Carnegie Mellon University
Software Engineering Institute

DANGERS OF AI FOR INSIDER RISK EVALUATION (DARE)

Austin Whisnant

October 2024

DOI: 10.1184/R1/27325569

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

1 Artificial Intelligence and Insider Risk

The goal of artificial intelligence (AI) is to build machines capable of simulating human thoughts and actions. There are many implementations of AI, and some of the most well-known are computer vision, robotics, machine learning (ML), and natural language processing (NLP). Each of these implementations uses different methods, inputs, features, parameters, and outputs, to solve a specific problem.

Typically, the problems being solved have a very narrow application, and the implementation chosen is highly tailored to the task at hand, such as the following examples:

- computer vision for detecting obstacles in the road [Janai 2020]
- NLP for speech recognition [Kamath 2019]
- deep learning for detecting breast cancer [Chan 2020]

This type of AI is called “narrow AI,” and it is programmed to operate within a predefined set of parameters, rules, and context. The models developed for narrow AI applications cannot be used for other tasks, even if the tasks are very similar. For example, a model programmed to tell the difference between images of dogs and cats would not be able to detect different dog breeds, just as a model programmed to detect customer bank fraud would not be able to predict bank employee fraud.

The categories of AI most relevant to the insider risk domain are ML and NLP. ML uses algorithms trained to find patterns in large datasets by analyzing different features or attributes of the data. For example, the dog versus cat model might analyze ear shape and muzzle length. Once the algorithm has been adjusted to find the correct patterns, it can then be used to predict, cluster, or classify additional data.

ML is split into three main methods of learning: supervised, unsupervised, and reinforcement. All three methods offer potential solutions for insider risk problems:

- **Supervised learning** requires labeled data for training (e.g., data labeled *dog* or *cat*).
- **Unsupervised learning** does not use labeled data; instead, it finds its own patterns or clusters in the data.

- **Reinforcement learning** enables the machine to learn by interacting with its environment and getting feedback about its decisions. It is commonly used to train models for games, such as chess or Go [Silver 2018].

Choosing an ML method is based on several factors, including the type of data available for training, the efficiency and scalability of the algorithm, the accuracy of the algorithm, potential ethical concerns, and legal requirements. All of these factors should be considered during the design phase discussed in Section 3.1.

Another commonly used category of AI is *deep learning*, which is a hierarchy of mini algorithms that comprise layers of nodes that take inputs from the previous layer and produce outputs for the next layer in the hierarchy until a set of useful features are found for solving the given problem. Deep learning can be found in many implementations and applications of AI, including computer vision, robotics, and NLP.

1.1 Examples

Although the applications of AI are nearly endless, the use of AI and ML in the insider risk domain is still limited, especially in the publicly reviewed academic space. In practice, the use of ML is quickly growing among some of the most common insider risk tools. The following examples from other domains share some similarities with insider risk, either in context, end goals, or methods used. These examples provide insight into possible solutions for insider risk, their challenges, and important lessons.

- *Online Exam Proctoring* – Many schools and certification proctoring companies now conduct online exam proctoring with at least partial assistance from AI. Leveraging AI decreases the need for human proctors and deters cheating on online exams, which have become much more common in a post-COVID world. Proctoring algorithms use computer vision and deep learning techniques to detect and analyze eye gaze, face recognition, body positioning, body movement, or a combination of those, along with monitoring of keystrokes, software, and/or browser tabs to detect potential cheating [Nigram 2021].
- *Vaccine Communication* – Google's Intelligent Vaccine Impact solution includes a sentiment analysis component [Schroeder 2020] that gathers data from a wide variety of sources to help governments and organizations understand how individuals and groups feel about vaccines and predict their intent to obtain them. The World Health Organization is also exploring ways to use AI techniques to influence communication about vaccines, intervene when disinformation is detected, and provide personalized vaccine education via chatbots when individuals have specific concerns.
- *Welfare Fraud Risk* – The city of Rotterdam used a supervised ML algorithm to generate a score for individuals who are most at risk of committing welfare fraud. The algorithm used 315 different attributes such as age, gender, language fluency, and number of children. The results were used to initiate fraud investigations against the top 10% of those with the highest risk scores [Braun 2023, Mehrotra 2023].

Each of these examples is similar to insider threat because its main goal is detecting problems that could have serious consequences. All of them recruit AI to reduce the burden on human analysts and alert decision makers to potential issues. Each example also has its own insights that may be valuable to the design of AI solutions for insider threat:

- *Online Exam Proctoring* – Methods for automated exam proctoring could be applied to insider threat via monitoring of subjects' actions on a computer or through video or web camera monitoring with an automated analysis of technical and behavioral cues. These methods also apply to monitoring the actions and identifying remote subjects whose environments cannot be controlled.
- *Vaccine Communication* – The advances in NLP used in vaccine communication can also be used to help insider threat detection via monitoring of subjects' chat, email, or phone communication. For example, sentiment analysis can look for negative attitudes or changes in behavior that might indicate problems at work or at home, or it could even detect an insider's intent to harm the organization or their co-workers. AI can potentially be used for automatic intervention when serious or imminent threats are detected.
- *Welfare Fraud Risk* – Fraud constitutes the largest category of insider incidents.[1] The techniques used in this example—analyzing both personal background and behavioral cues—apply directly to insider risk. In fact, this method of risk scoring is already a common technique used for insider risk detection and research.

1.2 State of AI for Insider Risk

The goal of insider risk analysis is to prevent an insider from leveraging their authorized access to take actions that will harm their organization. Insider risk is a particularly difficult domain for detecting harmful action because employees are usually authorized to take many actions and the system detecting actions may interpret them as a regular part of the employee's job and performed without malicious intent.

For example, it is very difficult to detect the difference between a bank employee opening an account for a new customer versus opening an account for a fraudulent business based solely on the action of opening the account. Detecting the fraud in this case would require correctly guessing the intent behind the action and monitoring many other sources besides account activity.

This difficultly in detecting harmful action is why some researchers and vendors focus on *prediction* (i.e., estimating the probably of future occurrence of insider threat) rather than *detection* (i.e., identifying existing patterns indicative of an issue). Prediction is often done via scoring the individual risk of insiders and is seen as a way to stop potential issues before damage occurs.

Organizations have huge amounts of data on employees and their activities, and these organizations have the potential to collect additional data specifically for monitoring purposes. This data includes technical data from day-to-day network activities (e.g., file access, web searching, email) and non-technical data (e.g., human resources [HR] data, badging records, performance

[1] According to CERT's MERIT repository, which tracks incidents prosecuted in U.S. federal court, fraud makes up 70% of all insider incidents.

evaluations). Additional monitoring data could include keystroke logging, video capture from webcams or security cameras, or audio from recorded phone calls. ML can help find patterns in these vast datasets, with deep learning in particular being able to build features out of seemingly innocuous or useless attributes. However, deep learning comes with certain caveats and challenges, which we discuss in Section 2.

Of the many potential implementations of AI, those most relevant to the field of insider risk are ML (including supervised and unsupervised methods), NLP, and computer vision. For example, computer vision has the potential to be particularly helpful in detecting insider actions in the physical domain (e.g., employees removing their employer's property or installing a hardware keylogger on a workstation).

NLP can be used in the insider risk domain to convert speech from monitored phone calls or videos that are posted to social media into text transcripts for further analysis. It can be used to sift through text (e.g., transcripts, chat logs, emails, and social media postings) for sentiment analysis or to detect changes in mood or state of mind [Wankhade 2022], and it can potentially be used for intent detection or automated intervention. NLP can also be used as an elevated form of keyword searching to help detect versions of keywords in idiomatic language. NLP applications for insider risk are quickly growing in the available tool set for insider risk analysis.

Previous academic research has analyzed insider threat applications using traditional supervised ML classification algorithms with reasonable results, but more recent literature focuses on using unsupervised or partially supervised deep learning to improve accuracy and scalability. The choice of algorithm largely depends on the data that is available for training and testing and whether the data is labeled or not (e.g., *insider threat* or *non-insider threat*).

Recent papers using deep learning algorithms have obtained insider threat detection results with 90% or higher accuracy on test datasets [Lu 2019]. The algorithms that researchers choose vary significantly in terms of their level of data granularity [Le 2020]:

- time windows for data capture (e.g., minutes, months, years)
- data types (e.g., technical, non-technical, host data, network data) [Foroughi 2018]
- choice of parameters (e.g., including a decay component with the model)

Given the complexity and variety of these options, some of the newest papers have begun to recommend an approach using multiple algorithms in a hierarchy, where the choice of algorithms is based on available data types, granularity level, and the output of previous algorithms in the hierarchy.

Figure 1 shows some of the key insider risk tools that insider risk analysis teams commonly use. These tools already incorporate some form of ML, though vendors do not usually disclose the specific methods and algorithms they use. Some tools that might include ML are user activity monitoring (UAM) for detection, user (entity) behavioral analytics (UBA/UEBA) for prediction and risk scoring, security information and event monitoring (SIEM) for correlation and detection, and data loss prevention (DLP) to detect data exfiltration. In practice, these systems do not have high accuracy right out of the box and require tuning to reduce false positives and negatives.

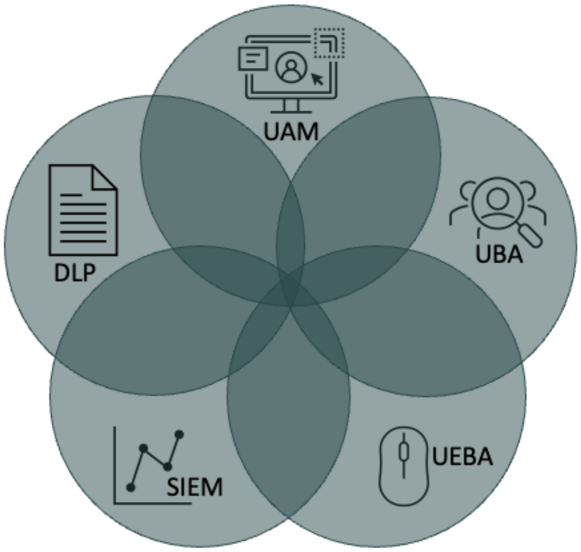

Figure 1: Insider Risk Tools That Commonly Use ML

2 Challenges with AI

Despite, and perhaps because of, significant advancements in AI in recent years, there are many challenges when it comes to implementing AI solutions in practice. We break these challenges into three categories: technical, ethical, and organizational.

2.1 Technical Challenges

Technical challenges of AI encompass both the AI implementation itself and the underlying infrastructure used to support it:

- *Availability of Training Data* – Most implementations of AI require vast amounts of training data. It can be challenging to find datasets large enough and detailed enough to be useful and, even when available, it can be difficult to obtain access due to privacy or regulatory limitations. Labeling data for supervised learning is often a time-consuming and tedious process that can be prone to inaccuracies.
- *Data Fidelity* – It can be difficult to ensure that training data represents the entire population. Easily accessible datasets may be skewed to certain populations or characteristics not representative of the total population. Datasets might be old, incomplete, lacking detail, or simply inaccurate.
- *Expertise* – New AI implementations and models require data scientists to choose, clean, and preprocess the data and AI experts to interpret the problem, then build, train, and test the models. Not all organizations have this expertise in house, and it can be expensive to hire a third party.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- *Accuracy* – Sometimes the solution presented is not accurate enough to be reliable when applied in the real world where the stakes can be high. The lack of accuracy could be due to a skewed or incomplete dataset, a poor model, or an inaccurate interpretation of the results. It can sometimes be difficult to increase accuracy due to limited expertise, limited computing power, the incorrect choice of data, or flawed preprocessing of training data.
- *Scalability* – If the size of the dataset increases, whether due to an increase in the population or an increase in usage, the underlying infrastructure may need to be upgraded and/or the model itself may need to be retrained. AI implementations can require significant processing power, storage infrastructure, and potentially specialized hardware. Retraining or retesting a model will likely require time and effort from data scientists or AI experts to preprocess the new training set or build a new model.
- *Data Security* – In many implementations, the data collected for training or analysis purposes needs to be kept secure. This security requirement is due to privacy or regulatory restrictions or for intellectual property or classification concerns.
- *Robustness* – Models should be designed to withstand "hacking" the system via poisoning or injecting inaccurate data. When appropriate, models should be kept secret to prevent gaming the system (e.g., avoiding keywords that the user knows the system is looking for).

2.2 Ethical Challenges

Besides being inherently important, ethical challenges in AI can affect legal liability, regulatory responsibility, and cultural or organizational response to a particular implementation. Ethical challenges are addressed at length in *IBM Artificial Intelligence Pillars* [IBM 2023] and the EU's *Ethics Guidelines for Trustworthy AI* [AI HLEG 2019b]. The following summarizes many of these challenges:

- *Privacy* – At its most basic level, privacy refers to the appropriate control over and accessibility of data collected and generated by AI systems. Privacy definitions vary by context and legal system, but these definitions can include principles such as consent, accountability, limited use, limited retention, safeguards, access, recourse, and deletion. Privacy goals and requirements should be determined and addressed before building the system. AI systems should also avoid using internally derived features that may violate privacy constraints.
- *Fairness* – AI systems should "ensure that individuals and groups are free from bias, discrimination, and stigmatization" [AI HLEG 2019b]. Unintentional bias can show up in the training data or the model itself in various forms, such as the underrepresentation of minority groups or different decision paths for different groups. Models and data should be carefully evaluated for these kinds of biases. In addition, AI system users and those impacted by its analysis should be able to contest predications or classifications made by the system, and the system should not impact individual autonomy.
- *Transparency* – The challenge of transparency encompasses the following:
  - *Traceability*: clear documentation of all datasets, methods, labels, analysis of results (particularly of any errors), testing and verification methods, and decisions made by the algorithm.

  - *Communication*: identifying the use of an AI system to those who might be affected by it and communicating the system’s appropriate uses, accuracy, and limitations.
- *Explainability* – Both the technical methods used by the algorithm and the decisions made by humans should be clearly reasoned and well documented. The technical methods should be understandable and reproducible by a human. Increasing the system’s explainability may come at the expense of accuracy, since deep learning algorithms in particular are often “black boxes” in terms of how they come to their conclusions.
- *Accountability* – AI systems should be auditable and minimize negative impacts. The developers of algorithms and the organizations that implement them are responsible for complying with ethical principles as well as all applicable laws and regulations. Accountability applies during all phases of the process: from data collection to model building to deployment.

2.3 Organizational Challenges

Even with a promising AI implementation and successful model, organizations may run into the following operational or policy challenges:

- *Cost* – Many AI solutions require vast amounts of data and significant processing power to build and train the models as well as processing power and storage to operationally run the models. Processing power and trained professionals to build the solutions can be expensive.
- *Oversight* – Conflicting legal opinions and the lack of regulatory oversight can make it difficult to implement an AI solution that clearly conforms to policy and legal requirements.
- *Change as a Constant* – Many organizations change, grow, and adapt on an ongoing basis, which leads to changes in data patterns. When data patterns change or new types of data are introduced, models are no longer as accurate and must be retrained or rebuilt.
- *Overtrust (aka Automation Bias)* – Human analysts who rely on the support of AI solutions tend to be biased toward the AI’s results. This can lead to lax oversight, poor decision making, or a loss of skills required to conduct manual analysis with non-AI methods [Ahmad 2023]. This phenomenon has been well documented in many areas, including by cancer radiologists [Dratsch 2023].
- *Expertise* – There can be a lack of expertise to adapt AI to the organization in terms of policy and technical requirements [AI HLEG 2019a]. Policy expertise must include understanding regulatory and privacy data collection and analysis requirements as they relate to AI. Technical expertise is needed to tune a model to the organization and set up the infrastructure to support the implementation.
- *Reusability* – AI implementations are trained (i.e., *tuned*) for a specific organization’s data to solve a specific challenge or accomplish a specific task. So, one organization cannot simply copy a model from another to reuse the AI implementation without tuning or retraining the model to its own organization.

2.4 Examples: Challenges and Lessons

The three example applications described previously have had their own challenges in real-world implementations. The challenges in the following list can provide valuable insight when evaluating AI implementations for insider risk:

- *Online Exam Monitoring* – Several universities have decided not to renew their contracts with online exam monitoring providers due to numerous issues [Bergmans 2021, Nigam 2021], including the following:
  - not successfully identifying cheaters with a high enough level of accuracy
  - test takers' concerns about who can access and use their data
  - potential legal issues related to privacy
  - racial bias when using web camera monitoring
  - effects of monitoring on test takers, including anxiety, stress, and lack of focus
  - invasiveness, including controlling the test taker's system and viewing their entire room
  - perceived unfairness by those taking the exams in person
  - overtrust in the AI results on the part of the human proctors
  - difficulty of challenging results when accused
- *Vaccine Communication* – This is one domain where several studies [Passanante 2023, Karinshak 2023, Lee 2023] analyzed the efficacy of using NLP AI for intent detection and intervention. The study results describe the following:
  - Users have a strong preference for being told they are interacting with AI.
  - Since some implementations do not have access to the subject's personal history and background due to privacy and ethical concerns, the algorithms are unable to provide tailored analysis and intervention, causing concerns about inaccurate recommendations.
  - Long-term effects of direct interaction with AI (in this case, with chatbots) have yet to be studied due to their relative novelty.
  - Sentiment analysis algorithms in other domains have had issues with bias [Poria 2023, Rozado 2020], particularly in scenarios where they are more likely to label text from men or ethnic minorities as negative.
  - Sentiment analysis algorithms in other domains have had inaccuracies with emotion detection [Mao 2023].
- *Welfare Fraud Risk* – There were technical, ethical, and legal issues that were specific to Rotterdam's implementation of AI:
  - There was bias in data collection (e.g., anonymous tips, investigation of specific zip codes).
  - The sample size used for training was too small.
  - Some data was overly subjective (e.g., from case worker notes).
  - Investigation of flagged individuals was extremely disruptive and invasive.
  - The training data lacked detail (e.g., *fraud* has a wide range of implications).
  - The model discriminated based on gender and ethnicity.
  - The legality of the algorithm's discrimination was unclear in this jurisdiction.

  - Results were purely predictive (i.e., subjects did not have to do anything wrong to be flagged).
  - Results were extremely difficult for flagged individuals to challenge.
  - The model did not meet accuracy or performance standards.

2.5 Challenges Specific to Insider Risk

The challenges in the previous section vary widely in their extent and impact across sectors, regions, and applications; however, many of these challenges also apply to the insider risk domain and are explained the following three subsections.

2.5.1 Technical

There are technical challenges specific to AI's implementation in the insider risk domain:

- There is a lack of real-world data for training and testing models, particularly in a public, rigorously reviewed way.
- There is a lack of ground truth in existing training data (i.e., it is not always clear when an insider action has gone undetected).
- Existing training data does not represent all types of organizations that may want to use AI for insider risk.
- Solutions need to be tuned to an organization before they can be used. This tuning takes time and requires expertise.
- Insider threat data is extremely skewed (i.e., there are very few insider incidents compared to non-insider incidents). This skew in the data limits the choices of available algorithms and changes how accuracy should be measured.
- *Low and slow* attacks are still less likely to be detected. Running algorithms over data with long time horizons causes problems with scalability, privacy, accuracy, and level of granularity in the data.
- Robustness can be affected through intentional or unintentional attacks. For example, an employee with a grudge against another employee can submit a false report to HR, or an information technology (IT) issue could cause data loss for a subset of employees or a subset of features (i.e., attributes).

2.5.2 Ethical

The significant power imbalance between employers and employees requires that the ethics of AI implementations in the insider risk domain be treated as critical. Results from these solutions can affect individuals' livelihoods, reputations, and mental health; these results can also have potential legal ramifications:

- Extensive privacy challenges, including legal constraints, prevent or discourage the collection of and access to certain types of data. These limitations can cause problems with accuracy and may also limit third-party auditing of the methods used.

- Monitoring employee actions and communications can have mental health impacts (e.g., cause stress and anxiety). This monitoring can also lead employees to hesitate to take actions required by their job role for fear of being marked as a risk. For example, an employee might not report travel as required because they think it might increase their risk score.
- Geographically diverse organizations may have different legal requirements in different regions due to varying privacy regulations. This variation can lead to varying models within the organization, which can cause inequities across different regions.
- There is no existing analysis of bias in current insider threat datasets or algorithms.[2] However, bias is known to exist in algorithms for related domains (e.g., policing, fraud detection, hiring).
- Many commercial off-the-shelf systems are not fully explainable or transparent.
- Predictive scoring systems in general, already commonly used in insider risk analysis, are specifically called out by AI ethics groups as problematic due to the challenges outlined in the welfare example from the previous section. Specifically, the European Union's high-level expert group on artificial intelligence (HLEG) warns that "a fully transparent procedure should be made available to citizens, including information on the process, purpose and methodology of the scoring […] mechanisms for challenging and rectifying the scores must be given. This is particularly important in situations where an asymmetry of power exists between the parties" [AI HLEG 2019a].

2.5.3 Organizational

- Since models are based on patterns in data, they have a limited ability to incorporate one-off events that may trigger insider risks (e.g., layoffs, performance reviews, changes in benefits).
- The solutions are expensive to purchase/build, run, and maintain. They require ongoing investment in infrastructure, expertise, and (in some cases) per-unit licensing costs.
- Lack of ground truth makes it difficult to quantitatively prove that an AI solution is worth the cost.
- Each organization must comply with the laws, regulations, and requirements in its jurisdiction. Examples of these include chain of custody requirements; the EU's General Data Protection Regulation (GDPR), California's Consumer Privacy Rights Act (CPRA), or other privacy laws; redress/grievance opportunities; anti-discrimination laws; and data retention regulations.

[2] To the best of our knowledge, as of publication of this report, such an analysis does not exist.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

3 Incorporating AI Carefully

Though the challenges with using AI for insider risk are extensive, we recognize that in many cases the risk of not having AI-enabled assistance is greater than the risks and challenges that come with using it. It is therefore important that members of the insider risk analysis community learn to carefully incorporate AI into their processes and tools. Unfortunately, it is not possible to recommend a specific, one-size-fits-all tool, algorithm, or data collection method that will work successfully across all organizations and applications. Choosing an AI-enabled insider risk solution depends on many factors, including the following factors specific to the organization:

- risk tolerance
- privacy regulations
- AI laws
- available data
- expertise
- budget

Equipped with information about these organization-specific factors, there is a pathway for making decisions about and when to incorporate AI into the organization's insider risk workflow. Figure 2 shows this pathway, which is explained in the paragraphs that follow.

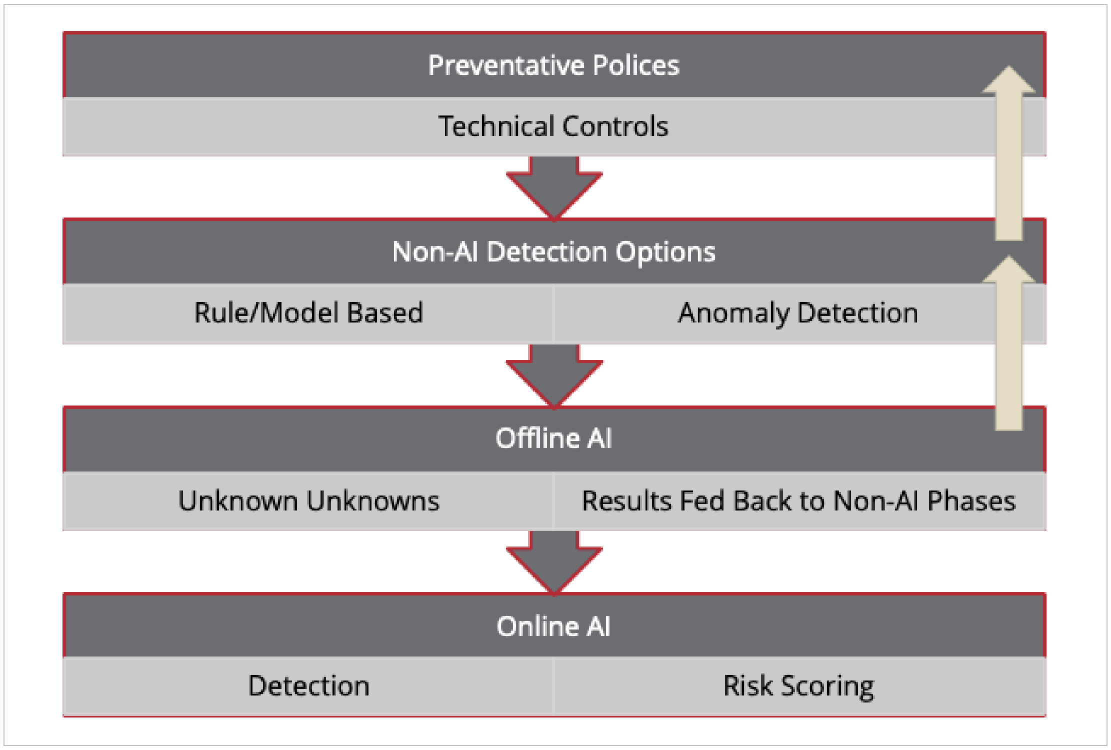

Figure 2: Pathway for Incorporating AI into Insider Risk Analysis

Given all of the challenges with AI systems, it is important to consider non-AI solutions to the problem of insider risk before turning to the more complex, expensive, and potentially detrimental AI-based alternatives. Insider risk programs should initially focus on prevention via technical and procedural intervention for enforcing policies. For example, organizations can turn off the ability to use Universal Serial Bus (USB) drives to copy files, or they can implement a separation-of-duties procedure that requires approval of two separate individuals to successfully create a new account.

Once prevention-based options are exhausted, organizations should explore non-AI options for detection (e.g., rule-based, anomaly, or model-based detection methods). For example, an organization could use DLP to detect outbound emails with labeled confidential documents attached or send an alert when there is an atypical pattern of file access. By working through these options, whether starting with a use case or a risk analysis, organizations can create useful solutions to existing issues.

AI solutions are most useful when finding unknown unknowns. Once an organization has exhausted rules and policies, instituted mechanisms to alert on anomalies, and implemented solutions for potential concerning use cases, AI systems can be added to help detect and characterize any residual issues. AI should not be applied to large amounts of unprocessed data simply to “see what sticks.” Doing so can eventually lead to inaccuracies, lack of transparency, over-weighted features, bias, and scalability issues. A correctly implemented AI solution for insider risk must carefully define the problem to be solved and curate the data that goes into the system.

To find unknown unknowns, first consider using the AI system in an offline mode, where data is collected and stored on a central system so that it can be run and rerun as needed without affecting real-time operations. Using this offline approach can identify patterns of behavior an organization has not yet considered, highlight new questions to be answered, identify new features or types of data that should be considered for use in other detection tools, and support requests for more resources on the cybersecurity or insider risk teams.

Approaching AI as an offline pattern detection solution also eases some of the ethical and infrastructure concerns raised in the previous section. For example, an organization might try a clustering method on a collection of data, which may lead to discovering that the duration of time a user is logged off during midday is a metric that can be used to identify concerning behavior. That feature can then be implemented in simpler tools using rule or pattern-based detection mechanisms.

Using AI systems in an online (i.e., live) mode should be considered only once the previous steps have been addressed to the best of the organization’s capabilities. Online AI systems are used to directly detect concerning behavior in real time or to predict potential insiders by calculating risk scores for individuals on an ongoing basis. The following sections provide guidelines for designing, implementing, and using AI systems for insider risk evaluation.

3.1 During Design

Using the following good design principles from the start will limit problems during deployment and minimize challenges down the road:

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- Start by establishing an internal team of stakeholders, including cybersecurity, legal, HR, IT, risk management, and insider risk teams as well as individuals who can represent employee interests.
- Establish third-party oversight from an external party or a separate internal review board to audit AI solutions and review changes during the lifecycle of the system.
- Define a goal for the system that has a positive outcome, such as “protect customer data from irresponsible or inappropriate use,” or “support the organization’s mission by ensuring the availability of information resources” or “reduce the workload of human insider risk analysts.”
- Determine whether to try to detect or predict insider threats. For example, “detect inappropriate use of customer account information” versus “predict employees most likely to misuse customer account information.” If possible, avoid risk scoring and predictive analytics. If not possible, use predictive scoring only in limited scenarios to support positive early intervention for specific critical use cases (e.g., workplace violence).
- Narrowly define the use case or problem being addressed. For example, “detect employees using the organization’s resources for personal projects” or “predict employees most likely to conduct physical workplace violence.”
- Ensure policies and procedures are in place ahead of time to appropriately handle output from the AI system. Know who will review the results and what they will do with them.
- Create plans for how to handle inaccurate or contested results. Ensure there is a process that enables employees to contest decisions.
- Use the system to support policy and technical changes across the board rather than to penalize or monitor individual employees.
- Create data access and cybersecurity plans for the entire system lifecycle to ensure privacy and integrity of the data and output against both internal and external attacks.
- Consider ways the system might be misused in the future and establish policies to prevent that. For example, risk scoring metrics should not be included in promotion or salary discussions. A system built to predict workplace violence should not be used to determine which employees are most likely to commit fraud.

3.2 During Development or Procurement

Whether developing a system in house, contracting externally for a new system, or purchasing an off-the-shelf system, incorporate the following guidelines into the process:

3.2.1 Metrics

- Understand the different metrics used for AI systems,[3] as well as cost, power usage, and processing speed. Do not base procurement or deployment decisions on the accuracy metric alone; precision or recall is usually far more important.
- Choose metrics that reflect the organization's desired goal for the system.
- Set a high bar for the results and performance of the system. Use lessons from other domains as well as any legal or regulatory guidance to help set that bar.
- Understand how much data and processing power will be needed to support the *full* solution to achieve your stated goal. Otherwise, the system may be limited to just a few capabilities of its full feature set.
- Ask similar organizations about their experiences implementing a new AI solution. Were there any surprise costs? How long did it take to tune (i.e., reduce false positives to an acceptable level)? Is it generating the results you expected?

3.2.2 Data

- Ensure quality and integrity of the data (i.e., know where the data comes from and review the data for inaccuracies, missing features, and biases prior to training).
- Use non-technical features (i.e., personal or behavioral features) sparingly or not at all. Use them only as supplements to subsequent investigations, not as part of the AI system.
- Consider using anonymized data that cannot be traced back to a particular user. The system can point analysts to an event, data type, or time window for further investigation, or it can alert to a gap in policy or technical controls.
- Training data should be as inclusive as possible (i.e., not just from one location, department, job role, or time window).
- Be careful about using *association* (i.e., tracing social networks or membership in groups) as a feature, since that can violate freedom of association laws. However, it might be useful to watch for *changes* in networks.
- Do not use discriminatory features such as gender or ethnicity. Carefully consider features that may act as proxies for discriminatory features (e.g., credit scores standing in for race or an NLP keying on phrases that correlate with gender or age) [Prince 2020].

3.2.3 Methods

- Choose algorithms with explainability in mind. Consider supervised or hierarchical methods over deep learning. Choose algorithms that can point to the data or pattern that triggered the alert.
- Discuss, document, and address any tradeoffs made between accuracy and explainability of the system, and why those tradeoffs fit the stated goal of the system.

[3] Common metrics related to AI accuracy are accuracy, precision, recall, f-1, and ROC curves. Ensure that key decision makers understand the differences between these metrics.

- Document all steps for data collection, training, and testing as well as any errors and all accuracy metrics.
- Identify the risks to the system (technical, ethical, or organizational) and find ways to mitigate them.
- Consider the role of fairness in your algorithm. For example, if it weighs its decision against employees emailing external organizations (to catch intellectual property theft), that may be unfair to employees whose job it is to regularly send external emails (e.g., the media relations team). Ensure the AI model or organizational policies account for the possibility of unfairness.
- If possible, use local-only decision making (i.e., do not send data back to the cloud for processing).

3.2.4 Testing

- Use a variety of test cases to understand the system's biases and limitations.
- If purchasing a solution, ask for a trial run to test on the organization's production network.
- Test which use cases the system will catch and which ones it will not catch. For example, a fraud incident looks very different than a workplace violence incident.
- Determine under what circumstances the tests are reproduceable and when they are not.
- Test whether the solution works across the entire user population, including vendors, temporary employees, interns, and contractors. If not, create a plan to handle the differences fairly or deal with the lack of coverage across certain subpopulations.

3.2.5 Compliance

- Ask legal counsel to help ensure compliance with all applicable laws and regulations, particularly for privacy, data storage and retention regulations, and employee rights.
- Ensure all AI methods, tools, and policies are reviewed and approved by the organization's legal counsel and other necessary parties (e.g., an internal review board, an external auditor).
- If purchasing an off-the-shelf solution, ask vendors detailed questions to obtain answers that address the guidelines in this section or ask for a third-party audit report.

3.3 During Deployment

When deploying the AI system, incorporate the following guidelines into the process:

- Verify that the implementation works in its production environment by reproducing the test results from the development environment if possible.
- Conduct a red team assessment against the system, if appropriate, based on the algorithm and the data used. For example, simulate the loss of a key data feature as the result of a network outage.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

- Consider using other metrics in addition to the number of users caught, such as the number of false positives, the number of policy or technical changes supported, or the number of employees referred to early counseling.
- Set up technical and policy controls to limit, monitor, and log changes to the system, including changes to the model and data collection.
- Place significant limits on who can access the underlying data or the output of the model. Closely monitor and track who accesses the data or output, when they access it, and why.
- Create a separation of duties between the analysts using system results and the analysts or data scientists collecting the data and training the system.
- Train all stakeholders on how the algorithm works, the data it uses, how it makes its decisions, its accuracy, and its limitations.

3.4 During Use

AI solutions require ongoing care and maintenance to ensure fairness and accuracy, so incorporate the following guidelines into the maintenance process:

- Baseline your AI solution against the results from your human analyst team (without assistance from AI). Analyze any differences carefully and implement changes to the analysis procedures or AI solution as appropriate.
- Use the AI solution to support the analysts rather than replace them.
- Test the AI solution regularly for bias, accuracy, and reproducibility. Determine if certain subpopulations are overrepresented or underrepresented in the output.
- Check for intentional or unintentional adversarial attacks against the system by pairing it with anomaly reporting as data is ingested to detect atypical data. Investigate the causes of atypical data.
- Use the AI solution only for its intended purpose, as defined during the design phase.
- Retrain the system when appropriate, document and report any changes, and involve stakeholders in decisions regarding changes to the system.
- Monitor for negative effects (e.g., anxiety, stress) on employees, particularly if using invasive monitoring (e.g., webcams, keystroke loggers, communications monitoring, detailed technical process tracking).
- Provide a way for employees to anonymously report concerns with the system.
- Take steps to prevent analysts' overreliance or overconfidence on the system. For example, have analysts routinely conduct investigations without the help of AI to teach them how AI works, and have them regularly review the errors and accuracy metrics of the system.
- Do not automatically take any actions based on the AI solution's output. For example, do not automatically deactivate an employee's account once their risk score goes above a certain number.
- Keep all historical data (not just the data used to trigger an alert) to support redress.
- If using predictive scoring, consider allowing employees to see their scores and the components that make up that score. This transparency allows employees to improve their

understanding and behavior, presents an opportunity for redress, and supports positive deterrence, similar to pulling a credit report.

- Use the results to support policy improvements and technical mitigations. Consider removing the AI solution (or a model for a specific goal) once the policy and related technical changes are in place.

4 Conclusion

Artificial intelligence has a definite place in insider risk analysis. Through thoughtful design and careful application, it is possible to use AI to support and enhance detecting insider risk effectively, ethically, and efficiently.

The following are some of the most exigent challenges in applying AI to the insider risk domain:

- a **lack of real-world, ground-truth data** for building models and for third-party auditing
- the **changing legality** of predictive algorithms for applications affecting human rights and liberties
- **lack of transparency and explainability** of certain types of algorithms
- **limited ability** to catch some of the most damaging stealthy insider attacks
- **cost** of purchasing, implementing, supporting, and maintaining the applications
- **lack of internal organizational expertise** for fully informed decision making

An awareness of these limitations and the others listed in the previous sections can go a long way in facilitating the thoughtful use of AI in an organization’s insider risk program.

The following are most crucial conditions for successfully implementing AI for insider risk:

1. Be specific about how and why AI is being used.
2. Use AI in a limited capacity after other options have been exhausted.
3. Be realistic about outcomes, cost, and effort.
4. Be involved in the decision-making process.

References

URLs are valid as of the publication date of this report.

[Ahmad 2023]

Ahmad, S. F.; Han, H.; Alam, M. M.; Rehmat, M. K.; Irshad, M.; Arraño-Muños, M.; & Ariza-Montes, A. Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and

Safety in Education. *Humanities and Social Sciences Communications*. Volume 10. Issue 1. June 2023. Pages 1–14. http://dx.doi.org/10.1057/s41599-023-01787-8

[AI HLEG 2019a]

High-Level Expert Group on Artificial Intelligence (AI HLEG). *A Definition of AI: Main Capabilities and Disciplines*. European Commission. April 2019. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[AI HLEG 2019b]

High-Level Expert Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI. *European Commission Website*. April 8, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[Bergmans 2021]

Bergmans, L.; Bouai, N.; Luttikhuis, M.; & Rensink, A. On the Efficacy of Online Proctoring Using Protorio. Pages 279–290. In *Proceedings of the 13th International Conference on Computer Supported Education - Volume 1: CSEDU*. April 2021. https://www.doi.org/10.5220/0010399602790290

[Braun 2023]

Braun, J.; Constantaras, E.; Aung, H.; Geiger, G.; Mehrotra, D.; & Howden, D. Suspicion Machine Methodology. *Lighthouse Reports Website*. March 2, 2023. www.lighthousereports.com/methodology/suspicion-machine

[Chan 2020]

Chan, H.; Samala, R. K.; & Hadjiiski, L. M. CAD and AI for Breast Cancer—Recent Development and Challenges. *The British Journal of Radiology.* Volume 93. Issue 1108. April 1, 2020. Page 20190580. https://doi.org/10.1259/bjr.20190580

[Dratsch 2023]

Dratsch, T.; Chen, X.; Rezazade, M.; Kloeckner, R.; Mähringer-Kunz, A.; Püsken, M.; Baeβler, B.; Sauer, S.; Maintz, D.; & Pinto dos Santos, D. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*. Volume 307. Issue 4. May 2023. https://doi.org/10.1148/radiol.222176

[Foroughi 2018]

Foroughi, F. & Luksch, P. Observation Measures to Profile User Security Behaviour. Pages 1–6. In *2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security).* June 2018. https://www.doi.org/10.1109/CyberSecPODS.2018.8560686

[IBM 2023]

IBM. IBM Artificial Intelligence Pillars. *IBM Website.* August 30, 2023. www.ibm.com/policy/ibm-artificial-intelligence-pillars/

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

[Janai 2020]
Janai, J.; Güney, F.; Behl, A.; & Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision*. Volume 12. Issue 1–3. July 6, 2020. Pages 1–308. https://doi.org/10.1561/0600000079

[Kamath 2019]
Kamath, U.; Liu, J.; Whitaker, J. *Deep Learning for NLP and Speech Recognition*. Springer. 2019. ISBN: 978-3030145958. https://doi.org/10.1007/978-3-030-14596-5

[Karinshak 2023]
Karinshak, E.; Liu, S. X.; Park, J. S.; & Hancock, J. T. Working with AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*. Volume 7. Issue CSCW1. Article Number: 116. April 16, 2023. Pages 1–29. https://doi.org/10.1145/3579592

[Le 2020]
Le, D. C. & Zincir-Heywood, N. Exploring Adversarial Properties of Insider Threat Detection. Pages 1–9. In *2020 IEEE Conference on Communications and Network Security (CNS)*. June 2020. https://www.doi.org/10.1109/CNS48642.2020.9162254

[Lee 2023]
Lee, K. Y.; Dabak, S. V.; Kong, V. H.; et al. Effectiveness of Chatbots on COVID Vaccine Confidence and Acceptance in Thailand, Hong Kong, and Singapore. *npj Digital Medicine*. Volume 6. Issue 96. May 25, 2023. https://doi.org/10.1038/s41746-023-00843-6

[Lu 2019]
Lu, J. & Wong, R. Insider Threat Detection with Long Short-Term Memory. Pages 1–10. In *Proceedings of the Australasian Computer Science Week Multiconference (ACSW '19)*. January 2019. https://doi.org/10.1145/3290688.3290692

[Mao 2023]
Mao, R.; Liu, Q.; He, K.; Li, W.; & Cambria, E. The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Transactions on Affective Computing*. Volume 14. Issue 3. July–Sept. 2023. Pages 1743–1753. https://www.doi.org/10.1109/TAFFC.2022.3204972

[Mehrotra 2023]
Mehrotra, D.; Constantaras, E.; Geiber, G.; Braun, J.; & Aung, H. Inside the Suspicion Machine. *Wired*. March 6, 2023. www.wired.com/story/welfare-state-algorithms/

[Nigam 2021]
Nigam, A.; Pasricha, R.; Singh, T.; & Churi, P. A Systematic Review on AI-Based Proctoring Systems: Past, Present, and Future. *Education and Information Technologies*. Volume 26. June 23, 2021. Pages 6421–6445. https://doi.org/10.1007/s10639-021-10597-x

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

[Passanante 2023]
Passanante A.; Pertwee, E.; Lin, L.; Lee, K. L. Y.; Wu, J. T.; & Larson, H. J. Conversational AI and Vaccine Communication: Systematic Review of the Evidence. *Journal of Medical Internet Research*. Volume 25. Article Number e42758. October 2023. https://doi.org/10.2196/42758

[Poria 2023]
Poria, S.; Hazarika, D.; Majumder, N.; & Mihalcia, R. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*. Volume 14. Issue 1. January-March 2023. Pages 108–132. https://doi.org/10.48550/arXiv.2005.00357

[Prince 2020]
Prince, A. & Schwarcz, D. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*. Volume 105. Issue 3. March 15, 2020. Pages 1257–1318. https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data

[Rozado 2020]
Rozado, D. Wide Range Screening of Algorithmic Bias in Word Embedding Models Using Large Sentiment Lexicons Reveals Underreported Bias Types. *PLOS ONE*. Volume 15. Issue 4. April 2020. Page e0231189. https://doi.org/10.48550/arXiv.1905.11985

[Schroeder 2020]
Schroeder, T. Introducing Google Cloud Sentiment Analysis: A Foundation for a Successful COVID-19 Vaccination Strategy. *Google Cloud Website*. December 17, 2020. https://cloud.google.com/blog/topics/public-sector/introducing-google-cloud-sentiment-analysis-foundation-successful-covid-19-vaccination-strategy

[Silver 2018]
Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; & Hassabis, D. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play. *Science*. Volume 362. Pages 1140–1144. https://doi.org/10.48550/arXiv.1712.01815

[Wankhade 2022]
Wankhade, M.; Rao, A. C. S.; & Kulkarni, C. A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artificial Intelligence Review*. Volume 55. Issue 7. Pages 5731–5780. February 7, 2022. https://doi.org/10.1007/s10462-022-10144-1

Legal Markings

Copyright 2024 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Requests for permission for non-licensed uses should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM24-1412

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu
Email: info@sei.cmu.edu