Securing AIML Systems in the Age of Information Warfare

Susanna Cox / CX7

Critical Alliance

April, 2022

Introduction

Concerns around heightened cybersecurity risk environments naturally include so-called *Hybrid Warfare* and, specifically, *Information Warfare* (IW) (Libicki 2020; Beskow 2020; Lin & Kerr 2017; Zannettou et al, *Proceedings of the 10th ACM Conference on Web Science* 2019; Connell & Vogler 2017). Increasing scrutiny around the cyber resilience of computer systems in a contested Information Warfare environment led in early 2022 to the March 15th passage and signing of bipartisan US legislation to require cyber incident reporting by organizations in what are deemed "critical infrastructure" sectors (Conger 2022, Conger 2022; The White House 2022). This cyber incident reporting legislation is not all-encompassing; rather it initiates the assembly of critical information about cyber attacks against US companies into a central repository that can be accessed by US intelligence and law enforcement agencies. At the time of writing it is unclear how specific reporting requirements will take shape.

This paper aims in part to provide a practical and immediately actionable framework to aid organizations in early compliance with regulatory requirements. Rather than survey the vast and constantly expanding cybersecurity field, this paper focuses on assessing and pragmatically improving the *cyber resilience of artificial intelligence/machine learning systems* within the contested *Information Warfare* environment, as well as formalizing and operationalizing the production of artifacts for regulatory compliance.

Artificial Intelligence systems are, at their core, information processing systems. As such they have particular vulnerabilities, and due to their propagation now in virtually every aspect of society (Chopra & Singh 2018; Ntoutsi et al 2020), a special set of potential downstream consequences for failure.

Mass manipulation of social media algorithms is now possible and in many cases even convenient for non-state actors (Beskow 2020; Mejías & Vokuev 2017). The use of bots, as well as surrogates (sometimes called "orange actors") in cyber warfare–potentially including state or criminal alliances–has been demonstrated effective in disrupting networks at scale (Im et al 2020; Shao et al, *PLoS ONE* 2018). Bots and surrogates may work to mask attribution in coordinated cyber campaigns (Keller et al 2019; Snyder et al 2020); the proliferation of AIML-backed platforms means that mass manipulation of social networks and their algorithms can cause unsuspecting individuals to become participants in a message's spread (Badawy, Lerman & Ferrara 2018; Ahmed 2021; DiResta 2021), further complicating the task of attribution. The rapid increase of production AIML systems for both offensive and defensive security purposes generates a *low-risk, high-reward* environment for potential attackers.

There has been much discussion around how algorithms used in social media and other platforms aid in the dissemination of certain messages over others, and how these features are exploitable in production by bad actors (Zannettou et al, *Companion Proceedings of The 2019 World Wide Web Conference* 2018; Shao et al, *Nature Communications* 2018), the probabilistic distributions of various types of messages most likely to be amplified by their relative position along a political spectrum (Chowdhury, Belli, and Lamar 2021, and Chowdhury, Belli, & Lamar 2021; Huszár et al 2022), and how this so-called *algorithmic amplification*, in conjunction with the lack of transparency from the platforms themselves with regard to algorithmic decisions in both design and use phases, is readily exploitable by maleficent actors (European Union Counter-Terrorism Coordinator 2020; Dumbrava 2021) with the potential to disrupt democratic norms (European Parliament 2021; Christodoulou & Iordanou 2021).

These discussions are rapidly moving beyond the theoretical, as illustrated by an increasing number of calls from diverse segments of society to legislate new regulations for AIML platforms (Lee, Resnick, & Barton 2019; Adesina, Kearns, & Roth 2020; Mökander & Axente 2021; European Parliament 2021) and hold organizations legally accountable for the messages they propagate (Kirtley 2022; Kornbluh 2022; Diamantis 2020; Barocas & Selbst 2016; Koshiyama et al 2021). There is also literature around effective network and cloud security applications of AIML; referencing the state-of-the-art or current technologies here would prove ineffective as these technologies are evolving rapidly.

For these reasons, a generalizable framework to allow organizations to probe their security dependencies and develop decision criteria for resource allocation *vis-a-vis* AIML information processing systems and Information Warfare (IW) is needed.

Frontier developments in AIML research, however--including the recent proliferation of ethics and bias frameworks--demonstrate that frameworks are not enough. Ethics and auditing frameworks are widely acknowledged as efforts to reduce the harm of AIML systems (Raji & Buolamwini 2019; Bandy 2021), but for many organizations, their implementation remains elusive (Jobin, Ienca & Vayena 2019; Stahl et al 2022). Practitioners often find broad latitude for interpretation in such systems, and organizations vary in their implementations (Ibáñez, Camacho & Olmeda 2021). Frameworks tend to be either too high-level and thus difficult to implement in practice, or on the opposite end of the spectrum, too specific to be generalizable (Morley et al, *Minds Mach* 2021); and while many frameworks exist, there is considerable overlap among them (Ryan et al 2021). Specifically, practitioners indicate difficulty with operationalizing bias-aware development due in part to the sheer number of high-level frameworks and the comparative lack of guidance on implementation (Morley et al, *AI & SOCIETY* 2021).

It is also important to consider that implementation is never free, and the cost of experimentation must be factored into the development of a bridge between the high-level/theoretical, and actionable production pipelines. Organizations run the risk of ethics-washing and other pitfalls if ethics frameworks are implemented poorly (Floridi 2019), rendering any such efforts potentially meaningless and the resources invested in their development wasted. With little guidance on practical implementation, teams are often left on their own to bridge this gap. Despite organizational, social, and even government interest in AIML ethics reform, the Pew Research Center warned that robust industrial implementation of AIML harm reduction could take years to come (2021).

The focus of this paper is thus both measuring and improving cyber resiliency of artificial intelligence/machine learning (AIML) systems in an adversarial Information Warfare (IW) environment, via actionable *Machine Learning Operations* (MLops) pipelines, in production. The specific aim of this paper is to provide a practical, workable blueprint for immediate steps AIML practitioners and organizations can take to both assess and increase the resilience of their systems to IW attack vectors, with an awareness and integration of MLOps best practices in a rapid-deployment environment.

MLOps-First Approach

MLOps *(machine learning operations) is "a set of standardized processes and technology capabilities for building, deploying, and operationalizing ML systems rapidly and reliably"* (Salama, Kazmierczak, and Schut 2021). MLOps is analogous to the DevOps and DataOps fields in its focus on rapid deployment and scalability through the formalization and operationalization of mission-critical data processing actions:

> *[MLOps] advocates formalizing and (when beneficial) automating critical steps of ML system construction. MLOps provides a set of standardized processes and technology capabilities for building, deploying, and operationalizing ML systems rapidly and reliably. MLOps supports ML development and deployment in the way that DevOps and DataOps support application engineering and data engineering (analytics). The difference is that when you deploy a web service, you care about resilience, queries per second, load balancing, and so on. When you deploy an ML model, you also need to worry about changes in the data, changes in the model, users trying to game the system, and so on (ibid).*

MLOps best practices, such as formalizing a process by which new models are reviewed before being deployed, increase both resilience and security of a system by providing a baseline to return to in case of a problem–and hopefully by ensuring that problematic models are never deployed in the first place.

Yet despite these and many other clear benefits, many practitioners in the industry remain slow to adopt MLOps best practices and continue to engage in ad hoc experimentation which is difficult to scale and often results in very little actual model deployment. A report by Deloitte published in the Harvard Business Review (Ronanki & Davenport 2017) listed integration issues (and the talent to remedy these) as chief bottlenecks in producing actionable AIML, and a 2020 McKinsey study found that standardized, scaleable development processes were a key differentiator in high-performing AIML organizations (Balakrishnan et al).

Conversely, larger organizations have had difficulty responding effectively to adversarial algorithmic manipulation in part due to the scale of their deployment and degree of automation, and automation of AIML systems is likely to increase as more stakeholders adopt an MLOps approach to development. The scale and speed of deployment of modern AIML systems–combined with the near-ubiquity of their adoption at nearly every layer of the socio-technological stack–indicates a need for urgency in the development of scalable, MLOps-intelligent cyber readiness assessment metrics.

MLOps is a relatively nascent discipline. At the time of this writing, there are many papers in the field, and current research often focuses on the unification and taxonomy of the varying methodologies and lexicons. However, there are few references to AIML system cyber resiliency, either direct methodology or evaluation. There is thus an urgency to holistically integrate cyber resilience metrics which are specifically tailored to AI applications, and MLOps-aware for ease of adoption in industry and government.

The cyber metrics framework presented in this paper is adapted in part from a report prepared by RAND Project AIR FORCE (PAF), a division of the RAND Corporation, and the U.S. Air Force's federally funded research and development center. The report, *Measuring Cybersecurity and Cyber Resiliency* (Snyder et al 2020), was originally created to draft a framework for scoring the cyber readiness of United States Air Force weapons systems across two distinct vectors, *cybersecurity* and *cyber resiliency*. This paper uses these terms similar to the Department of Defense definitions referenced in the PAF report: referring to a system's ability to withstand attack (*cybersecurity*), versus its ability to recover to a stable state post-incident (*cyber resilience*).

Artificial intelligence systems are not necessarily weapons systems, although some are. However, even those in use in non-defense applications are arguably so embedded in critical socio-technical systems that assessing their ability to be resilient to information warfare (IW) attack vectors is crucial to national security. This paper aims to apply lessons from the RAND PAF weapons systems cyber readiness scoring research to create a baseline cyber/IW readiness scoring framework for AI systems in production.

Machine learning/artificial intelligence systems, referred to in this paper as *AIML*, rely on information–data–to make what humans might term *inference*. The extent to which AI systems are “intelligent” is the extent to which they have “learned” (in other words, *been trained on*) some subset of data, *and* the extent to which the selected data subset mirrors real life. Without good data, there is no possibility for optimal AI, and AI systems affect critical aspects of everyday citizens’ lives on levels too numerous to list here in detail, from military to transportation, banking, and finance. Information Warfare can also be understood as Data Warfare; thus AI practitioners must be increasingly vigilant in heightened IW threat environments.

Red vs Blue Formulation

This paper utilizes the Red/Blue *attack/defend* formulation typical in cybersecurity settings. While this framework may be unfamiliar to AIML practitioners, the two-team model provides useful game-theoretical analysis points in an adversarial security environment.

In the two-team model, Red represents the offensive/attacking team, and Blue represents the defensive/organizational security team. Additional color teams for AIML Ops have been proposed with potential benefits for organizational and managerial analysis (Kalin, Noever & Ciolino 2021); these are outside the scope of this paper, which refers exclusively to the Red/Blue security formulation.

Red’s Attack Path

The RAND PAF report defines a high-level attack path for Red involving four basic, non-sequential attributes of any successful Red offensive–or in the language of the report, “actions” that Red must carry out in order to complete a successful attack. These attributes are *access*, *knowledge*, *capability*, and *impact*:

> *At the highest level, Red must do four things: (1) access the system in question (access) and (2) know enough about it to execute an attack (knowledge), and (3) have the resources and capability to carry out the attack (capability) and (4) create an effect that has significant negative mission repercussions to the defender (impact)* (Snyder et al 2020).

This analysis gives rise to a defined attack path for Red with a Boolean relational structure among the elements: at the highest level, the elements are connected by Boolean *and* statements, indicating that Red must complete all actions/attributes in order to be successful at disrupting Blue’s mission operations; Boolean relationships become more varied at increasing levels of decompositional granularity (fig 1).

While these four high-level categories must all be achieved for Red to claim a successful attack against Blue, within each category, Red has a series of options, with potentially differing relationships at various levels of decomposition. For example, within the access metacategory, Red may achieve this requirement through any of several means, such as through supply chain access or via insider threat. Because Red generally needs only one viable vector to achieve access, these potential vectors are connected by Boolean *or* statements. This is contrasted with the *knowledge* metacategory, where although Red may attain knowledge of Blue’s systems through any of a number of means–connected by Boolean or statements–there is an additional constraint, labeled *currency*, which Red must satisfy in order to achieve success in this category: *“Red needs to act on knowledge it has collected before Blue makes any relevant changes to the targeted weapon system”* (Snyder et al 2020). While vectors for Red to gain knowledge such as publicly shared information via open-source intelligence (OSINT) and exfiltration (such as espionage) are connected by statements indicating that Red can utilize any one of these, all knowledge must maintain *currency* to be used effectively against Blue. In other words, Red may gain publicly available knowledge via OSINT or privately held (secret) knowledge via exfiltration, *and* that knowledge must be current, in order to impact Blue.

Properly applied, this relational framework can suggest areas of potential vulnerability to Red attacks, as well as paths of most efficient resource allocation for Blue:

> *This framework emerges from the fact that some actions that Red must take must all be done, and, in other cases, Red has options. The four high-level actions…are linked by Boolean and statements because each must happen or the attack is thwarted. Because Red must do all four to some degree, reducing the likelihood of success of one or more of them increases the likelihood of interrupting the attack path. The greater the number of these four actions that the*

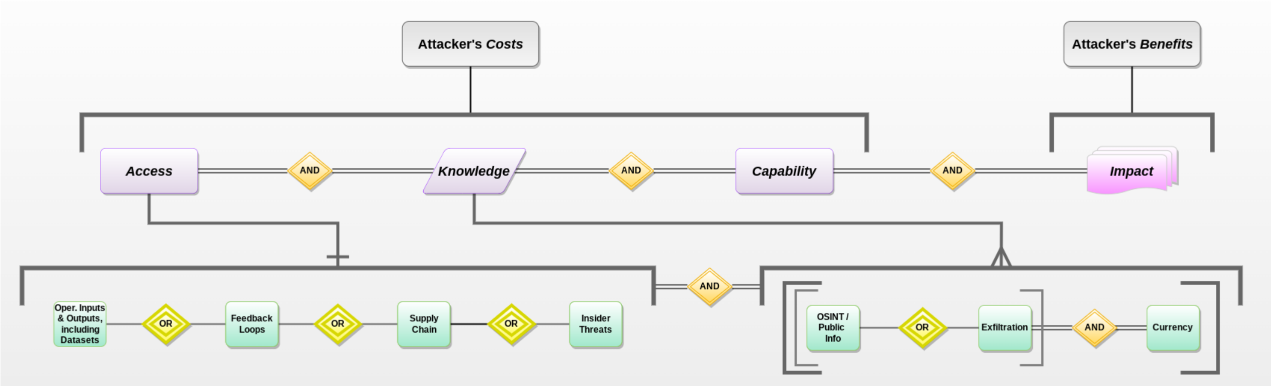


Fig 1: Boolean relationships among elements of Red’s attack vectors

defender can inhibit, and the greater the confidence in inhibiting each, the more difficult the job of the attacker and the more survivable and effective the system is in that cyber environment (ibid.).

The strength of this system as applied to algorithmic models & artificial intelligence systems lies in the Boolean relationships among Red's attack paths and methods. The framework is designed to allow a thorough analysis of any system through the decomposition of Red attack vectors while avoiding being either overly prescriptive or quantitative in a field where applicable technologies, and the bleeding edge itself, are constantly shifting *(ibid.)*.

Special Considerations for Red-Blue Attach Path Analysis of AIML Systems

Red's four requirements for successful attack–*access*, *knowledge*, *capability*, and *impact*–take on new significance when applied to AIML systems in an IW context. For example, an attacker attempting to exert influence over a public-facing system may gain *knowledge* of the models that power it by probing the system to test for patterns in behavior (Chase, Ghosh & Mahloujifar 2021); proprietary knowledge of the models' code is not needed. Public manipulation of social media platforms' algorithms to self-promote is a non-malicious example of how entities may influence a model's outputs without direct access to the model itself.

Another category in which the calculus differs significantly from more traditional cybersecurity attack surfaces is the *access* requirement. Because AIML systems are only as intelligent as the data on which they are trained, an attacker does not need access to models themselves to impact their outputs–only to their datasets. When these datasets are created from publicly available sources–particularly non-curated, mass-scraped–Red's *access* is defacto & definitionally guaranteed. This consideration applies to AIML systems that utilize one-time or continuous web scraping, as well as those using models pre-trained on public datasets. Continuous data scraping entails its own vulnerabilities due to the dynamic nature of the data being harvested; when the system interacts with the data it scrapes, the potential for algorithmic feedback amplifies Red attack vectors. This potential attack vector amplification is discussed in greater detail below; however for now it should be understood that any potential algorithmic feedback loop grants Red *access* to the model itself by allowing Red to interact with–and in some cases, co-create–the training data.

While Red's potential *impact* with IW algorithmic attacks has already been discussed elsewhere in great detail, an important and often-overlooked aspect of AIML attack surfaces is the interplay of Red's *knowledge*, *access*, and *capability*–and how these work to bolster one another.

A final dimension to consider is the availability of AI to Red to amplify and aid in attacks. Examples include scripts using GANs to create photos of people who never existed in order to propagate fake social media accounts, and AI-enabled mass content- & persona-management systems that can exert artificial/unintended influence over systems.

The OODA Loop & Game Theoretical Modeling of Information Warfare Scenarios

The Boolean *and* relationship between Red's various means of gaining either secret or publicly held knowledge of a system, and the currency of said knowledge, indicates a need for Blue to remain within Red's *observe*, *orient*, *decide*, and *act* (OODA) loop (fig 2).

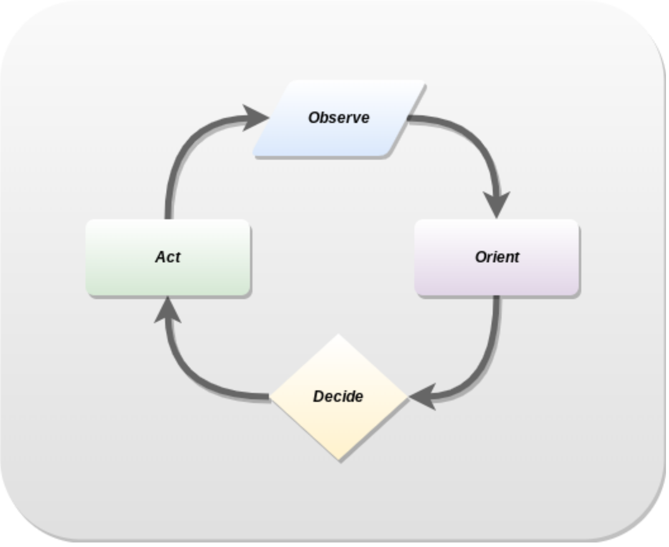


Fig 2: The OODA Loop

The OODA loop is an information processing and decision framework used in military, law enforcement, and other tactical mission operations; it has demonstrated applications in game-theoretical modeling of Information Warfare scenarios (Jormakka & Mölsä 2005). The OODA loop–as an information processing tool–is particularly suited for application to mass information processing systems in an IW setting. This section discusses the intersystem relationship between artificial intelligence systems and their artifacts, and the OODA loop in an adversarial security setting.

In a Red vs. Blue adversarial security framing, Red must gather, filter, make decisions, and act on information promptly to impact Blue. Artificial intelligence systems may enhance Red's ability to complete each of these steps, via (as examples) enhanced breadth and/or volume of information collection capabilities, faster information processing, and similar AIML-assisted reduction of time constraints. AIML systems may also amplify the impact of Red actions against Blue through network or other means.

Within the *knowledge* requirement, Red must expend resources to gather information through either OSINT or exfiltration, *and* this information must be *current* enough for Red to utilize successfully against Blue. In this example, Red's capabilities in the domain may be offset by a number of Blue actions, such as changing system details frequently so that Red's hypothetical acquired system-related information is not timely enough to be useful. In this way, Blue remains within Red's OODA loop: Red cannot gather, filter, analyze, and apply data with enough speed to have a significant *impact* on Blue. Red has expended resources with no gain.

However, the potential for artificial intelligence systems to aid in the processing of mass quantities of data, and thus provide a boost in resource capability in one or more of Red's *gather*, *filter*, *analyze*, and *apply* requirements, may change the calculus significantly for Blue defenders. Artificial Intelligence applications for filtering and analysis of both structured and unstructured data speed these tasks up enormously. In some cases, the gains are not just speed, but in the new ability to process and derive meaningful information–rapidly–in quantities of data previously unimaginable. Artificial intelligence thus has current applications in both exfiltration and OSINT intelligence gathering which Blue must take seriously.

It should be noted that information gathering and processing are not the sole domains of Red; Blue may also utilize AIML systems in its intelligence-gathering (INT) systems. In this respect, artificial intelligence systems may significantly impact the adversarial OODA loop, for better or for worse, on behalf of either side.

This paper notes an additional special use case of the OODA loop with regard to AIML applications in the Information Warfare (IW) context. The self-reinforcing nature of some algorithmic applications, as well as targeted disinformation attacks, may contribute to tightening Red's OODA loop in ways that can rapidly become prohibitively difficult for Blue to overcome. For example, algorithmic reinforcement loops could open a potential Red attack vector in AI-driven security systems via the use of adversarial means to gradually train an anomaly detection system to ignore malicious activity.

Disinformation campaigns are a particularly salient example of the potential to exploit algorithmic reinforcement within the IW context. Exploiting social media algorithms to spread disinformation rapidly can place Red in control of the information battlefield, and tighten Red's OODA loop beyond Blue's ability to regain narrative control or otherwise respond in a timely, and thus effective, manner. Interconnections among social graphs interact with algorithmic feed rankings resulting in the decentralized and unpredictable spread, where even unsuspecting individuals become participant nodes (Mejías & Vokuev 2017; Badawy, Lerman & Ferrara 2019). This amplification can not only make attribution in the case of bad actors impossible, but it can tighten the OODA loop in favor of attackers–as Red's targeted messaging is amplified, memed, and interacted with via other in-system mechanisms–quickly and in real-time.

Feedback loops, in which models directly interact with the data and/or the populations generating data, have emerged as a particular concern for AIML systems with regards to ethics, fairness, and accountability, as well as long-term model efficacy (Chaslot 2019; Sîrbu et al 2019; Liu et al 2020; Heidari, Nanda & Gummadi 2019; Mansoury et al. 2020). Feedback loops are also a special security concern for AIML systems, as they can dramatically increase Red's *access* to Blue's training data, and thus to Blue's models themselves. In a scenario where Blue's models are trained on some form of scraped or otherwise publicly-generated data with which they do not interact (or interact with only passively), Red's potential ability to gain access to Blue's training data, and thus influence Blue's models, is already a considerable attack vector–and this access increases as Red is increasingly aware of Blue's data sources.

However, Red's *access* potential via the data vector is increased when Blue's models interact with the data on which they are trained. As two high-level examples, consider (1) a social media feed that is responsive to both trending topics (aggregation of mass social data), and a particular user's engagement history, and (2) an AIML-driven security application trained on historical data to detect malware (or any anomalies in some type of network traffic).

In the case of (1) the social network, data is aggregated en masse from users and stored in a database; data about an individual user's preferences are similarly stored. These data are retrieved and combined by the social network's data scientists to model the likelihood that showing a particular topic to a specific user will generate engagement (vs, it can be assumed, the likelihood of causing the user to disengage with the platform). Topics or media are presented to users, users engage with these media, and the data on engagement is collected and fed back into the model. Because Red actions can *create* data on which the models are then trained, Red potentially has increasingly direct *access* to the models themselves (Jagielski et al 2021; Suya et al 2020;).

Consider case (2), a security application dynamically trained on historical data to detect anomalies suggesting malware. Model re-training may be continuous, periodic, or triggered by detected model decay. If data are not checked thoroughly for drift or the source of decay (i.e. data- or concept-drift) datasets might be generated from new historical data, which, over time, can introduce potentially malicious bias into the model (Severi et al 2020; Yang et al 2022), making a subsequent attack less likely to be flagged–and more likely to succeed (Zheng & Baochun 2021). If the system interacts with the public/gives Red access in a way that allows Red to probe its behavior, thus increasing Red's *knowledge*, the speed with which this "slow-shift" attack can occur may escalate dramatically (Severi, Meyer & Coull 2021).

In the case of both (1) and (2), Red's OODA loop is tightened due to near-direct access to Blue's models via the training data–datasets of which Blue may have inadvertently made Red a co-creator. When Red is made a potential collaborator in Blue's model training, Blue loses the upper hand in any adversarial AIML scenario. Increasing levels of data awareness on the part of Red grants increasing *access* and efficiency of attack (Deng et al 2020), and may even cause data leakage (Chase, Ghosh & Mahloujifar 2021) leading to the exfiltration of sensitive user or system data. Put differently, giving Red access to the training data is giving Red the keys to the system.

These two cases additionally illustrate a special consideration with regard to bias: while bias is often thought of as a purely ethical issue, in an Information Warfare scenario it is arguably an even greater security risk. Malicious biasing of models is no longer a future hypothetical, and thus AIML practitioners must take bias seriously as a security concern–via instituting, formalizing, and operationalizing mitigating processes.

Additionally, due to the proliferation of very large language models and other Natural Language Processing (NLProc) systems in industrial applications, and their susceptibility to malicious data poisoning attacks (Chen et al 2020; Kurita, Michel & Neubig 2020; Chen et al 2021; Wallace et al 2020; Yang et al 2021) versus their relative lack of algorithmic auditing (Bender et al 2021), it is possible to imagine a coordinated large-scale attack vector wherein a sufficiently motivated and AI-enabled attacker could disseminate lexical disinformation at such scale as to produce noticeable downstream effects within the models, and subsequently, their applications (Bagdasaryan & Shmatikov 2021). While these potential attacks might seem far-fetched, their feasibility has been demonstrated many times over in the literature. It should be further noted that the scale of these models' adoption, combined with that of the models themselves, makes auditing their vulnerability–or potential social impact–nearly impossible under current industry or regulatory standards.

Finally, it should be noted that additional attack vectors exist for Red; these, along with more detailed attack descriptions, have been omitted here for security reasons. Organizations are encouraged to decompose Red's Boolean attack path to assess specific IW risks within their AIML systems.

Special Considerations for Red-Blue Attach Path Analysis of AIML Systems

"Don't Make the Perfect the Enemy of the Good."
- Voltaire

The RAND Project Air Force (PAF) report recommends scoring cybersecurity and cyber resiliency using a maturity index, modeled off similar cyber metric maturity indices published by the US Department of Energy and the US Department of Homeland Security (2014). The use of an index is intended to avoid overly precise (and potentially misleading) quantifications in favor of a range that reflects the subjective nature of the evaluations (Snyder et al 2020).

The PAF report stresses both the importance of skilled workers in creating assessments and in the artifacts produced at the working level:

> *The most important product of the assessment at the working level would be artifacts that document what is being done to counter each of the Red actions. It is at this level where the technical and operational details lie. This level must do the most detailed assessment to document the Blue countermeasures and their qualities against Red. (24)*

This focus on careful documentation of the assessment process itself is complementary to an MLOps approach to AIML development, where models and their accompanying artifacts are stored in repositories as part of the model governance process.

The PAF report also gives the following questions to aid in assessment; the deeper the following questions are assessed in the affirmative, the more mature Blue's cyber response systems are judged to be:

- *Has a baseline trusted state been defined for each system to return to?*
- *Is continuous monitoring done?*
- *Have resilient methods been identified to return to this state?*
- *Can these actions respond within Red's OODA loop?*
- *Is the recovery time adequate for mission needs?*
- *Have these measures been implemented?*
- *Have they been exercised and tested?*
- *Have they been found to be adequate?*

AIML systems present unique challenges & potential pitfalls for security in a contested Information Warfare environment. These challenges begin at model development and data acquisition and continue throughout the AIML application lifecycle. To meet these challenges, the above questions are adapted at the conclusion of this paper to provide a cogent and widely applicable cyber resilience maturity assessment for AIML systems.

In the following sections, this paper outlines processes and artifacts specific to AIML systems that can enhance the maturity of an organization's AIML cyber response. These processes and artifacts include audits, Failure Modes and Effects Analyses (FMEAs), adversarial testing, datasheets & model cards, and a security/resilience portfolio to be checked into model repos as models are developed and iteratively improved.

Audits: a First Defense for AIML Systems

Audits are an important tool in the Blue security arsenal. Robust AIML auditing protocols provide critical insight into and documentation of the development process, data sources, regulatory compliance pitfalls, possible value misalignment, and, pertinent to this paper, potential Red attack vectors. Audits should be considered a primary Blue countermeasure against Red for AIML systems in an adversarial environment.

This paper draws from the SMACTR framework for internal audits (Raji et al 2020). The SMACTR framework was designed by researchers at top institutions; SMACTR has the benefit of being specifically tailored to AIML systems in a production setting.

The creators of the SMACTR framework acknowledge that organizational and system requirements will vary, and so any framework for evaluation must be flexible enough to accommodate these differing needs and use cases. Accordingly, teams can and should carefully evaluate their needs and the potential applications of the final product when determining which steps to prioritize in the auditing process:

> *Though not covered here, an equally important process is determining what systems to audit and why. Each industry has a way to judge what requires a full audit, but that process is discretionary and dependent on a range of contextual factors pertinent to the industry, the organization, audit team resourcing, and the case at hand. Risk prioritization and the necessary variance in scrutiny is a separately interesting and rich research topic on its own. The process outlined below can be applied in full or in a lighter-weight formulation, depending on the level of assessment desired* (Raji et al 2020, 38).

While this paper recommends that organizations take the most comprehensive approach to SMACTR that is institutionally feasible, sometimes a full audit is not possible, or certain aspects of the auditing framework may not apply to the AIML system in question. Additionally, AIML systems in an adversarial IW environment present specific challenges with regard to model monitoring and maintenance which require more than a one-time assessment, and implementation of a full-scale audit for every model iteration could prove burdensome.

For these reasons, several processes and artifacts from the SMACTR auditing framework which are uniquely applicable to an adversarial security setting are given below. While certainly not a substitute for auditing, application of these processes can provide enhanced cyber resilience capabilities to an organization in the absence of a full-scale audit. This paper also recommends the adoption of these processes at critical junctions in the MLOps workflow, regardless of audit status.

Failure Modes and Effects Analysis (FMEA)

Failure Modes and Effects Analysis (FMEA) is a fault avoidance technique in the aerospace industry and safety engineering as a whole (Stamatis 2003); its applicability to software engineering has been demonstrated (Reifer 1979; Nguyen 2001; Ozarin & Siracusa 2003; Ozarin 2008). FMEA is a “methodical and systematic risk management approach that examines a proposed design or technology for foreseeable failures” (Raji et al 2020, 36). The purpose of the FMEA is a systematic documentation of potential engineering pitfalls, unintended consequences, and other risks that might be associated with a system. The FMEA process includes analysis of known potential failure points, conducting literature reviews and interviews with relevant stakeholders, and collecting extant technical documentation (*ibid*).

This documentation of the known risks, literature, and processes involved in the development of an AIML system is beneficial to several stakeholders, as well as crucial to the production of other artifacts described in this framework. Critically, the FMEA provides a carefully documented and ideally clearer picture of system requirements and challenges to all members of engineering and management teams. While a full auditing process requires an FMEA, even in a scenario in which a full audit is not required or feasible, the FMEA informs later adversarial testing (Raji et al 2020, 41).

Adversarial Testing

Adversarial testing is a common practice in software development and network security administration, and has demonstrated effectiveness in finding potential data poisoning attack vectors (Steinhardt, Koh & Liang 2017). In the AIML cyber resilience framework presented here, adversarial testing begins with IW risks and other issues documented in the FMEA (Raji et al 2020, 41).

Adversarial Testing of AIML takes a methodical approach in attempts to nudge the system into “bad” behavior. Testing may take a variety of forms and approaches; these are based on potential failure points identified in the FMEA and documented in artifacts included in the system’s production documentation. This process is distinct from, and complementary to, adversarial pixel-manipulation attacks, and should be employed as part of a diverse testing arsenal drawing from the FMEA:

> *...direct non-statistical testing uses tailored inputs to the model to see if they result in undesirable outputs…This is distinct from adversarially attacking a model with human-imperceptible pixel manipulations to trick the model into misidentifying previously learned outputs, but these approaches can be complementary. This approach is more generally defined—as encompassing a range of input options to try in an active attempt to fool the system and incite identified failure modes from the FMEA* (Raji et al 2020, 41).

While critical to the initial assessment of an AIML product or system, adversarial testing should continue throughout the system lifecycle as part of development best practices:

> A*dditionally, proactive adversarial testing of already-launched products can be a best practice for*

the lifecycle management of released systems. The FMEA should be updated with these results, and the relative changes to risks assessed (ibid.).

For this reason, this paper recommends integrating adversarial testing for data, bias, behavior, ethical, and other unintended risks into existing MLOps workflows. The frequency of re-testing should depend on organizational needs and the nature of risk associated with the model in the FMEA. Systems identified in the FMEA process to have a larger attack surface, potentially magnified social consequences for failure, or feedback loops increasing Red's *access* and/or potential *impact*, will require more frequent testing. Subsequent FMEA documents should be updated with new adversarial testing results.

Datasheets & Model Cards

Datasheets in the MLOps workflow are a method for thoughtful creation and documenting of datasets; they are analogous to methods in the electronics hardware industry, where datasheets describe the operating characteristics, recommended uses test results, and so forth of a particular component (Raji et al 2020, 41). Decomposition of Red attack vectors makes the benefits of datasheets for physical and even software components clear. Organizations cannot create all components of any modern AIML system in-house, whether physical or virtual, and thus must rely on a trusted, standardized mechanism for communicating critical information about the components they outsource. While datasheets are standard in other mission-critical sectors, leading AIML practitioners to note their adoption in the field of datasets has yet to become widespread (*ibid*), creating a potential security gap between organizations that implement datasheets into the MLOps workflow, and those which do not.

In a cyber security/resilience framing, datasheets document critical potential failure points and attack surfaces and provide data for adversarial testing. This paper urges AIML practitioners to consider a model or system's security assessment incomplete without having completed datasheet(s) for all datasets as thoroughly as possible and checked in as mandatory artifacts in the model development workflow.

The framework laid out by Gebru et al (2018) provides questions for data preparers to answer, as well as a workflow to guide the process. These questions span six areas of the dataset creation process: *motivation*;*composition*; *pre-processing/cleaning/labeling; uses; distribution*; and *maintenance*. In an IW scenario, it is possible to imagine Red attack vectors within each of these areas. Because AIML systems are only as good as the data on which they are trained, and because many AIML systems rely on data to which the general public, including Red attackers, may have some form of access, documenting dataset creation as thoroughly as possible should be considered a baseline preparedness step for AIML cyber resilience.

Datasheets are complimented by *model cards* (Raji et al 2020, 41, Mitchell et al 2019). Model cards are an artifact created as part of the model documentation process. AIML systems should not be considered properly scoped for security vulnerabilities without full documentation, including model cards checked into repositories as part of model governance procedures.

The importance of model cards in an IW scenario becomes apparent when considering the information they are intended to convey:

Model cards serve to disclose information about a trained machine learning model. This includes how it was built, what assumptions were made during its development, what type of model behavior different cultural, demographic, or phenotypic population groups may experience, and an evaluation of how well the model performs with respect to those groups (Mitchell et al 2019).

As argued in previous sections of this paper, in a contested Information Warfare environment bias is not only an ethical concern but a practical one that can open AIML systems to malicious manipulation. Feedback loops amplify these concerns. In the IW environment, model cards fulfill the critical role of documenting where these potential attack vectors may lie.

In addition, properly implemented model cards also serve to document an original model state to which a system may return after an attack. The benefits of integrating model cards into the MLOps workflow to an organization's overall cyber resilience strategy are clear.

Blueprint: an MLOps approach

Operationalizing AIML systems into production can be challenging for organizations for a variety of reasons; these may be organizational, ranging from talent shortages to lack of development operations support in the institutional culture (Ronanki & Davenport 2017; Balakrishnan et al 2020); or directly related to the nature of AIML development itself; challenges identified in the literature include the customization and reuse of models, managing model component modularity (avoiding component entanglement), and data acquisition and management (Karamitsos, Albarhami & Apostolopoulos 2020; Amershi et al 2019). Factors in the research and publication environments such as field nascence and non-traditional publication venues (owed in large part to the scale and speed of research) affect teams' ability to stay current in the state-of-the-art (Mart'inez-Fern'andez et al 2021). An Amazon Research paper listed model retraining decisions and adversarial scenarios as specific examples of pertinent challenges in AIML production deployment (Schelter et al 2018). The Google Practitioner's Guide to MLOps gives several factors considered complexities specific to AIML engineering. Among these are: *Preparing and maintaining high-quality data for training ML models; Tracking models in production to detect performance* *degradation; Performing ongoing experimentation of new data sources, ML algorithms, and hyperparameters, and then tracking these experiments; Maintaining the veracity of models by continuously retraining them on fresh data; and most notably, Handling concerns about model fairness and adversarial attacks*.

The whitepaper additionally gives several insights into the state of AIML development. In short, successful deployment of AIML systems remains elusive for many organizations:

Despite the growing recognition of AI/ML as a crucial pillar of digital transformation, successful deployments and effective operations are a bottleneck for getting value from AI. Only one in two organizations has moved beyond pilots and proofs of concept. Moreover, 72% of a cohort of organizations that began AI pilots before 2019 have not been able to deploy even a single application in production…models don't make it into production, and if they do, they break because they fail to adapt to changes in the environment (Salama, Kazmierczak, and Schut 2021).

The tendency of teams to do experimentation primarily through ad hoc work, combined with factors hindering the adoption of best practices from software engineering to the AIML development environment, are significant contributors to the AIML deployment bottleneck:

…ML systems cannot be built in an ad hoc manner, isolated from other IT initiatives like DataOps and DevOps. They also cannot be built without adopting and applying sound software engineering practices, while taking into account the factors that make operationalizing ML different from operationalizing other types of software (ibid).

Integrating MLOps processes into organizational workflows does not just increase the likelihood of successful deployment, it can also help organizations manage risk:

Organizations need an automated and streamlined ML process. This process does not just help the organization successfully deploy ML models in production. It also helps manage risk when organizations scale the number of ML applications to more use cases in changing environments, and it helps ensure that the applications are still in line with business goals (ibid).

The tiered solutions proposed in this paper–beginning with appropriate documentation in the form of model cards and datasheets for datasets, and progressing in maturity to SMACTR audit, periodic adversarial testing, and triggers for increased scrutiny in case of model decay–are designed to be implemented in a streamlined MLOps workflow for greater cyber resilience with production scalability. Increased alignment with organizational goals and stated public values is a separate benefit of proper AIML system operationalization.

Reproducible results are a cornerstone of AIML experimentation and engineering for system stability and resilience, but software engineering best practices must be adapted to the AIML development workflow. Thus the potential contribution of an MLOps approach to overall AIML system security is high.

Key MLOps Processes: *Prototyping & Training Operationalization, Continuous Monitoring & Retraining, & Model Governance*

Rather than serve as an exhaustive guide to implementing MLOps workflows, this paper focuses on integrating best practices for AIML cyber resiliency into key processes within the MLOps pipeline: p*rototyping, training operationalization, & model governance; continuous training,* and *continuous monitoring*. Areas of focus within these processes include parallel development and testing, support for lineage analysis, and, in particular, the development of an IW security asset portfolio including triggers for model performance review. Finally, this paper introduces a formalized workflow for security analysis and systems for implementation.

There are of course additional avenues for increased security and system resilience which will vary among organizations. Practitioners are encouraged to adapt the fullest suite of MLOps best practices that are both relevant to and feasible for their organization's needs.

Prototyping, Training Operationalization & Model Governance

This paper proposes the development of a model security asset portfolio and security protocols for Information Warfare (IW) scenarios. This asset portfolio is organized around and integrated within three core MLOps capabilities, beginning in the *model prototyping* stage with a parallel development process consisting of AIML experimentation, coupled with scoping of security attack surfaces in the FMEA, adversarial testing, and datasheet documentation. Optimally, this process also includes a full audit using the SMACTR framework for AIML. However, this paper recommends that AIML systems should not be considered fully documented for IW security purposes without at a minimum FMEA, adversarial testing, model cards, and datasheets for training data.

AIML prototyping in most organizations may be represented as a development cycle (fig 3). Outputs of the AIML experimentation/prototyping cycle typically include artifacts such as notebooks & other systems for tracking experiments, hyperparameters, and configurations for model training, and metrics checked into a metadata repository; trained model(s) checked into a model repository; as well as code and other configurations for a model training pipeline. Importantly, the result of this phase is not a trained model, but rather a formalized model pipeline for later training operationalization.

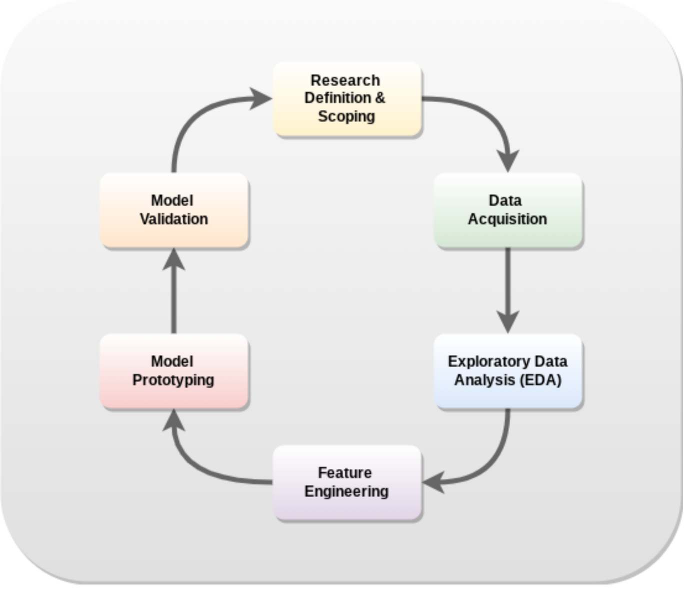


Fig 3: Canonical AIML Development Cycle

This paper recommends an additional output of the experimentation process, developed in parallel with model prototyping: a model security portfolio containing three key assets. The first is a set of potential attack vectors documented from the FMEA and subsequent adversarial testing. The second artifact within the proposed model security portfolio is a formalized set of baseline triggers for model performance review drawn from the model's FMEA and adversarial testing results, as well as any other pertinent documentation arising from the auditing process (if one is performed). These should be indicators of potential IW attacks drawn from the attack vectors discovered in the FMEA and adversarial testing processes. The third recommended security asset is a set of formalized and operationalizable comparison threshold points for the automated monitoring process, implemented in case of *any* triggered model decay. Similarly, these thresholds are indicators of potential IW attacks discovered in the prototyping process and act as a redundant system to the FMEA/testing-derived model decay triggers. Automation of these security checks within the model monitoring pipeline is intended to minimize human review costs. Figure 4 shows the AIML development cycle unrolled, with recommended mitigation and documentation processes in parallel.

The model security asset portfolio is checked in along with other prototyping process artifacts and implemented within the *continuous monitoring* process. Triggers for review, as well as comparison thresholds, should be considered key security assets within the model portfolio and should be treated accordingly.

The role of the *model governance* process is to ensure that models are not released into production without requisite documentation, including the FMEA/adversarial testing-derived security portfolio described above, and in the figure below.

Continuous Monitoring & Training Pipelines

Within a canonical MLOps continuous monitoring & training system there exist numerous opportunities to operationalize state-of-the-art IW vector detection and defense mechanisms. Anomaly detection (Paudice et al 2018), along with varying methods for detection and mitigation such as gradient shaping (Hong et al 2020); adversarial training (Geiping et al 2021; Vijaykeerthy et al 2019); activation clustering (Chen et al 2019); and reverse engineering (Dong et al 2021) have been demonstrated viable against data poisoning backdoor attacks. Specific methodologies should be explored and elucidated in the FMEA and adversarial testing phases of prototyping. Novel attack vectors require robust and constantly evolving defenses, highlighting the importance of operationalizing the referencing and updating of the model security asset portfolio.

Continuous Monitoring

Within the *continuous monitoring* pipeline, a key asset of the model IW security portfolio is operationalized: FMEA- & testing-derived triggers for model decay review, which work in tandem with a failsafe, consisting of a secondary set of thresholds for data evaluation similarly adapted from the FMEA and adversarial testing phases of model prototyping. The secondary evaluation thresholds are referenced during the *data validation* phase of the *continuous training* pipeline. Failing either step of this two-phase validation triggers a human review.

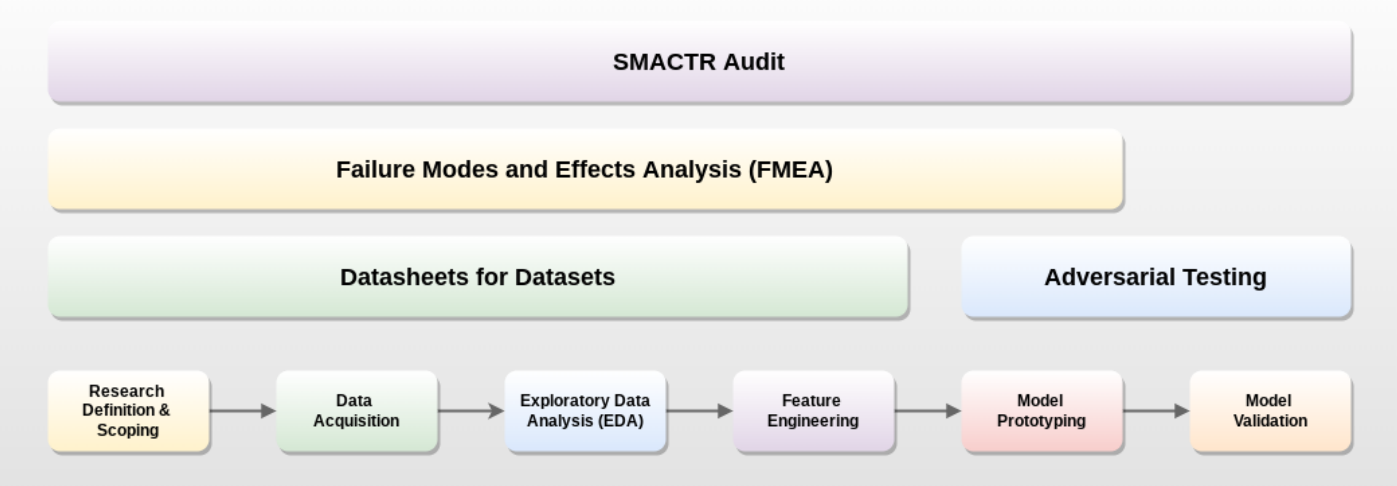


Fig 4: Parallel development processes for AIML model prototyping and security portfolio

Continuous Training

Depending on their use case, most models require retraining at some point. The frequency with which models are trained on new data depends on project requirements. Model retraining may be scheduled, triggered by some event or data threshold, or instantiated manually on an ad hoc basis. Regardless of how retraining is initiated, model scalability and effectiveness in production necessitate a training pipeline to automate retraining and track the necessary artifacts to support *lineage analysis*–the process of tracking a particular model back to its originating dataset, supporting documentation, and metadata–as well as other important model governance and accountability processes.

Organizations can improve system security via consistent, formalized pipeline metadata tracking in several ways. These include the ability to debug systems, which are crucially aided by the consistent availability of tracked metadata. Another benefit to teams is model reproducibility, which plays a critical role in model scalability and the ability to deploy to production. The ability to reproduce, scale, and debug models in production are bedrock security practices for AIML deployment.

Lineage analysis can also play a critical role in model security in an IW environment. The ability to track a model back to the dataset on which it was trained, as well as relevant metadata (including pipeline evaluations, data transformation steps, hyperparameters, and other configurations), are vital elements within the incident analysis workflow described below.

Properly implemented metadata tracking for lineage analysis also supports the feasibility of implementation for organizations. It may not be practical to create a new datasheet for every new model retraining iteration. Implementing IW security measures may become much more feasible when dataset documentation and tracking begin with detailed datasheets and are supplemented by metadata tracking capabilities already present in a well-implemented training pipeline.

Malicious IW attacks may manifest in data systems in different ways, depending on the AIML system specifics and use case. Detailed attack vectors have been omitted here for security and scope. Practitioners are encouraged to decompose the attack paths of their data sources and other relevant system concerns using the Boolean system described above. It is important to note that while specific attack surfaces will vary among (and even within) organizations, all teams can benefit from robust model and dataset documentation.

For IW security purposes this paper assumes MLOps best practices in place for training pipelines with robust lineage analysis capabilities and instead focuses on integrating an IW incident response workflow into a specific instantiation case of the training pipeline: retraining that is triggered by detected model performance degradation. Model retraining that is manually initiated or scheduled on a recurring basis can certainly utilize the workflow as well, supported by formalized documentation and lineage analysis capabilities.

In a typical AIML pipeline, the retraining process may begin with a trigger from the *continuous monitoring* system. Training data is extracted from dataset and feature repositories; data is then validated to check for corruption, schema or distribution skews, etc.

This data validation stage is relevant to the security process for several reasons. Models are only as good as the data they are trained on, and data can change over time in many ways, including shifting distributions, or the addition or loss of features. Without proper metadata/artifact and dataset tracking, this can be especially difficult to notice throughout multiple training iterations; this is why lineage analysis capabilities are a critical tool for teams when models perform sub-optimally. Checking the data should thus be a go-to step when evaluating model performance decline.

In an IW environment, data shifts might also be evidence of a malicious attack. Beyond typical model performance triggers implemented in standard MLOps systems, this paper recommends adding the use of triggers derived from the FMEA and adversarial testing performed during the parallel experimentation/development process as an added layer of security for AIML systems. Also recommended is the introduction of a set of FMEA- and adversarial testing-derived comparison points to the retraining pipeline and its *data validation* subprocess.

Automation and Reuse of Model Assets for Increased Implementation Feasibility

The addition of FMEA/adversarial testing-derived triggers to the model monitoring pipeline is intended to increase the feasibility of implementation for organizations. In many cases it may not be practical to implement a full human review of models at every retraining junction; doing so could potentially be so burdensome on organizations as to limit the scope of adoption. For these reasons this workflow utilizes the “built-in” functionality of retraining triggers in most AIML pipelines, re-using the FMEA and adversarial testing results to aid in incident response automation.

Full automation of this process is impractical for several reasons. Security analysis requires skilled workers; this is unlikely to change. However, automating the flags that invoke expensive human labor to the greatest extent possible can reduce overall costs, while increasing adoption feasibility. Full human review of the FMEA and adversarial test results, as well as lineage analysis, is triggered only after pertinent data checks have been run with backup systems.

Because the organizational cost of false negatives in IW security scenarios may be substantially higher than for false positives, detection of potential malicious incidents is prioritized, to reduce human review costs via primary automation. In this pipeline, built-in data validation functionality thus acts as a gatekeeper for both additional human review and additional model training.

Incident Analysis Workflow

This paper recommends the following workflow as a starting point for MLOps security integration and incident response. Organizations are encouraged to adopt recommendations that apply to and are feasible for their specific use case.

The process begins in prototyping, where the model IW security asset portfolio is produced in a parallel development process which includes the development of an FMEA and adversarial testing. These results are used to derive the three main assets within the portfolio: a set of potential attack vectors documented from the FMEA and subsequent adversarial testing; a formalized set of baseline triggers for model performance review drawn from the same source(s); and a set of formalized and operationalizable comparison threshold points for the automated monitoring process, to be implemented in case of *any* triggered model decay.

The model security portfolio materials are checked into the metadata repository; triggers and comparison points are added to relevant processes within the *continuous monitoring and training* pipelines. These include evidence for specific attack vectors discovered in the FMEA and adversarial testing processes during prototyping.

Triggers are added to the continuous monitoring pipeline which monitors for model decay. Comparison thresholds are integrated into the data validation workflow of the *continuous training* pipeline (fig 5) on the following page.

The *continuous monitoring pipeline* loads inference logs, IW security assets, model baseline statistics and reference schema, and other artifacts. The monitoring engine compares model performance metrics with thresholds from the security portfolio. If criteria are met for a potential IW security incident, security review is initiated and relevant parties notified via email, chat, or other internal notification system.

If criteria are not met for a potential IW security incident, the monitoring process continues with checks for schema skews, distribution or concept drift, and/or any applicable model performance/decay metrics. In the case of retraining triggers separate from the security portfolio assets, the *continuous training* pipeline is initiated. Retraining triggers are documented in the model portfolio (fig 6) on the following page.

The *continuous training* system requests relevant model artifacts and metadata from data stores, including the model security asset portfolio and threshold comparison points for data validation. Current model training data is evaluated against the validation thresholds. If IW security incident thresholds are not met, model training continues, and incident documentation is added to the model metadata. If IW security incident criteria are met, the system proceeds to alert model owners to potential system problems, triggering a human security review (fig 7) contained on the next page.

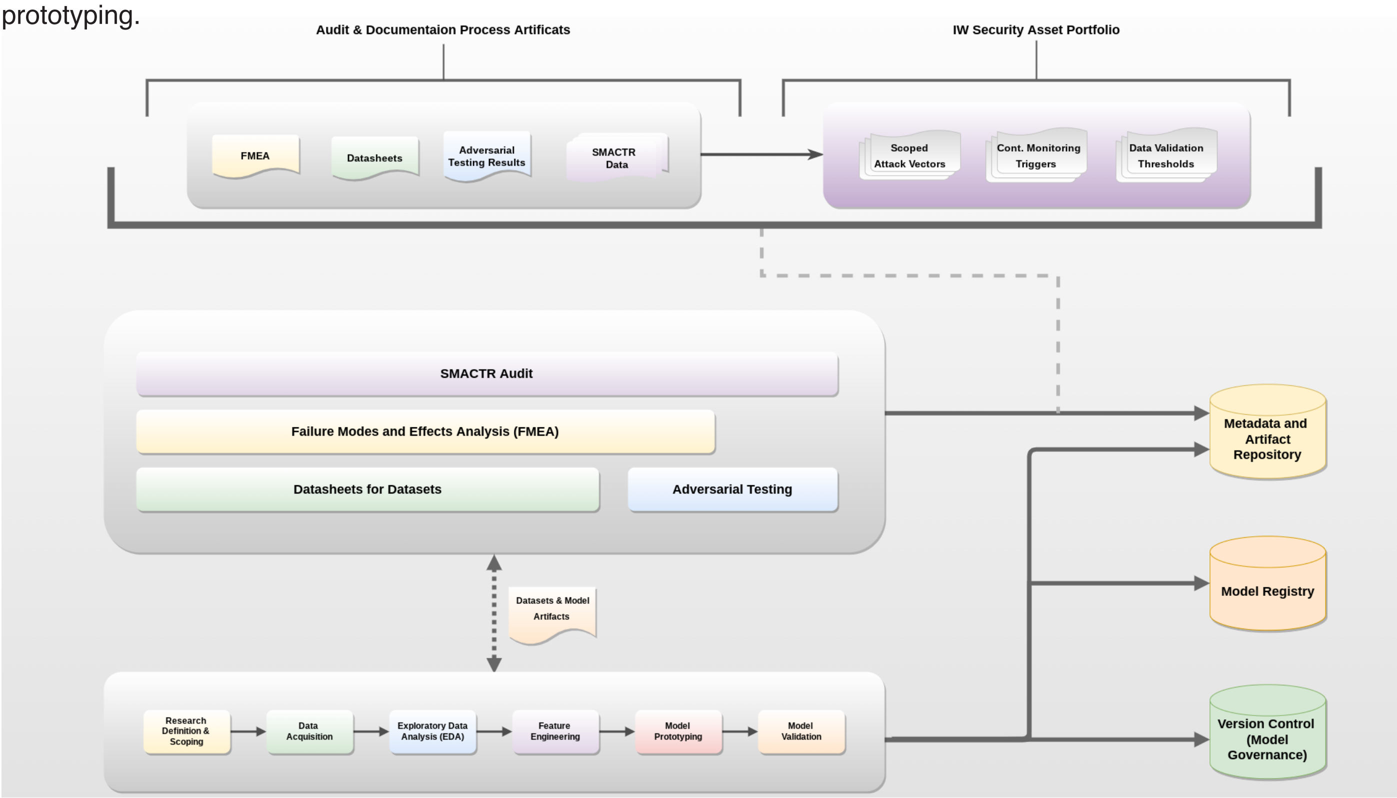

Audit & Documentaion Process Artificats
IW Security Asset Portfolio
FMEA
Datasheets
Adversarial Testing Results
SMACTR Data
Scoped Attack Vectors
Cont. Monitoring Triggers
Data Validation Thresholds
SMACTR Audit
Failure Modes and Effects Analysis (FMEA)
Datasheets for Datasets
Adversarial Testing
Metadata and Artifact Repository
Model Registry
Datasets & Model Artifacts
Research Definition & Scoping
Data Acquisition
Exploratory Data Analysis (EDA)
Feature Engineering
Model Prototyping
Model Validation
Version Control (Model Governance)

Fig 5: Outputs of the AIML prototyping and security development processes. Model security portfolio is generated in prototyping, and checked into the model metadata/artifact repository

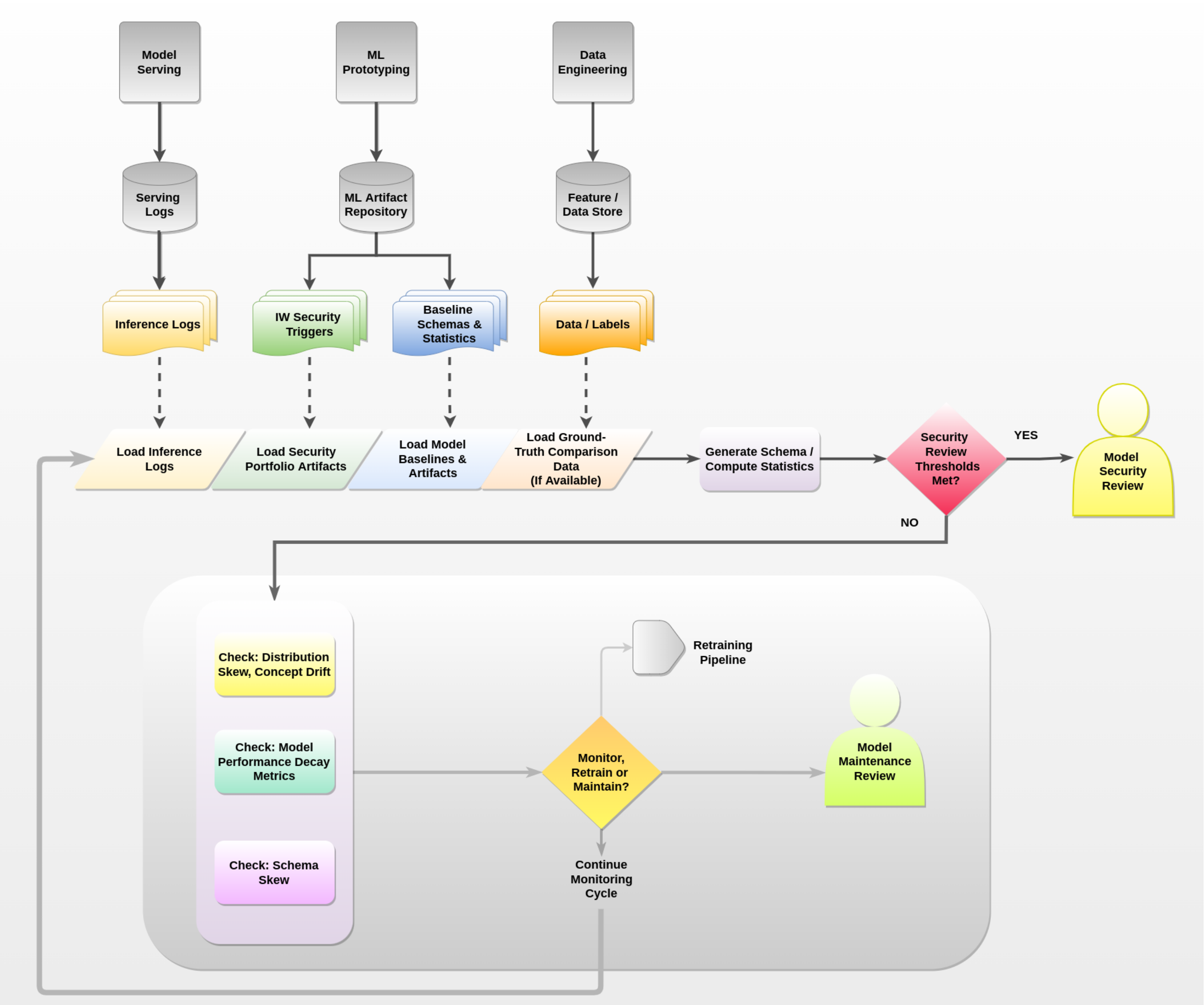

Model Serving
ML Prototyping
Data Engineering
Serving Logs
ML Artifact Repository
Feature / Data Store
Inference Logs
IW Security Triggers
Baseline Schemas & Statistics
Data / Labels
Load Inference Logs
Load Security Portfolio Artifacts
Load Model Baselines & Artifacts
Load Ground-Truth Comparison Data (If Available)
Generate Schema / Compute Statistics
Security Review Thresholds Met?
YES
NO
Model Security Review
Check: Distribution Skew, Concept Drift
Check: Model Performance Decay Metrics
Check: Schema Skew
Retraining Pipeline
Monitor, Retrain or Maintain?
Model Maintenance Review
Continue Monitoring Cycle

Fig 6: Continuous monitoring pipeline with model security checkpoint

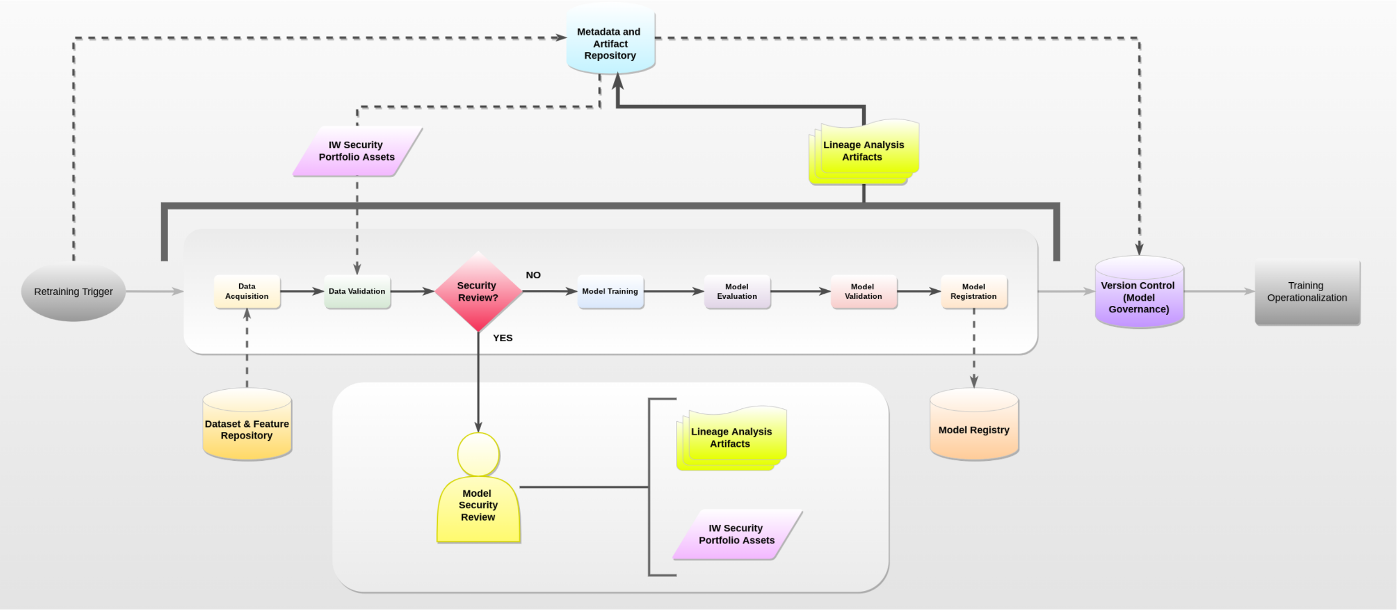

Metadata and Artifact Repository
IW Security Portfolio Assets
Lineage Analysis Artifacts
Retraining Trigger
Data Acquisition
Data Validation
Security Review?
NO
YES
Model Training
Model Evaluation
Model Validation
Model Registration
Version Control (Model Governance)
Training Operationalization
Dataset & Feature Repository
Model Registry
Model Security Review
Lineage Analysis Artifacts
IW Security Portfolio Assets

Fig 7: Continuous training pipeline with backup human review trigger implemented within the canonical data validation checkpoint

Human security review integrates model datasheets, lineage analysis (if the model has been retrained since datasheet production), security portfolio assets including FMEA and adversarial testing results, and SMACTR audit documentation if available. All documentation and review results are added to the model security portfolio. Incident documentation is submitted for regulatory compliance where applicable.

A Maturity Index for AIML Information Warfare Security Preparedness & Resilience

For AIML systems in a contested Information Warfare environment, this paper recommends a dual maturity index, building off the recommendations of the RAND Project Air Force *Measuring Cybersecurity and Cyber Resilience* paper, and including maturity level indexing specific to AIML systems. Both indices are given side-by-side below (fig 8).

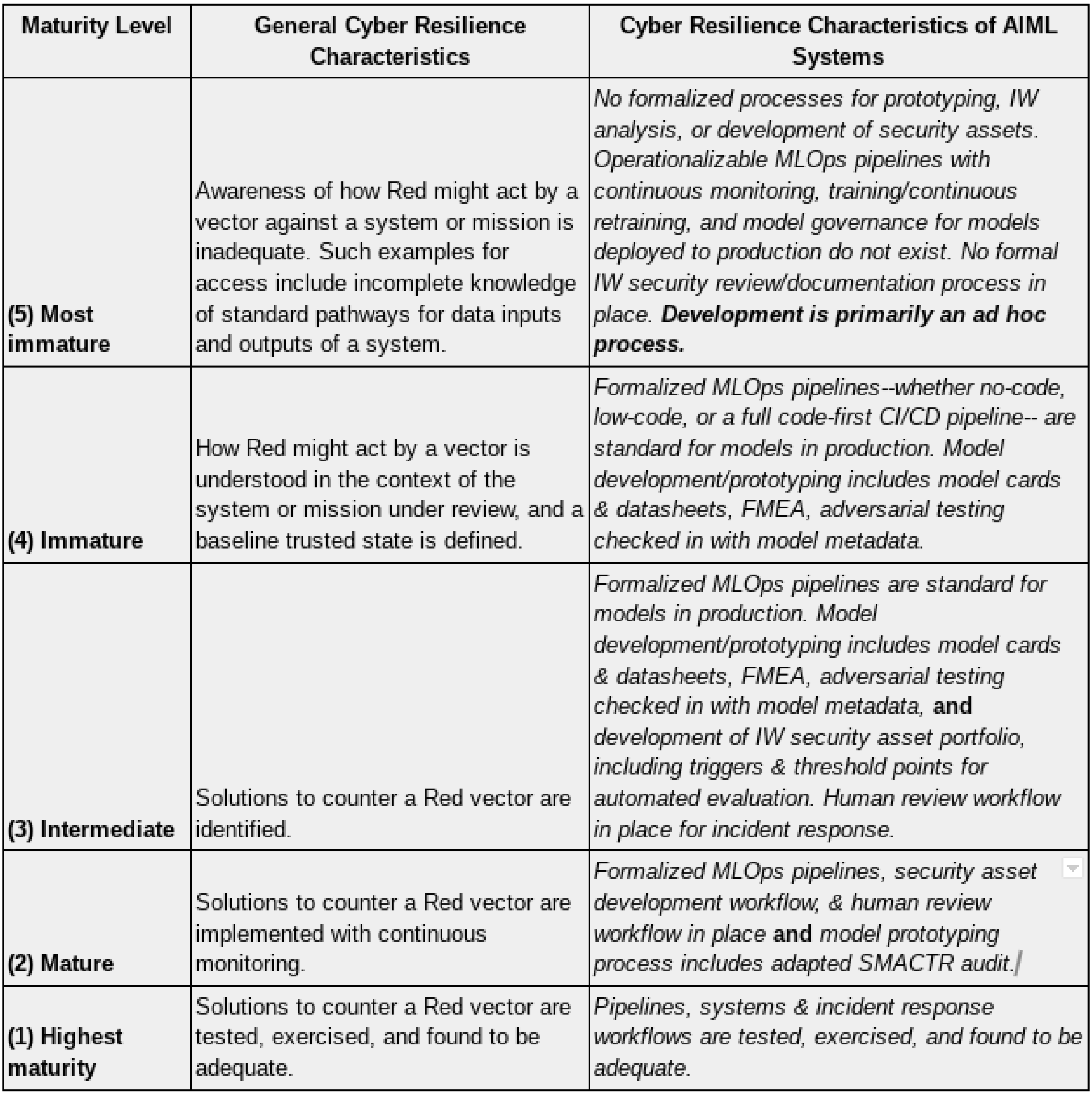
Finally, this paper recommends integrating the following AIML system-specific questions into the maturity assessment process. The more deeply these questions are answered in the affirmative, the more mature the system:

Question Group 1: Has a baseline trusted state been defined for each system to return to?

Additional [G1] Questions for AIML Systems: Do datasheets for datasets and model cards exist? Are there robust, formalized pipelines for metadata tracking to support lineage analysis?

Question Group 2: Is continuous monitoring done?

Additional [G2] Questions for AIML Systems: Is there a no-code or code-forward CI/CD pipeline, including continuous monitoring with security triggers and data validation thresholds within the continuous



Maturity Level	General Cyber Resilience Characteristics	Cyber Resilience Characteristics of AIML Systems
(5) Most immature	Awareness of how Red might act by a vector against a system or mission is inadequate. Such examples for access include incomplete knowledge of standard pathways for data inputs and outputs of a system.	*No formalized processes for prototyping, IW analysis, or development of security assets. Operationalizable MLOps pipelines with continuous monitoring, training/continuous retraining, and model governance for models deployed to production do not exist. No formal IW security review/documentation process in place.* ***Development is primarily an ad hoc process.***
(4) Immature	How Red might act by a vector is understood in the context of the system or mission under review, and a baseline trusted state is defined.	*Formalized MLOps pipelines--whether no-code, low-code, or a full code-first CI/CD pipeline-- are standard for models in production. Model development/prototyping includes model cards & datasheets, FMEA, adversarial testing checked in with model metadata.*
(3) Intermediate	Solutions to counter a Red vector are identified.	*Formalized MLOps pipelines are standard for models in production. Model development/prototyping includes model cards & datasheets, FMEA, adversarial testing checked in with model metadata,* **and** *development of IW security asset portfolio, including triggers & threshold points for automated evaluation. Human review workflow in place for incident response.*
(2) Mature	Solutions to counter a Red vector are implemented with continuous monitoring.	*Formalized MLOps pipelines, security asset development workflow, & human review workflow in place* **and** *model prototyping process includes adapted SMACTR audit.*
(1) Highest maturity	Solutions to counter a Red vector are tested, exercised, and found to be adequate.	*Pipelines, systems & incident response workflows are tested, exercised, and found to be adequate.*

training process, in place for models in production?

Question Group 3: Have resilient methods been identified to return to this state?

Additional [G3] Questions for AIML Systems: Has the organization formalized an incident response workflow (such as the one recommended in this paper)?

Question Group 4: Can these actions respond within Red's OODA loop?

Additional [G4] Questions for AIML Systems: What special OODA considerations exist within the system? Are there feedback loops which might tighten Red's OODA loop? How have these been addressed?

Question Group 5: Is the recovery time adequate for mission needs?

Additional [G5] Questions for AIML Systems: What are the consequences for losing control of data? Do these consequences become amplified over time? Is the projected recovery time acceptable within this collateral context?

Question Group 6: Have these measures been implemented?

Additional [G6] Questions for AIML Systems: Are formalized prototyping, development, deployment, and monitoring systems in place?

Question Group 7: Have they been exercised and tested?

Additional [G7] Questions for AIML Systems: Have prototyping, development, deployment, and monitoring workflows been tested for workability and resilience? Have model security portfolio assets and human review workflows been formalized and tested in the production environment?

Question Group 8: Have they been found to be adequate?

Conclusion

Concerns around cybersecurity vis-a-vis Information Warfare have taken a pronounced role in research and public policy. While governments are increasing regulatory compliance requirements, and researchers push out papers demonstrating actionable attacks against critical AIML infrastructure systems, organizations struggle with implementing MLOps best practices and security workflows. Moreover, there exist few resources bridging the gap between high-level frameworks and practical implementation.

Practical advice for assessing cyber resilience in the contested IW environment includes the use of maturity indices versus overly specific quantitative measures, and decomposing possible attack vectors using a Boolean attack path structure.

With robust model and dataset documentation, such as the security asset portfolio described above, in concert with pipeline support for robust model lineage analysis, opportunities exist for automation of IW/data security checks within a canonical MLOps pipeline. Integrating these assets into the MLOps security workflow re-utilizes artifacts created in regulatory/ethical compliance, maximizes automation with redundant systems, and minimizes the need for costly human review.

Lastly, within this framework there exist numerous avenues for customization and adoption to meet the demands of regulatory compliance at production scale. Integration and testing of specific frameworks against use cases and/or security scenarios offers a fascinating avenue for future investigation. As interest in AIML security grows, novel applications–and novel attacks to match–will continue to provide researchers and practitioners with new challenges, at the speed of information.

References

Adesina, Olubukola S., Michael Kearns, and Aaron Roth. 2020. “Ethical algorithm design should guide technology regulation.” Brookings. https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/.

Ahmed, Saifuddin. “Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size.” *Telematics Informatics* 57 (2021): 101508.

Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi and Thomas Zimmermann. “Software Engineering for Machine Learning: A Case Study.” *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (2019): 291-300.

Badawy, Adam, Emilio Ferrara and Kristina Lerman. “Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign.” *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018): 258-265.

Badawy, Adam, Kristina Lerman and Emilio Ferrara. “Who Falls for Online Political Manipulation?” *Companion Proceedings of The 2019 World Wide Web Conference* (2019): n. Pag.

Bagdasaryan, Eugene and Vitaly Shmatikov. “Spinning Language Models for Propaganda-As-A-Service.” *ArXiv* abs/2112.05224 (2021): n. Pag.

Balakrishnan, Tara, Michael Chui, Bryce Hall, and Nikolaus Henke. 2020. “Global survey: The state of AI in 2020.” McKinsey. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020.

Bandy, Jack. “Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits.” *ArXiv* abs/2102.04256 (2021): n. Pag.

Bender, Emily & Gebru, Timnit & McMillan-Major, Angelina & Shmitchell, Shmargaret. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?.” 610-623. 10.1145/3442188.3445922.

Beskow, David M.. “Finding and Characterizing Information Warfare Campaigns.” (2020).

Chase, Melissa, Esha Ghosh and Saeed Mahloujifar. “Property Inference From Poisoning.” *IACR Cryptol. ePrint Arch.* 2021 (2021): 99.

Chaslot, Guillaume. 2019. “The Toxic Potential of YouTube's Feedback Loop.” WIRED. https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/.

Christodoulou, Eleni and Kalypso Iordanou. “Democracy Under Attack: Challenges of Addressing Ethical Issues of AI and Big Data for More Democratic Digital Media and Societies.” *Frontiers in Political Science* (2021).

Chen, Bryant, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Ben Edwards, Taesung Lee, Ian Molloy and B. Srivastava. “Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering.” *ArXiv* abs/1811.03728 (2019): n. Pag.

Chen, Kangjie, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li and Chun Fan. “BadPre: Task-agnostic Backdoor Attacks to Pre-trained NLP Foundation Models.” *ArXiv* abs/2110.02467 (2021): n. Pag.

Chen, Xiaoyi, A. Salem, Michael Backes, Shiqing Ma and Yang Zhang. “BadNL: Backdoor Attacks Against NLP Models.” *ArXiv* abs/2006.01043 (2020): n. Pag.

Chopra, Amit K. and Munindar P. Singh. “Sociotechnical Systems and Ethics in the Large.” *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018): n. Pag.

Chowdhury, Rumman, Luca Belli, and Donna Lamar. 2021. “Examining algorithmic amplification of political content on Twitter.” Twitter Blog. https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent.

Conger, Kate. 2022. “With Eye to Russia, Biden Administration Asks Companies to Report Cyberattacks.” The New York Times. https://www.nytimes.com/2022/03/23/us/politics/biden-russia-cyberattacks.html.

Connell, Michael, and Sarah Vogler. 2017. “Russia's Approach to Cyber Warfare.” CNA.org. https://www.cna.org/CNA_files/PDF/DOP-2016-U-014231-1Rev.pdf.

Deng, Samuel, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody and Abhradeep Thakurta. “A Separation Result Between Data-oblivious and Data-aware Poisoning Attacks.” (2020).

Diamantis, Mihailis E.. “Algorithms Acting Badly: A Solution from Corporate Law.” *SSRN Electronic Journal* (2020): n. Pag.

DiResta, Renée. 2021. “It's Not Misinformation. It's Amplified Propaganda.” *The Atlantic*, October 9, 2021. https://www.theatlantic.com/ideas/archive/2021/10/disinformation-propaganda-amplification-ampliganda/620334/.

Dong, Yinpeng, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su and Jun Zhu. “Black-box Detection of Backdoor Attacks with Limited Information and Data.” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021): 16462-16471.

Dumbrava, Costica, and European Parliamentary Research Service. 2021. “Key risks posed by social media to democracy.” European Parliament. https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698845/EPRS_IDA(2021)698845_EN.pdf.

European Parliament. 2021. *The Impact of Disinformation on Democratic Processes and Human Rights in the World*. Brussels: European Parliament DROI Subcommittee. 10.2861/59161.

European Union Counter-Terrorism Coordinator. 2020. “The role of algorithmic amplification in promoting violent and extremist content and its dissemination on platforms and social media.” Open Data. https://data.consilium.europa.eu/doc/document/ST-12735-2020-INIT/en/pdf.

Floridi, L.. “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical.” *Philosophy & Technology* 32 (2019): 185-193.

Gebru, Timnit & Morgenstern, Jamie & Vecchione, Briana & Vaughan, Jennifer & Wallach, Hanna & Daumeé, III & Crawford, Kate. (2018). “Datasheets for Datasets.” Communications of the ACM. 64. 10.1145/3458723.

Geiping, Jonas, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller and Tom Goldstein. “What Doesn't Kill You Makes You Robust(er): How to Adversarially Train against Data Poisoning.” (2021).

Heidari, Hoda, Vedant Nanda and Krishna P. Gummadi. “On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning.” *ICML* (2019).

Hong, Sanghyun, Varun Chandrasekaran, Yigitcan Kaya, Tudor Dumitras and Nicolas Papernot. “On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping.” *ArXiv* abs/2002.11497 (2020): n. Pag.

Huszár, Ferenc & Ktena, Sofia Ira & O’Brien, Conor & Belli, Luca & Schlaikjer, Andrew & Hardt, Moritz. (2022). “Algorithmic amplification of politics on Twitter.” Proceedings of the National Academy of Sciences. 119. e2025334119. 10.1073/pnas.2025334119.

Ibáñez, Javier Camacho and Mónica Villas Olmeda. “Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study.” *AI & SOCIETY* (2021): n. Pag.

Im, Jane, Eshwar Chandrasekharan, John Singer artist Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens and Eric Gilbert. “Still out there: Modeling and Identifying Russian Troll Accounts on Twitter.” *12th ACM Conference on Web Science* (2020): n. Pag.

Jagielski, Matthew, Giorgio Severi, Niklas Pousette Harger and Alina Oprea. “Subpopulation Data Poisoning Attacks.” *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (2021): n. Pag.

Jobin, Anna, Marcello Ienca and Effy Vayena. “The global landscape of AI ethics guidelines.” *Nature Machine Intelligence* (2019): 1-11.

Jormakka, Jorma and Jarmo Mölsä. “Modelling Information Warfare as a Game.” (2005).

Kalin, Josh, David Noever and Matthew Ciolino. “Color Teams for Machine Learning Development.” *ArXiv* abs/2110.10601 (2021): n. Pag.

Karamitsos, Ioannis, Saeed Albarhami and Charalampos Apostolopoulos. “Applying DevOps Practices of Continuous Automation for Machine Learning.” *Inf.* 11 (2020): 363.

Keller, Franziska Barbara, David Schoch, Sebastian Stier and JungHwan Yang. “Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign.” *Political Communication* 37 (2019): 256 - 280.

Kirtley, Jane E. 2022. “Liability for Amplification of Disinformation: A Law of Unintended Consequences?.” American Constitution Society. https://www.acslaw.org/expertforum/liability-for-amplification-of-disinformation-a-law-of-unintended-consequences/.

Kornbluh, Karen. 2022. “Disinformation, Radicalization, and Algorithmic Amplification: What Steps Can Congress Take?” Just Security. https://www.justsecurity.org/79995/disinformation-radicalization-and-algorithmic-amplification-what-steps-can-congress-take/.

Koshiyama, Adriano Soares, Emre Kazim, Philip C. Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, Janet Adams, Christina Hitrova, Jeremy Barnett, Parashkev Nachev, David Barber, Tomas Chamorro-Premuzic, Konstantin Klemmer, Miro Gregorovic, Shakeel Ahmad Khan and Elizabeth Lomas. “Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms.” *Software Engineering eJournal* (2021): n. Pag.

Kurita, Keita, Paul Michel and Graham Neubig. “Weight Poisoning Attacks on Pretrained Models.” *ArXiv* abs/2004.06660 (2020): n. Pag.

Lee, Nicole T., Paul Resnick, and Genie Barton. 2019. “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms.” Brookings. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Libicki, Martin C.. “The convergence of information warfare.” (2020).

Lin, Herbert S. and Jaclyn A. Kerr. “On Cyber-Enabled Information/Influence Warfare and Manipulation.” (2017).

Liu, Lydia T., Ashia C. Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs and Jennifer T. Chayes. “The disparate equilibria of algorithmic decision making when individuals invest rationally.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020): n. Pag.

Mansoury, Masoud, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. “Feedback Loop and Bias Amplification in Recommender Systems.” *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (October), 2145-2148. 10.1145/3340531.3412152.

Mart'inez-Fern'andez, Silverio, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer and Stefan Wagner. “Software Engineering for AI-Based Systems: A Survey.” *ArXiv* abs/2105.01984 (2021): n. Pag.

Mejías, Ulises Ali and Nikolai E. Vokuev. “Disinformation and the media: the case of Russia and Ukraine.” *Media, Culture & Society* 39 (2017): 1027 - 1042.

Mitchell, Margaret & Wu, Simone & Zaldivar, Andrew & Barnes, Parker & Vasserman, Lucy & Hutchinson, Ben & Spitzer, Elena & Raji, Inioluwa & Gebru, Timnit. (2019). “Model Cards for Model Reporting.” 220-229. 10.1145/3287560.3287596.

Mökander, Jakob, Jessica Morley, Mariarosaria Taddeo and L. Floridi. “Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations.” *Science and Engineering Ethics* 27 (2021): n. Pag.

Mökander, Jakob and Marian Axente. “Ethics-Based Auditing of Automated Decision-Making Systems: Intervention Points and Policy Implications.” *ArXiv* abs/2111.04380 (2021): n. Pag.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander and L. Floridi. “Ethics as a service: a pragmatic operationalisation of AI Ethics.” *Minds Mach.* 31 (2021): 239-256.

Morley, Jessica, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi and L. Floridi. “Operationalising AI ethics: barriers, enablers and next steps.” *AI & SOCIETY* (2021): n. Pag.

Nguyen, Dong. “Failure modes and effects analysis for software reliability.” *Annual Reliability and Maintainability Symposium. 2001 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.01CH37179)* (2001): 219-222.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina E. Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernández, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis and Steffen Staab. “Bias in data-driven artificial intelligence systems—An introductory survey.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2020): n. Pag.

Ozarin, Nathaniel W. and M. Siracusa. “A process for failure modes and effects analysis of computer software.” *Annual Reliability and Maintainability Symposium, 2003.* (2003): 365-370.

Ozarin, Nathaniel W.. “The Role of Software Failure Modes and Effects Analysis for Interfaces in Safety-and Mission-Critical Systems.” *2008 2nd Annual IEEE Systems Conference* (2008): 1-8.

Paudice, Andrea, Luis Muñoz-González, András György and Emil C. Lupu. “Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection.” *ArXiv* abs/1802.03041 (2018): n. Pag.

Pew Research Center. “Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm in the Next Decade.” June 16, 2021.

Raji, Inioluwa Deborah and Joy Buolamwini. “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products.” *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019): n. Pag.

Raji, Inioluwa & Smart, Andrew & White, Rebecca & Mitchell, Margaret & Gebru, Timnit & Hutchinson, Ben & Smith-Loud, Jamila & Theron, Daniel & Barnes, Parker. (2020). “Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing.” 33-44. 10.1145/3351095.3372873.

Reifer, Donald J.. “Software Failure Modes and Effects Analysis.” *IEEE Transactions on Reliability* R-28 (1979): 247-249.

Ronanki, Rajeev, and Thomas Davenport. 2017. “Artificial Intelligence for the Real World – A Business Perspective.” Deloitte. https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/hbr-report-artificial-intelligence-for-the-real-world.html.

Ryan, Mark, Josephina Antoniou, Laurence D. Brooks, Tilimbe Jiya, Kevin Macnish and Bernd Carsten Stahl. “Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality.” *Science and Engineering Ethics* 27 (2021): n. Pag.

Salama, Khalid, Jarek Kazmierczak, and Donna Schut. 2021. “Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning.” Google Cloud. https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf.

Schelter, Sebastian, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert and Gyuri Szarvas. “On Challenges in Machine Learning Model Management.” *IEEE Data Eng. Bull.* 41 (2018): 5-15.

Severi, Giorgio, Jim Meyer, Scott E. Coull and Alina Oprea. “Exploring Backdoor Poisoning Attacks Against Malware Classifiers.” *ArXiv* abs/2003.01031 (2020): n. Pag.

Severi, Giorgio, Jim Meyer and Scott E. Coull. “Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers.” *USENIX Security Symposium* (2021).

Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini and Filippo Menczer. “The spread of low-credibility content by social bots.” *Nature Communications* 9 (2018): n. Pag.

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer and Giovanni Luca Ciampaglia. “Anatomy of an online misinformation network.” *PLoS ONE* 13 (2018): n. Pag.

Sîrbu, Alina, Dino Pedreschi, Fosca Giannotti and János Kertész. “Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model.” *PLoS ONE* 14 (2019): n. Pag.

Snyder, Don, Lauren A. Mayer, Guy Weichenberg, Danielle C. Tarraf, Bernard Fox, Myron Hura, Suzanne Genc, and Jonathan W. Welburn. “Measuring Cybersecurity and Cyber Resiliency.” Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RR2703.html. Also available in print form.

Stahl, Bernd Carsten, Josephina Antoniou, Mark Ryan, Kevin Macnish and Tilimbe Jiya. “Organisational responses to the ethical issues of artificial intelligence.” *AI & SOCIETY* 37 (2022): 23-37.

Stamatis, D. H. 2003. *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. N.p.: ASQ Quality Press.

Steinhardt, Jacob, Pang Wei Koh and Percy Liang. “Certified Defenses for Data Poisoning Attacks.” *NIPS* (2017).

Suya, Fnu, Saeed Mahloujifar, David Evans and Yuan Tian. “Model-Targeted Poisoning Attacks: Provable Convergence and Certified Bounds.” *ArXiv* abs/2006.16469 (2020): n. Pag.

U.S. Department of Defense. “DOD Dictionary of Military and Associated Terms.” Washington, D.C. July 2017.

U.S. Department of Energy and U.S. Department of Homeland Security. “Cybersecurity Capability Maturity Model (C2M2).” Version 1.1, February 2014.

Vijaykeerthy, Deepak, Anshuman Suri, Sameep Mehta and Ponnurangam Kumaraguru. “Hardening Deep Neural Networks via Adversarial Model Cascades.” *2019 International Joint Conference on Neural Networks (IJCNN)* (2019): 1-8.

Wallace, Eric, Tony Zhao, Shi Feng and Sameer Singh. “Customizing Triggers with Concealed Data Poisoning.” *ArXiv* abs/2010.12563 (2020): n. Pag.

The White House. 2022. “FACT SHEET: Act Now to Protect Against Potential Cyberattacks.” The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/21/fact-sheet-act-now-to-protect-against-potential-cyberattacks/.

Yang, Limin, Zhi Gang Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro and Gang Wang. “Jigsaw Puzzle: Selective Backdoor Attack to Subvert Malware Classifiers.” *ArXiv* abs/2202.05470 (2022): n. Pag.

Yang, Wenkai, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun and Bin He. “Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models.” *ArXiv* abs/2103.15543 (2021): n. Pag.

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini and Jeremy Blackburn. “Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web.” *Companion Proceedings of The 2019 World Wide Web Conference* (2019): n. Pag.

Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini and Jeremy Blackburn. “Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls.” *Proceedings of the 10th ACM Conference on Web Science* (2019): n. Pag.

Zheng, Tianhang and Baochun Li. “First-Order Efficient General-Purpose Clean-Label Data Poisoning.” *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications* (2021): 1-10.