



12th Annual Insider Risk Management Symposium

JUNE 12, 2025

**Carnegie
Mellon
University**
Software
Engineering
Institute

Document Markings

Copyright 2025 Carnegie Mellon University.

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.

References herein to any specific entity, product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

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.

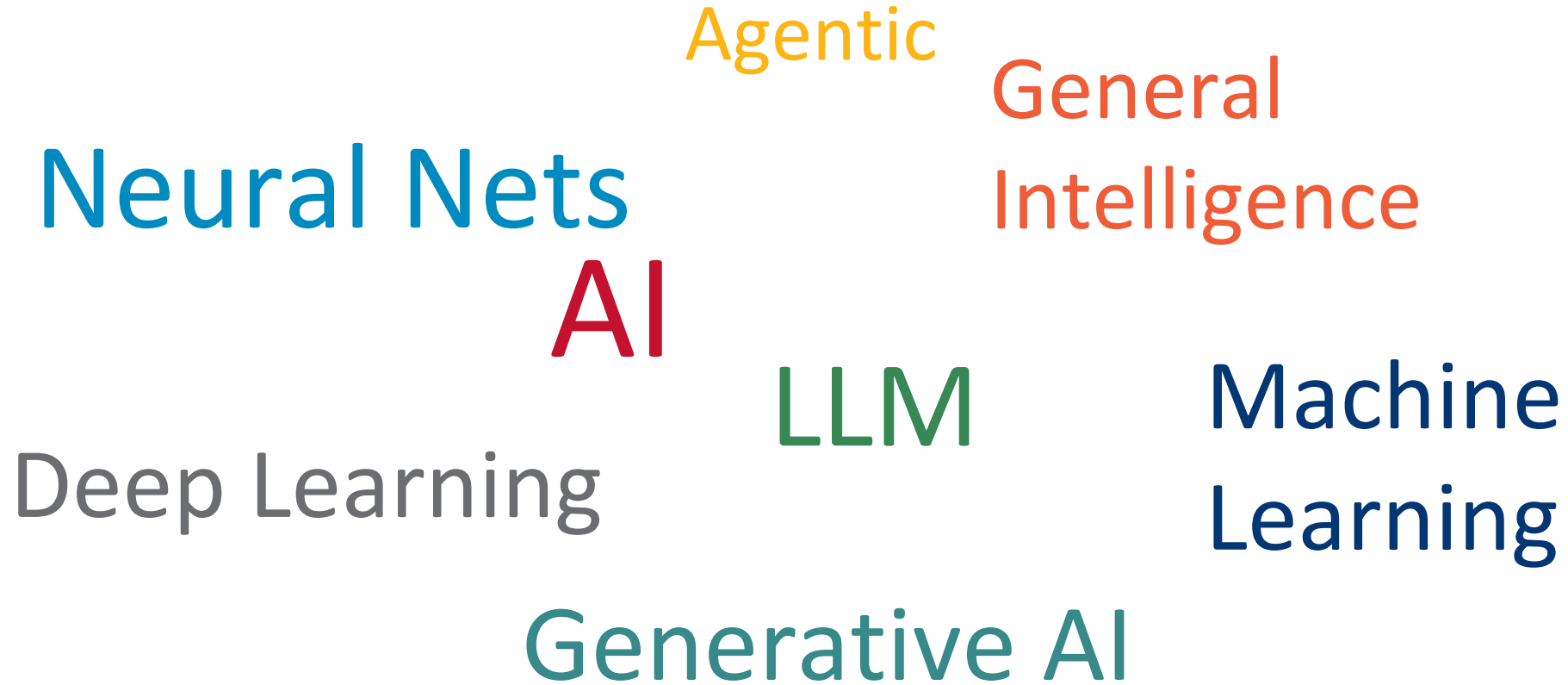
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

CERT® and Carnegie Mellon® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

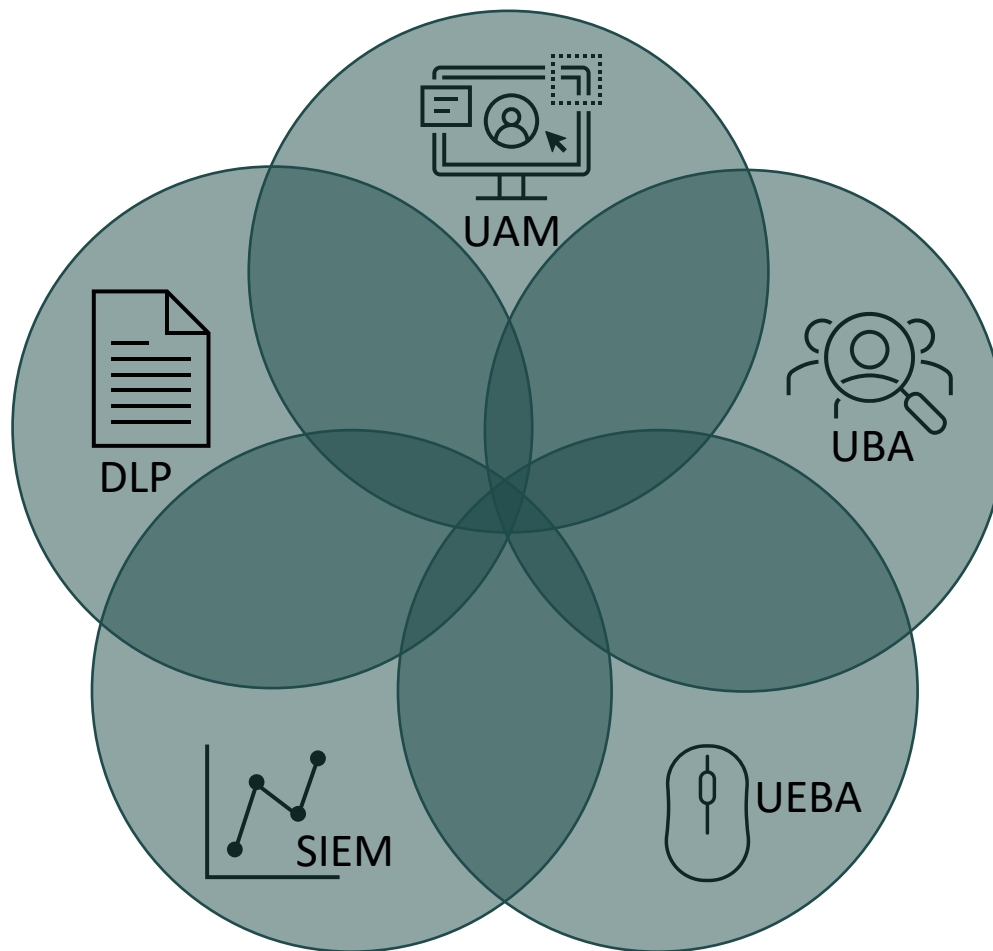
DM25-0807

Artificial Intelligence



State of AI for Insider Threat

- Classification of threats
- Anomaly detection
- Predictive analytics
- Risk scoring
- Text analysis
- Image recognition



An Insider Threat-Adjacent Example

The city of Rotterdam used a supervised machine learning **algorithm** to generate a **score** for **individuals most at risk** of committing welfare fraud. The algorithm used 315 different attributes such as age, gender, language fluency, number of children, etc. The results were used to initiate fraud **investigations against the top 10%** of highest risk scores.

Inside the Suspicion Machine

Obscure government algorithms are making life-changing decisions about millions of people around the world. Here, for the first time, we reveal how one of these systems works.

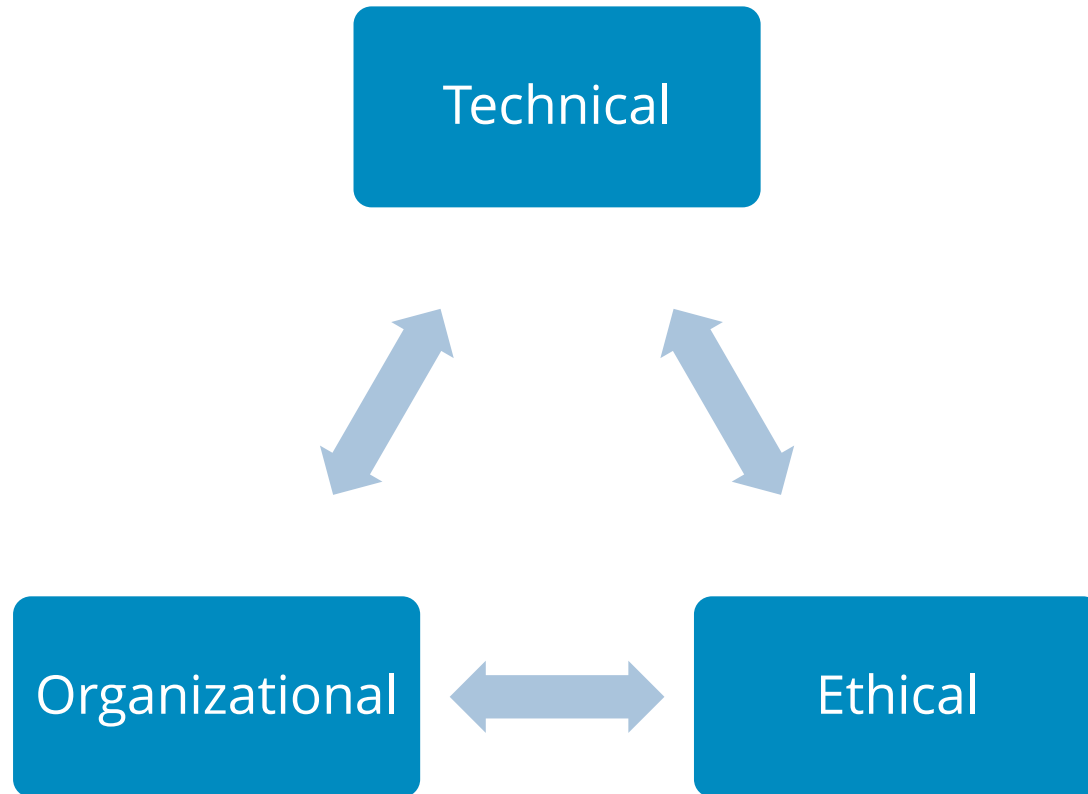
EVA CONSTANTARAS, GABRIEL GEIGER, JUSTIN-CASIMIR BRAUN, DHRUV MEHROTRA, HTET AUNG
MAR 6, 2023 7:00 AM

<https://www.wired.com/story/welfare-state-algorithms/>

<https://www.lighthousereports.com/methodology/suspicion-machine/>

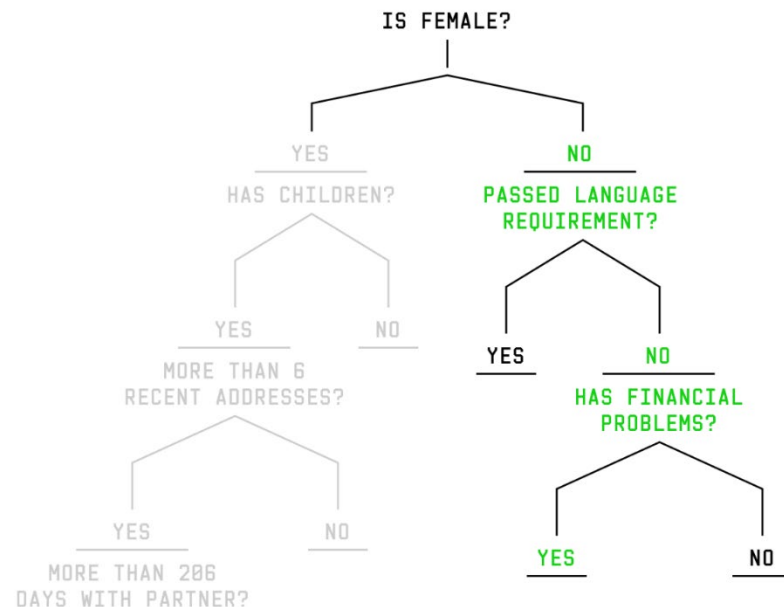
30,000 families were wrongly accused of fraud

Challenges



Rotterdam's Challenges in Predictive Scoring

- Potential **bias** in the data collection
- The **sample size** used for training was too small
- Some data was overly **subjective**
- **Invasive**, both during the data collection and investigation phases
- **Lack of detail** in the training data
- **Model discriminated** based on gender and ethnicity
- **Legality** of discrimination on the part of an algorithm was unclear in this jurisdiction
- Results were purely **predictive**
- Results were extremely difficult for flagged individuals **to challenge**
- The model did not meet good **accuracy** or performance standards



Technical



Data Availability



Data Fidelity



Expertise



Accuracy



Scalability

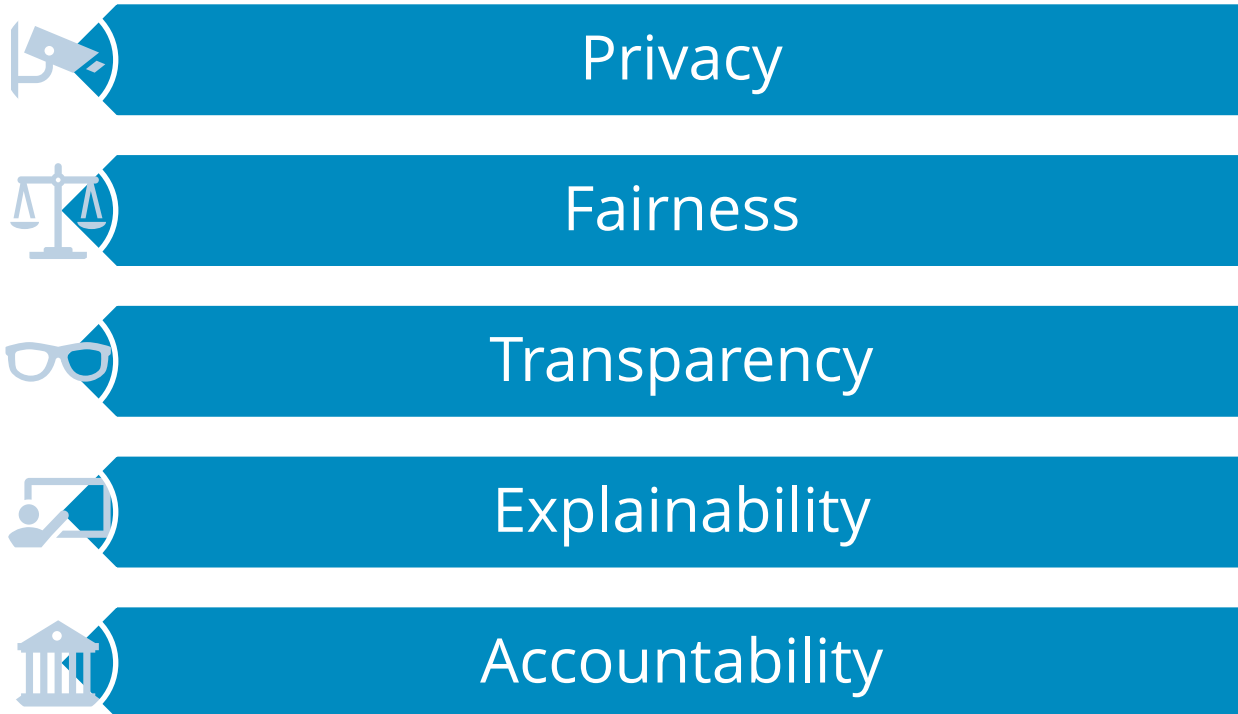


Data Security



Robustness

Ethical



<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Organizational



Cost



Oversight



Overtrust



Reusability



Change



Expertise

What About Insider Threat?

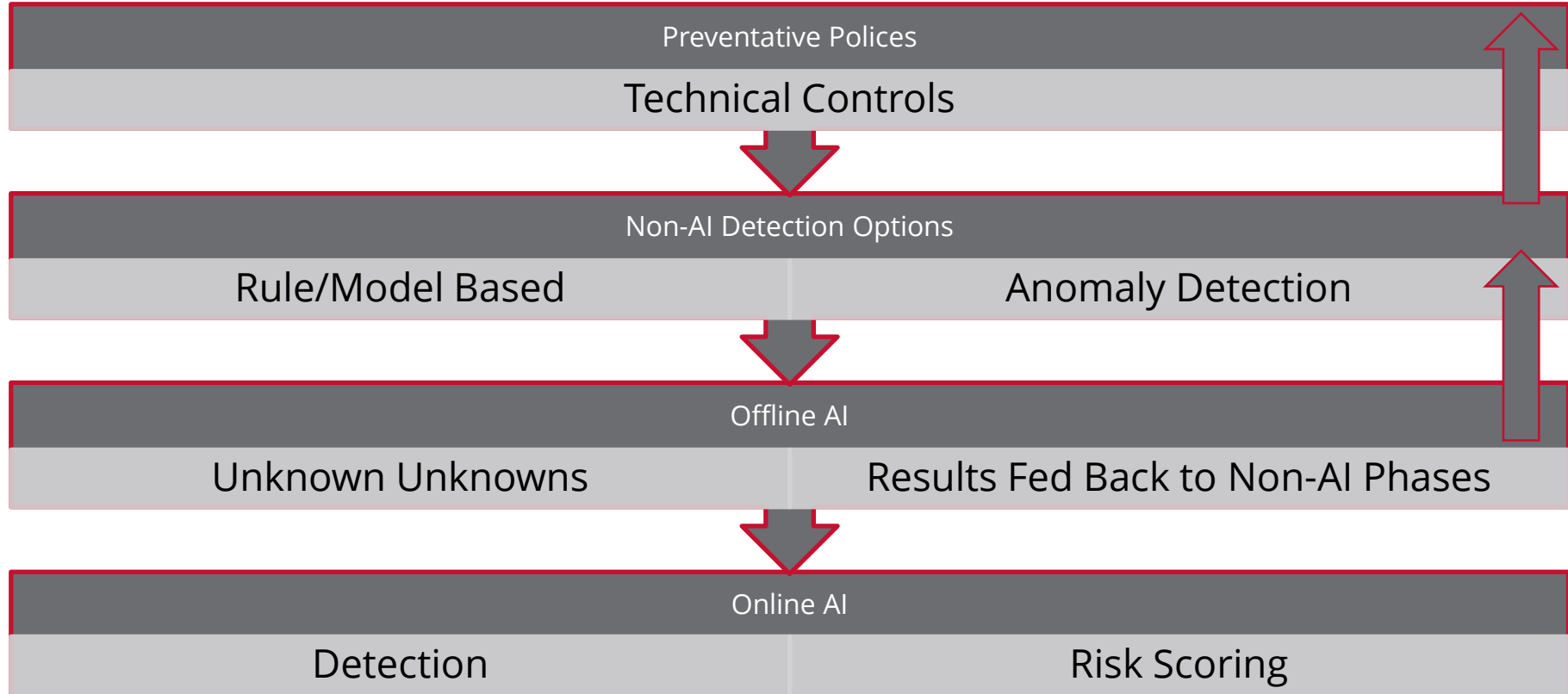
- There exists a **lack of real-world data** for training and testing models, particularly for third party auditing
- There is a **lack of ground truth** in existing training data – i.e. it is not always clear whether an action or event was actually an insider action that went undetected.
- Existing training data is **not representative** of 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
- The data for this domain is extremely **skewed** – i.e. there are very few insider incidents compared to non-insider incidents. This impacts how we chose and train algorithms.
- **“Low and slow”** attacks are still less likely to be caught

The Right Data and Algorithm...

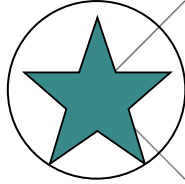
...depends on your specific:

- Use case
- Risk appetite
- Applicable privacy regulations
- Applicable AI laws
- Available data
- Expertise
- Budget

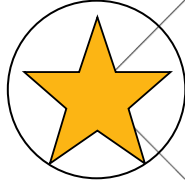
Doing it Right



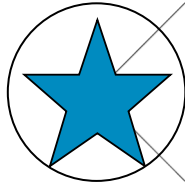
AI Has Its Place



Use AI in a limited capacity after other options have been exhausted



Be realistic about outcomes, cost, and effort



Be specific about how and why AI is being used



Be involved in the decision process

Questions



Contact



Austin Whisnant
Senior Researcher
Insider Risk, CERT, SEI, CMU

Email: abwhisnant@sei.cmu.edu

Some Guidelines

