RSAConference2019
San Francisco | March 4 – 8 | Moscone Center

SESSION ID: MLAI-T09

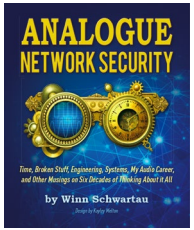
Ethical Bias in AI-Based Security Systems: The Big Data Disconnect

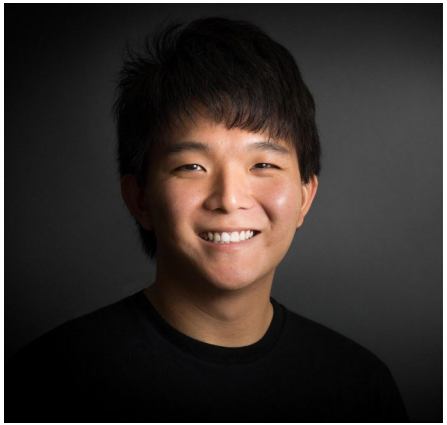
Winn Schwartau
Founder, Winn Schwartau, LLC

Clarence Chio
Co-founder, CTO, Unit21

About Winn & Clarence

- Security Guy since 1993
- 16 Books
- 1,000s of talks/articles
- Controversial (I Hope!)
- Creator/Author
  Analogue Network Security &
  Measureable Security
- Electronic Pearl Harbor (1991)

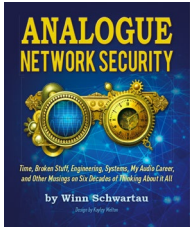




- Security guy since 2013
- 1 Book
- 50s of talks/articles
- “Machine Learning & Security” (O’Reilly, 2018)
- Founded company using AI to fight money laundering

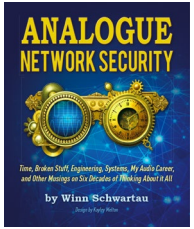
Agenda

1. Goals/Takeaways/Questions to Answer
2. Some truths about AI
3. Bias Everywhere
4. Trolleyological Conundra
5. Should we, can we trust AI in security?
6. Bias in statistical learning
7. Case studies
8. A trust model for AI
9. What can you do?



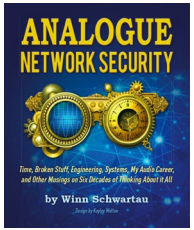
1. Goals/Takeaways/Questions to Answer

1. As AI focuses on security, what is the real value?
2. Do you know how AI really works?
3. Do you understand bias and 'honest' datasets?
4. Can you expect the same answers reliably?
5. What makes an AI bot racist?
6. What do you need to ask your AI vendors
7. What is AI good for?

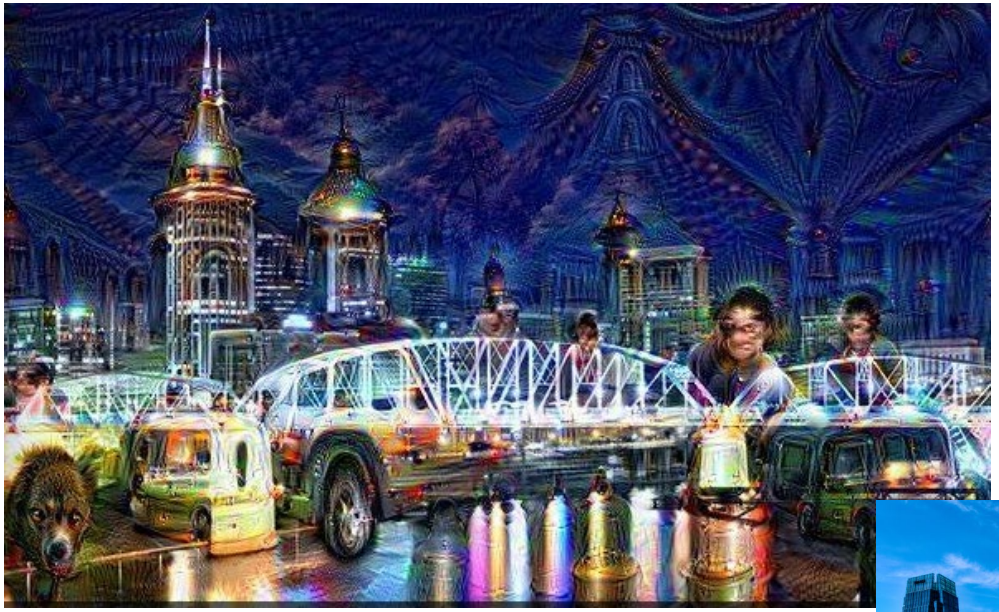


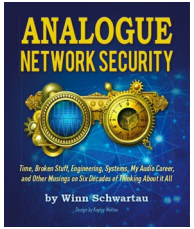
2. Some truths about AI

- AI is not absolute. It is NOT deterministic.
- AI is probabilistic. (Fuzzy, analogue, not binary)
- No one, not even Data Scientists or AI designers, know how an AI system arrives at its conclusions.
- There is no way to ask an AI, “How did you arrive at that answer?” (See XAI later)
- AI is entirely bias (data set) sensitive.
- Can your vendor promise more?

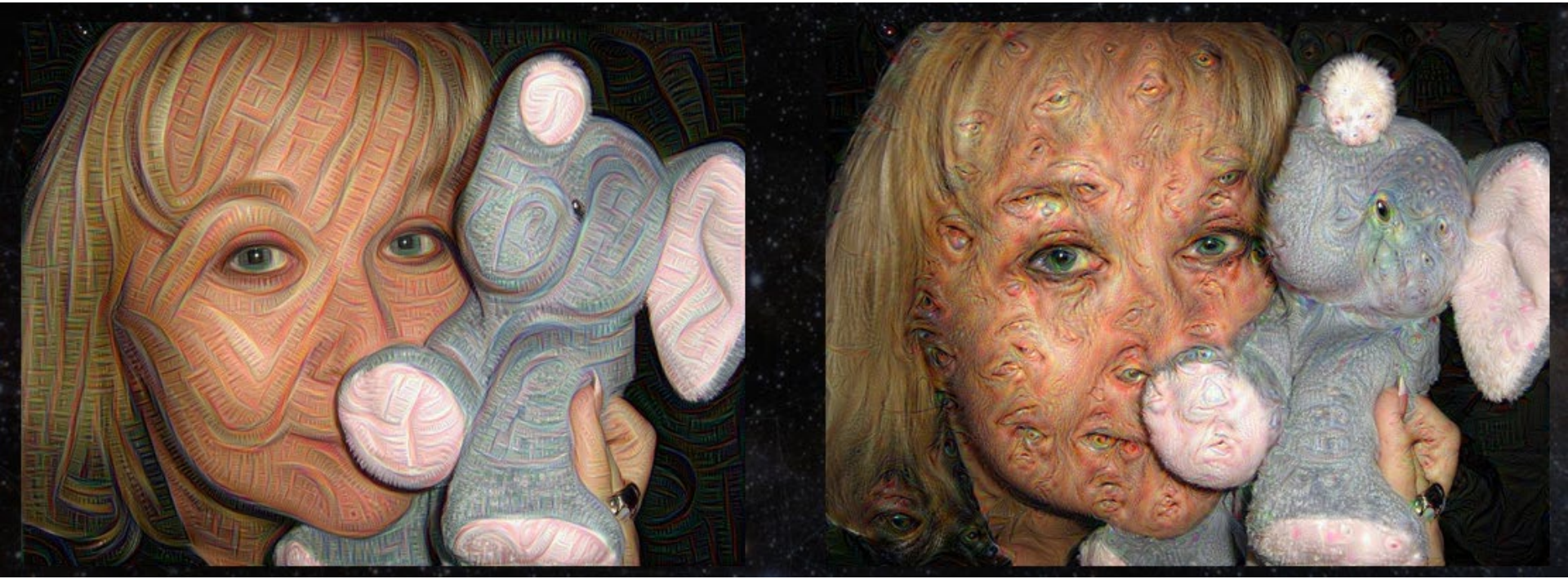


Nashville – Through AI eyes. True or Not-True?



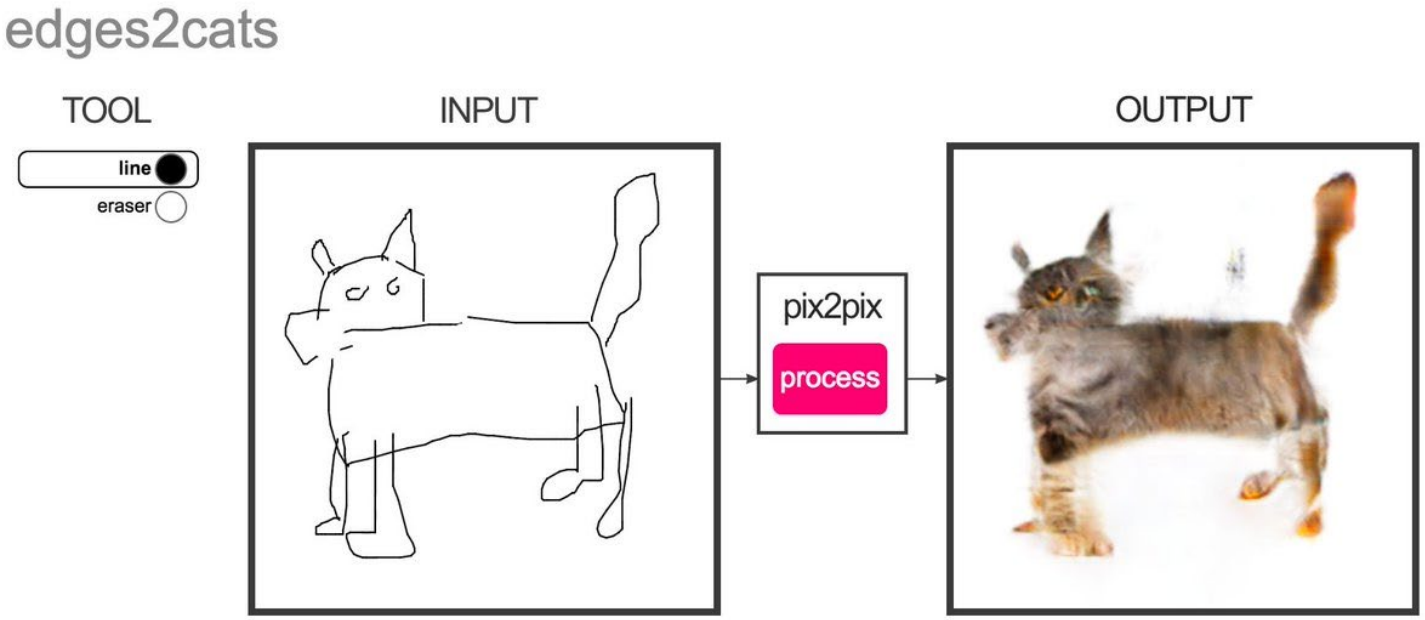


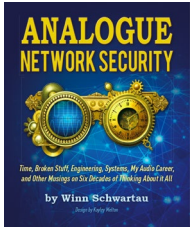
My wife – as seen by AI on LSD (Deep Thought)



3. Bias

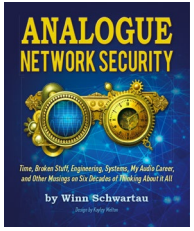
- Garbage in, garbage out





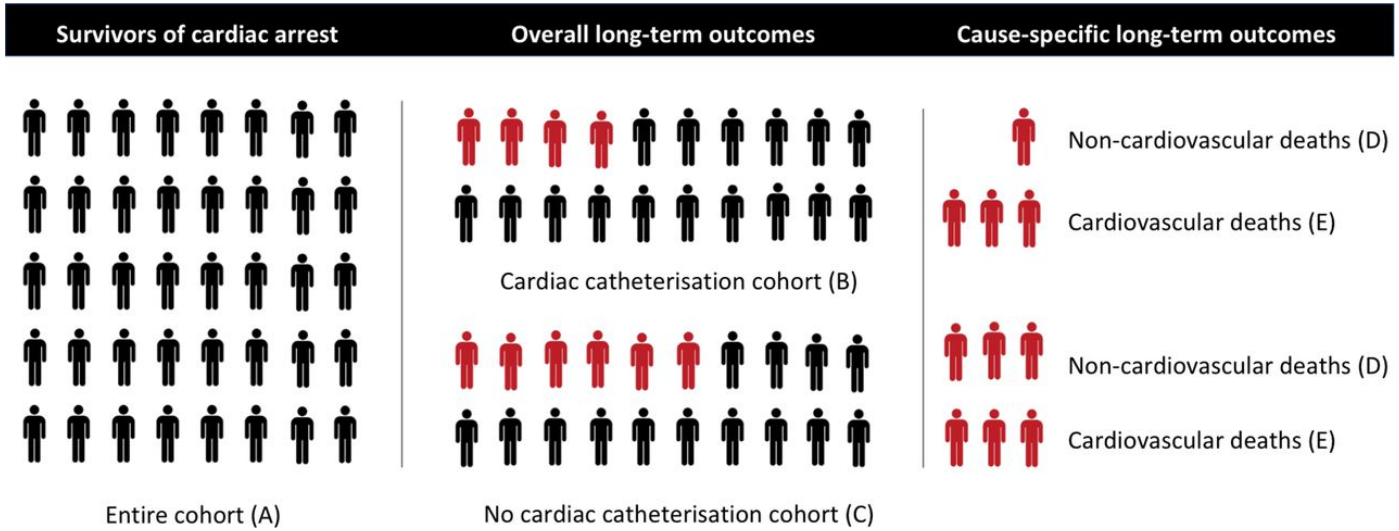
3. Bias - DATA IS EVERYTHING

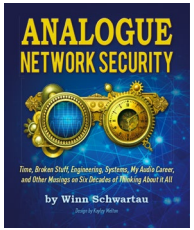
- Data quality dictates the trajectory of any autonomous algorithm's development
- Data collection and procurement processes are designed by humans, and humans are inherently biased
- Most attempts by implementors to detect and measure bias are again biased
- Undetected bias in data can have completely unexpected and catastrophic effects, especially when fed through statistical learning systems



3. Bias - Types

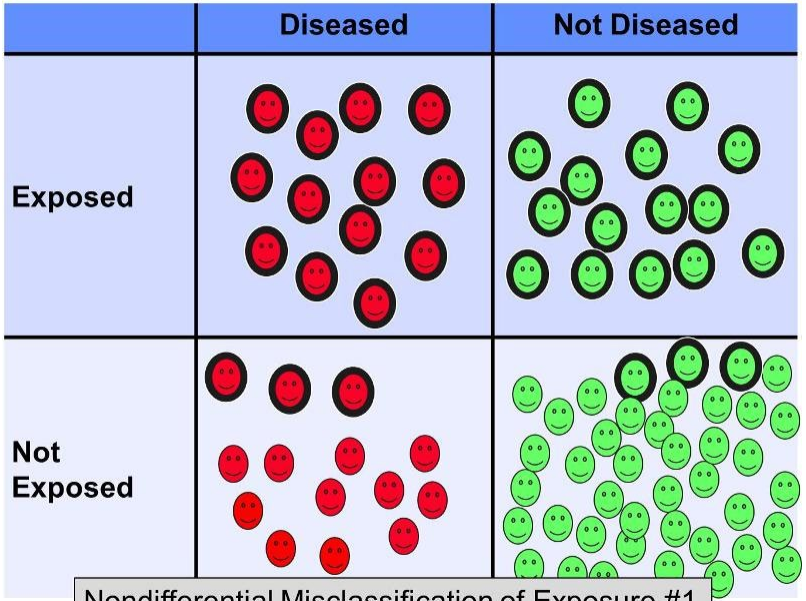
Selection bias



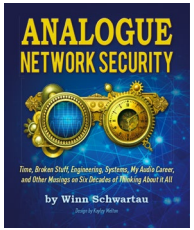


3. Bias - Types

Misclassification bias a.k.a. observational bias



Nondifferential Misclassification of Exposure #1



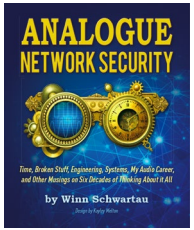
3. Bias - Types

Confounding bias

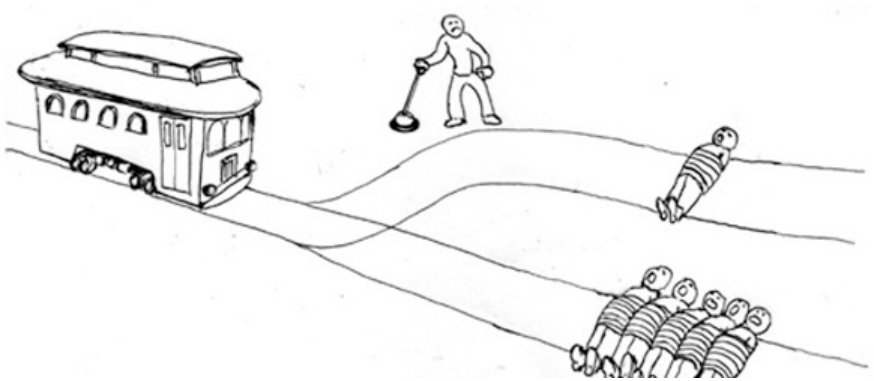
Reliable
Precise
Lack of Random Error

Reliable and Valid

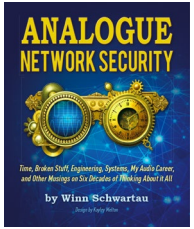
Valid
Lack of Systematic Error



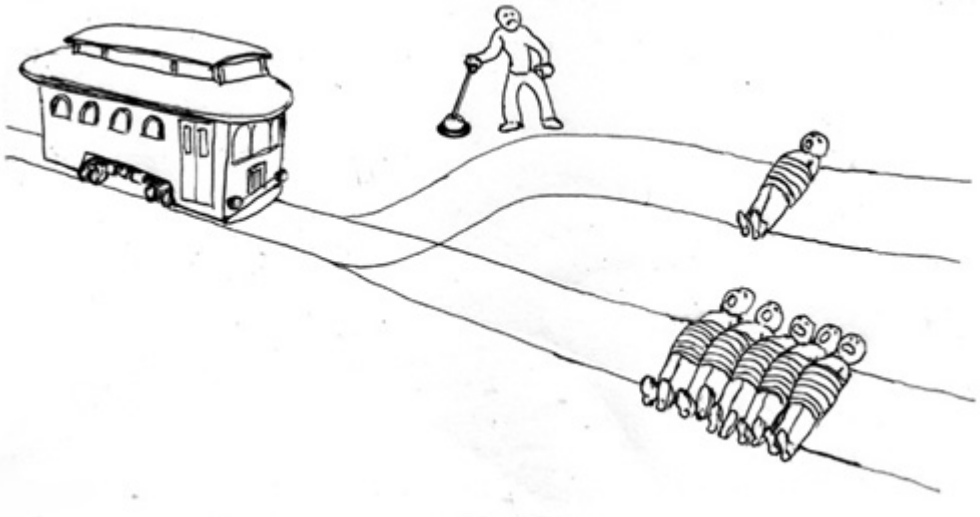
3. Trolleyological Conundra

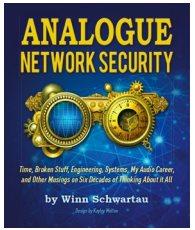


- Ethical concerns
- In programming – a solution can be hard wired into code – if death1 < death2, choose death1.
- Unless death1 is you, or a family member AND death2 is not.
- BUT – in machine learning systems, they follow actual human behavior



Can we leave these decisions to AI?



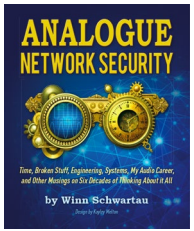


The Kobayashi Maru of AI?

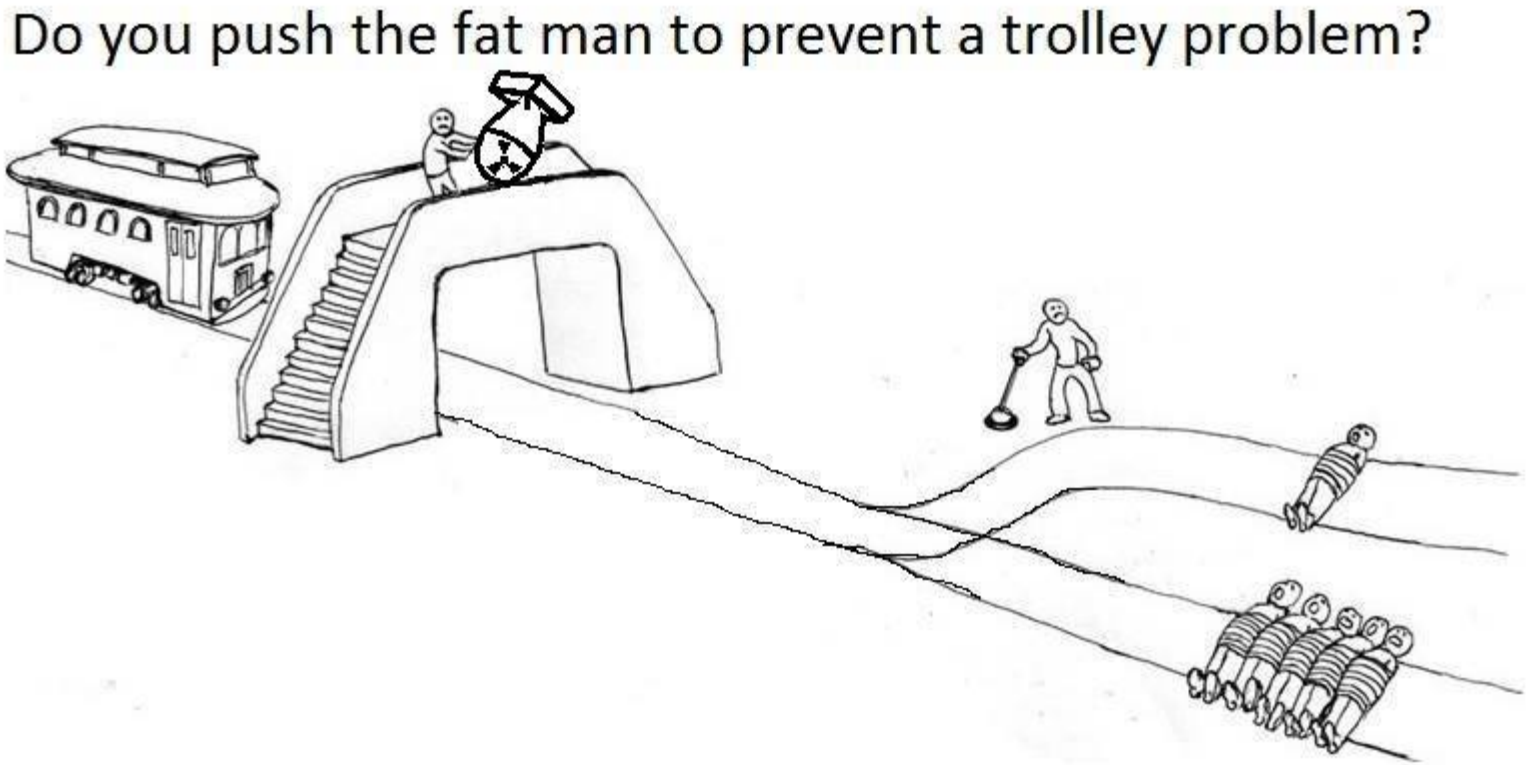
The trolley problem of

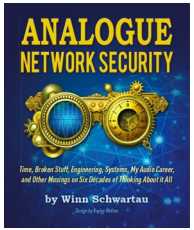
Isaac Asimov's first law of robotics

1) A robot may not injure a human being or, through inaction, allow a human being to come to harm



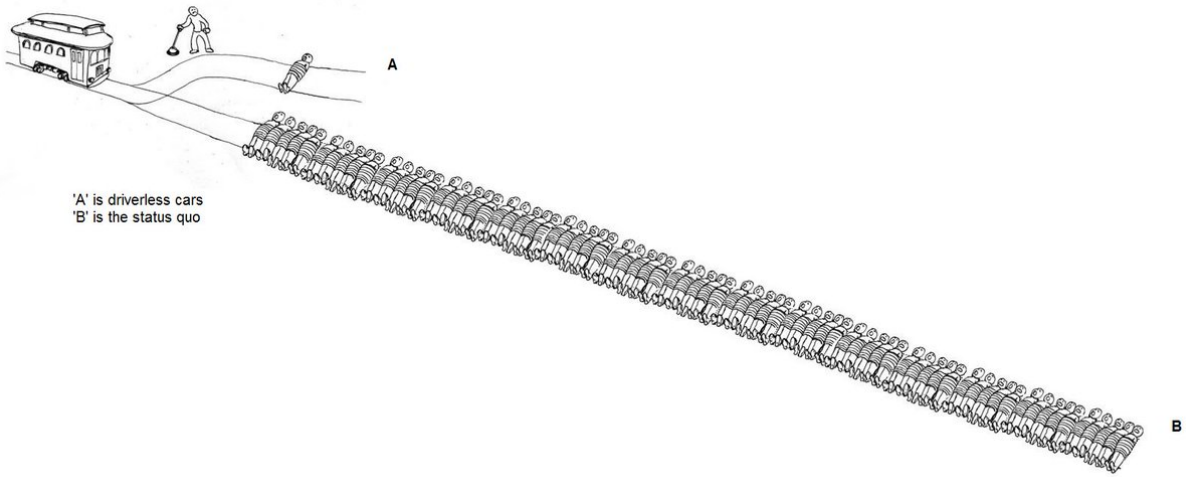
Kirk-like answer to Kobayashi Maru

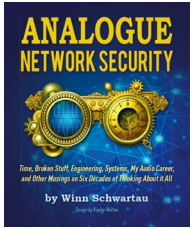




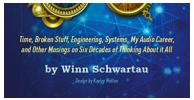
AI is not Zero-Sum

- Sometimes there's no good answer, and these constructs like the "Trolley problem" all seem informative, but none of them are able to capture all the nuances of the problem

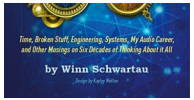




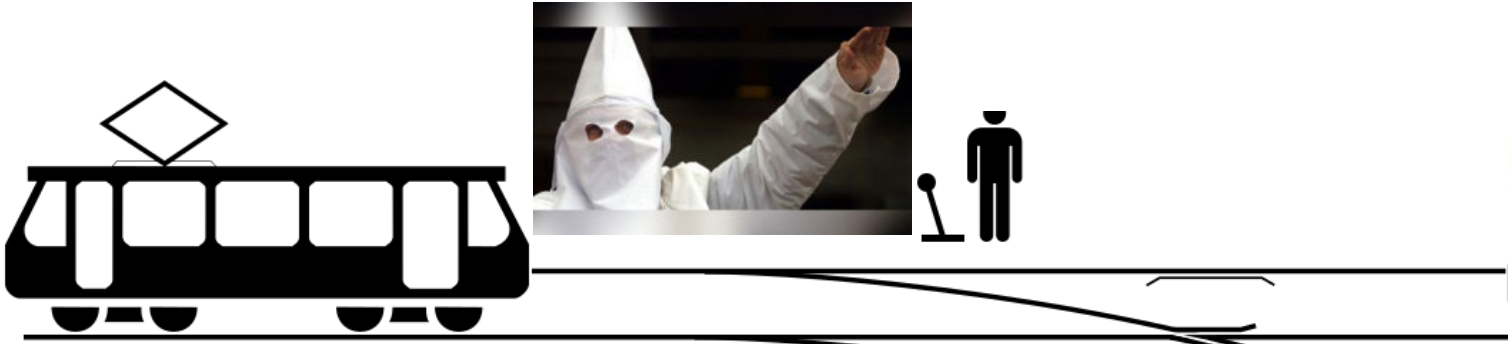
The Ethical Decisions that Anthro-Cyber-Kinetic AI Security Must Answer

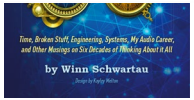


Same Ethical Decision in NYC? Alabama? UK? SA?

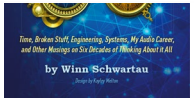


Who is Setting the Bias?

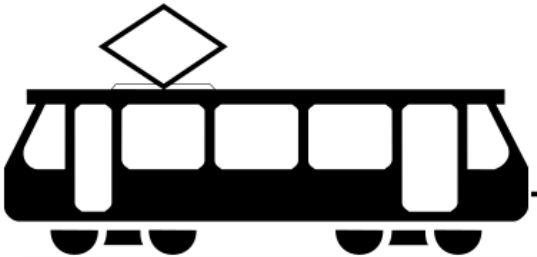


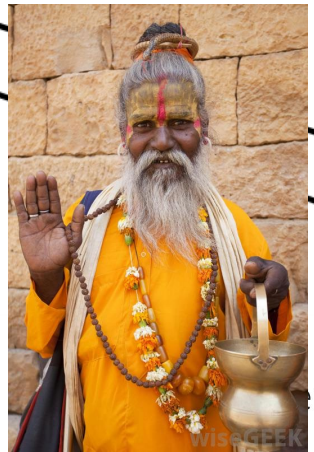


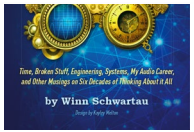
Who is Setting the Bias?



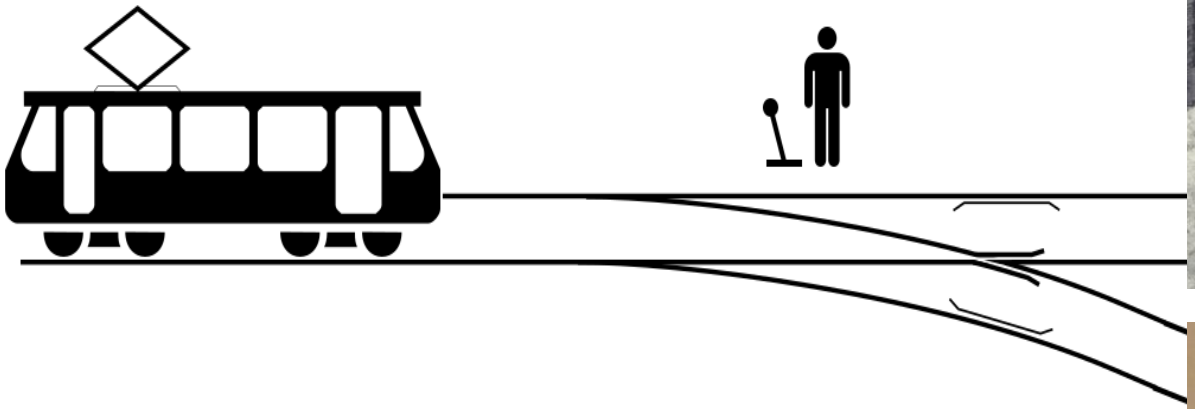
Cultural Bias in Anthro-Cyber-Kinetic AI Decision Matrices: Same Everywhere?

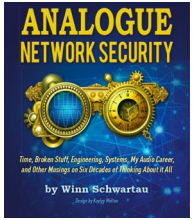




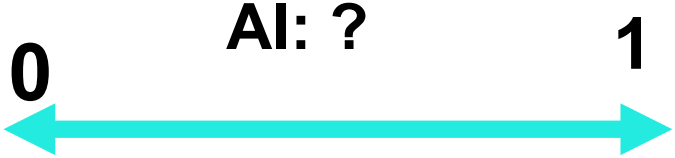


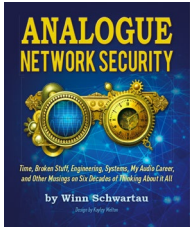
Same Decision Everywhere?
Or… is it Location Dependent?





Bias:
Degree of Trust You Choose?
Who is the Bias 'Decider'

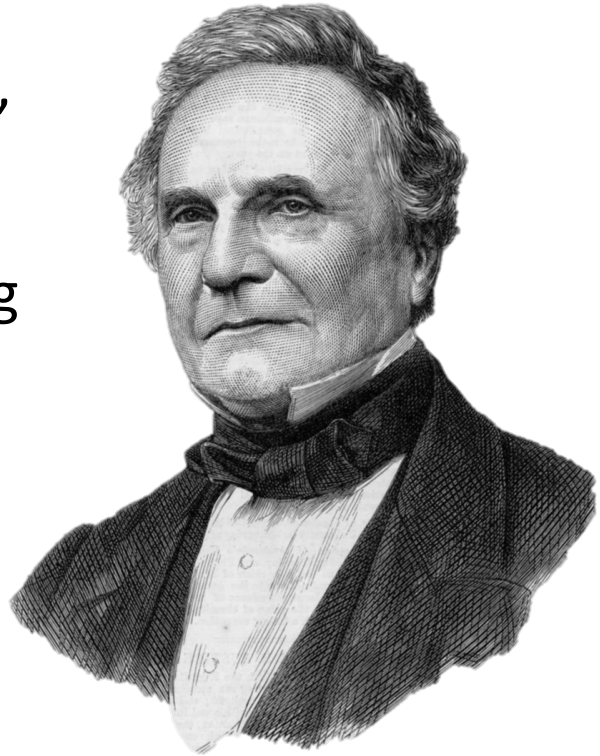


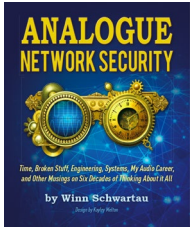


Two Centuries of Computer Bias

"On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?'" wrote computing pioneer Charles Babbage in 1864.

And thus, the fundamental software principle of 'garbage in, garbage out' was born.

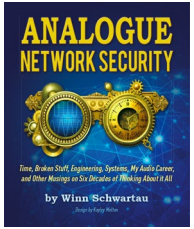




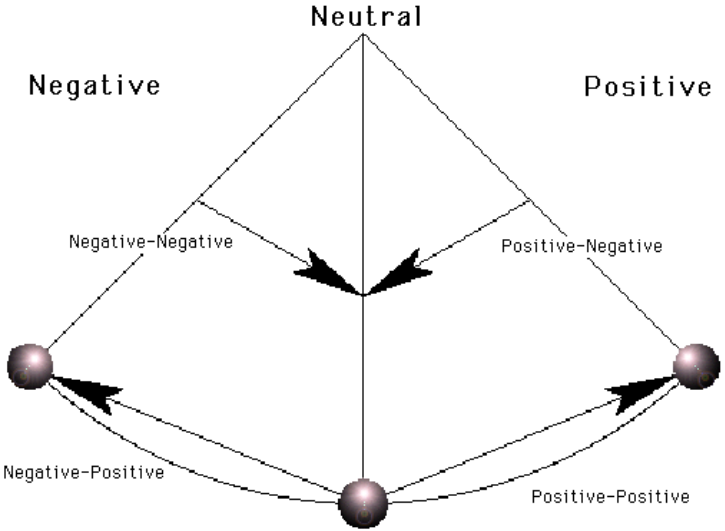
Today: AI Bias

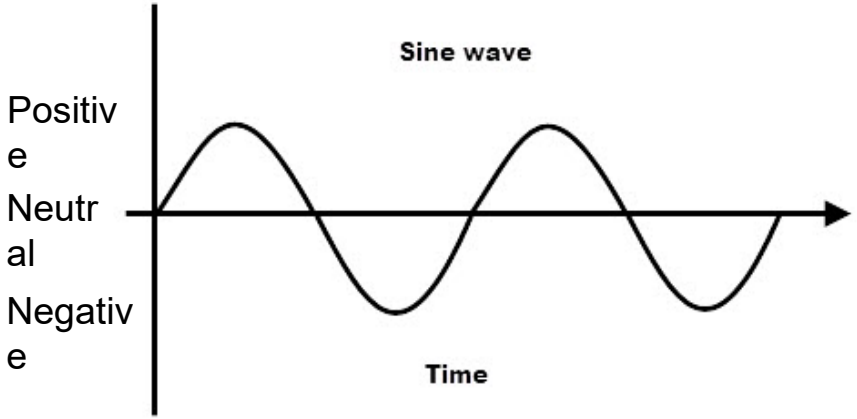
"'Pray, Mr. Chio, if you put into the machine the same data, at many different times, when other questions are asked of the machine in intervening times, will the same answers come out?'"

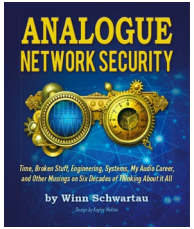
And thus, the fundamental AI principle of 'we don't know for every case' was born. (Bias change, unknown out.)



Bias Shifts Over Time (Analogue function)



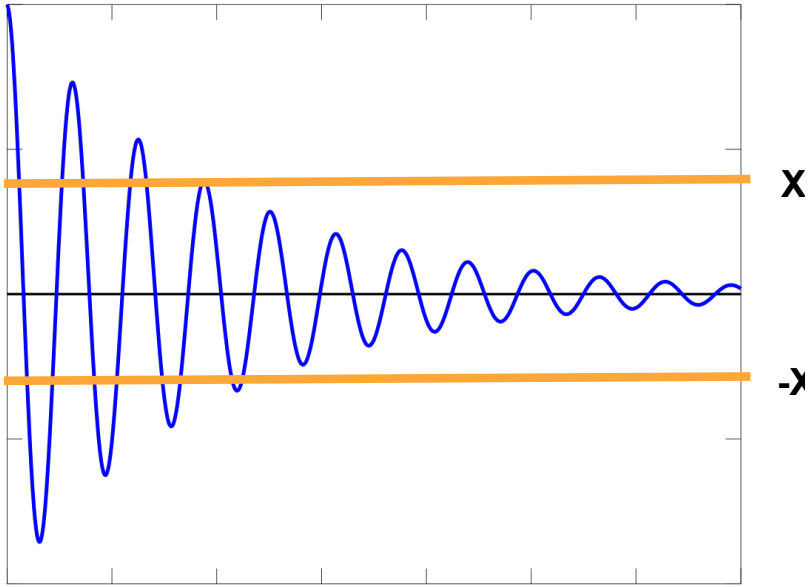


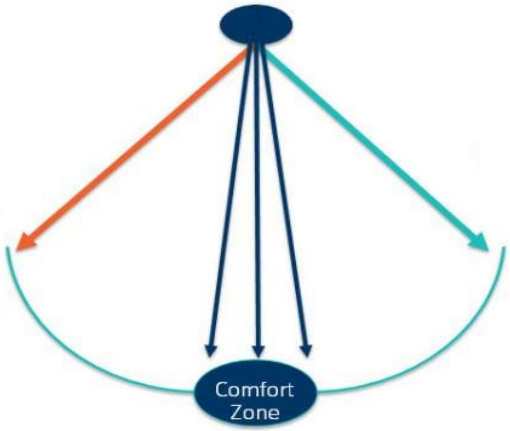


- The Goal is to minimize the value/swing of the pendulum (Bias)
- Lower the upper and lower extremes of the sine wave (Bias)

The Goal of Bias is Damping the Swings

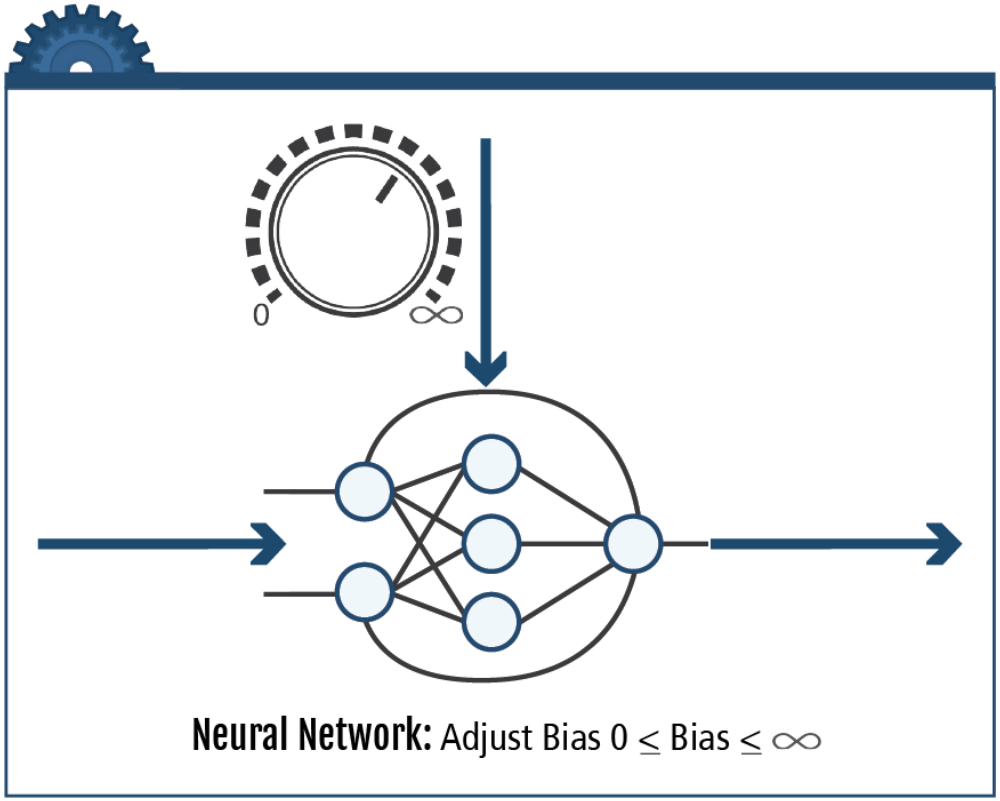
A Bias swing > |X| should trigger a response

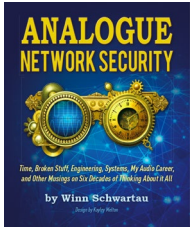




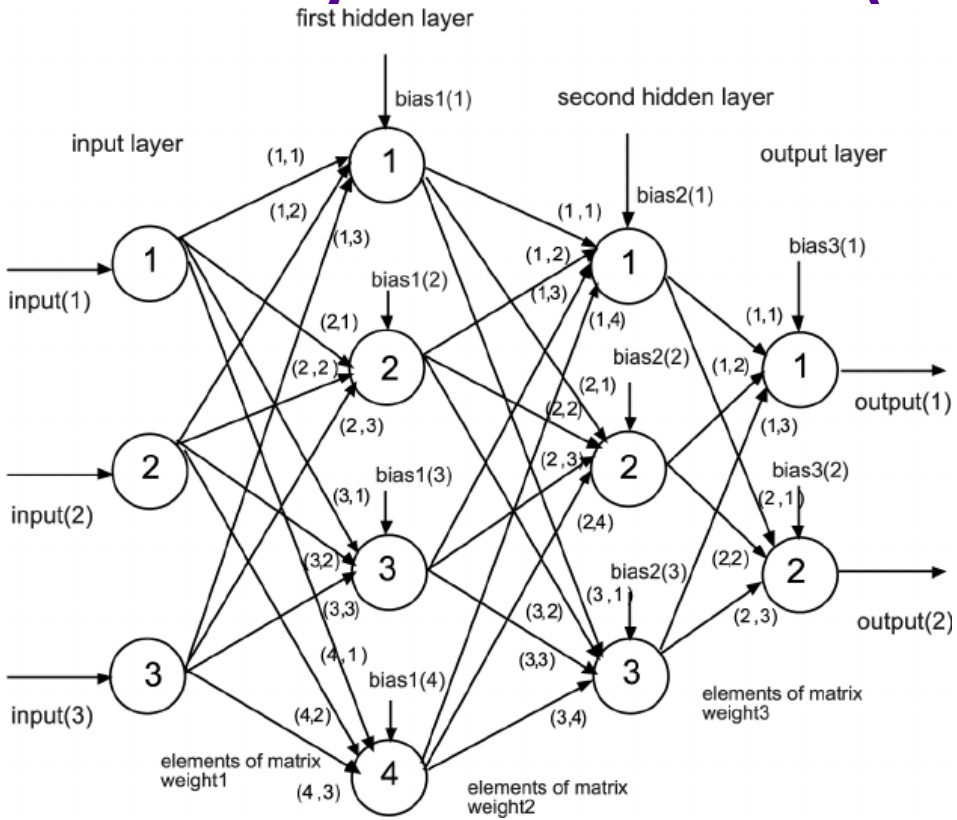
The Bias swings are reduced over time

Bias Control: Setting Min-Max D/R Conditions

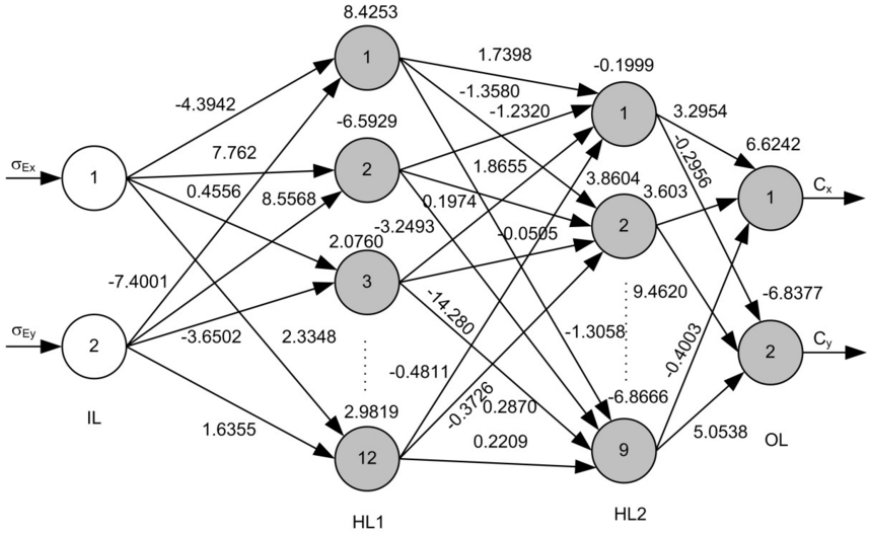




Multi-Layer Bias in Neural (DL) Networks

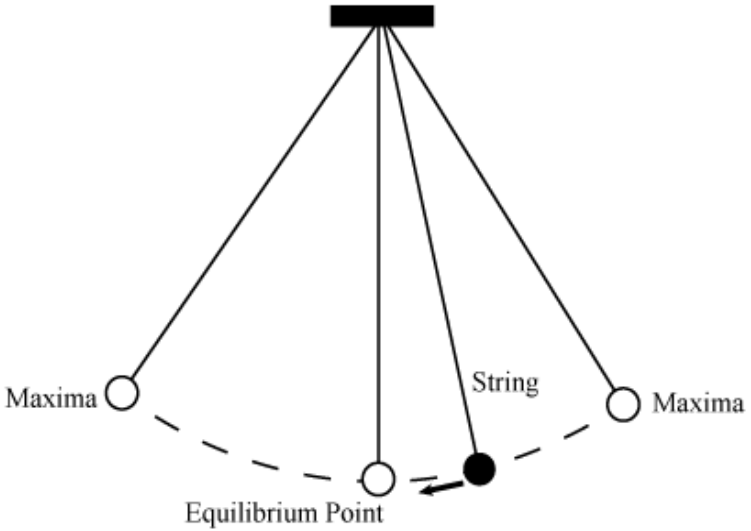


Measured, a la
Analogue Network Security



Bias in Data Sets: Looking for Neutrality

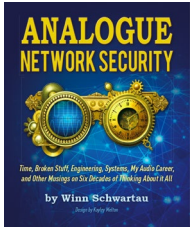
- HR: Human bias for/against interviewees
  - If the input data are biased – say, consisting of mostly young white males (our 'garbage in,' as it were), then who will the AI recommend? You guessed it: mostly young white males (predictably, the 'garbage out').
- Research Studies: Neutral participants?
- Male/Female dominance?
- Race/Color/Creed biases?
  - If the training sets aren't really that diverse, any face that deviates too much from the established norm will be harder to detect.
- Reflected in data aggregation
  - Sample company size, location, culture



Responsible Training: Bias → 0

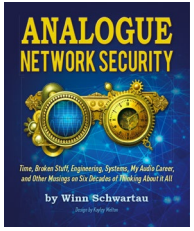
The most powerful algorithms being used today "*haven't been optimized for any definition of fairness*".

Assumptive Bias



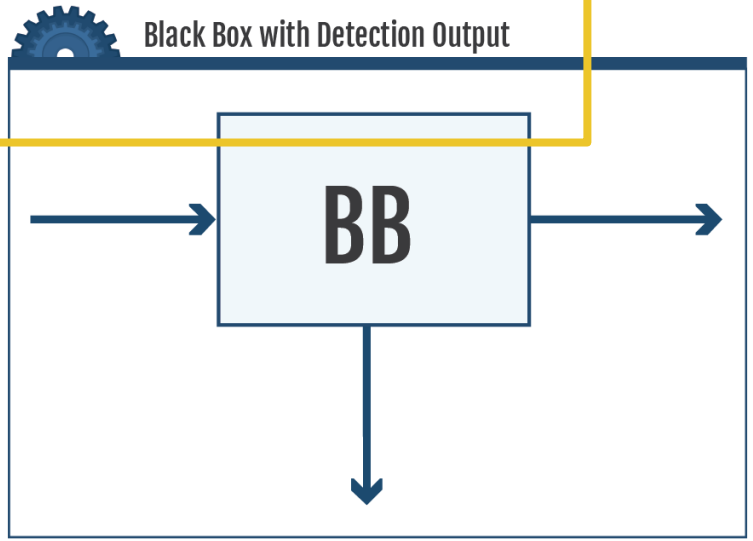
What is Artificial Intelligence?

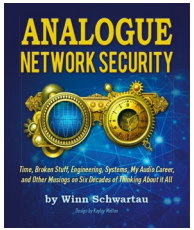
- The tradition definition is much too broad to draw a risk assessment.
- What is and isn't true AI? What constitutes an actual threat?
  - GOFAI (Symbolic manipulation)
  - Adaptive Neural Networks
  - Cognitive Simulation (based on psychological research)
  - Self-modifying algorithms
  - Data mining, clustering, and synthesis
  - …other forms
- We often conflate AI and AGI… Artificial General Intelligence.



AI as a Black Box

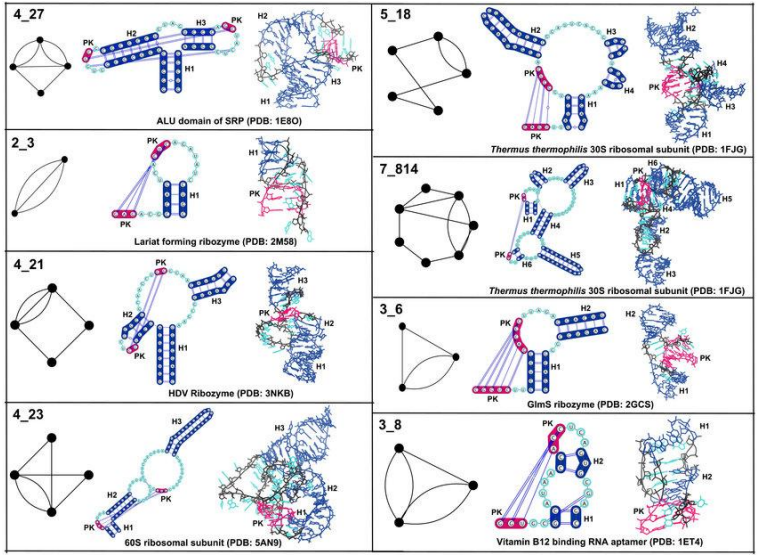
- Would we accept a C-Suiter saying, “Here’s the answer. But I can’t tell you why I made that decision.
- Unless we know how the AI arrives at an answer, can we trust it?

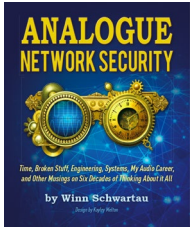




Poisoning Data Sets: A Topological Approach

1. Trolls on Social Media
2. Repetitive fake news, lies, distortions
3. Ignorance of bias & neutrality
4. No re-vetting/balancing
5. Criminal input to existing systems
6. Adversarial AI attacks
7. Hack the AI (algorithm/data)





https://www.bcg.com/publications/2018/artificial-intelligence-threat-cybersecurity-solution.aspx

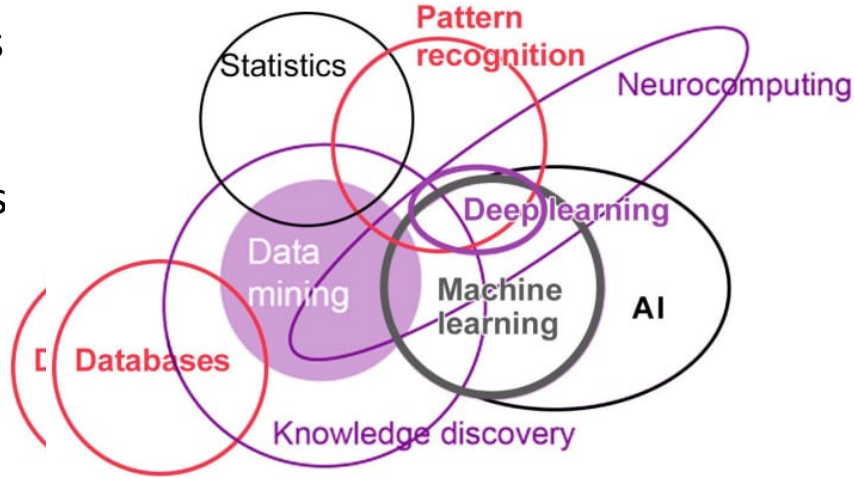
Introducing Bias - Unintentional

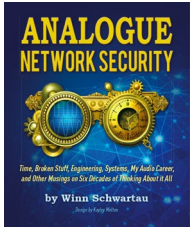
1. Confirmation Bias (I like that answer, therefore I trust it.
2. Availability: It's the only data I could find.
3. Executive Override (He's the smartest person in the room.)
4. Emergent bias (positive feedback)
5. Errors from misconfigurations

Bias in Data Sets & Statistical Learning

Bias in data sets

- Well balanced data sets are scarce
- Influenced by developers' cultural bias
- Can be caused by
  - sampling bias (selection/exclusion bias
  - observer-expectancy effects
  - Label inaccuracy
  - Missing data & interpolation strategies

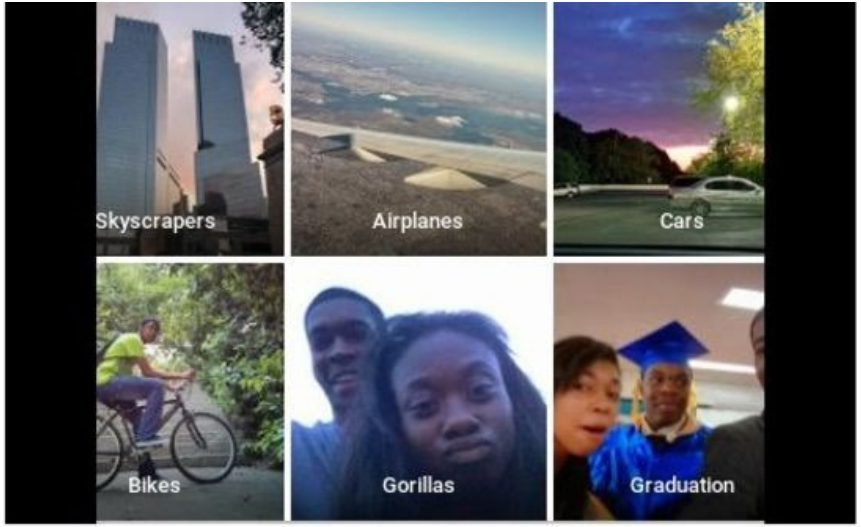


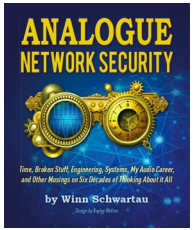


RSAConference2019

Case study A:
Closed-box Education & Racist Algorithms

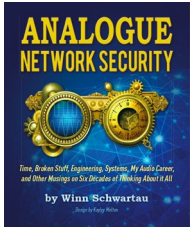
Case study A:
Closed-box Education & Racist Algorithms





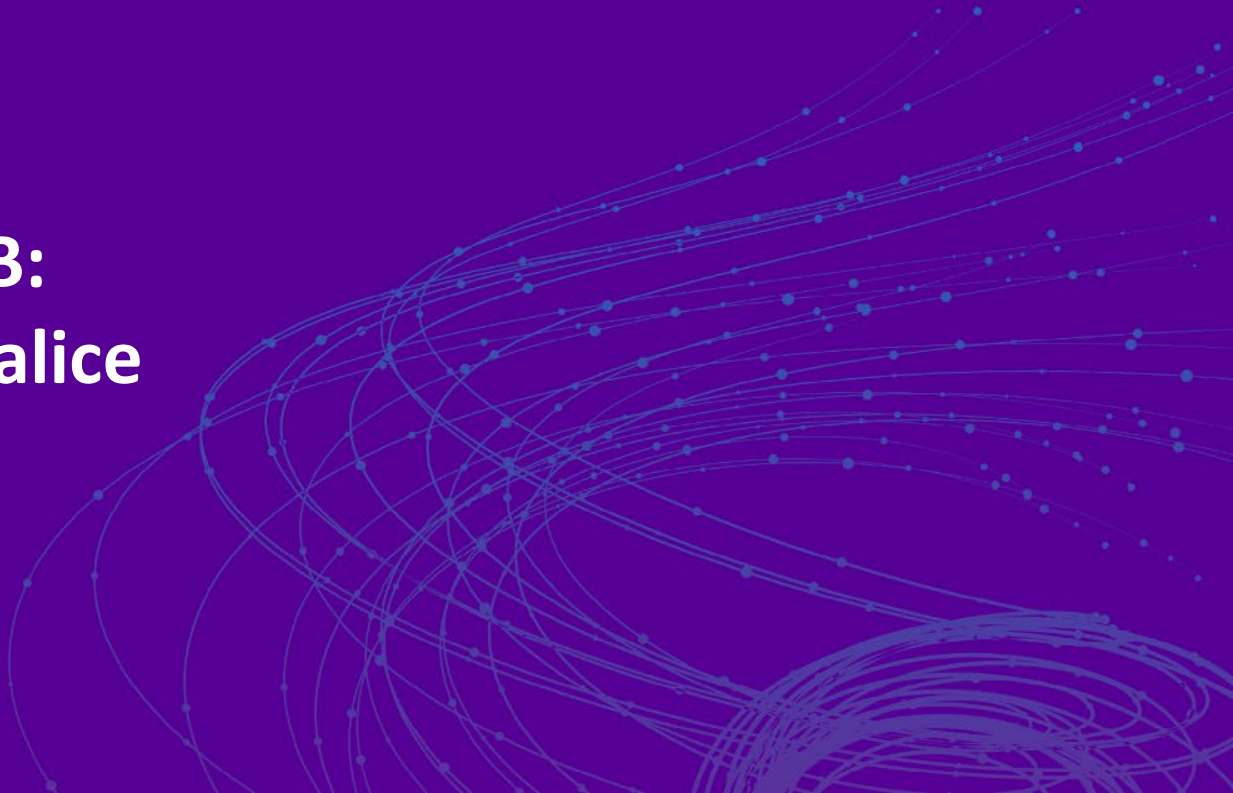
Case study A: Closed-box Education & Racist Algorithms

- Cultural and societal biases seep into statistical learning algorithms in a scary way
- The fix is **really, really** expensive
- In the race to build the most powerful autonomous AI & data systems, data quality is frequently sacrificed in exchange for highly-scalable ways of extracting data.
- Unhealthy consensus: More data trumps bad algorithms
- Fundamental problem is related to messaging & society's expectation of AI systems

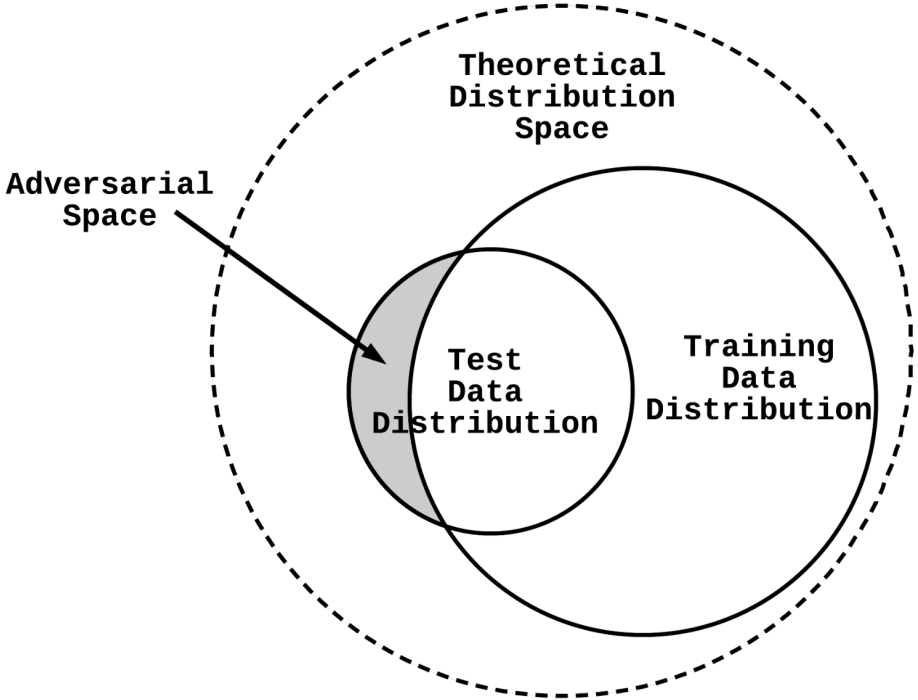


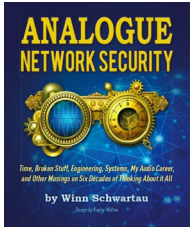
RSAConference2019

Case study B: Targeted Malice

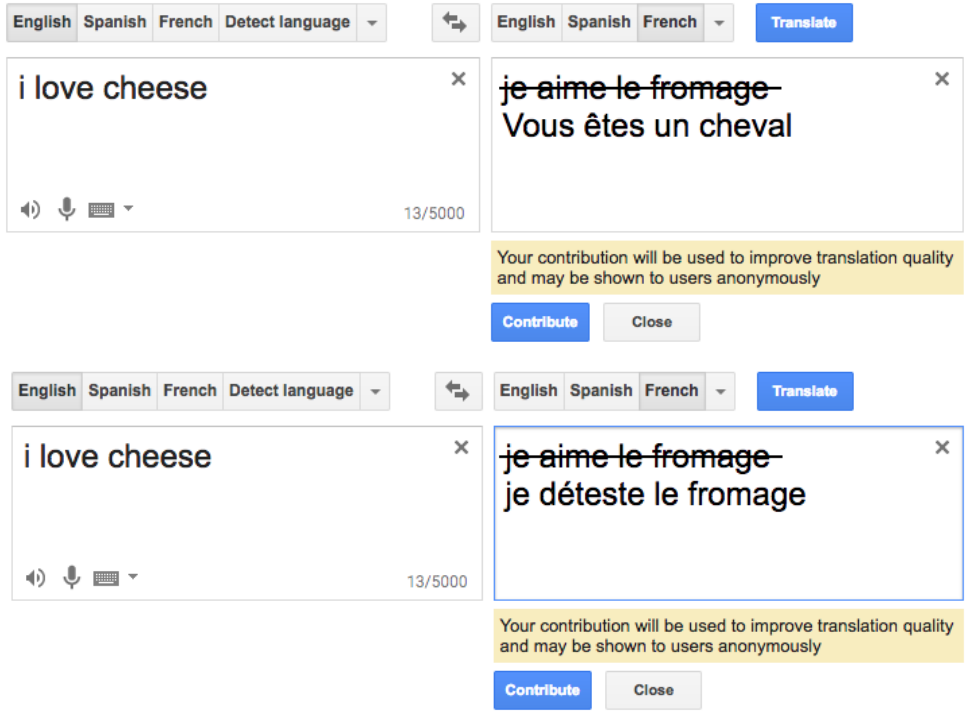


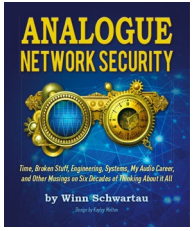
Case study B: Targeted malice





Case study B: Targeted malice





Case study B: Targeted malice

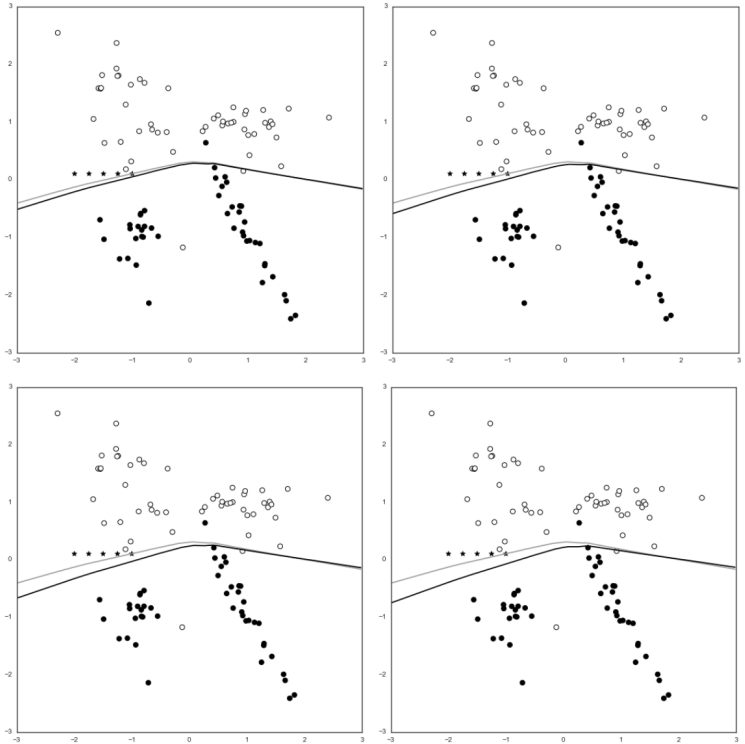
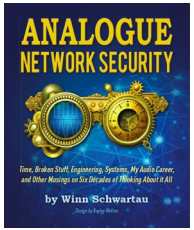
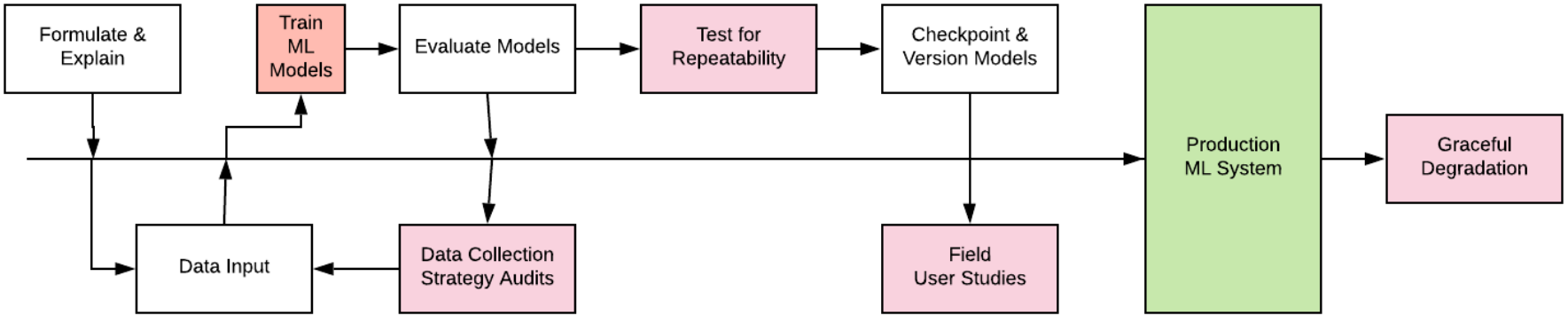
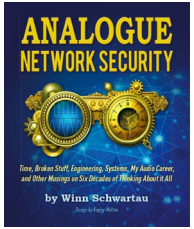


Figure 8.8 Shifted decision boundaries after 2x, 3x, 4x, 5x partial fitting of 5 chaff points (10%,15%, 20%, 25% attack traffic - top-left, top-right, bottom-left, bottom-right)



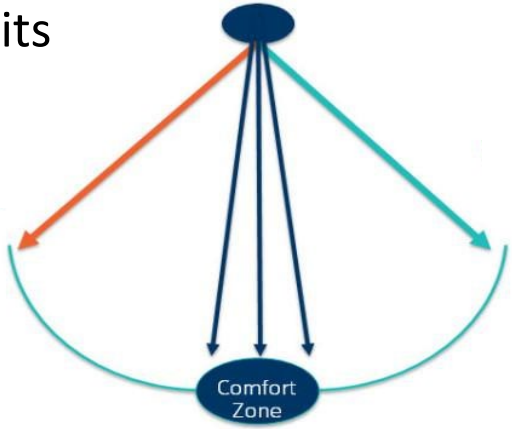
A Bias Evaluation Framework

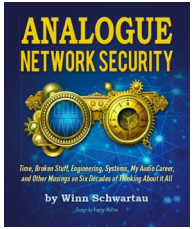




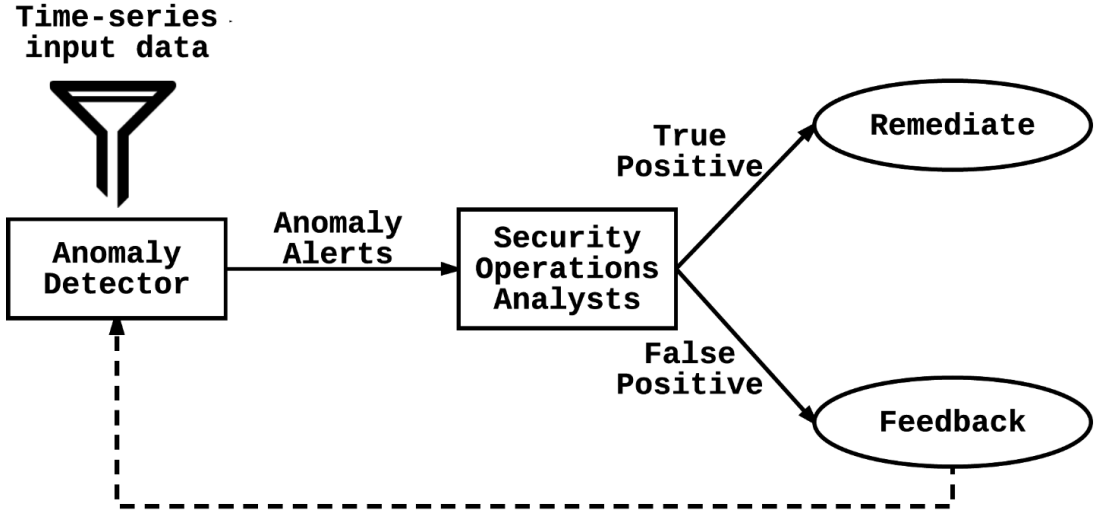
A Bias Evaluation Framework

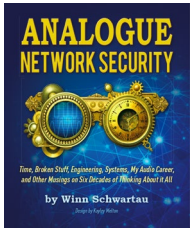
1. Explain
2. Tune input data
3. Independent data collection strategy audits
4. Maintain trained models
5. Repeatability
6. Checkpoint and version models
7. Evaluate in limited user studies
8. Graceful degradation of services





A Trust Model for AI





XAI – Explainable AI
The Right to Understand (a la GDPR)

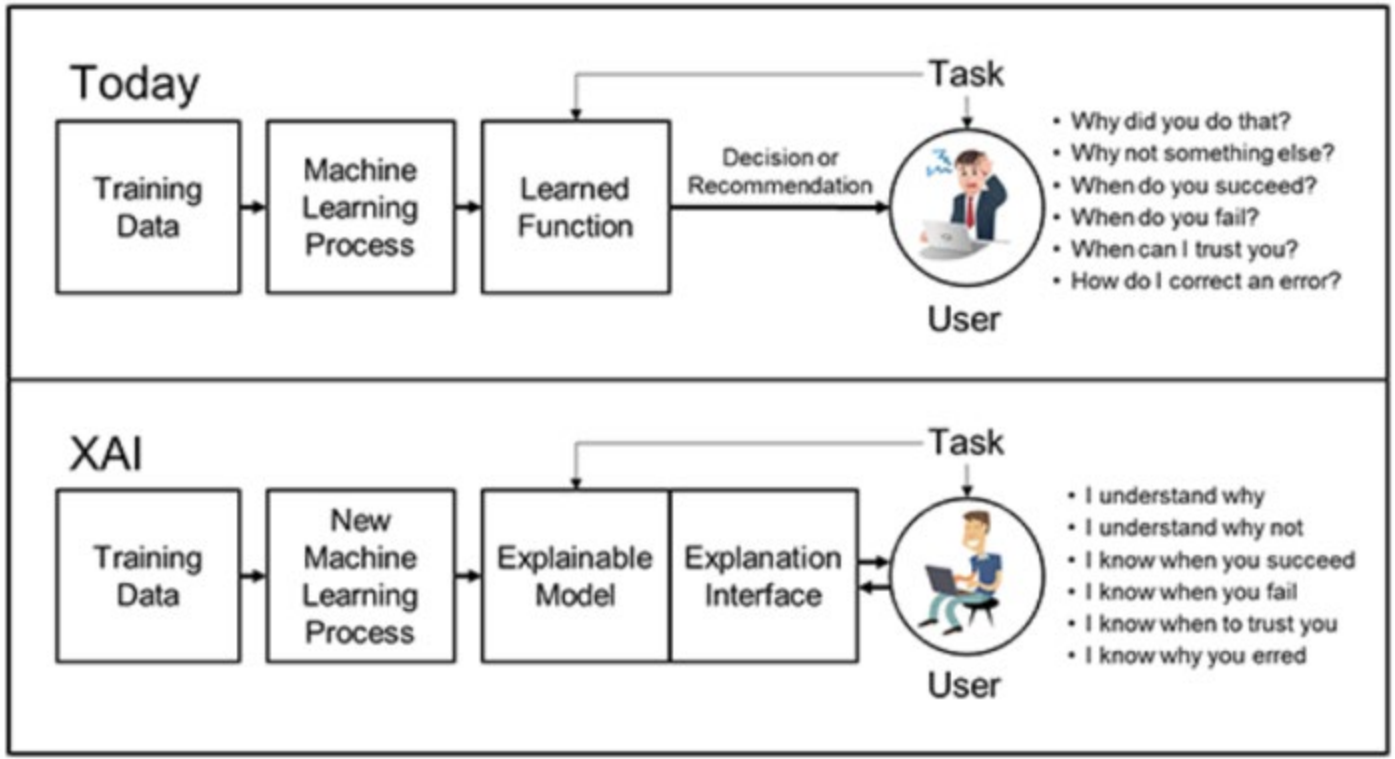


Figure 1: XAI Concept

AI to detect biased AI algorithms

MIT researchers show how to detect and address AI bias without loss in accuracy

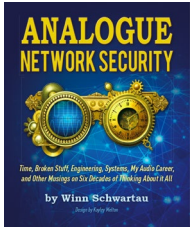
KHARI JOHNSON @KHARIJOHNSON NOVEMBER 16, 2018 10:00 AM

IBM launches tool aimed at detecting AI bias

By Zoe Kleinman
Technology reporter, BBC News

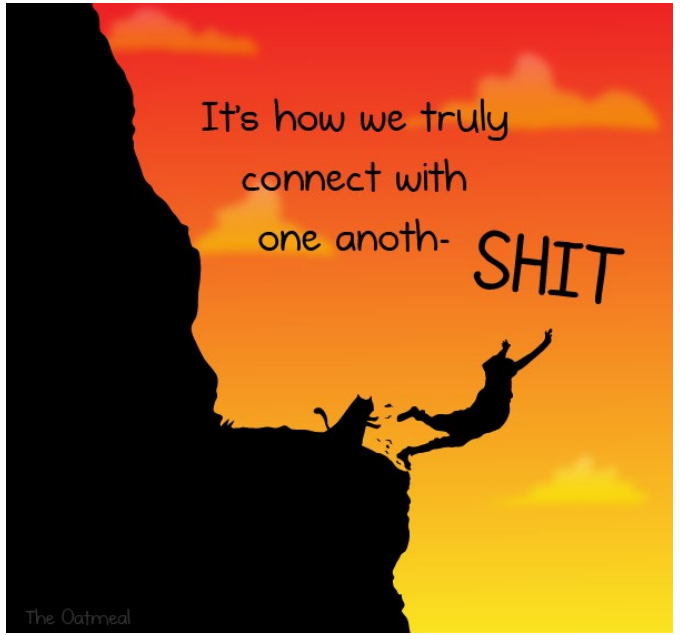
Microsoft is creating an oracle for catching biased AI algorithms

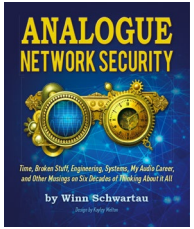
As more people use artificial intelligence, they will need tools that detect unfairness in the underlying algorithms.



Should we, can we trust AI in security? How much should we?

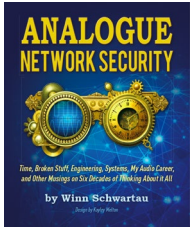
- Intentionally poisoned data sets? (How would we know?
- Unintentional poisoning from biased 3rd party data sets.
- The answer doesn't make sense. Then what? Trust the AI or…?
- Do we automate responses to AI decisions?





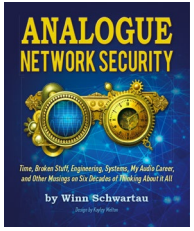
Key Takeaways #1: What is AI good for?

1. When approximate answers are good enough
2. Language translation; Picking your music, purchasing suggestions, etc. When you have oversight and final approval.
3. Some cyber-kinetic systems (with no autonomous life-death decisions)
4. Medical diagnosis (with human oversight)
5. Security – with oversight. Analyzes massive amounts of data
6. Finds patterns and trends we might not notice or might ignore
7. Looks for the Ghost in the Machine – things that come in below the radar
8. Statistically distinguishes between “normal” and abnormal behavior
9. Develop challenge/response on its own
10. Resilient against some forms of social engineering



Key Takeaways #2:
What Should You Ask of AI Security Vendors?

1. Make them show you how answers and results are arrived at?
2. Show you why Their AI is better than the Other-Guy's AI.
3. Demonstrate that a given set of inputs will consistently give a constant set of outputs (In MSA - master service agreement and Warranty).
4. Demonstrate the initial bias conditions and how additional experience will change those biases and decision outputs.
5. Explain how they use XAI; or, if they don't, why not?
6. Decide if current AI is worth the investment, uncertainty, and possible ethical errors with unknown ramification.



RSAConference2019

Questions?

Follow up with Clarence or Winn anytime.

Winn@TheSecurityAwarenessCompany.Com

mail@cchio.org

Some extra slides…

TBC