RSA Conference 2019

San Francisco | March 4 – 8 | Moscone Center

BETTER.

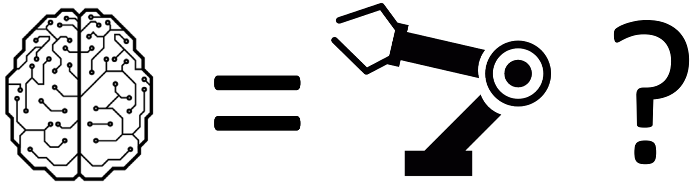
SESSION ID: MLAI-T06

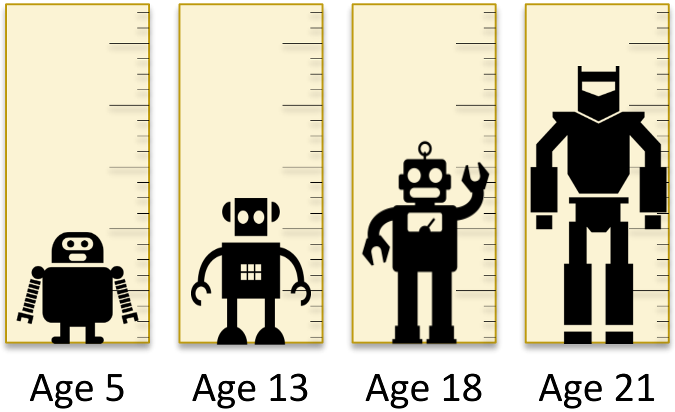
Lessons Learned in Automating Decision Making: Pitfalls and Opportunities

Sounil Yu

@sounilyu

Questions to ponder





- How is AI/ML distinct from automation?
- How mature are our AI/ML and automated decision making capabilities? How mature do they need to be for security?
- What can we learn from failed cases of automated decision making in security?
- What guardrails should be considered for automated decision making until sufficient maturity is achieved?

Framework #1: Modified OODA Loop

Sensing
- Sensors
- Raw Telemetry
- Big Data

Sense Making
- Analytics
- Artificial Intelligence
- Machine Learning

Decision Making
- Defined Processes
- Courses of Action
- Orchestration

Acting
- Execution
- Robotic Automation
- Response Scripts

Artificial intelligence capabilities are distinct from robotic automation capabilities

Framework #2: DARPA's Perspective on AI

https://www.darpa.mil/about-us/darpa-perspective-on-ai

Notional intelligence scale

Perceiving RICH, COMPLEX AND SUBTLE INFORMATION ABOUT THE OUTSIDE WORLD TO UNDERSTAND WHAT'S GOING ON

Learning WITHIN AN ENVIRONMENT AND ADAPTING TO ITS CONDITIONS AND SITUATIONS BASED ON WHAT IS PERCEIVED

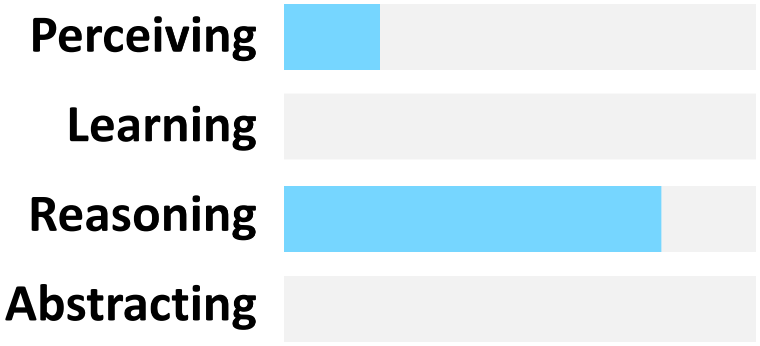
Reasoning TO PLAN / DECIDE BASED ON A SET OF PRESCRIBED OR IMPLIED RULES AND UNDERSTANDING WHY

Abstracting BY TAKING KNOWLEDGE OF ONE DOMAIN AND APPLYING TO OTHER DOMAINS TO CREATE NEW MEANINGS

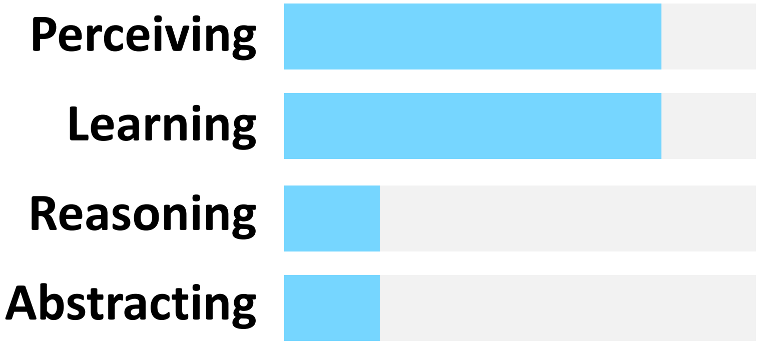
Human Level

DARPA's Perspective on AI

https://www.darpa.mil/about-us/darpa-perspective-on-ai

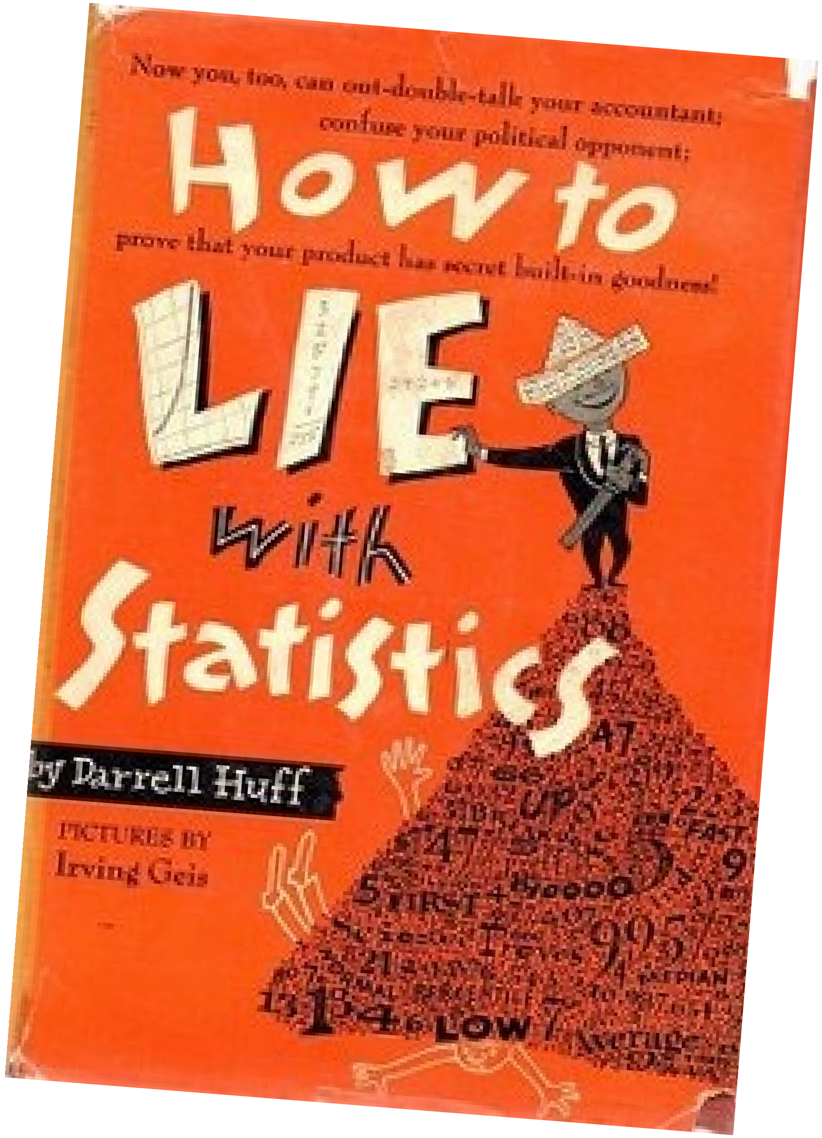


- **First Wave of AI – Handcrafted Knowledge**
  - Enables reasoning over **narrowly defined problems**
  - No learning capability and **poor handling of uncertainty**
  - Examples: Turbotax, Chess, Logistics, DARPA Cyber Grand Challenge
  - First generation SIEMs



- **Second Wave of AI – Statistical Learning**
  - Nuanced classification and prediction capabilities
  - **No contextual capability and minimal reasoning capabilities**
  - Examples: Voice recognition, Face recognition, DARPA Grand Challenge – Self Driving Cars
  - Statistically impressive, individually unreliable
  - Current generation SIEMs

s/Machine Learning|Deep Neural Networks/Statistics/g

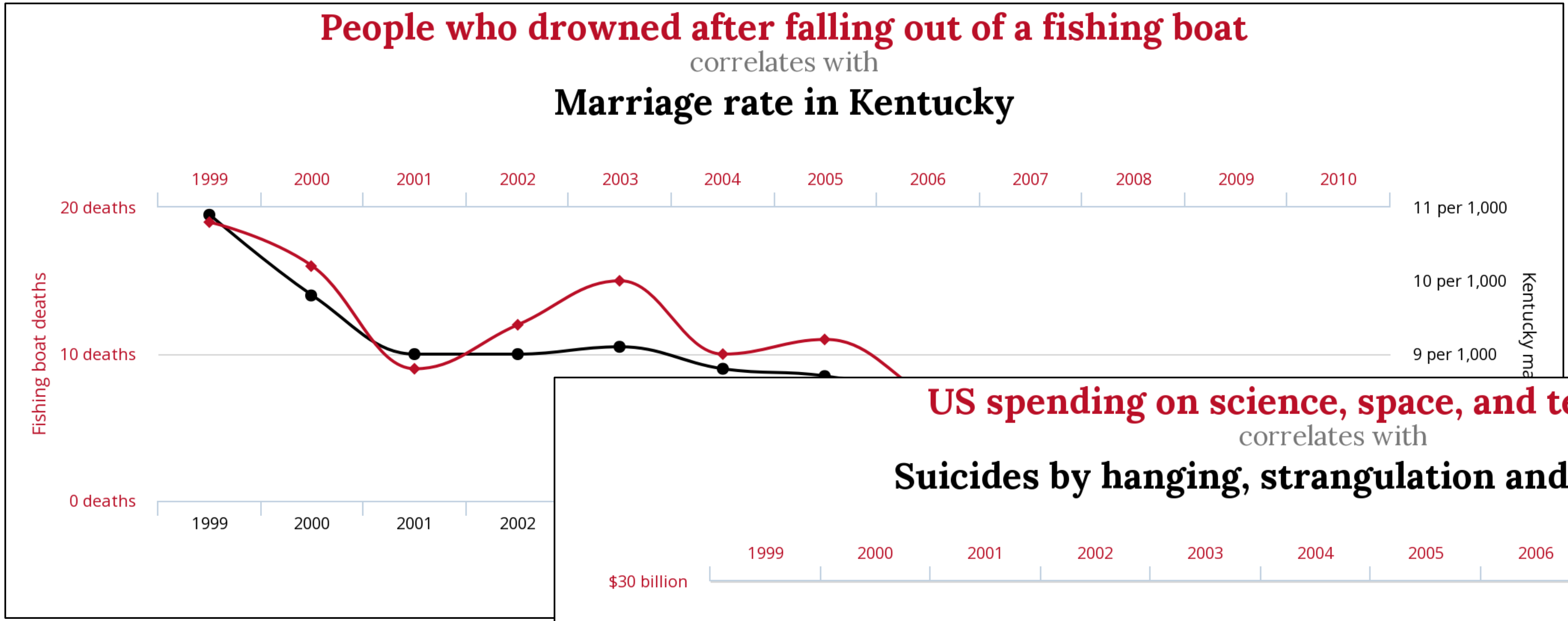


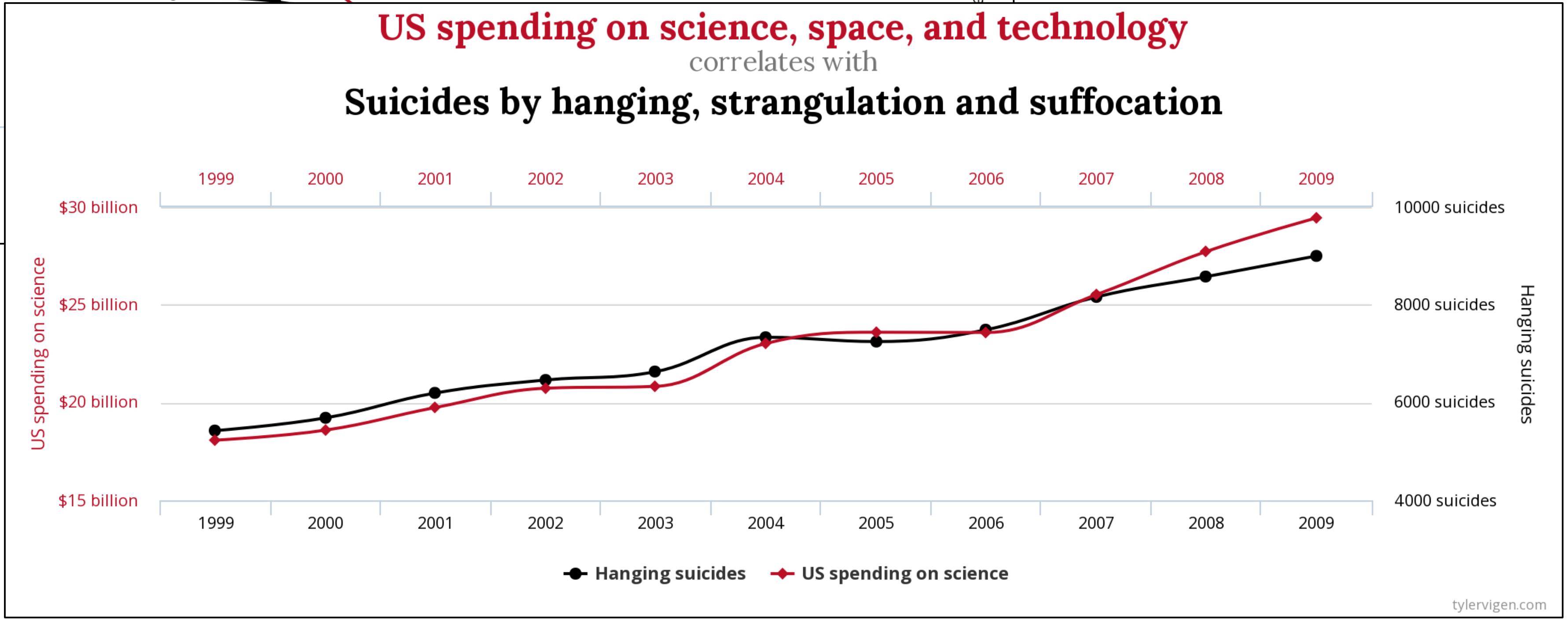
83.7% of all statistics are entirely made up

Sounil Yu

Correlation, not Causation

http://www.tylervigen.com/spurious-correlations



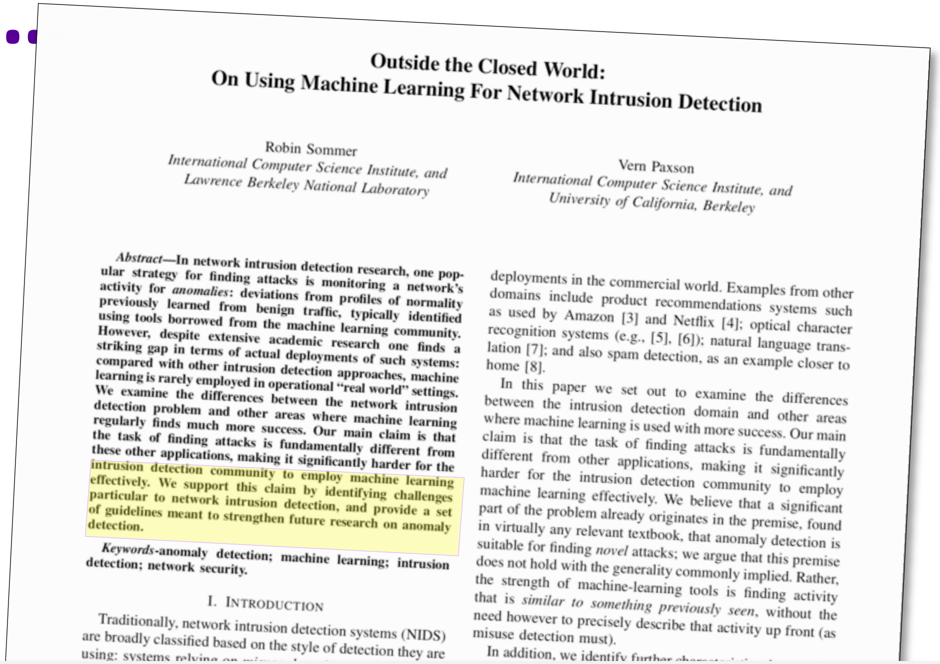


Why statistical-based machine learning and neural networks DO NOT work for security..

Outside the Closed World: On Using Machine Learning for Network Intrusion Detection

Robin Sommer, Vern Paxson, 2010

https://www.icsi.berkeley.edu/icsi/node/4511



Outside the Closed World:
On Using Machine Learning For Network Intrusion Detection

Robin Sommer
International Computer Science Institute, and
Lawrence Berkeley National Laboratory

Vern Paxson
International Computer Science Institute, and
University of California, Berkeley

Abstract—In network intrusion detection research, one popular strategy for finding attacks is monitoring a network's activity for *anomalies*: deviations from profiles of normality previously learned from benign traffic, typically identified using tools borrowed from the machine learning community. However, despite extensive academic research one finds a striking gap in terms of actual deployments of such systems: compared with other intrusion detection approaches, machine learning is rarely employed in operational "real world" settings. We examine the differences between the network intrusion detection problem and other areas where machine learning regularly finds much more success. Our main claim is that the task of finding attacks is fundamentally different from these other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We support this claim by identifying challenges particular to network intrusion detection, and provide a set of guidelines meant to strengthen future research on anomaly detection.

Keywords-anomaly detection; machine learning; intrusion detection; network security.

I. Introduction

Traditionally, network intrusion detection systems (NIDS) are broadly classified based on the style of detection they are

deployments in the commercial world. Examples from other domains include product recommendations systems such as used by Amazon [3] and Netflix [4]; optical character recognition systems (e.g., [5], [6]); natural language translation [7]; and also spam detection, as an example closer to home [8].

In this paper we set out to examine the differences between the intrusion detection domain and other areas where machine learning is used with more success. Our main claim is that the task of finding attacks is fundamentally different from other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We believe that a significant part of the problem already originates in the premise, found in virtually any relevant textbook, that anomaly detection is suitable for finding *novel* attacks; we argue that this premise does not hold with the generality commonly implied. Rather, the strength of machine-learning tools is finding activity that is *similar to something previously seen*, without the need however to precisely describe that activity up front (as misuse detection must).

In addition, we identify further

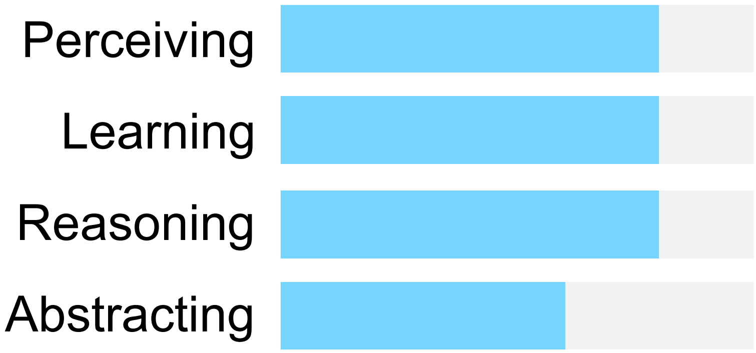
- Bounded vs unbounded environments
- Inviolable rules vs shifting rules
- Human adversaries deliberately try to shift the rules (i.e., novel attacks)

"…[ML is generally not] suitable for finding ***novel*** attacks … Rather, the strength of machine-learning tools is finding activity that is ***similar to something previously seen***…"

"Our main claim is that the task of **finding attacks is fundamentally different** from these other applications, **making it significantly harder** for the intrusion detection community **to employ machine learning effectively**."

DARPA's Perspective on AI continued

https://www.darpa.mil/about-us/darpa-perspective-on-ai



- **Third Wave of AI – Contextual Adaptation**
  - Systems construct **explanatory models** for classes of real world phenomena
  - Models **explain decisions** (cause and effect)
  - **Understand why and why not**
    ➔ leads to an understanding of when the system will succeed or fail
    ➔ leads to when to trust and why mistakes are made

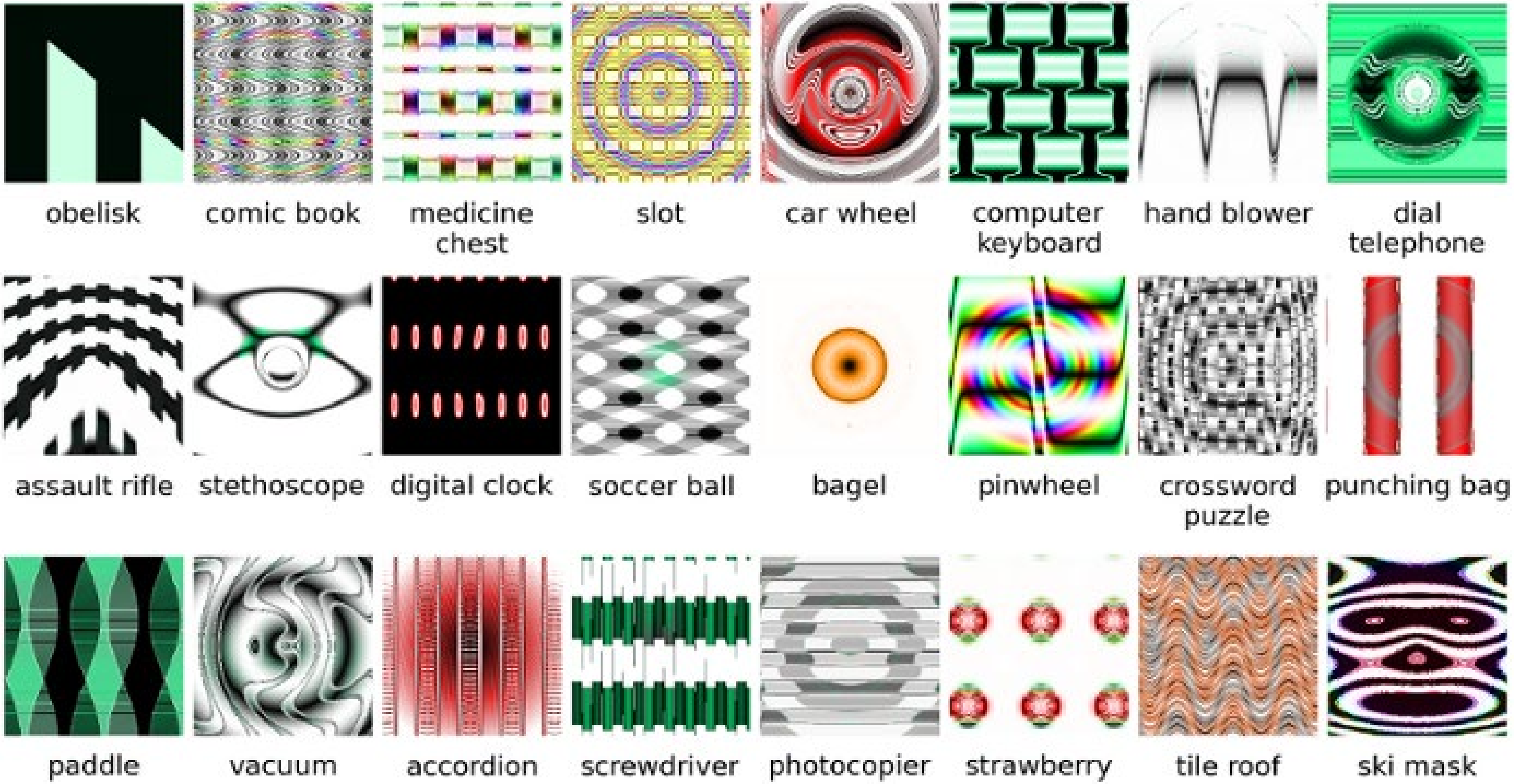
Source: Robust Physical-World Attacks on Machine Learning Models By Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song

Classified 100% of the time as…

How do I know this really is a stop sign?

Explainable model:
- Red
- Octagonal
- At intersections

Why Do We Need Better Decision Making?
… Because of Deliberate Attempts to Fool Sensing and Sense Making…



Source: Nguyen A, Yosinski J, Clune J.
Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.
In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015

Framework #3: Classical Education Trivium

Rhetoric

Convincing and persuading
Bearing fruit in wisdom
Applying and integrating subjects

Dialectic/Logic

Investigating the truth of opinions
Gaining in understanding
Explaining “why” and “how”

Grammar

Structures and rules
Soaking in knowledge
Memorizing a broad base of facts

Remember:
According to the DARPA Framework, this capability emerges in the Third Wave

The “Age” of Machine Learning

Rhetoric
Convincing and persuading
Bearing fruit in wisdom
Applying and integrating subjects

Dialectic/Logic
Investigating the truth of opinions
Gaining in understanding
Explaining “why” and “how”

Grammar
Structure and rules
Soaking in knowledge
Memorizing a broad base of facts

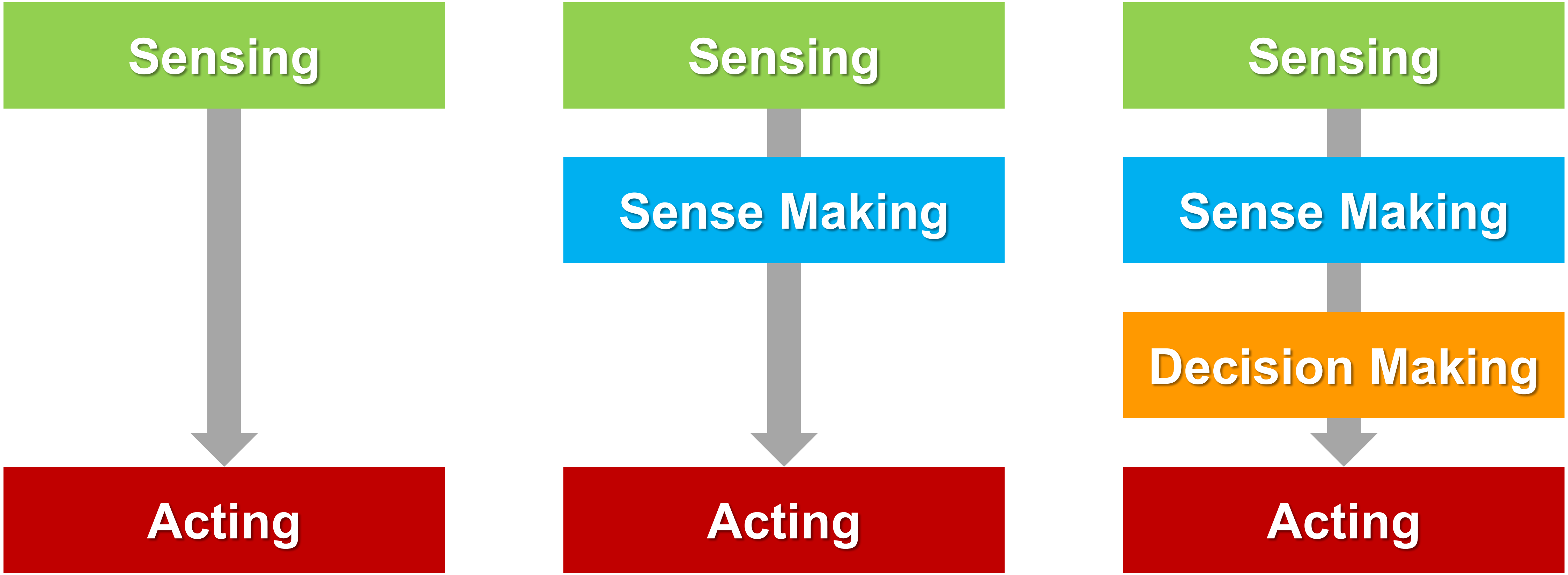
<table>
<tr><th>Grades K-6
Elementary School</th><th>Grades 7-9
Junior High School</th><th>Grades 10-12
High School</th></tr>
<tr><td>Rhetoric</td><td>Rhetoric</td><td rowspan="2">Rhetoric</td></tr>
<tr><td>Dialectic</td><td rowspan="2">Dialectic</td></tr>
<tr><td rowspan="2">Grammar</td><td>Dialectic</td></tr>
<tr><td>Grammar</td><td>Grammar</td></tr>
</table>

Machine learning has not left this stage yet

Children vs Fully Grown Adults

	Children	Fully Grown Adults
Sensing	**Questionable Inputs** • Peers, Memes, Social Media	**Better Filters and Prioritization** • Knowledge and Wisdom • Discernment • Model Based Perception
Sense Making	**"Algorithms"** • Irrational • Illogical	**"Algorithms"** • Rational and logical • Explainable
Decision Making	**Amygdala** • Fear • Emotion • Impulse • Aggression • Instinct	**Prefrontal Cortex** • Coordinates and adjusts complex behavior • Controls impulse • Prioritizes competing and simultaneous inputs
Acting	**Fully Developed at Birth**	**Fully Developed at Age 25**

Lessons Learned when Skipping Steps



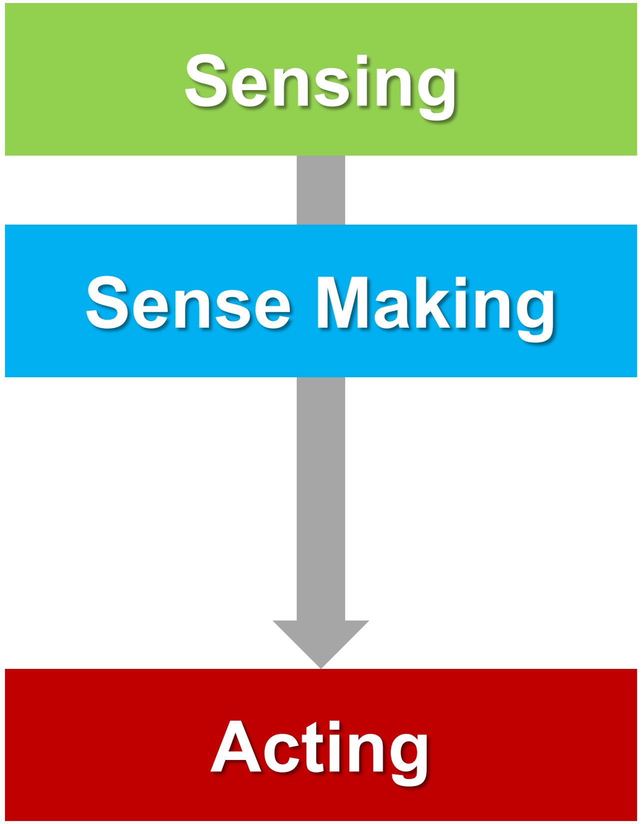
Lessons Learned: Reflexive Stimulus – Response

Sensing

Acting

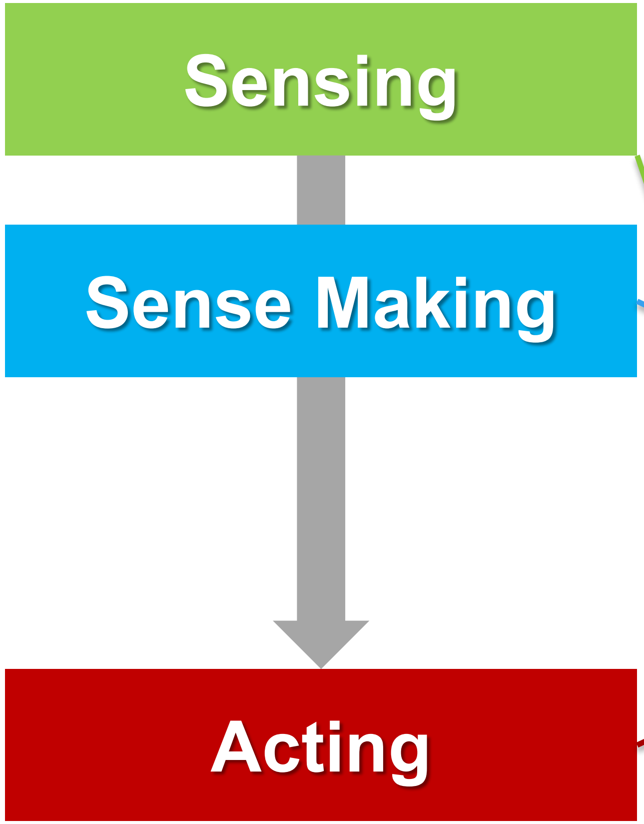
- Threat Intel Sharing
  - Blocking google.com
  - Null routing 0.0.0.0/0
- Automated patching
  - NotPetya
  - Windows 10 1809 – The "I hope you made a backup" Update
- Guardrails:
  - Ensure sensor **sources are trustworthy and reliable**
  - Apply actions that are **narrowly scoped**
  - Have a **kill switch ready** if it goes beyond the scope
  - Make the action immediately **reversible**

Lessons Learned: Conditional response based on analysis, enrichment, and regression testing



- Threat Intel Sharing
  - Would this new regex pattern create **false positives** by matching on anything else over the past 30 days?
  - Is there a **unanimous verdict** based on enrichment from multiple other sources? (i.e., what does VirusTotal say?)
- Automated patching after regression testing
  - Do all systems in the testbed continue to **operate as expected** after the patch?
  - Did any applications **stop working** after the patch?

Lessons Learned: Conditional response based on analysis, enrichment, and regression testing

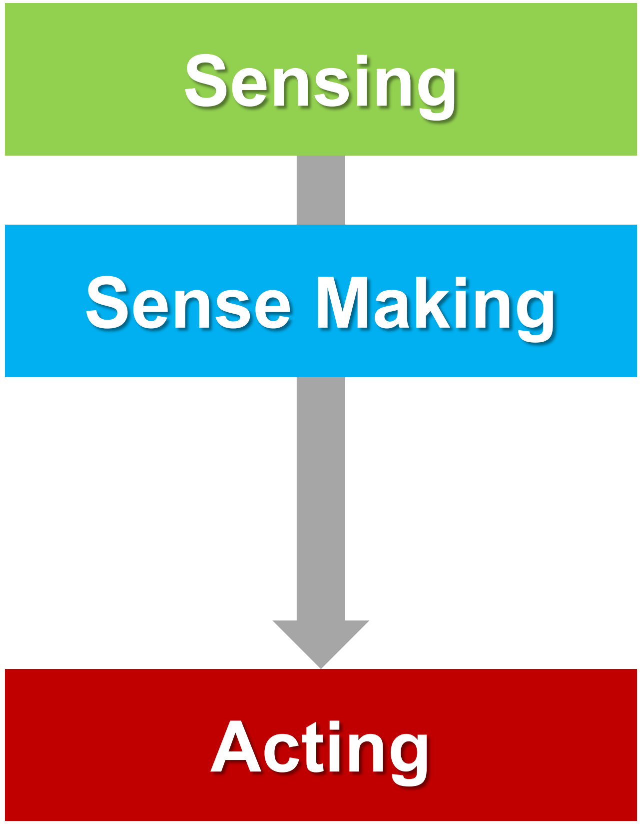


The CHASE program seeks to develop automated tools to detect and characterize novel attack vectors, collect the right contextual data, and disseminate protective measures both within and across enterprises

> "The automation process has to leave a trail of logic behind decisions so humans can follow it up,"
>
> Sam Hamilton, Chief Scientist
> BAE Systems, Cyber Tech Group

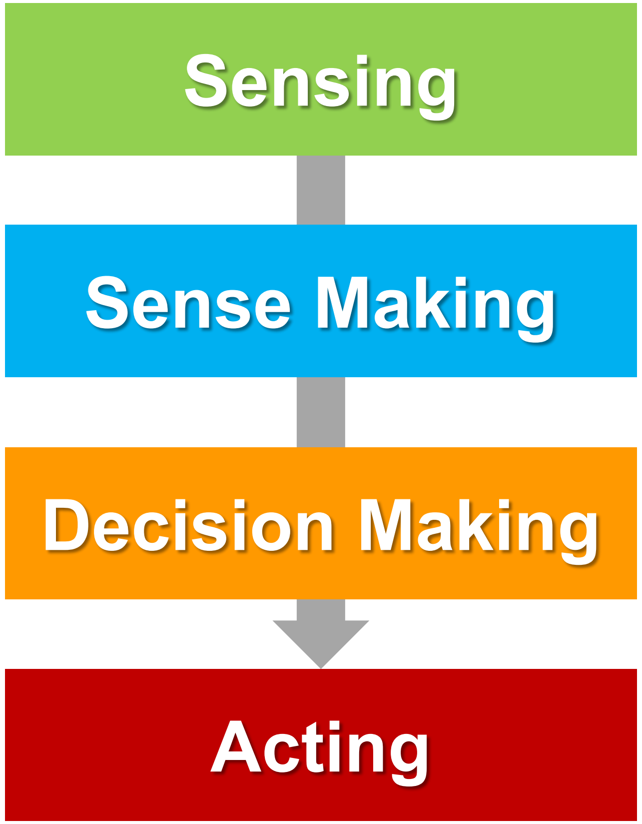
Sources:
https://www.darpa.mil/program/cyber-hunting-at-scale
https://defensesystems.com/articles/2018/08/17/bae-cyber-ai-tool.aspx

Lessons Learned: Conditional response based on analysis, enrichment, and regression testing



- Automating funds transfers requests from emails to an electronic transfer system
  - Assumed that emails were only coming from legitimate sources
  - Assumed that copying action was purely mechanical and didn’t involve any further analysis or thought beyond mapping email content to the proper fields
- Guardrails
  - Use regression testing to ensure the outcomes are **fully deterministic**
  - **Validate assumptions** that no other decision making is actually needed or happening
  - Ensure entire process is **well documented** and **understood** by the operators

Lessons Learned: Conditional response based on business considerations



- Bank of Valletta shuts down all of its operations after hackers broke into its systems and shifted funds overseas
- DoD responds to the Code Red worm by disconnecting NIPRNet from the Internet, resulting in the Army Corps of Engineers not being able to control the locks on the Mississippi River
- Guardrails:
  - **Pre-establish thresholds** where the costs of inaction are worse than the negative repercussions of action
  - Pre-determine **authorities** for actions and **accountabilities** for outcomes

Comparing Frameworks… How Mature is AI/ML Today?

OODA Loop	DARPA	Classical Education
Sensing	Perceiving	Grammar Stage
Sense Making	Learning	Dialectic Stage
Decision Making	Reasoning	Rhetoric Stage
Acting	Abstracting	

Summary of Guardrails: When might it be okay to enable automated decision-making?

Sensor Diversity
Ensure sensor sources are trustworthy and reliable based on multiple sources of truth

Algorithmic Integrity
Ensure entire process and all assumptions are well documented and understood by the operators

Bounded Conditions
Ensure decisions are highly deterministic and narrowly scoped using regression testing

Brakes and Reverse Gear
Have a kill switch ready if it goes beyond the scope and make the action immediately reversible

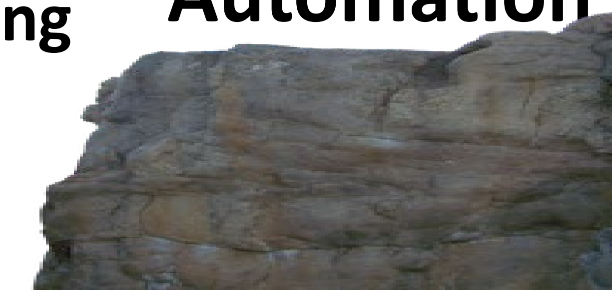
Established Thresholds
Know when the costs of inaction are worse than the negative repercussions of action

Authorities and Accountabilities
Pre-establish authorities for taking action and accountabilities for outcomes

AI/ML

Create a conscious mental chasm that you deliberately choose to cross when enabling automated decision-making

Robotic Automation



“Apply” Slide

- Within the next month
  - Start inventorying capabilities broken out by Sensing, Sense-Making, Decision-Making, Acting
  - Determine where automated decision-making may be happening within those capabilities
- Within the next 90 days
  - Review potential guardrails for automated decision-making
- Within the next year
  - Establish governance processes to ensure that systems with automated decision-making stay within those guardrails

For further reading

- DARPA’s Perspective on Artificial Intelligence
https://www.darpa.mil/about-us/darpa-perspective-on-ai
- AI is now so complex its creators can’t trust why it makes decisions
https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/
- Automation Should Be Like Iron Man, Not Ultron
https://queue.acm.org/detail.cfm?id=2841313
- How can we be sure AI will behave? Perhaps by watching it argue with itself.
https://www.technologyreview.com/s/611069/how-can-we-be-sure-ai-will-behave-perhaps-by-watching-it-argue-with-itself/
- AI is more powerful than ever. How do we hold it accountable?
https://www.washingtonpost.com/outlook/ai-is-more-powerful-than-ever-how-do-we-hold-it-accountable/2018/03/20/e867b98a-2705-11e8-bc72-077aa4dab9ef_story.html
- Artificial Intelligence Has A Problem With Bias, Here's How To Tackle It
https://www.forbes.com/sites/bernardmarr/2019/01/29/3-steps-to-tackle-the-problem-of-bias-in-artificial-intelligence/
- The case against understanding why AI makes decisions
https://qz.com/1192977/the-case-against-understanding-why-ai-makes-decisions/
- When computers decide: European Recommendations on Machine-Learned Automated Decision Making
https://dl.acm.org/citation.cfm?id=3185595
- Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR
https://arxiv.org/abs/1711.00399
- Explanation in Artificial Intelligence: Insights from the Social Sciences
https://arxiv.org/abs/1706.07269

RSAConference2019

Questions?

@sounilyu