforce with high-quality security-conscious data engineers, who, by necessity, are the vanguard for ensuring the public's trust in technological innovations. There exist clear challenges to providing such effective training. The deep learning field is highly technical, and the underlying concepts and principles can be abstract and difficult to master. A learner can be intimidated by the high theoretical threshold when the subject is taught in the traditional lecture-heavy and mathematics-laden manner. The instructor faces a pedagogical barrier of their own: the deep learning field is rapidly evolving, and instructors must continually revise their course topics, materials, and examples to maintain pace with innovations in the industry. Finally, time and energy expended on the technical requirements to deploy labs may heighten the learners' frustration, and become a new, though inadvertent, stumbling block against efficacy.

To address the aforementioned challenges, a team of researchers across three universities is developing a series of easy-to-implement experiential learning activities, through which prospective data engineers increase their awareness of potential vulnerabilities in deep learning models and develop skills in building secure applications on their own. The labs, the most recent addition to the Eureka Labs curriculum (www.eurekalabs.net), are organized around three important secure uses of deep learning models: Integrity, Confidentiality and Equity (ICE). This article will discuss the conceptual framework on which the ICE labs were developed, an overview of the topics addressed through these labs, and implications for implementation of AI labs for computer science instructors and other practitioners.

II. Pedagogical Framework

The ICE series of Eureka Labs are designed based on the cycle of experiential learning [2], which is built on the premise that first-hand experience with data, code, and AI models can facilitate better learning (see Figure 1). Eureka Labs are designed to provide context and relevance to hypothetical concepts by situating them within authentic scenarios to which learners can relate rather than placing primacy on abstract theoretical principles detached from realistic situations. In essence, the labs leverage what students can see or experience in order to explain what they cannot see or experience first-hand. Through carefully thought-out questions and prompts that draw learners' attention to key concepts and principles, they concurrently make observations and connect what they are experiencing to foundational knowledge and skills within each lab. Learners cultivate their ability to transfer their knowledge and skills to subsequent contexts when they are able to disentangle the mathematical and theoretical principles from the scenario in which they are embedded [3].

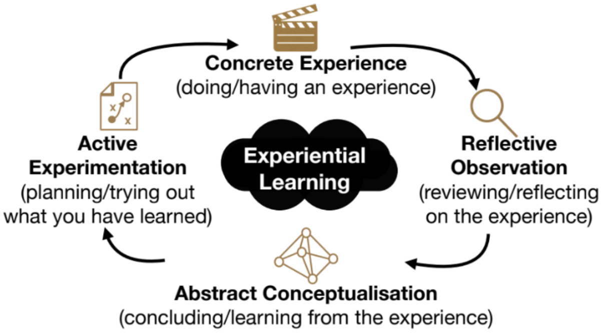
Concrete Experience
(doing/having an experience)
Reflective Observation
(reviewing/reflecting on the experience)
Abstract Conceptualisation
(concluding/learning from the experience)
Active Experimentation
(planning/trying out what you have learned)
Experiential Learning

Fig. 1. Kolb's Cycle of Experiential Learning

Hands-on, exploratory learning is of particular importance when it comes to students acquiring and developing their understanding and skills around AI and deep learning models. The security of a deep learning model is necessarily measured by its invulnerability against potential attacks. However, vulnerabilities remain hidden until an attack occurs. In teaching about security, therefore, it is important to point out not only the secure features in a model, but also to disclose the weaknesses that might be exposed without them.

This tension between teaching the architecture that makes deep neural networks possible and their potential weaknesses is a characteristic of cybersecurity education. According to Hamman and Hopkinson [4], cybersecurity education should include a blend of analytical thinking, where students understand the logical and mathematical concepts driving the deep neural network, and "adversarial thinking," where students place themselves in the mind of a hacker. Adversarial thinking relies on the other two aspects of Sternberg's [5] triarchic theory of cognitive processing. Analytic thinking comprises one prong of this theory, but cognitive processing also utilizes creative and practical thinking. Creative thinking is the ability to make unique connections and see the world in original ways, and practical thinking involves the ability to plan, strategize, and accomplish goals 4. Cybersecurity education lends itself to experiential labs because it blends the analytical thinking required to understand algorithms and deep learning models with the creative and practical thinking a hacker would employ in order to exploit and subvert its vulnerabilities.

The AI security labs described below begin with an examination of how certain features of a deep learning model might be exploited from an adversarial viewpoint. Once the potential vulnerabilities of a model are known, students will understand not only the reason behind the security features,

DOI: 10.32727/8.2023.34

but they will also comprehend how they can be tailored in response to possible exploitation. On a deeper level, students will also learn that security comes with a cost. Although the activities are dubbed as "attack" and "defense" for succinctness, cybersecurity should not be perceived as a black-and-white, one-dimensional world. The ICE labs allow students to explore the relationship between levels of security and their costs on a continuum of exchanges. New forms of attacks lead to new responses, which often make applications and services more cumbersome by adding layers of security [6]. Future programmers and data engineers must build and apply AI models with security and their inherent vulnerabilities in mind in order to balance complex security measures with a seamless, efficient user experience.

III. Design Principles

The ICE labs are hands-on learning experiences that take the students on an interactive journey where they observe facts, ask questions, gather information, investigate clues, research answers, and implement algorithms with appropriate hyper-parameters in executable programs. Realistic environments and intuitive interactions support the mental models necessary to acquire new knowledge. Students use hints and clues to deconstruct the results of security-related analyses, and select and implement appropriate analytical strategies to modify the variables and obtain different results based on adjusted factors.

To summarize, the ICE labs strive to make abstract concepts tangible, encourage learning in a non-lecture format, expose the students to methodologies in action, and infuse intrigue and authenticity into otherwise dry code. Throughout these immersive experiences, the ICE labs maintain a clear association between concrete activities and the learning goals. In addition, the labs can be implemented smoothly in a variety of settings and across multiple platforms without excessive preparatory work that detracts from the learning goals. These principles are elaborated below.

A. Discover through Experience

Active involvement (e.g., hands-on activities) and security relevance (e.g., real-world problems) hold the potential to inspire learners and sustain their interest in AI learning experiences. ICE labs are designed intentionally to engage students in meaningful, problem-solving activities that can be completed either individually or in groups through discussion and collaboration. The scenarios presented in these labs require students to formulate a hypothesis, carry out meaningful research, analyze data, derive conclusions, and translate solutions into implementations. The labs are carefully sequenced in order to allow students to build their foundational knowledge and skills before progressing on to more complex tasks.

B. Student Interest and Engagement

Capturing and retaining students' interest and attention lie at the heart of teaching. If students are solely motivated by external rewards such as grades or checking items off a list, they may stay moderately engaged, and in some cases external rewards can lead to anxiety and task avoidance [7]. In order to cultivate the deep, sustained attention necessary to complete challenging complex problems, learning activities must arouse interest through mystery, a driving question, or the cognitive dissonance that comes from unexpected outcomes. While some students may stay engaged due to their personal interest in a topic, many students do not have prior personal interest in AI and deep learning models, and they will need to be invited into the learning activities through intentional, thoughtful attention to the aforementioned situational factors. The ICE labs create anticipation and help students focus on the learning activities through the use of driving questions centered on realistic applications of AI.

C. Affordable and Scalable Implementation

Finally, the ICE labs are built on a versatile platform that can be installed and operated by users of varying levels of expertise and on multiple operating systems. In order to make the learning environment as efficient as possible, three conditions must be met. First, the technical and installation demands of the learning resources and platform must be accessible for instructors and learners with varying degrees of expertise. Second, the cost of purchasing necessary equipment or software needs to be reasonable. Finally, the required preparation time (such as installation and configuration) needs to be manageable within the time-frame of a lab setting.

Modifying and debugging the lab activities should not be labor-intensive. The ICE labs are hosted on a flexible and extensible container-based virtualization computing platform, which can save setup time and costs. Data, code and configurations are coupled in each set of activities, which makes all of the lab resources easy access. Each lab also includes detailed supporting documents for prospective users.

IV. ICE Series

The ICE series of labs is built around experiments and strategies pertaining to AI model integrity, model confidentiality, and model equity. In each of the labs, students have the opportunity to manipulate AI deep neural network models, and they are provided strategies to address each security risk. The three phases of these labs are discussed below.

A. Integrity

When an adversary is able to manipulate or inject data into a model's training dataset, it can change the way the underlying algorithm learns and trick the model into making incorrect predictions. On a practical level, this could range from incorrectly including images of cats in an image query for dogs, to autonomous vehicles incorrectly recognizing Stop signs or One Way signs on city streets. The Integrity series of labs is centered on three types of attacks: adversarial example, poisoning attacks, and backdoor attacks. Each of these learning activities allows students to experiment with a few prominent attacks on the integrity of various learning models and deploy corresponding countermeasures.

A recently discovered vulnerability of deep neural networks, termed as “adversarial example attack” [8] [9], can seriously degrade prediction accuracy by inserting purposefully crafted imperceptible modifications on inputs. In the Integrity activity sets, students experiment with generating various adversarial examples to launch attacks on neural networks with different structures, then they apply mainstream defense measures to counter such attacks.

A poisoning attack is an attempt to affect the overall accuracy of the learning model by inserting adversarial samples into the training dataset. The poisoning attack learning activities help students to become aware of techniques that can facilitate and defend data poisoning attacks on deep learning models. Students experiment with “poisoning” training models for deep neural networks and reduce perturbation through cropping, scaling, and/or compressing training data.

Deep neural networks have been proven vulnerable to backdoor attacks, where hidden features (i.e., patterns) are covertly trained to a learning model [10]. The hidden features can only be activated by certain specific inputs (called triggers), and so they can trick the model into producing unexpected behavior [11]. In the backdoor activities, students learn how to insert hidden features to a learning model and employ triggers to deceive the model into making an erroneous decision, such as misclassifying spam. Students will also learn how to identify the backdoor vulnerability and apply proper defense mechanisms, such as applying a generative adversarial network (GAN) to lower or remove perturbation at both the data level and model level.

B. Confidentiality

When building a deep learning model, representation (e.g., features and statistics) of training data will be inevitably captured inside the model’s hyperparameters, which introduces the risk of giving away sensitive information. Recently, the Italian data protection authority (DPA) issued an order to block ChatGPT because the AI model was unlawfully processing personal data [12]. An adversary can adopt various techniques (e.g., membership inference attacks) [13] [14], and model inversion attacks [15] [16], to exploit the vulnerability. Serious privacy breaches and violations can happen if these techniques are successful. In this learning activity series, students will experiment with a few prominent attacks on deep learning models and their countermeasures.

Activities of individual users such as their purchase orders, health records, and locations are commonly used as training data by many companies for their learning models. Although direct access to the training dataset is forbidden to outsiders, individual records are still vulnerable to membership inference attacks [13] [14] (i.e., the existence of such records in the model’s training dataset). Such attacks can result in a privacy breach because they reverse an anonymization and obfuscation on the dataset. In the Confidentiality lab activity series, learners become aware of two possible sources of leakages from a training model: i) The background knowledge of an attacker with querying abilities; ii) Overfitting of the training dataset.

The possibility of learning sensitive information from linear classifiers by abusing the adversarial access to a classifier is first shown in [16]. Its subsequent work 14 further demonstrates the capability of reconstructing a human face image of the training dataset via a unique identifier. This attack is termed as the model inversion attack, which threatens privacy-sensitive applications built upon learning models that capture statistical facts between input features and output labels.

C. Equity

Deep learning algorithms can be biased – it has been shown that they can discriminate, reinforce prejudices, and polarize opinions. The data used to train a deep learning algorithm are finite. Therefore, bias can arise from the choice of training and test data, if they do not adequately represent the true population. AI models are made by people. Bias is a reflection of the data algorithm authors choose to use, as well as their data blending methods, model construction practices, and how results are applied and interpreted [17]. Indeed, AI processes are driven by human judgments.

Often, such biases can be introduced inadvertently to a learning model through such factors as the selection of datasets or the underlying cost function. In the hands-on Confidentiality activities, students will explore how subjective human or societal biases may emerge in the seemingly objective world of deep learning algorithms, and how to prevent them.

Deep learning algorithms have been widely used by banks in credit assessment systems. Studies have exposed racial bias in these systems [18]. In Equity labs, students are made aware of potential biases that can be introduced by data samples and countermeasures to eliminate these biases. For example, students can train a few credit assessment models based on their selection of n factors such as income, zip code, occupation, gender, and race. Then, they will examine the assessment results of these models and reflect on the possible causes of differences among them.

Students are asked to identify one cause of interest rates, say the racial factor, and experiment with two mitigation measures. The first measure is to divide the original dataset based on race information and train a model for each sub-dataset. Since there is no race information in the sub-datasets, the racial bias can be significantly reduced. Students will also reflect on why the racial bias is not completely removed (for instance, indirect racial information from other factors such as zip code). The second measure is to select balanced samples for each race group (e.g., the same number of individuals with high and low incomes) from the original dataset to form a new dataset. Since each race group now has a better representation, the resulting model is expected to be less biased.

In yet another lab, not related to finance and lending algorithms, students observe a different type of bias, algorithmic bias, through the board game of Go. AI-driven versions of Go utilize an AlphaZero-type algorithm [19], which has proven to be highly successful. Nevertheless, the same team that are responsible for the ICE labs have shown that such algorithms contain algorithmic bias that prevent it from making optimal plays [20]. Such an algorithm heavily relies on a cost function associated with winrate, which is a sort of expected value indicating the likelihood of a win. Consequently, it has a highly interesting tendency of deviating from optimal lines of play in order to protect a win or to avoid a loss, even when the optimal plays are "known" to the algorithm. For each activity, two sequences of plays for the same game position will be given to the students, one simple yet sub-optimal, the other optimal but complicated. If winning the game requires the optimal line, the AlphaZero-type model has no trouble finding it. However, the algorithm will consistently choose the simpler line of play if doing so already wins the game! Similarly, examples where the algorithm chooses aggressive but suboptimal plays to avoid a certain loss will be given. After observing the impact of the cost function, the students are encouraged to adjust certain parameters to either induce or prevent this algorithmic bias.

V. Technological Platform

There exist many popular platforms for performing deep learning related computation tasks. For example, Google Colab (research.google.com/colaboratory) is a free Jupyter based environment that allows users to create Jupyter programming notebooks to write and execute Python (and other Python-based third-party tools) in a web browser. Since many modules (such as Pandas, PyTorch, Tensorflow, and Keras) come preinstalled within Google Colab, there is little need to install additional modules to run code. Colab uses the computing power of Google servers instead of a user's local machine, which greatly reduces computer hardware requirements. Nonetheless, datasets need to be uploaded to Google Drive (a Google account is required) and authenticated to be used by Google Colab, which can be a daunting task for non-technical users. Another caveat is that all Google Colab notebooks are saved in the cloud by default, which may not be a preferred feature for privacy-sensitive users.

Another popular platform is Anaconda (www.anaconda.com), where users can perform Python/R based machine learning on a single machine with thousands of available open-source packages and libraries. Anaconda can have multiple environments with different versions of Python and supporting libraries. This way, a version mismatch can be avoided and is not affected by existing packages and libraries of the operating system. However, users will need to conscientiously manage specific versions of libraries with their dependencies and environments. Unfortunately, Anaconda is notorious for its slow start as it places various expensive overhead on computer hardware (such as CPU and RAMs). The initial installation of Anaconda can be confusing and time consuming.

The ICE labs are enabled by recent advances in virtualization technology. Container-based virtualization (www.docker.com/resources/what-container/) has become enormously popular over the last few years, which offers many desirable features for serving as the platform for our learning activities. A container sits on top of a host OS (e.g., MacOS, Linux or Windows) on a local machine or the cloud. Typically, a container is composed of just the application, which makes it exceptionally lightweight in action. The encapsulation of application operating code means that there are no guest OS environment variables or library dependencies to manage. Moreover, containers naturally support source code and dataset integration, which is vital for performing deep learning computation. Specific versions of software tools and environments can be "frozen" in a container so that consistent results can be delivered. Lastly, a container is not tied to any specific hardware infrastructure so that it can run on most systems without requiring code changes. The ICE labs are designed and built upon customized containers for carrying out the learning activities. The lab containers are shipped in a generic format so that they can be easily adopted either on local computers or cloud platforms. The lab containers and associated materials are freely available to the public.

VI. Conclusion

According to James [21], Cybersecurity statistics indicate that there are 2,200 cyber attacks per day, with a cyber attack happening every 39 seconds on average. In the US, a data breach costs an average of 9.44 million dollars, and cybercrime is predicted to cost 8 trillion dollars by 2023. These attacks equate to lost income, productivity, revenue, privacy, and personal information, and the frequency, severity, and complexity of these attacks will only continue to increase.

Robots and AI are expected to permeate our daily lives by 2025. This could have huge implications on several business sectors, most notably healthcare, customer service and logistics [22]. Despite the many ways AI can increase efficiency, convenience, and reduce the instances of human error, it is not without some risk and concern. As with other significant technologies that have had an impact on human civilization, the development and deployment of AI may proceed at a rate far faster than our ability to understand all its effects, which may lead to undesirable and unintended consequences [23].

In order to prepare prospective AI programmers and cybersecurity professionals to protect our Nation's critical cyberinfrastructure and the future of AI development, they must be equipped with sound principles, current trends and strategies, creativity, and insight into adversarial thinking. By providing computer science instructors and students with innovative labs built upon proven experiential learning design principles using flexible, lightweight virtualization environments, the ICE Labs is launching the next generation of AI professionals into the future.

References

[1] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," Cancer Letters, vol. 471, pp. 61-71, 2020. http://dx.doi.org/10.1016/j.canlet.2019.12.007

[2] D. Kolb, "Experiential Learning: Experience as the Source of Learning and Development," vol. 1, Upper Saddle River, NJ: Prentice Hall, 1984.

[3] T. J. Nokes and D. M. Belenky, "Mechanisms of knowledge transfer," Thinking and Reasoning, vol. 15, no. 1, pp. 1-36, 2009. http://dx.doi.org/10.1080/13546780802490186

[4] S. T. Hamman and K. M. Hopkinson, ”Teaching adversarial thinking for cybersecurity,” Journal of The Colloquium for Information Systems Security Education, vol. 4, no. 1, pp. 19-19, Oct. 2016.

[5] R. J. Sternberg, ”The Triarchic Mind,” New York: Penguin Books, 1988.

[6] B. Morel, ”Artificial intelligence and the future of cybersecurity,” in Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, Oct. 2011, pp. 93-98. http://dx.doi.org/10.1145/2046684.2046699

[7] K. Chamberlin, M. Yasué, and I. C. A. Chiang, ”The impact of grades on student motivation,” Active Learning in Higher Education, 2018. http://dx.doi.org/10.1177/1469787418819728

[8] I. Goodfellow, J. Shlens, and C. Szegedy, ”Explaining and harnessing adversarial examples,” in International Conference on Learning Representations, 2015.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, ”Intriguing properties of neural networks,” in International Conference on Learning Representations, 2014.

[10] J. Dumford and W. Scheirer, ”Backdooring convolutional neural networks via targeted weight perturbations,” in 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1-9, IEEE, 2020. http://dx.doi.org/10.1109/IJCB48548.2020.9304875

[11] M. Alberti, V. Pondenkandath, M. Wursch, M. Bouillon, M. Seuret, R. Ingold, and M. Liwicki, ”Are you tampering with my data?” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 1-18, IEEE, 2018. http://dx.doi.org/10.1007/978-3-030-11012-3-25

[12] N. Lomas, ”Italy orders ChatGPT blocked citing data protection concerns,” in TechCrunch, March 31, 2023 [Online]. Available: https://techcrunch.com/2023/03/31/chatgpt-blocked-italy/

[13] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, ”Membership inference attacks on machine learning: A survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 54, no. 2, pp. 1-36, 2021. http://dx.doi.org/10.1145/3523273

[14] P. Irolla and G. Châtel, ”Demystifying the membership inference attack,” in 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), pp. 1-7, IEEE, 2019. http://dx.doi.org/10.1109/CMI48017.2019.8962136

[15] M. Fredrikson, S. Jha, and T. Ristenpart, ”Model inversion attacks that exploit confidence information and basic countermeasures,” in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS ’15), pp. 1322-1333, Association for Computing Machinery, 2015. http://dx.doi.org/10.1145/2810103.2813677

[16] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, ”Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing,” in Proceedings of the 23rd USENIX Conference on Security Symposium, SEC’14, pp. 17-32, USENIX Association, 2014.

[17] G. S. Nelson, ”Bias in artificial intelligence,” North Carolina Medical Journal, vol. 80, no. 4, pp. 220-222, 2019. http://dx.doi.org/10.18043/ncm.80.4.220

[18] K. Riyazahmed, ”AI in finance: Needs attention to bias,” Annual Research Journal of SCMS, Pune, vol. 11, no. 1, pp. 1-8, 2023.

[19] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, et al., ”A general reinforcement learning algorithm that masters chess, shogi, and go through self-play,” Science, vol. 362, no. 6419, pp. 1140-1144, 2018. http://dx.doi.org/10.1126/science.aar6404

[20] K. N. Ze-Li Dou, Liran Ma and K. X. Nguyen. ”Paradox of alphazero: Strategic vs. optimal plays,” In Proceedings of the 39th IEEE International Performance Computing and Communications Conference (IPCCC’20), 2020. http://dx.doi.org/10.1109/IPCCC50635.2020.9391562

[21] N. James, “160 Cybersecurity statistics 2023.” Astra, September 13, 2023 [Online]. Available: https://www.getastra.com/blog/security-audit/cyber-security-statistics.

[22] A. Stahl, ”The rise of artificial intelligence: Will robots actually replace people?” Forbes, May 3, 2022. [Online]. Available: https://www.forbes.com/sites/ashleystahl/2022/05/03/the-rise-of-artificial-intelligence-will-robots-actually-replace-people/?sh=1dcb4cfe3299

[23] B. Reed, ”The danger of blindly embracing the rise of AI,” The Guardian, April 3, 2023. [Online]. Available: https://www.theguardian.com/technology/2023/apr/03/the-danger-of-blindly-embracing-the-rise-of-ai