Human Factors in Robots, Drones and Unmanned Systems, Vol. 93, 2023, 171–182
https://doi.org/10.54941/ahfe1003760

AI Trust Framework and Maturity Model: Improving Metrics for Evaluating Security & Trust in Autonomous Human Machine Teams & Systems

Michael Mylrea[1] and Nikki Robinson[2]

[1]University of Miami, Institute of Data Science and Computing, Coral Gables, FL 33146 USA
[2]Capitol Technology University, Laurel, ME 20708, United States

ABSTRACT

AI innovation is advancing like wild fire. Advances in functionality, accessibility and performance of user friendly Large Language Models (LLMs) and Generative Artificial Intelligence (GAI) solutions have also increased adoption of AI. These gains have been accompanied by new opportunities (e.g. markets, services, insights). In order for AI gains to be ethical, sustainable and resilient to malicious exploitation by bad actors, its power must be guided by ethics. Adoption of secure and trustworthy AI frameworks are needed to create ethical guardrails in the design and management of AI. In absence, small sparks of general artificial intelligence may soon kindle into innovative flames (e.g., human-level capabilities, brain-computer interfaces, sentient machines) that are difficult to secure, trust and control. The following research examines opportunities and challenges for an AI Trust Framework and Maturity Model (AI-TMM) to improve metrics for evaluating trust and security in AI. These improved metrics may help improve trust and establish ethical guardrails in the design, management and governance of AI and other emerging technologies.

Keywords: AI trust framework maturity model, Non-proliferation, Cybersecurity, Trust, Resilience, Privacy, Autonomy, Generative AI, Large language models, Artificial intelligence, Machine learning, ChatGPT

INTRODUCTION

Improving metrics and the process for evaluating security and trust in AI may help optimize performance of Autonomous Human Machine Teams & Systems (A-HMT-S). Trust in AHMT-S involves the perception of dependability, competence, transparency, and effective communication between humans and machines (Mylrea 2023). This article highlights how the AI-Trust Framework and Maturity Model (AI-TMM) improves metrics for evaluating security, transparency, explainability in AI. For this study, the modular framework incorporates Google's SAIF metrics and adds a maturity model methodology to measure security and trust in a large language model (LLM) chat bot. Validation of AI-TMM's efficacy highlights that is provides a repeatable and explainable evaluation to facilitate the design and

management of Trustworthy AI. This research highlighted a number of security and trust gaps in existing AI solutions as well as mitigations that include designing and managing AI with more ethical guardrails. Providing a user friendly, accessible framework to address these gaps is critical to reducing the risks of exploiting and manipulating AI to harm civilization.

Securing AI through its lifecycle, from development to deployment, is required to improve trust between human-machine teams (Mylrea, 2022). This requires improved integrity of data lineage, monitoring detailed records of the origin, transformations, and movement of data throughout its lifecycle. Improved access control policies can also help ensure that only authorized individuals or systems can interact with the AI model and its associated data (Hansen and Venables 2023). As teams apply AI advances to improve autonomy and effectiveness of critical systems, from autonomous defence platforms to AI driven bioreactors, improved security and trust requires improved metrics to evaluate systems. Are these systems performing as expected? Are they making secure and dependable decisions, and effectively fulfilling their assigned roles within the team? (Lee and See 2004). Security and trust metrics should be holistic and include people, process and technology considerations around dependability and transparency (Mylrea 2023). Effective visualization and communication of the proposed AI-TMM also requires acknowledging the diverse needs and perceptions of human operators in the loop (Mosqueira-Rey et al. 2023).

Trust is critical to improve performance between human-machine teams. It is fostered through factors such as consistent and reliable performance, transparent decision-making processes, clear and comprehensible channels of communication, shared decision-making, productive collaboration, and a shared sense of responsibility and accountability (Aldridge and Bethel, 2023). While hints of general intelligence have been highlighted in some GAI and LLM outputs, there is a misconception that AI has human level intelligence. Effective application of an AI-TMM requires understanding the strengths and weaknesses, capabilities and weaknesses in how human-machine teams collaborate, communicate and compete (Mylrea 2023a).

SECURE AI FRAMEWORK (SAIF)

Cyber threats exploit human and machine vulnerabilities to reduce trust. To help mitigate this gap and improve trust between A-HMT-S, industry has developed various comprehensive frameworks to ensure the security of its AI systems (NIST Cybersecurity Framework). Google recently launched a Secure AI Framework (SAIF), focusing on proactive measures to identify and mitigate potential security risks associated with AI. It includes multiple layers of security, such as secure data handling, robust development and deployment processes, continuous monitoring, and collaboration with external security researchers. These are critical security measures that can help improve trust in AI (Google, 2023). Combining AI-TMM's maturity model approach with SAIF may help improve responses to threats that are complex, non-linear and evolving by providing a repeatable, iterative process to assess and mitigate trust gaps.

Table 1. SAIF security and trust principles, goals & gaps.

Security & Trust Principles	Goals	Security & Trust Gaps Targeted
Resilience	Enhance robust security measures across the AI ecosystem to ensure a strong foundation.	AI development lifecycles include systems and data that lack basic encryption, authentication and secure by design principles
Transparency	Extend detection and response to bring AI into an organization's threat model	Lack of threat modelling, identification and detection vulnerabilities and threats targeting critical data inputs and outputs of AI systems
Reliability	Employ automated defence mechanisms that can respond to existing and emerging threats.	Lack of secure by design systems that automate defence mechanisms and neutralize emerging threats.
Explainability & Consistency	Establish standardized security controls at the platform level to maintain consistent security practices throughout the organization.	AI controls are not always harmonized and consistently applied through the lifecycle creating gaps in trust and security. Google has been extending these protections through various secure-by-default protections (e.g., Vertex AI, Security AI Workbench, Perspective API)
Agility	Modify controls to tailor mitigations and establish agile feedback loops for the deployment of AI technologies.	AI systems lack continuous monitoring, detection and protection from attacks such as poisoning training data and algorithms
Responsible Use	Consider the risks associated with AI systems within the broader context of surrounding business processes.	AI systems often lack contextualization for responsible use. Improved examination of business use and application, including anomaly detection, operational behaviour monitoring and data lineage is needed.

SAIF facilitates harmonization of security controls by drawing from established security best practices, such as supply chain control, testing, and thorough review processes. SAIF takes into account emerging trends and risks associated specifically with AI systems and provides an effective approach to respond (Hansen and Venables 2023). The core elements of SAIF include the following security and trust principles, goals and gaps (see Table 1).

COMBINING AI TRUST FRAMEWORK & MATURITY MODEL (AI-TFMM) WITH THE SECURE AI FRAMEWORK (SAIF)

To improve SAIF's metrics of evaluation, applicability, and repeatability a maturity model methodology is applied by incorporating AI-TMM (see Table 2). Maturity models utilize weighting and measurement techniques to assess specific controls and enhance repeatability. This approach is particularly beneficial when evaluating performance in situations where the

Table 2. Applying a maturity model approach to the secure AI framework (SAIF).

Metrics of Evaluation with SAIF	Improved Metrics of Evaluation By Applying AI-TFMM
Level of maturity is measured concerning security principles. Creates susceptibility to binary yes or no answers	Level of maturity is measured with both security and trust principles and includes various levels of maturity
Outlines the desired state or goal for security principles, but could potentially be limiting in targeting improvements not included in the framework	Outlines the desired state or goal for security and trust principles in a way that provides a clear roadmap for improvement and evaluating those gains with weighted metrics
Takes measures to prevent unethical development and/or application of AI	Improves quantification of measures to prevent unethical development and/or unethical application of AI
Facilitates communication between internal and external stakeholders. However, without clear weights or maturity indicator levels this can create ambiguity and uncertainty	Improves fidelity of communication between internal and external stakeholders

adoption of security controls or privacy measures cannot be simplified into a binary pass or fail outcome (Mylrea et al. 2017). Moreover, human and organizational end users have different levels of resource constraints in realizing their security and trust goals. Instead of having a yes or no metric of evaluation, the AI-TMM provides different levels of maturity to help plan and execute resource allocation. Thus, managers, developers and other stakeholders can apply AI-TMM to better manage, design and govern AI applications in their enterprise. Moreover, its modular construct enables end users to easily incorporate other frameworks to realize their own security, governance, risk and compliance goals.

The following table highlights how combining AI-TMM with Google's Secure AI Framework (SAIF) help improve its metrics of evaluating security and trust in AI. Improves AI-TMM's maturity model methodology of security assessment and produces outputs that are intuitive and easy to communicate to internal and external stakeholders. Combining the frameworks also helps create common taxonomy for AI stakeholders to harmonize security and trust controls.

AI-TMM's methodology is underpinned by a maturity model with four Maturity Indicator Levels (MIL), MIL0 through MIL3, which apply independently to each domain principle (Mylrea 2023). MILs are applied to specific controls for each of the AI Trust pillars that make up the framework (see Figure 1).

AI-TMM is modular and readily incorporates others frameworks and regulatory security and compliance controls. Applying its maturity model methodology to SAIF would add a maturity indicator level (MIL) as well

Figure 1: AI trust framework key pillars (Mylrea 2023b).

as a more holistic approach in examining control from a people, process and technology perspective. As we advance towards general artificial intelligence, ethical guardrails will become increasingly important. Trust in A-HMT-S will increasingly require both humans and machines to trust each other. Thus, it is important to add a holistic people, process, and technology approach as well as a maturity model methodology. A brief explanation of the four Maturity Indicator Levels (MIL):

- **Fully Implemented:** Provides a Maturity Indicator Level Score or Weight of 3: Requires that the control is documented, managed and continuously validated through testing.
- **Largely Implemented:** Provides a Maturity Indicator Level Score or Weight of 2. Requires that the control is documented and actively managed by a human in the loop, but is *not* continuously validated through testing.
- **Partially Implemented:** Provides a Maturity Indicator Level Score or Weight of 1. Requires that the control is documented, but not actively managed by a human in the loop, nor continuously validated through testing.
- **Nothing Implemented:** Provides a Maturity Indicator Level Score or Weight of 0. Suggests no documentation, management or testing of a control.

The levels of maturity indicators (MILs) are applied independently to each principal domain, allowing AI-TMM users to operate at different MIL ratings for different domains. This means that an organization may be functioning at MIL2 in one domain, MIL3 in another domain, and MIL0 in a third domain. The MILs within each domain are cumulative, meaning that in order to attain a specific MIL in a domain, the organization must fulfil all the practices within that level and its preceding level(s). For instance, to achieve MIL2 in a domain, the organization must perform all the practices in MIL1 and MIL2. Likewise, to reach MIL3, the organization would need to complete all the practices in MIL1, MIL2, and MIL3 (Mylrea et al. 2017) (Mylrea 2023). Improving maturity level of critical controls can improve trust and security

Figure 2: AI-TMM methodology (Mylrea 2023).

for A-HMT-S. However, due to diversity of resources, goals and even potential business impacts if a gap is exploited, optimal MIL levels will vary across organizations.

USE CASE: APPLYING AI-TFMM TO ASSESSMENT THE MATURIY OF ACME COMPANY'S AI/ML MODEL RISK MANAGEMENT

The following use case applies the AI-TMM methodology to an illustrative use case at ACME CO, a fictitious AI technology company that a released a popular, new LLM chat bot. Regulators, consumers and privacy advocates have questioned the security and trust of ACME's algorithms design and management. Large industry players have prohibited the use of ACME's solutions because of the lack of access controls to prevent sensitive data loss as well as issues related to transparency of data lineage and explainability on how the algorithm is arriving at its conclusions. In response, ACME hired a third party to apply AI-TMM to assess and evaluate where there are critical trust gaps in the design and management of their chat bot. The Chief Information Security Officer inquired how easily the framework could include security controls from SAIF to complement their own security goals and their Google Cloud environment.

The AI-TMM methodology requires the following steps (See figure 2). Typically, a subset of controls from all 7 trust pillars (See figure 1) are evaluated, however that is beyond the scope of this use case. It is also important to note that this is not a static process. Security is not an end state, but a continuing process of improvements that help foster a culture of security. To realize that goal effectively, it is improve to include a multidisciplinary set of stakeholders (Mylrea 2017).

Step 1: Perform Evaluation: First, perform an evaluation based on the desired framework controls. While this use case tests four controls that are

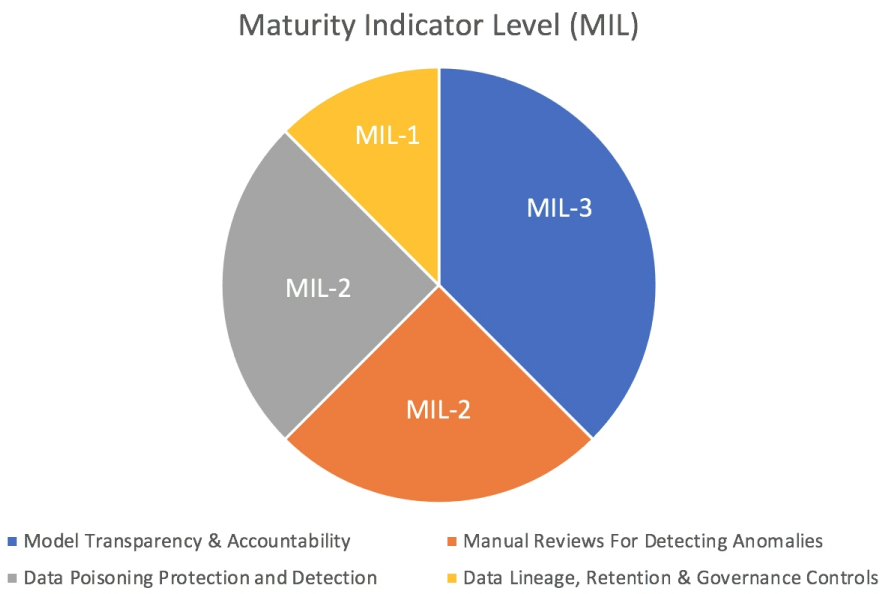

Figure 3: Maturity indicator level assessment of ACME's AI data management controls.

part of SAIF's AI/ML model risk management controls, future applications could include NIST AI Risk Management Framework and the ISO/IEC 42001 AI Management System Standard, which is the industry's first certification standard for AI. These standards draw heavily from the security principles outlined in the NIST Cybersecurity Framework and the ISO/IEC 27001 Security Management System (NIST 2021).

The scope of the evaluation demonstrated is limited to SAIF's AI/ML Model Risk Management controls. AI-TMM's methodology is applied to the illustrative use case examining the controls of ACME's chat bot.

- ***Model Transparency & Accountability:*** ACME has documented, managed and continuously validated through testing
  - ***Fully Implemented - Maturity Indicator Level*** = 3
- ***Manual Reviews For Detecting Anomalies:*** ACME documented this control which is actively managed by a human in the loop, but is *not* continuously validated through testing.
  - ***Largely Implemented - Maturity Indicator Level*** = 2
- ***Data Poisoning Protection and Detection:*** ACME documented this control which is actively managed by a human in the loop, but is *not* continuously validated through testing.
  - ***Largely Implemented - Maturity Indicator Level*** = 2

Table 3. Applying applies AI-TMM metrics of evaluation to an illustrative use case that leverages Google's SAIF controls for AI model risk management.

Secure Ai Framework (Saif) Control: Ai/Ml Model Risk Management	Maturity Indicator Level (MIL) Score	Potential Mitigations To Improve Trust & Security
Model Transparency & Accountability	3	**Trust Mitigation:** Perform continuous assessment and audits on AI algorithms, data inputs, decision-making processes, and outputs. These audits help verify the transparency and fairness of the AI models, identify potential biases or ethical concerns, and ensure compliance with relevant regulations and standards (Floridi et al. 2018). **Security Mitigation**: Improved access controls, such as employing authentication, authorization, and privilege management techniques, can help improve maturity of access policies, ensuring that only authorized individuals or systems can interact with the AI model and its associated data. This helps prevent unauthorized access, tampering, or misuse of the AI systems (Hansen and Venables 2023).
Manual Reviews For Detecting Anomalies	2	**Trust Mitigation:** XAI techniques such as rule-based systems, decision trees, or attention mechanisms can be applied to provide interpretable explanations, helping users, regulators, and stakeholders gain insights into the AI model's behaviour and build trust in its capabilities (Rosenfeld 2021). (Rudin 2019). **Security:** Ensuring the data's integrity, authenticity, and quality, organizations can minimize the risks of malicious data manipulations or adversarial attacks that aim to deceive the anomaly detection system. Proper data validation and sanitization techniques, such as outlier detection, data cleansing, and data integrity checks, help maintain the accuracy and reliability of the anomaly detection model, improving its security (Ahmed, Naeem, & Ghafoor, 2020)
Data Poisoning Protection and Detection	2	**Trust Mitigation:** Algorithmic transparency and auditing enable organizations and users to have a better understanding of how AI models operate, instilling trust and facilitating the detection of anomalous behaviours or malicious manipulations (Doshi-Velez & Kim, 2017). **Security:** Data quality monitoring and anomaly detection help mitigate the risks of using compromised or poisoned data in AI training, enhancing the security and reliability of the AI system (Garcia-Gasulla et al., 2020).

Continued

Table 3. Continued.

Secure Ai Framework (Saif) Control: Ai/Ml Model Risk Management	Maturity Indicator Level (MIL) Score	Potential Mitigations To Improve Trust & Security
Data Lineage, Retention & Governance Controls	1	**Trust Mitigation:** Improving data lineage documentation by capturing and maintaining detailed records of the origin, transformations, and movement of data throughout its lifecycle. By providing clear and transparent information about the data's lineage, including its sources, processing steps, and potential modifications, organizations can enhance trust in the data's accuracy, reliability, and compliance with governance policies (Lenzini et al., 2021). **Security:** Improving security of data lineage, retention, and governance controls via improved access controls and encryption mechanisms. Enforcing strong access controls, organizations can maintain the integrity and confidentiality of data lineage information. Additionally, encrypting sensitive data throughout its lifecycle, including during retention, storage, and transmission, provides an extra layer of protection against unauthorized access or data breaches (Khan, Khan, & Karim, 2019).

- ***Data Lineage, Retention And Governance Controls:*** ACME documented this control, but it is *not* actively managed by a human in the loop and it is *not* continuously validated through testing.
  - ***Partially Implemented - Maturity Indicator Level*** = 1

Step 2: Analyse Identified Gaps: Consider gaps in the context of organizational goals as well as potential impacts if those gaps or vulnerabilities are exploited. AI systems are applied to a diverse set of use cases, from anomaly detection in high assurance systems to improving efficiencies for menial tasks. For this reason, it is important to better understand security and trust frameworks in the context of resource availability, business impact, security, safety and risk.

Step 3: Prioritize and Plan: List gaps and potential consequences. Note organizational constraints. If a particular business impact or risk is unacceptable it is important to prioritize and plan effective allocation of resources to reduce associated risks. Once actions are identified to address gaps it is important to perform a cost-benefit analysis (CBA) on actions and priorities.

Step 4: Implement Plans: Based on AI-TFMM application to SAIF, the metrics of evaluation may facilitate more effective resource allocation to buy down risk in a measurable and repeatable way. Security and trust in AI are

not static. It is important to track progress and revaluation periodically in response to changes.

Improving metrics for AI security and trust evaluation is crucial for assessing effectiveness, enabling benchmarking and best practices, promoting transparency, and fostering trust among stakeholders (Liu et al., 2021). AI-TMM improves robustness of evaluation metrics to empower organizations applying ethical guardrails to AI. Well-defined controls and a user friendly process of evaluation facilitates identification of trust gaps and mitigations. These metrics also provide a standardized framework for comparing different AI systems and approaches, facilitating benchmarking and promoting best practices in the field. As shown in the use case, AI-TMM quickly highlighted opportunities to improve ACME's security and trust via improved transparency and accountability of it chat bot.

AI-TMM combined with other ethical AI frameworks, such as SAIF, can also empower stakeholders, including users and regulators to assess the level of security and trust in a given AI system. This fosters trust, promotes responsible AI deployment, and helps ensure that AI technologies are developed and used in a manner that aligns with ethical principles and societal expectations. When applying these frameworks with a maturity model methodology, it is important to align with business objectives and the organization's ethical AI strategy. Pursuing the highest MIL in all domains may not be the most advantageous approach. Instead, companies should carefully evaluate the costs and benefits associated with achieving a particular MIL. It is crucial to document and address any areas where gaps in ethical principles exist, employing appropriate mitigation strategies (Mylrea 2023).

CONCLUSION

This research examined how the AI Trust Maturity Model Framework (Mylrea 2023) can be applied to improve trust in A-HMT-S. An illustrative use case highlighted a number of opportunities and challenges to improve metrics of evaluation for trust in AI systems. Establishing a framework to improve measurements of security and trust in AI is needed to secure, sustain and advance the performance of A-HMT-S. It is important that these frameworks are supported by methodologies that are modular and can include the security and compliance controls required by their industry. This research highlighted how AI-TMM could easily incorporate Google's Secure AI Framework (Hansen and Venables 2023) to focus on specific security controls.

Future research should examine security and trust challenges with other emerging technologies, such as cyber and nuclear weapons. What other frameworks can be incorporated into the AI-TMM to improve ethical use and safeguards of emerging technology? What worked? What failed? What can we apply to harness AI to improve civilization? What can be learned from nuclear non-proliferation? Mutually assured destruction deterred abuse in part because the cost inevitably outweighed the opportunity. However, AI is introducing new opportunities for malicious actors, from social engineering with deep fakes to cyber weapon development and zero day exploit enumeration with LLMs. Unlike nuclear weapons which have signatures in their

development and delivery that can often be attributed to certain refinement locations and technology developers, AI adversary can easily obfuscate their identity. How can the data fuelling AI and it's ecosystem be better protected? Can AI protect itself via secure by design and ethical use guardrails?

Certainly, improved metrics of evaluation in security and trust can help in the design, management and governance of AI and other emerging technologies. Applying AI-TMM to an AI use case highlighted both opportunities and gaps of applying a maturity model methodology. This user friendly evaluation of trust and security metrics can incorporate other related framework to help shape the development of future algorithms that are safe and beneficial to humanity. Applying the AI-TMM to real world use cases through the AI lifecycle can also help improve AI safety and encourage ethical regulatory oversight of AI development. While the assessment was conducted on a fictitious corporation, recent real world incidents have revealed gaps in transparency, repeatability, and other ethical principles in a number of AI solutions (Syme 2023).

Gaps that undermine trust between human machines teams jeopardize the long term sustainability and gains that AI can bring to make the world a better place. AI-TMM improve the ability to measure, evaluate and mitigate these gaps. This is timely as AI innovation is rapidly advancing. New opportunities for AI are on the horizon to give impetus to cyber, physical, biological convergence that will blur the lines between humans and machines. When manipulating biological systems that are complex, stochastic and difficult to understand, secure and control (Mylrea et al. 2022), ethical guardrails are imperative. With the appropriate safeguards, we could prolong and improve our lives and civilization. Getting it wrong by applying black box AI solutions to drive synthetic biology production could inadvertently unleash a biological weapon of mass destruction – this should not be an option (Mylrea 2023b).

ACKNOWLEDGMENT

Dr William Lawless for his ongoing encouragement and mentorship through the Association for the Advance of Artificial Intelligence and Stanford University's Spring Symposium. Dr Nikki Robinson, my PhD dissertation Chair, for sharing her expertise in securing emerging technologies.

REFERENCES

Ahmed, I., Naeem, M., & Ghafoor, A. (2020). A review of anomaly detection in supervised and unsupervised machine learning paradigms. Concurrency and Computation: Practice and Experience, 32(24), e5582.

Aldridge, A. L., & Bethel, C. L. (2023, March). M-OAT Shared Meta- Model Framework for Effective Collaborative Human-Autonomy Teaming. In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (pp. 663–666).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

Garcia-Gasulla, D., Ahmadi, S., Cano, A., & De Capitani di Vimercati, S. (2020). Poisoning attacks against machine learning: A survey. arXiv preprint.arXiv:2006.08190.

Hansen and Venables. (2023). Introducing Google's Secure AI Framework. Accessed at https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/

J. D. Lee and K. A. See, 2004, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, 46(1) (2004), 50–80.

Lawless, W. F., Sofge, D. A., Lofaro, D., & Mittu, R. (2023). Interdisciplinary approaches to the structure and performance of interdependent autonomous human machine teams and systems. Frontiers in Physics, 11, 136.

Lenzini, G., Etalle, S., Huygens, C., & Verheul, E. (2021). Data lineage: what it is and how to use it. Data Intelligence, 3(1), 50–58.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56(4), 3005–3054.

Mylrea, M., & Robinson, N. (2023). AI Trust Framework and Maturity Model: Improving Security, Ethics and Trust in AI. Cybersecurity and Innovative Technology Journal, 1(1), 1–15.

Mylrea, M. (2023, May). Breaking Bad: Biological Internet of Things. S4 Conference on ICS Security, Miami, FL. Accessed here https://www.youtube.com/watch?v=8tskxSq592A

Mylrea, M., Fracchia, C., Grimes, H., Austad, W., Shannon, G., Reid, B., & Case, N. (2022). BioSecure Digital Twin: Manufacturing Innovation and Cybersecurity Resilience. In Engineering Artificially Intelligent Systems (pp. 53–72). Springer, Cham.

Mylrea, M., Nielsen, M., John, J., & Abbaszadeh, M. (2021). Digital Twin Industrial Immune System: AI-driven Cybersecurity for Critical Infrastructures. Systems Engineering and Artificial Intelligence, 197–212.

Mylrea, M., Gourisetti, S. N. G., & Nicholls, A. (2017). An introduction to buildings cybersecurity framework. In *2017 IEEE symposium series on computational intelligence (SSCI)* (pp. 1–7). IEEE.

National Institute of Standards and Technology. (2021). NIST AI Metrics Suite. Retrieved from https://pages.nist.gov/ai-metrics/

Rosenfeld, A. (2021). Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems* (pp. 45–50).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Syme, P. Apple is working on its own AI large language model and restricting employees from using ChatGPT over privacy concerns. Business Insider. May 19, 2023.