Threats and Risks to AI:
The Challenge

NATHAN HAMIEL
SENIOR DIRECTOR OF RESEARCH

KUDELSKI SECURITY

Challenges

- Different perspectives and priorities
- Responsibilities not defined
- Legacy processes and tooling
- Lack of AI knowledge on the security team
- Messy and complex world
- Governance not implemented
- Model implementations are highly specific
- Regulatory requirements
- Improper project definitions
- Lack of appropriate benchmarks

KUDELSKI SECURITY

Nathan Hamiel

SENIOR DIRECTOR OF RESEARCH

Security of Emerging Technologies

International Public Speaker

Black Hat Review Board Member

Track Lead: AI, ML, and Data Science

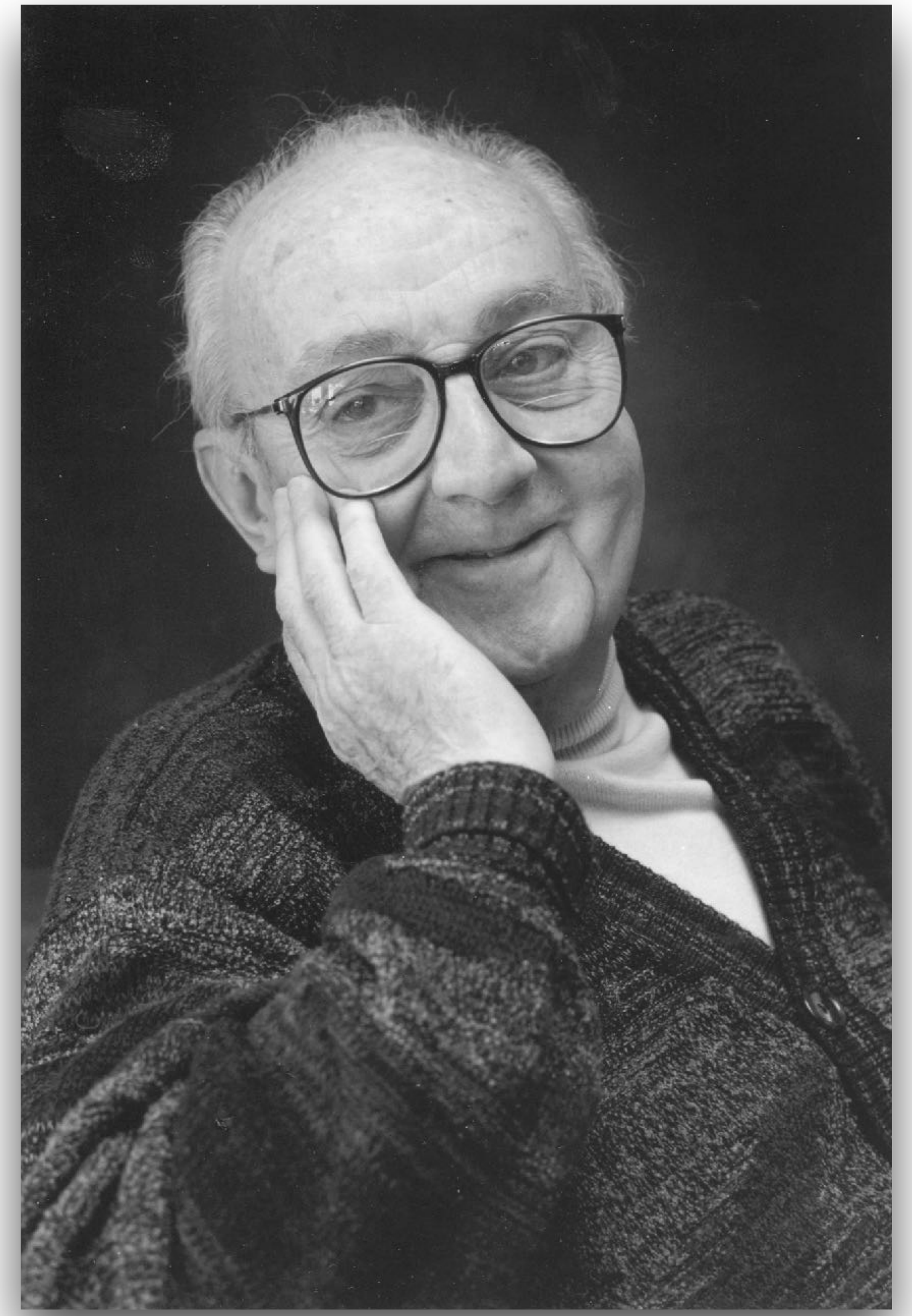
KUDELSKI SECURITY

https://kudelskisecurity.com

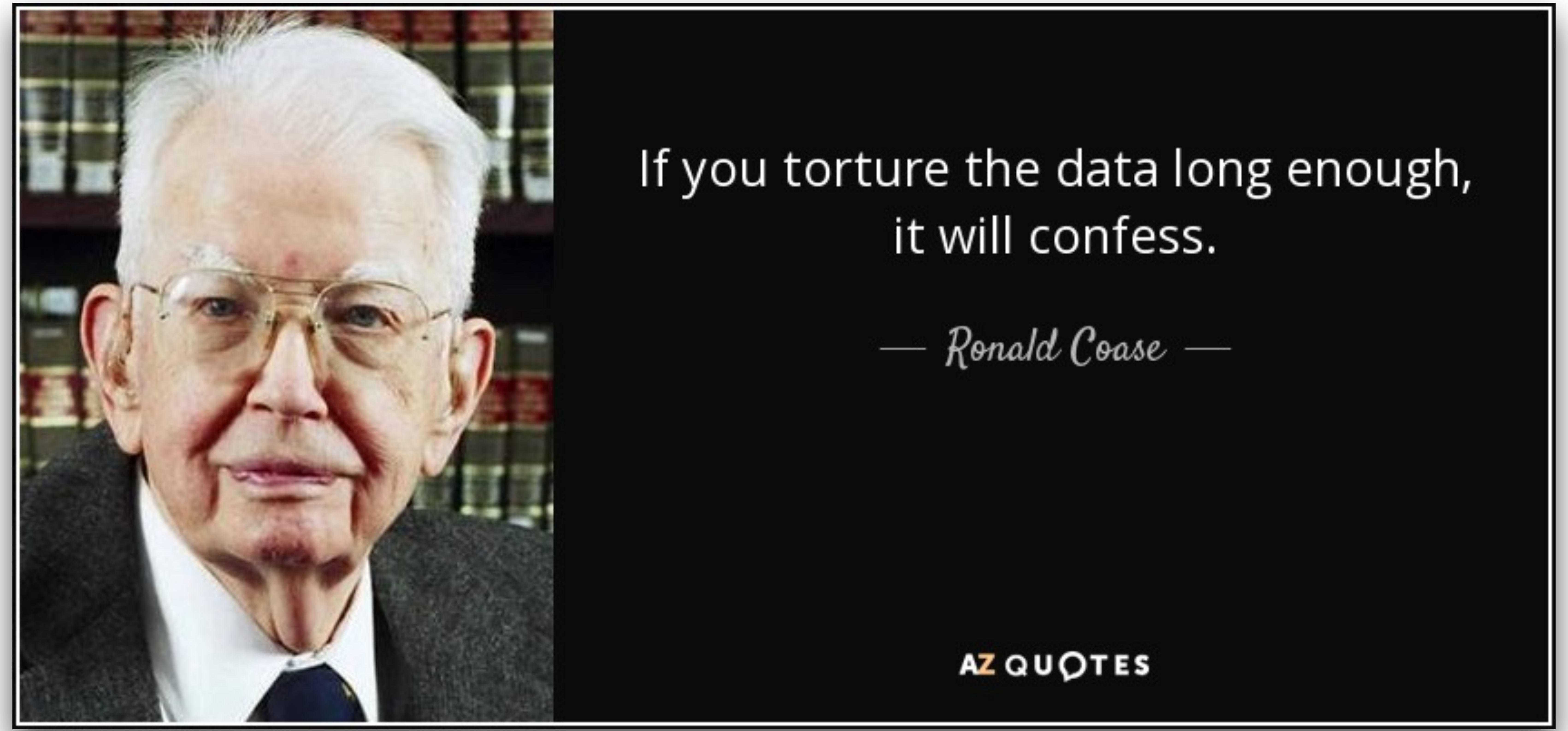
Models

"All models are wrong, but some are useful."

-George Box



Data



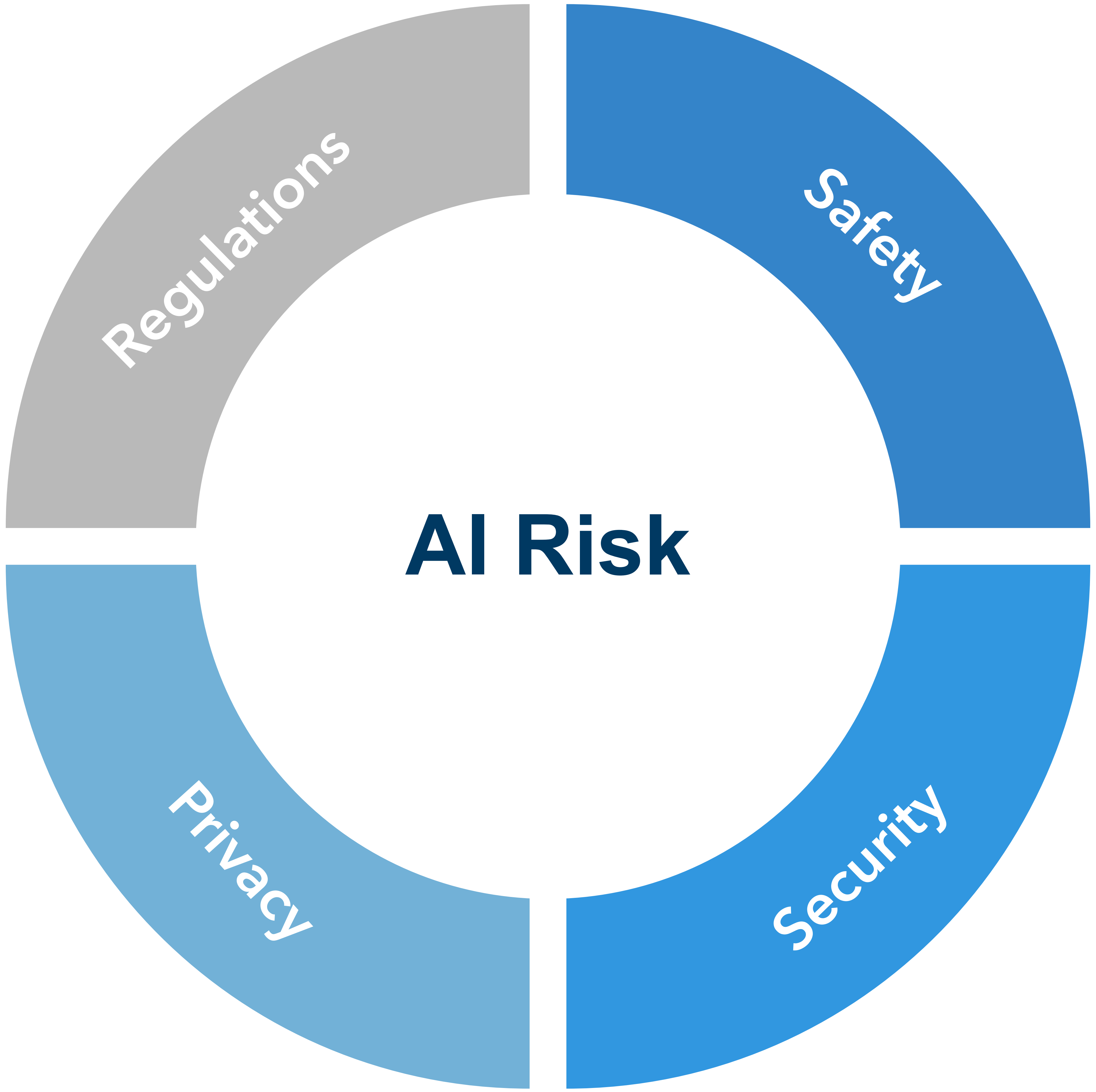
KUDELSKI SECURITY

Fundamental and Applied Research

Risk

The Real World Is Messy

AI Risk and Challenges



What Makes AI Risky?

- Poorly defined problem / Goals
- Lack of explicit programming logic
- Data
- Lack of visibility and explainability in some approaches
- Uncertainty
- Lack of appropriate benchmarks
- Concept Drift and Data Drift
- Legacy tools and process that don't align

What Bad AI Really Looks Like

TOM SIMONITE BUSINESS 07.10.2020 07:00 AM

Meet the Secret Algorithm That's Keeping Students Out of College

The International Baccalaureate program canceled its high-stakes exam because of Covid-19. The formula it used to "predict" scores puzzles students and teachers.

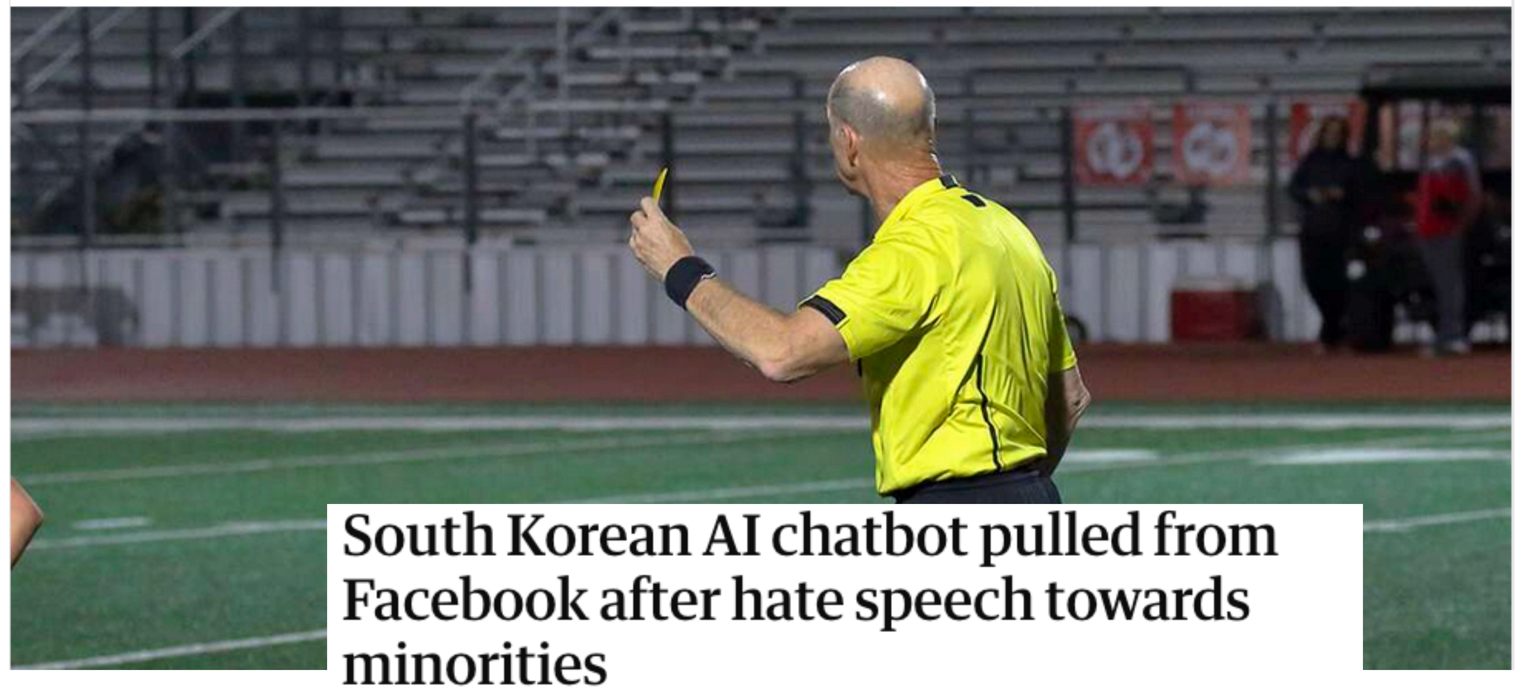
What Happens When Computer Programs Automatically Cut Benefits That Disabled People Rely on to Survive

October 21, 2020 / Lydia X. Z. Brown

University of Miami Reportedly Used Facial Recognition to Discipline Student Protesters

HEADLINE OCT 16, 2020

AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball



Why some onions were too sexy for Facebook

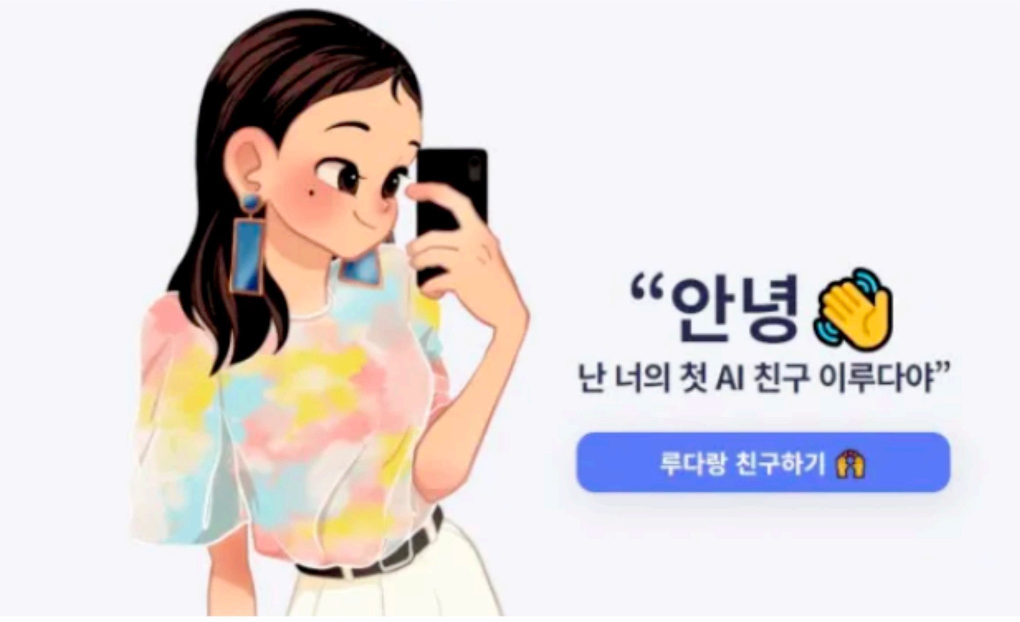
8 October

First death in a self-driving car happens in a Tesla

JUNE 30, 2016 / 6:29 PM / AP

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

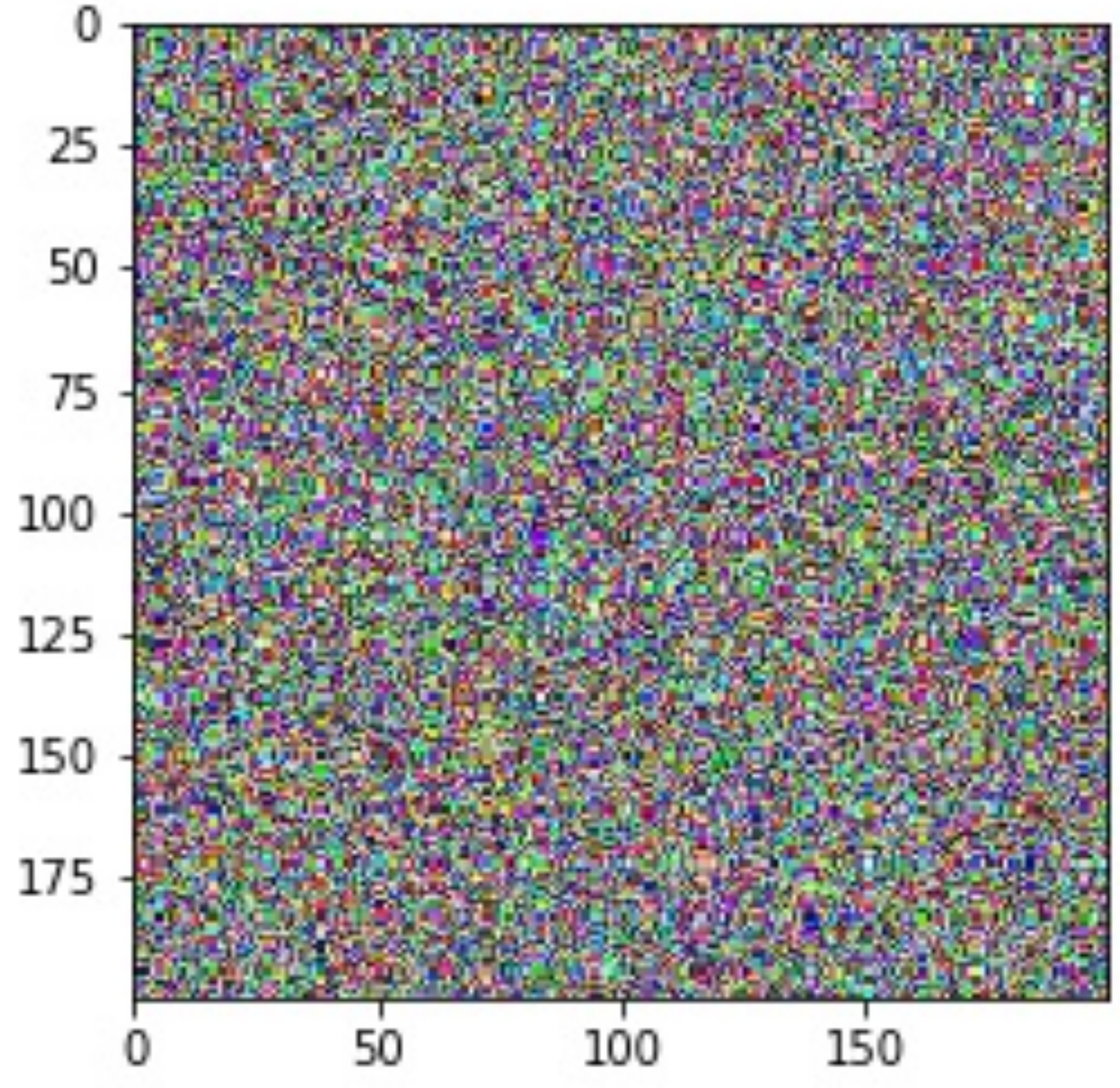
Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media

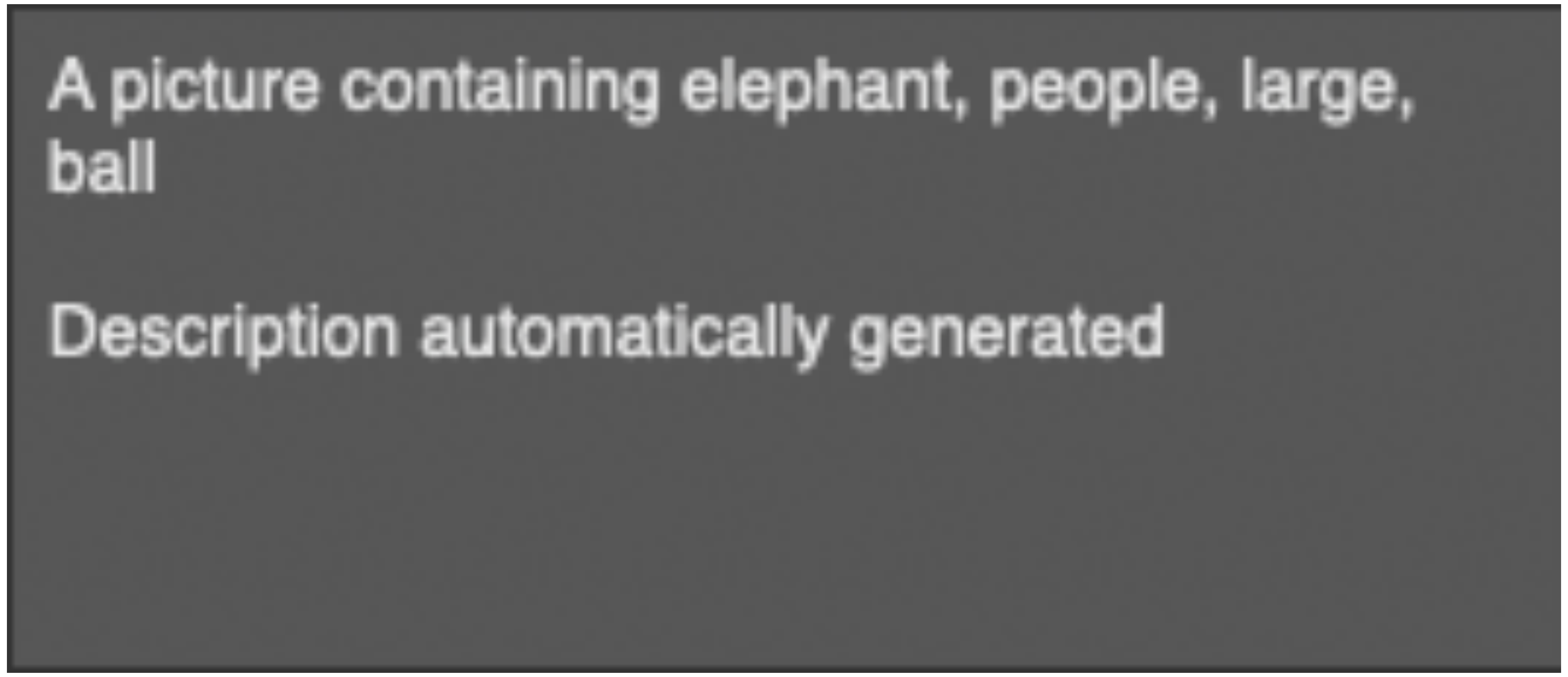


Man wrongfully arrested due to facial recognition software talks about 'humiliating' experience

KUDELSKI SECURITY

Uncertainty



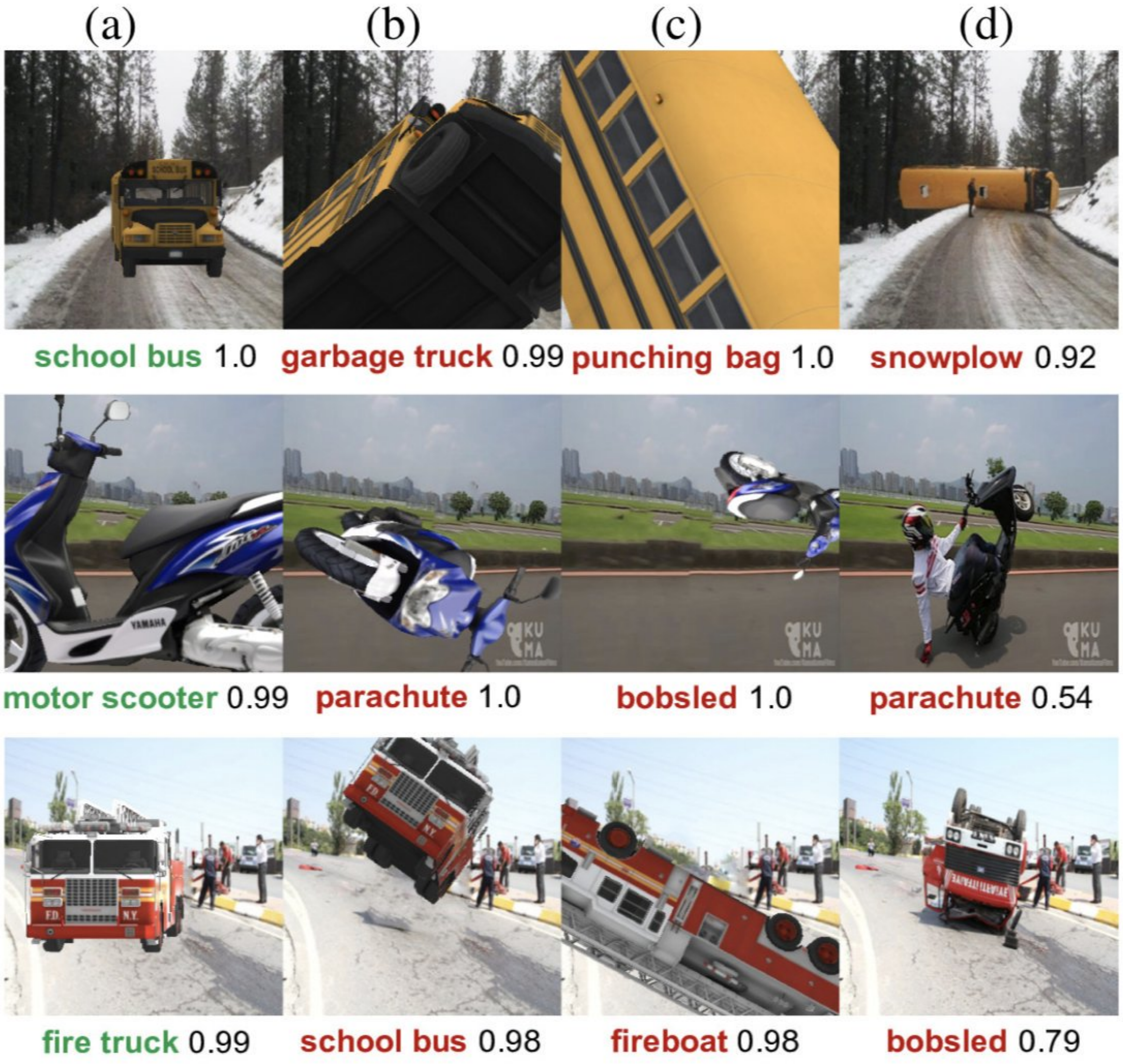


https://research.kudelskisecurity.com/2020/07/23/fooling-neural-networks-with-noise/

Fragility

Cutting Edge

Attacks!!!



Alcorn, et al., 2019

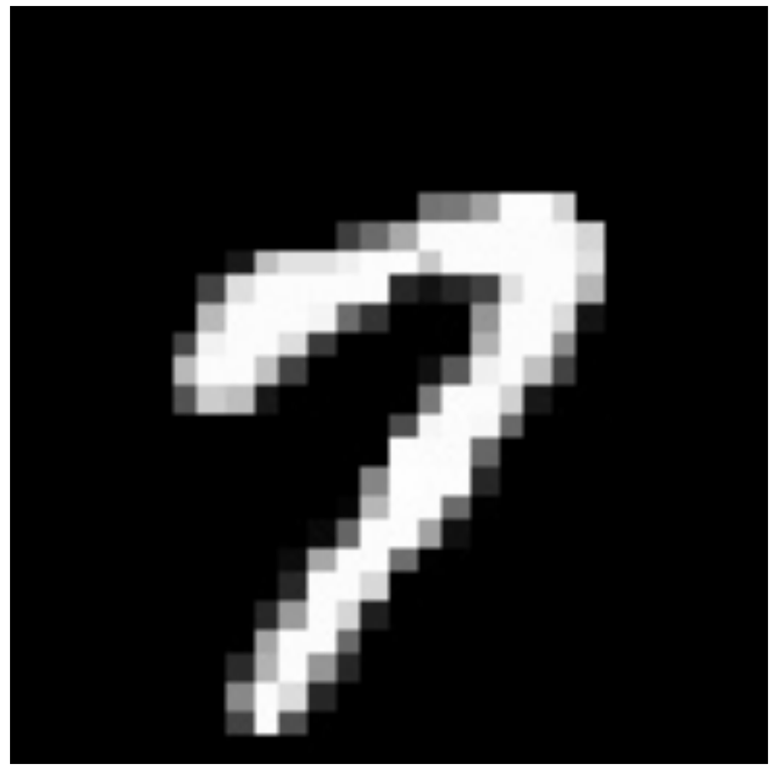
KUDELSKI SECURITY

Health and Safety

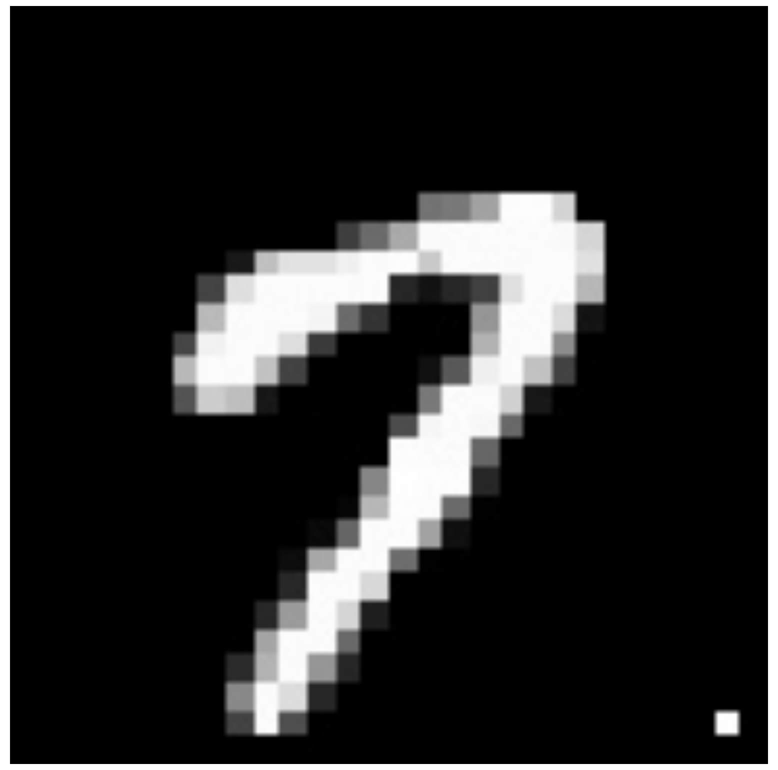
Network	Classification	Score
vgg16	cannon	0.3462
resnet18	tractor	0.2012
alexnet	tank	0.4665
densenet	thresher	0.1893
Inception	motor_scooter	0.5318

https://research.kudelskisecurity.com/2020/07/23/fooling-neural-networks-with-rotation/

Model Backdoors



Original Image



Single-Pixel Backdoor

Pattern Backdoor

Gu, et al., 2019

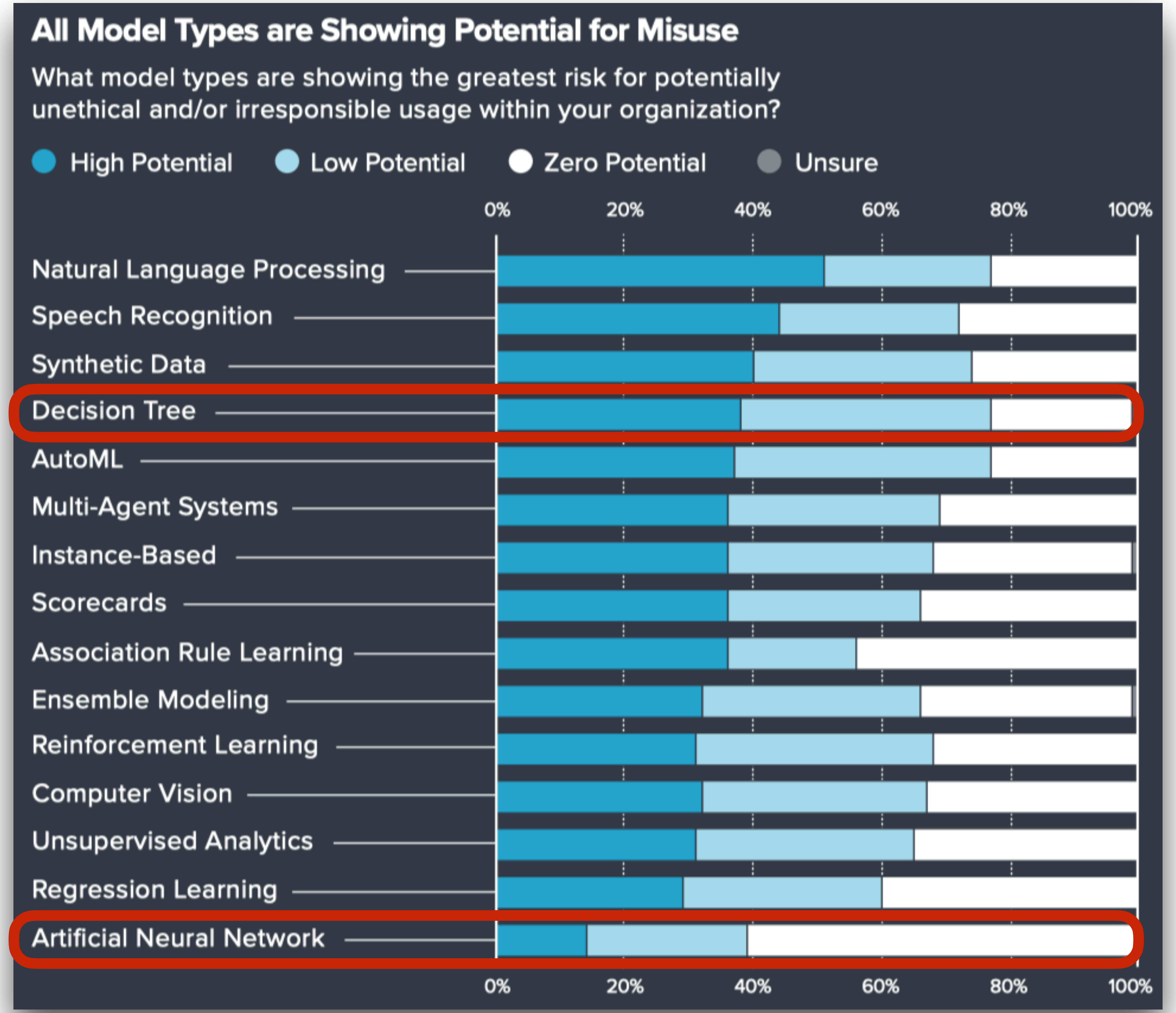
Supply Chain Issues

- You could inherit all of the issues of the previous model
- Attackers can exploit lack of visibility
- Model sharing and reuse is encouraged
  - How do you know when there is a problem
  - How do updates happen
- Use trusted sources

https://research.kudelskisecurity.com/2020/10/29/building-a-simple-neural-network-backdoor/

Risk Perspectives and Confusion

- 65% of respondents's companies can't explain how decisions or predictions are made
- 73% have struggled to get executive support for prioritizing AI ethics and Responsible AI practices



The State of Responsible AI: 2021

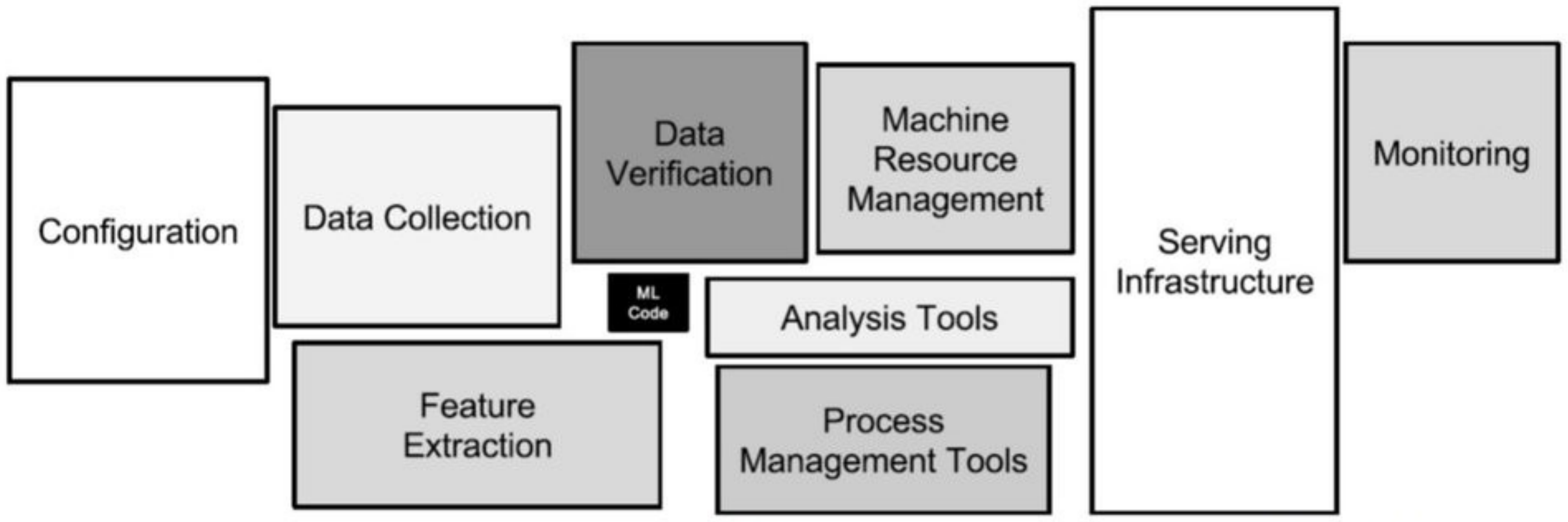
https://www.fico.com/en/latest-thinking/market-research/state-responsible-ai-2021

Security

- We live in an increasingly customized and specific world
- There isn’t a typical attacker process or kill chain to interrupt
- Security is often misaligned and out of the loop
- Security lacks expertise in the AI/ML area
  - We need to ensure that threats are identified during the development lifecycle and risk mitigated to acceptable levels
  - We apply proper testing to systems

Attack Surface

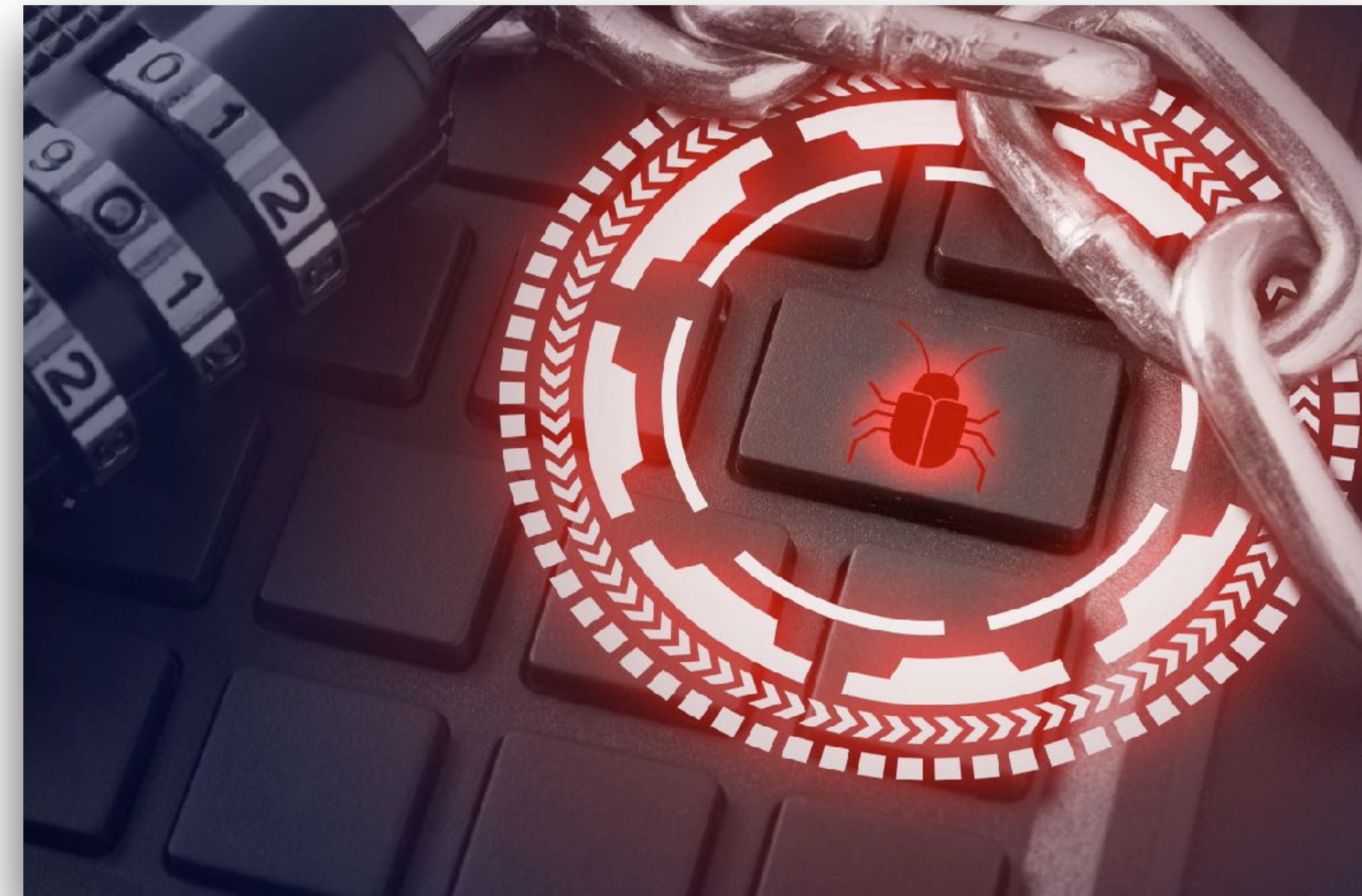
- Model
- Processes
- Hosting infrastructure
- Training data



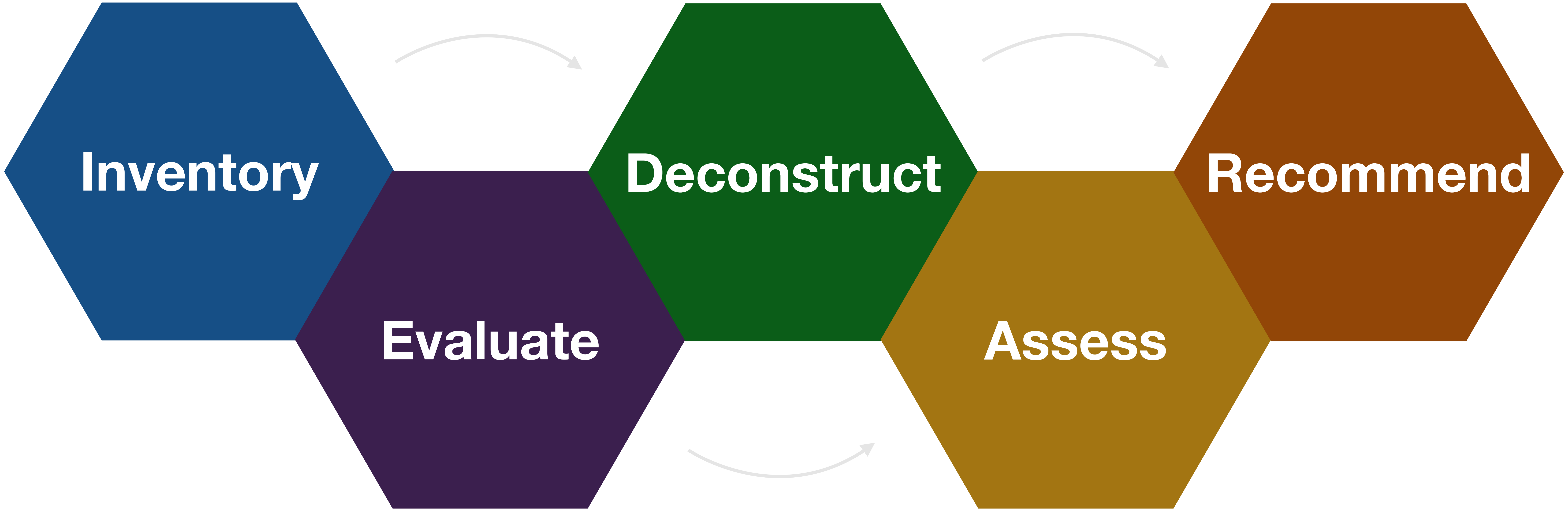
Sculley, et al., 2015

Attacks

- Traditional Platforms
  - Applications, Cloud, IoT, Sensors, etc.
- Some ML Specific Attacks
  - Model Evasion
  - Model Poisoning
  - Membership Inference
  - Model Theft / Functional Extraction



Assessing Risks



Quick Risk Eval

- What does the system do?
- Does it support a critical business process?
- Was it trained on sensitive data?
- How exposed is it going to be?
- What would happen if the system failed?
- Could the system be misused?
- Does it fall under any regulatory compliance?

Security Testing

- Spin up
- Encompass traditional and model specific approaches
- Define a goal and think like an attacker
- ML attacks are situational
  - Manipulate features and modify inputs based on ML approach
  - Observe outputs
  - Repeat
- A little coding, a little skill, and a little luck

Evaluate Attacks and Defenses

- More attacks and proposed defenses are coming
- Build a way to evaluate both attacks and defenses
  - Separate security testing pipeline
  - Integrate tooling
  - Evaluate effectiveness and impact

Be Careful

- Choices to use or not use a control should be purposeful
  - Robustness training
- Many recommendations may affect performance and accuracy
  - Fully homomorphic encryption
  - Defensive distillation
  - Feature squeezing
- Start with the basics and move on if necessary

Resources

- Failure Modes in Machine Learning
  - https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
  - https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml
- Adversarial Threat Matrix
  - https://github.com/mitre/advmlthreatmatrix
- Counterfit
  - https://github.com/Azure/counterfit/
- Adversarial Robustness Toolbox
  - https://developer.ibm.com/technologies/analytics/projects/adversarial-robustness-toolbox/

Resources

- ISO Standard (Future)
- NIST (Future)
- ENISA - AI Cybersecurity Challenges
  - https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
- ENISA - Securing Machine Learning Algorithms
  - https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

Contact

Nathan Hamiel

nathan.hamiel @ kudelskisecurity.com

Twitter: @nathanhamiel

LinkedIn

https://research.kudelskisecurity.com