Addressing the Vectors for Attack on Artificial Intelligence Systems Used in Clinical Healthcare through a Robust Regulatory Framework: A Survey

Benjamin Clark

I. Introduction and Overview

Artificial intelligence has captivated the current interest of the general public and academics alike, bringing closer attention to previously unexplored aspects of these algorithms, such as how they have been implemented into critical infrastructure, ways they can be secured through technical defensive measures, and how they can best be regulated to reduce risk of harm. This paper will discuss vulnerabilities common to artificial intelligence systems used in clinical healthcare and how bad actors exploit them before weighing the merits of current regulatory frameworks proposed by the U.S. and other nations for how they address the cybersecurity threats of these systems.

Primarily, artificial intelligence systems used in clinical research and healthcare settings involve either machine learning or deep learning algorithms.[1] Machine learning algorithms automatically learn and improve themselves without needing to be specifically programmed for each intended function.[2] However, these algorithms require that input data be pre-labeled by programmers to train algorithms to associate input features and best predict the labels for output, which involves some degree of human intervention.[3] The presence of humans in this process is referred to as "supervised machine learning" and is most often observed in systems used for diagnostics and medical imaging, in which physicians set markers for specific diagnoses as the labels and algorithms are able to categorize an image as a diagnosis based off the image's characteristics.[4] Similarly, deep learning is a subset of machine learning characterized by its "neural network" structure in which input data is transmitted through an

[1] Graham S, et al. *Artificial Intelligence for Mental Health and Mental Illnesses: an Overview*. CURR PSYCHIATRY REP. 2019 Nov 7;21(11):116. doi: 10.1007/s11920-019-1094-0.

[2] *Id.*

[3] *Id.*

[4] *Id.*

algorithm through input, output, and "hidden" layers to identify patterns in data.[5] Deep learning algorithms differ from those that utilize machine learning in that they require no human intervention after being trained; instead, deep learning algorithms process unlabeled data by determining what input is most important to create its own labels.[6]

As mentioned previously, these forms of artificial intelligence are often used for medical imaging for the purposes of identifying medical conditions such as cancer or other cell irregularities, which implicates risks surrounding how medical decisions made on the recommendation of artificial intelligence systems that have been compromised by an adversarial attack might affect patient health. Further, machine learning algorithms are commonly implemented in medical billing, creating financial incentives for bad actors to exploit vulnerabilities within these systems.[7] While there are clear risks associated with artificial intelligence systems that call for oversight into their cybersecurity hygiene, the current state of legislative action in the U.S. is not up to speed with how prevalent these algorithms have become in healthcare settings.[8] There are currently limited mandatory technical standards for artificial intelligence development at the federal level, and while recent executive orders have recognized a need for such standards, they have lacked specificity regarding what cybersecurity practices should be required of developers.[9] Given this context, we look to the regulatory frameworks put forward by foreign powers such as the EU and Singapore for perspective on how the vulnerabilities of artificial intelligence systems have

[5] *Id.*
[6] *Id.*
[7] Finlayson, SG, et al., *Adversarial attacks on medical machine learning.* SCIENCE. Mar 2019. 22;363(6433):1287-1289. doi:10.1126/science.aaw4399.
[8] *Id.*
[9] Comiter M, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It.*, 16, Harvard Kennedy School. Aug 2019.

been addressed. These frameworks can inform how policymakers and regulatory agencies in the U.S. approach the growing need to adequately secure artificial intelligence systems and protect against the risks to our critical infrastructure should they be exploited by bad actors.

II. Investigating Attacks on Artificial Intelligence Systems

a. What are the vulnerabilities of artificial intelligence systems?

Ironically, what makes artificial intelligence algorithms so useful and reliable in application, their ability to learn patterns, can also be what makes them susceptible to attack. These patterns identified by artificial intelligence systems, while excellent for making diagnoses or identifying abnormalities in images, are somewhat rigid, which makes them easier to disrupt. Bad actors are able to exploit the fixed and somewhat predictable nature of machine learning algorithms through inputting data that does not conform with the patterns they have identified.[10] Similarly, because machine learning and deep learning algorithms require training with very large datasets to yield consistent and accurate output, bad actors are given a specific target and know to direct attacks at the training datasets.[11] Influencing or corrupting the training data corrupts the entire system, making this a desirable vector for attack.[12] Further, developers of artificial intelligence systems have observed what is referred to as the "black-box" phenomenon, in which they do not understand the mechanisms by which these systems create their output or how data is transmitted specifically, because these algorithms are not being specifically programmed to identify certain patterns in data.[13] This phenomenon makes it difficult for developers and operators of artificial intelligence systems to know if their models

[10] *Id.*
[11] *Id.*
[12] *Id.*
[13] Graham S, et al. *Artificial Intelligence for Mental Health and Mental Illnesses: an Overview*. CURR PSYCHIATRY REP. 2019 Nov 7;21(11):116. doi: 10.1007/s11920-019-1094-0.

have been exploited or compromised by bad actors, which, as a result, encourages adversarial attacks.[14] It's these primary vulnerabilities that have been identified and addressed by policymakers in the existing regulatory frameworks for artificial intelligence systems– U.S. policymakers should similarly consider these vulnerabilities in drafting legislation.

b. *Types of Artificial Intelligence Attacks*

Cyberattacks on artificial intelligence systems generally fall into two categories: input attacks and data poisoning attacks.[15] Input attacks occur when data, such as an image, is fed into a system that is already trained to identify specific labels.[16] Bad actors then alter this image in subtle ways to trick the system into labelling the image incorrectly.[17] If the patterns of the input are inconsistent enough with the variations seen in the dataset during training, then the output of the algorithm may be inaccurate.[18]￼ Input attacks are characterized by two factors: perceivability and format.[19] Perceivability considers whether the attack is visible to a reasonable person, as some input attacks involve small, unobservable changes to the pixels of an image or even rotating the image slightly to disrupt an artificial intelligence system.[20]￼ Format asks whether the input is physical or digital, as this can impact how an attack is able to be identified and contained.[21] ￼Input attacks are particularly worrisome because they do not require the bad actor to corrupt the artificial intelligence system prior to introducing the altered

[14] Comiter, *supra* note 7, at 16.
[15] *Id.*
[16] *Id.*
[17] *Id.*
[18] Lohn A. *Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity.* Center for Security and Emerging Technology. Dec 2020.
[19] *Id.*
[20] Comiter, *supra* note 13, at 18.
[21] *Id.*

data, making it even harder for developers to identify if and how their systems have been compromised.[22]

Data poisoning attacks occur when bad actors corrupt the datasets that developers use during the training process for their models.[23] These attacks can corrupt systems by targeting specific input features or labels within an algorithm to yield a desired output or causing the model to "learn" a backdoor for a future cyberattack.[24] While there a variety of categories of data poisoning attacks, attacks that insert manipulated target labels present a tougher challenge in terms of the resilience of artificial systems, in that they can be harder to correct once the system has learned to incorrectly label input data.[25] Such attacks are referred to as "backdoor attacks" and disrupt artificial intelligence systems because the legitimate input used by developers during the learning process does not contain the manipulated target pattern that was created by bad actors with previously processed data and cannot be categorized accurately due to this.[26] However, backdoor attacks can be easier to identify, as input that has been corrupted or "poisoned" would appear as clearly mislabeled to those training the algorithm increasing the likelihood the attack could be stopped before it corrupts the system as a whole.[27] Despite this, data poisoning attacks pose real risks to artificial intelligence systems, particularly those that regularly process large amounts of data during training such as those typically seen in medical imaging.[28]

[22] *Id.*
[23] *Id.* at 28
[24] *Id.*
[25] Selvakkumar, Arawinkumaar, *Addressing Adversarial Machine Learning Attacks in Smart Healthcare Perspectives*, SCHOOL OF COMPUTER SCIENCE, Queensland University of Technology, 16 Dec 2021.
[26] *Id.*
[27] *Id.*
[28] *Id.*

Another way that cyberattacks on artificial intelligence systems can be categorized is by attacks that target one of the three core aspects of cybersecurity: confidentiality, integrity, and availability. Integrity attacks are those in which bad actors manipulate data to corrupt an artificial intelligence system and cause it to incorrectly identify input.[29] This category would encompass the types of input attacks and data poisoning attacks described above. Similarly, confidentiality attacks occur when bad actors seek to exfiltrate data processed by artificial intelligence systems that is located in the hidden layers of a model's neural network.[30] Availability attacks are those that affect the speed of functioning for machine learning models, or their ability to function at all, but are not seen as commonly as there are fewer incentives for bad actors with this type of attack.[31] This is because the model would be rendered unusable and could not be exploited for future gain.[32] Regardless of how attacks on artificial intelligence systems can be categorized, the vectors for attack remain the same, warranting further regulatory oversight into how these vulnerabilities can be better identified and protected against cyberattacks.

c. How are the vulnerabilities of artificial intelligence exploited by bad actors?

There is a particularly pressing need to regulate the cybersecurity measures of artificial intelligence systems, as they are not only susceptible to attack but can also be used to create cyberattacks on other models. Bad actors can operate artificial intelligence algorithms that use optimization theory to minimize or maximize the value of mathematic functions and identify the vulnerabilities in the algorithm that would give them a greatest ability to gain access to and

[29] Lohn, *supra* note 18 at 5.
[30] *Id.*
[31] *Id.*
[32] *Id.*

influence over the system.[33] Software that incorporates these optimization algorithms is often publicly available, making it incredibly easy for bad actors to develop and fine-tune input attacks on artificial intelligence systems.[34] Similarly, bad actors might also make use of Generative Adversarial Networks (GANs), which are machine learning systems that can learn to reproduce a specific distribution of data to exploit vulnerabilities in artificial intelligence systems[35]. Further, if bad actors gain access to the training dataset used to develop a particular system or find a sufficiently similar dataset, then they can create their own copy of the system and develop cyberattacks against the original system.[36]

As evidenced by these examples, there are easily accessible ways to infiltrate artificial intelligence systems, but these vulnerabilities can be exacerbated by policies and procedures put in place by those that operate them. In the context of clinical healthcare, artificial intelligence systems are often treated as commodities rather than sensitive assets and often lack the degree of cybersecurity that would be expected of many critical infrastructure technologies, even though they are being used to process protected health information (PHI) as defined by the HIPAA Security Rule[37]. Although artificial intelligence systems are incredibly useful tools for healthcare practitioners to use, they are still merely digital files and are susceptible to unauthorized access by bad actors just like any other file on a computer[38]. Because of this, they should be protected by the same, if not stricter, cybersecurity standards for current data processing technologies.

[33] Finlayson, SG, et al., *Adversarial attacks on medical machine learning*. Science. Mar 2019. 22;363(6433):1287-1289. doi:10.1126/science.aaw4399.
[34] Comiter at 25.
[35] *Id.*
[36] *Id.*
[37] *Id. See* 45 C.F.R. § § 164.302-164.318.
[38] Lohn at 14.

One study, focusing on the application of a machine learning model to diagnose various skin conditions, introduced adversarial images to the model in a controlled input attack to test the accuracy of its output[39]. The study involved the researchers inputting an image of a benign mole into a trained model, which accurately categorized and diagnosed the image as benign. However, the researchers then utilized an optimization algorithm to identify the best and smallest manipulations of the image to have the most disruptive effect, which amounted to miniscule alterations of select pixels in the image[40]. Even though these changes were invisible to the human eye, the model categorized and diagnosed the image of the benign mole as malignant with 100% confidence[41]. The study also observed that simpler manipulations than this, including turning an image to a specific angle, had a similar negative effect on the model's ability to yield accurate output[42].

As stated previously, poisoning attacks occur when bad actors input incorrectly labeled data into the datasets used for training, but these attacks can take place at various stages of the training process.[43] To provide context for a typical data poisoning attack, consider the same example of an imaging system used to diagnose skin cancer. In this situation, a bad actor could change an image correctly labeled as a benign mole to an image that actually depicted a malignant mole, which would cause the artificial intelligence system to fail in correctly identifying and diagnosing the skin cancer. The system in this case would fail because it would learn to identify images of benign moles as malignant during the training process.[44] If such an

[39] Finlayson, SG, et al., Adversarial attacks on medical machine learning. Science. Mar 2019. 22;363(6433):1287-1289. doi:10.1126/science.aaw4399.

[40] *Id.*

[41] *Id.*

[42] *Id.*

[43] Comiter *supra* note 34 at 28.

[44] *See id.*

attack were to be carried out, it could result in patients being misdiagnosed and being recommended ineffective treatment plans, causing serious risk of injury or bodily harm. Likewise, artificial intelligence systems are commonly used for medical billing, which presents incentives for bad actors to corrupt models used for imaging to misdiagnose certain conditions and fraudulently bill for unnecessary medical costs.[45] These cybersecurity risks and their potential adverse effects pose a serious threat to the U.S. healthcare system and should be weighed against the benefits of their use by policymakers.

Poisoning attacks can also occur during data collection, as bad actors can influence artificial intelligence systems to collect data that would cause the system to identify patterns and create labels advantageous to the attackers.[46] This type of attack can occur during the data collection process if a bad actor is able to learn how and from what sources the developers of an artificial intelligence system are collecting data by gaining access to this information directly or through an initial, more traditional cyberattack, which would open a backdoor for the poisoning attack.[47] By knowing what data is being collected and where it is being collected from, bad actors can influence the collection process to train an artificial intelligence system to accept manipulated data in the future.[48]

Despite the existence of these vectors for attacking artificial intelligence systems, there are various technical defensive measures that can be incorporated into these models to effectively prevent and mitigate the effects of both data poisoning and input attacks. One such measure is adversarial training, which involves developers intentionally training their model with

[45] Finlayson *supra* note 39 at 2.
[46] Comiter *supra* note 34 at 28.
[47] *Id.*
[48] *See id.*

adversarial images to be able to identify and appropriately categorize those that might be introduced by bad actors either during or after the training process.[49] Not only does adversarial training protect artificial intelligence systems against potential data poisoning or input attacks, but it also makes these systems generally more accurate, resulting in greater reliability of output while also reducing the risks of negative patient outcomes in the context of clinical healthcare.[50] The model examined in a similar study to the one discussed previously was found to be resistant to adversarial examples after being trained with adversarial input, producing low error rates typical of systems that had not been attacked with adversarial input.[51] The same study also found that another technical defensive measure, randomization, could be similarly effective against adversarial input attacks.[52] This measure involves randomizing the adversarial effects of manipulated input, through methods such as inputting images at random sizes, to decrease the effects that an input attack might have on an artificial intelligence system.[53] While it does not prevent input attacks entirely, randomization can reduce the impact that adversarial examples have on the model.[54]

Additionally, some have suggested that extracting what is called a "fingerprint hash" of the data could help developers of artificial intelligence systems identify and mitigate the effects of input attacks and data poisoning attacks.[55] By comparing the hash of a system that has not been attacked to the hash captured after manipulated data is inputted into a system, developers would

[49] Selvakkumar at 6.
[50] Comiter *supra* note 34 at 28
[51] *Id.*
[52] *Id.*
[53] *Id.*
[54] *Id.*
[55] ***See*** Sara Kaviani et al., Adversarial Attacks and Defenses on AI in Medical Imaging Informatics: A Survey, 198 EXPERT SYS. WITH APPLICATIONS, Mar. 19, 2022, at [pincite].

be able to see if that data had been altered by a bad actor.[56] However, this defensive measure would require that the computer systems of the U.S. healthcare sector have the capacity to store these hashes, which would currently not be possible.[57] Further, some technical experts have proposed the practice of federated learning as a potential defensive measure against poisoning attacks during data collection and the general training process.[58] Federated learning involves the training of multiple, smaller models on separate devices that are then combined to form a final artificial intelligence system, as opposed to using one large dataset to train one model.[59] However, with this method, attackers could direct an attack on one device and poison the data processed by the algorithm on that device, which would then have an effect on the final artificial intelligence system.[60]

III. Legislative Proposals and Existing Regulatory Frameworks Addressing Artificial Intelligence System Cybersecurity

a. Legislative Developments in the U.S.

Like much of the law in the U.S., the approach to regulating artificial intelligence systems has been piecemeal, looking to the standards set by organizations in different sectors to develop a national industry best-practice for all developers of artificial intelligence systems. Legislative developments have been outpaced by the growing use of artificial intelligence in critical infrastructure, which has resulted in a real need for mandatory technical safeguards to protect against input attacks and data poisoning attacks.

[56] *See id.*
[57] *Id.*
[58] Comiter *supra* note 34 at 30.
[59] *Id*. at 31.
[60] *Id.*

A key development in the process of creating a regulatory framework occurred when President Joe Biden issued an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence in October 2023, which "requires robust, reliable, repeatable, and standardized evaluations of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use.[61]" This general standard seems to echo the language of other U.S. cybersecurity laws, but as with these existing laws, developers of artificial intelligence systems might have trouble understanding the federal government's interpretation of what is "robust" or "appropriate" enough.[62] The executive order also touches on the existence of cybersecurity vulnerabilities in artificial intelligence systems, as well as the risks these systems pose to critical infrastructure, as evidenced by its call for regular testing and monitoring of artificial intelligence systems.[63] Likewise, the executive order requires that artificial intelligence systems be "resilient against misuse or dangerous modifications, are ethically developed and operated in a secure manner, and are compliant with applicable Federal laws and policies," which indicates requirements for technical defensive measures such as adversarial training or randomization as well as the recognition of existing sectoral laws.[64] While these executive orders may lack specificity in regard to the vectors for attack common to artificial intelligence systems, they provide excellent insight into how the U.S. government is approaching the regulation of these systems that are implemented in critical infrastructure such as healthcare.

[61] Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023). (T1.2)
[62] *See generally* LabMD, Inc. v. Fed. Trade Comm'n, 894 F.3d 1221 (11th Cir. 2018).
[63] *See* 88 Fed. Reg. 75191
[64] *Id.*

Specifically, this executive order suggests that the National Institute for Standards and Technology's ("NIST") AI Risk Management Framework will be the skeleton for a more complete legislative proposal in the future.[65] This framework takes a risk-based approach to regulating artificial intelligence systems but does not go as far to distinguish categories of high risks and low risks like other regulatory frameworks.[66] Instead, it provides that risks are highly contextual, which indicates that artificial intelligence systems used in critical infrastructure may require more robust cybersecurity measures.[67] Further, NIST's framework requires that systems be both secure and resilient, defining resilience as the ability of artificial intelligence systems to "withstand unexpected adverse events or unexpected changes in their environment or use or . . . maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary.[68]" The framework also enumerates specific adversarial techniques that can be leveraged against artificial intelligence systems, such as adversarial examples, data poisoning, and the exfiltration of models and training data.[69] In this way, NIST's guidance serves to make developers of artificial intelligence systems aware of common vectors for attack, increasing the likelihood that they will appropriately secure their models against such attacks. Further, NIST explains that resilience "goes beyond the provenance of the data to encompass unexpected or adversarial use," which not only places a particular importance on proper data collection practices during and after training but also recognizes the need for technical defensive measures.[70]

[65] *Id.*
[66] NIST 100-1, 1.2.1
[67] *Id.* at 1.2.2.
[68] *Id.* at 3.3.
[69] *Id.*
[70] *Id.*

Regarding security, NIST explicates that security is achieved if an artificial intelligence system "can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use," pointing to its general standards for cybersecurity and risk management that are not specific to artificial intelligence.[71] However, compliance with NIST's AI Risk Management Framework is voluntary and does not require that developers of artificial intelligence systems implement any specific technical defensive measures to combat cyberattacks.[72] That said, President Biden's previously discussed executive order charges NIST, among other regulatory agencies, to develop more cohesive guidelines, including standards for cybersecurity, for developers of artificial intelligence systems in different sectors.

For example, in the healthcare sector, The Office of Information Security under the Department of Health and Human Services has issued guidance for developers of artificial intelligence systems, pointing them to the Adversarial Threat Landscape for Artificial-Intelligence Systems ("ATLAS") framework, which is described as "a taxonomy for adversarial tactics and techniques" that are commonly leveraged against systems used for medical imaging and diagnostics.[73]" The ATLAS framework explains what vulnerabilities exist in these models and how bad actors are able to exploit them, ranging from how they gain initial access to a system to how the integrity of these systems can be eroded.[74] Further, each entry includes case studies where the specific type of cyberattack at issue was carried out, as well as potential technical mitigation strategies for that adversarial technique.[75] For example,

[71] *Id.*
[72] *See id.*
[73] *Artificial Intelligence, Cybersecurity, and the Healthcare Sector*, 46, Office of Information Security. July 13, 2023.
[74] Poison Training Data, MITRE, https://atlas.mitre.org/techniques/AML.T0020/ (last visited Dec. 18, 2023).
[75] *See id.*

one such entry focuses on the poisoning of training data but also includes links to related entries that explain how backdoor triggers are inputted into machine learning models to enable that poisoning attack, which creates a clear and understandable timeline for how these attacks occur.[76] While the ATLAS framework is non-binding on developers of artificial intelligence systems, it is an excellent resource for U.S. policymakers to understand the various cybersecurity vulnerabilities of these systems, which could help them outline implementation specifications for technical defensive measures in a future legislative proposal.

b. Singapore's Model Artificial Intelligence Governance Framework

Similar to the U.S., Singapore has not yet chosen to specifically regulate artificial intelligence systems at the federal level.[77] However, one of the nation's regulatory agencies, Infocomm Media Development Authority (IMDA), has issued extensive guidance in the form of its Model Artificial Intelligence Governance Framework on how these systems should be developed with cybersecurity principles in mind.[78] Unlike much of the cybersecurity laws in the U.S. and EU, Singapore's proposal is sector-agnostic, serving as a "baseline for organizations in any sector to adopt," while acknowledging that certain sectors could choose to promulgate additional rules or requirements depending on the need for further technical measures and the use of their artificial intelligence systems.[79] Likewise, the framework is technology-agnostic and does not focus on specific types of artificial intelligence such as machine learning or deep learning algorithms and applies to all types of data collection and

[76] *See id.*

[77] *Model Artificial Intelligence Governance Framework*, Second Edition, 10, Infocomm Media Development Authority. Personal Data Protection Commission. Jan. 2020.

[78] *Id.*

[79] *Id.*

storage methods.[80] This core principle of the Model Artificial Intelligence Governance Framework addresses the risks of data poisoning and input attacks, as it ensures that there are standards for the data practices of all artificial intelligence systems, increases the ability of developers that follow the framework's recommendations to identify, and prevents manipulations to the datasets AI systems use. Singapore's proposal also highlights the problem of the black-box phenomenon, acknowledging the cybersecurity vulnerabilities that can stem from this lack of insight into how artificial intelligence systems learn[81].

With these principles in mind, Singapore's regulatory framework proposes a general risk management structure to inform the conduct of developers and operators of artificial intelligence systems alike. These risk management strategies include conducting a risk assessment at the outset of system development that considers the severity of, and potential for, an adverse impact on individuals " whose data might be processed by an artificial intelligence.[82] The framework provides an example for how the company Mastercard implemented risk management procedures and internal controls into their operations, highlighting how they designated a Chief Data Privacy Officer role to conduct risk assessments and a Chief Information Security Officer to implement security by design principles in the development of artificial intelligence systems.[83] Even though these recommendations are more administrative than technical, they still reflect the importance of appropriately securing artificial intelligence systems to reduce risk of harm, particularly when they are implemented in critical infrastructure.

[80] *Id.*
[81] *Id.*
[82] *Id.* at 22.
[83] *Id*. at 27.

While the Model Artificial Intelligence Governance Framework incorporates aspects of the risk-based approaches of most U.S. cybersecurity laws and the EU's Artificial Intelligence Act, it recommends varying degrees of cybersecurity measures based on the level of human involvement in the development and use of artificial intelligence systems. One such category of human involvement described in Singapore's framework is the "human-in-the-loop" approach, in which the artificial intelligence simply provides recommendations and does not have full decision-making ability, leaving the human user of the system with full control.[84] Most uses of artificial intelligence in clinical healthcare, such as medical imaging and diagnostics, would fall into this "human-in-the-loop" category. For example, a physician would be able to use a trained machine learning model to accurately categorize a specific medical condition, but the system would still need the physician's command to make a given decision, in this case a diagnosis. In short, artificial intelligence systems are simply used to inform human decision-making rather than make the decisions without human oversight under this approach.

Conversely, in the "human-out-of-the-loop" approach, artificial intelligence systems have the ability to make decisions that go beyond just recommendations made to a human operator.[85] In clinical healthcare setting, these systems could be used for patient management, as an artificial intelligence system could be used to identify where potential disruptions to patient triage could occur and would give this output to a "solver model" to find solutions without the need for human intervention.[86] Singapore's regulatory framework also introduces a "human-over-the-loop" approach, in which the operator of an artificial intelligence system merely

[84] *Id*. at 30.
[85] *Id*.
[86] *Id*.

serves a supervisory role, such as for system monitoring.[87] Interestingly, the framework also suggests that systems that monitor the performance and accuracy of artificial intelligence system output can be autonomous themselves and incorporate some degree of machine learning, raising questions about how the monitoring system could be protected when it is already playing a role in protecting another artificial intelligence system.[88]

In acknowledgment of the particular importance of data to the training of artificial intelligence models, Singapore's regulatory framework delves into standards for data governance both in models that are being developed and those in active use.[89] The structure of these standards is analogous to the U.S. HIPAA Security Rule in that the framework sets the baseline standards for administrative, procedural and technical safeguards while including implementation specifications to further explain requirements to developers and operators of artificial intelligence systems.[90] In the context of data governance, the Model Artificial Intelligence Governance Framework requires that there be appropriate policies and procedures in place to ensure the quality of data.[91] The framework then specifies that this should be accomplished through regular monitoring and review of systems currently in use in order to identify vulnerabilities and let developers know where data practices should be reformed.[92] Additionally, it recommends specific administrative safeguards, such as training for those operating artificial intelligence systems to interpret output and identify potential manipulations in data.[93] In the context of clinical healthcare, training physicians that are informing their

[87] *Id.*
[88] *Id.* at 24.
[89] *See id.*
[90] *See* 45 CFR § 164.308(a)(1)
[91] *Id.*
[92] *Id.*
[93] *Id.*

patient care decisions based on the recommendations of an artificial intelligence model is an important aspect to mitigating the risks associated with this technology. While Singapore's proposal does not address data poisoning or input attacks specifically, these data governance measures would impact the ability of bad actors to manipulate datasets used during the training process.

c. European Union's Artificial Intelligence Act

The European Union's Artificial Intelligence Act presents one of the most comprehensive risk-based regulatory frameworks in terms of how it addresses cybersecurity vulnerabilities in algorithms.[94] Similar to Singapore's Model Artificial Intelligence Governance Framework, the EU regulation sorts various types of artificial intelligence algorithms into categories based on risk, including unacceptable risk, high risk, limited risk, and low risk.[95] Algorithms that are used to operate critical infrastructure such as healthcare are placed in the high risk category, which requires that providers of these systems first register their systems in an EU-wide database prior to marketing them to consumers or using them in practice.[96] The machine learning algorithms commonly used for medical imaging and diagnostics would be subject to these obligations, which would not only provide another level of oversight for new technologies being used in high-risk settings but also make it easier to identify and address potential vulnerabilities in these systems. Article 15 of the EU Artificial Intelligence Act

[94] *See* Proposal For A Regulation Of The European Parliament And Of The Council Of April 21, 2021 Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

[95] *See id.*

[96] Tambiama Madiega . *Briefing on EU Legislation in Progress: Artificial intelligence act.* EUROPEAN PARLIAMENTARY RESEARCH SERVICE. (June 2023), Artificial intelligence act (europa.eu)

addresses cybersecurity concerns directly, setting higher technical standards for higher risk models.[97]

Specifically, the law requires that high risk artificial intelligence systems have an "appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle," mirroring the "reasonable" and "appropriate" cybersecurity standards set by frameworks such as the HIPAA Security Rule and the FTC's Safeguards Rule in the U.S.[98] The EU's proposal provides that all systems in this risk category have backup models or fail-safe plans to ensure that critical infrastructure operations are not disrupted by potential cyberattacks.[99] Further, the EU's legislation addresses how it will fit into existing sectoral regulatory frameworks, as it explains that artificial intelligence systems will also be governed by sectoral regulations that apply to their use, such as rules regarding the safety of medical devices.[100]. The Artificial Intelligence Act does not preempt existing laws that might even require that stricter technical safeguards be put in place by developers. Given this principle, policymakers in the U.S. could look to existing cybersecurity regulations like the HIPAA Security Rule to use consistent and recognizable language for developers to understand the requirements and leave space for sectoral laws that could regulate specific applications of artificial intelligence systems more efficiently.

What makes the EU's regulatory framework different from those that have been proposed by other countries (such as Singapore or the web of non-binding documents in the U.S.) is that

[97] *See* Proposal For A Regulation Of The European Parliament And Of The Council Of April 21, 2021 Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

[98] *Id.*

[99] *Id.*

[100] *Id.*

it requires developers of artificial intelligence systems to implement specific technical defensive measures to combat cyberattacks. Not only does it require that cybersecurity be prioritized in the development and use of artificial intelligence, but it is one of the only current pieces of legislation that will soon be binding and address specific cybersecurity vulnerabilities that bad actors commonly exploit. This framework also strikes a balance in setting standards for preventative and defensive measures to be taken against cyberattacks. It provides that high risk systems be "resilient" against unauthorized third-party access, acknowledging the need for defensive measures to combat the effects of potential cyberattacks.[101] The Artificial Intelligence Act enumerates specific vulnerabilities that should be guarded against with heightened technical safeguards, including defensive measures to prevent data poisoning and adversarial input attacks.[102] Meanwhile, Article 15's particular focus on data governance serves an important role in preventing data poisoning attacks, as it requires that "[t]raining, validation and testing data sets shall be relevant, representative, free of errors and complete."[103] While this standard was likely intended to address issues such as algorithm bias, the requirement for consistent monitoring of training datasets provides developers more opportunities to identify attempts by bad actors to corrupt their models through vectors in data. Legislative proposals in the U.S. should consider including a provision such as this into a potential regulatory framework, as training data plays such a crucial role in the implementation and success of artificial intelligence systems in critical infrastructure.

d. An Alternative Approach

[101] *See* Proposal For A Regulation Of The European Parliament And Of The Council Of April 21, 2021 Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

[102] *Id.* at art. 15.

[103] *Id.*

Some have argued that an approach other than the EU and Singapore's risk-based proposals could be taken to regulate artificial intelligence systems and address cybersecurity concerns.[104] Law professor Orly Lobel argues that a regulatory framework for artificial intelligence should include a balancing test between the anticipated negative impacts of artificial intelligence systems and opportunities for their positive impacts, which contrasts the more traditional risk-based approach of the EU that places greater restrictions on higher-risk models regardless of their potential or current use for good.[105] In the same way, Lobel's argument centers the rights of those operating artificial intelligence systems and those whose data is being used to train those systems to a potential legislative approach by U.S. policymakers.[106] He frames the human-out-of-the-loop approach artificial intelligence systems, as outlined in Singapore's Model Artificial Intelligence Governance Framework, as a right to automated decision-making.[107] Further, Lobel advocates for a right to data maximization, as artificial intelligence systems often require very large and representative sets of data to accurately create labels, categorize the data, and yield consistently reliable output.[108] He argues that regulatory frameworks that limit the data that can be used both during and after the training process of machine learning and deep learning models can create more security issues for developers moving forward.[109] This approach comes into stark contrast with the data governance standards set in the EU's Artificial Intelligence Act and Singapore's regulatory framework, which

[104] Orly Lobel, *The Law of AI for Good*, January 26, 2023. UNIVERSITY OF SAN DIEGO SCHOOL OF LAW. San Diego Legal Studies Paper No. 23-001 , http://dx.doi.org/10.2139/ssrn.4338862
[105] *Id.* at 45
[106] *Id.* at 48
[107] *Id.*
[108] *Id.*
[109] *Id.* at 50.

prioritize data minimization and increased security measures for data collection and storage.[110] While Lobel prioritizes the development of artificial intelligence systems, his argument largely sidesteps the specific cybersecurity vulnerabilities that are common to these systems– particularly poisoning attacks that target the data collection processes of developers. Under his approach, training data would be even harder for developers to monitor and identify potential manipulations by bad actors, as larger datasets would provide more opportunities for these bad actors to corrupt the model and go unnoticed. Regulatory frameworks with a focus on risk management seem to lend themselves better to the existing sectoral cybersecurity laws in the U.S. that often require that cybersecurity measures be reasonable and appropriate for a particular context. Despite this, Lobel's focus on the rights of data subjects and his call for a balancing test between the risk of harm and potential for good of artificial intelligence systems is extremely valuable and could be used to inform a legislative proposal in the U.S.

IV. Takeaways

Given the increasing and novel uses of artificial intelligence in the context of clinical healthcare and medical research, among other sectors of critical infrastructure, one must question how legislative bodies both within and outside the U.S. plan to address the real cybersecurity risks these systems pose. Current regulatory frameworks such as the EU's Artificial Intelligence Act and Singapore's Model Artificial Intelligence Governance Act weigh these risks and require that developers of artificial intelligence systems be able to identify the specific vectors for attack within their systems and take appropriate measures to secure them. While these frameworks require that technical defense measures be implemented

[110] *See generally* Proposal For A Regulation Of The European Parliament And Of The Council Of April 21, 2021 Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

with varying degrees of specificity, they set important standards that can be further explained through rules promulgated by regulatory agencies in the future. The approach currently underway in the U.S. seems to mirror those of other foreign powers and conforms with existing sectoral cybersecurity laws. As with any other area of law, it will be interesting to see how explicitly U.S. policymakers and regulators require developers of artificial intelligence systems to prevent and defend against particular cyberattacks such as input or data poisoning attacks, or if their legislative proposal will leave the specifics of implementation in the hands of developers.