RSAConference2020

San Francisco | February 24 – 28 | Moscone Center

HUMAN ELEMENT

SESSION ID: MLAI-W11

All That Glitters? Debunking Fool's Marketing of ML and AI

Diana Kelley

Cybersecurity Field CTO
Microsoft
@dianakelley14

Dr. Char Sample

Chief Research Scientist
Cybercore Division at Idaho National Laboratory
@stillchar

Agenda

- Bright Shiny Objects – Terminology Clarifications
- All about that Data
- Cognition and Bias
- Sifting out the Fool's Gold
  - Five Questions to Ask Vendors

AI vs ML

- Marketing often uses the terms interchangeably
- Getting technical though…
  - AI:
    - Turing test, being “human”, auto-translation, solving problems
    - The application of what was learned
  - ML: Mathematical models, better than humans?
    - Lessons from events
    - Prioritization of those lessons

AI vs ML

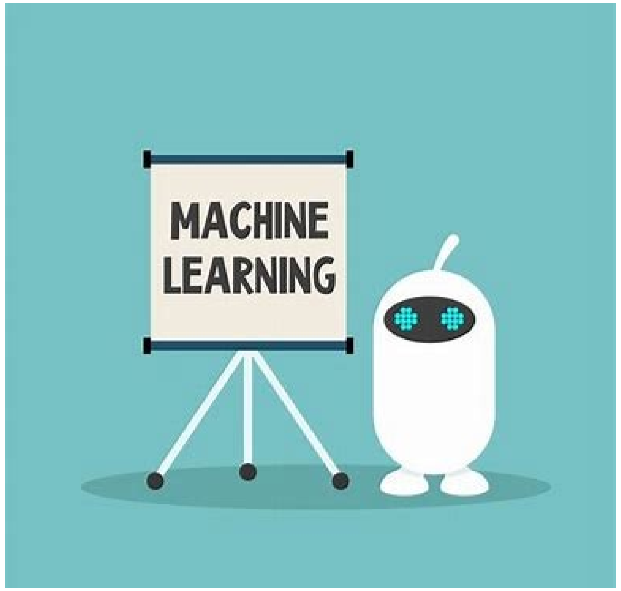
- Term coined by Arthur Samuel, defined by Tom Mitchell:

*"A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in **T**, as measured by **P**, improves with experience **E**."*

How ML "Learns"

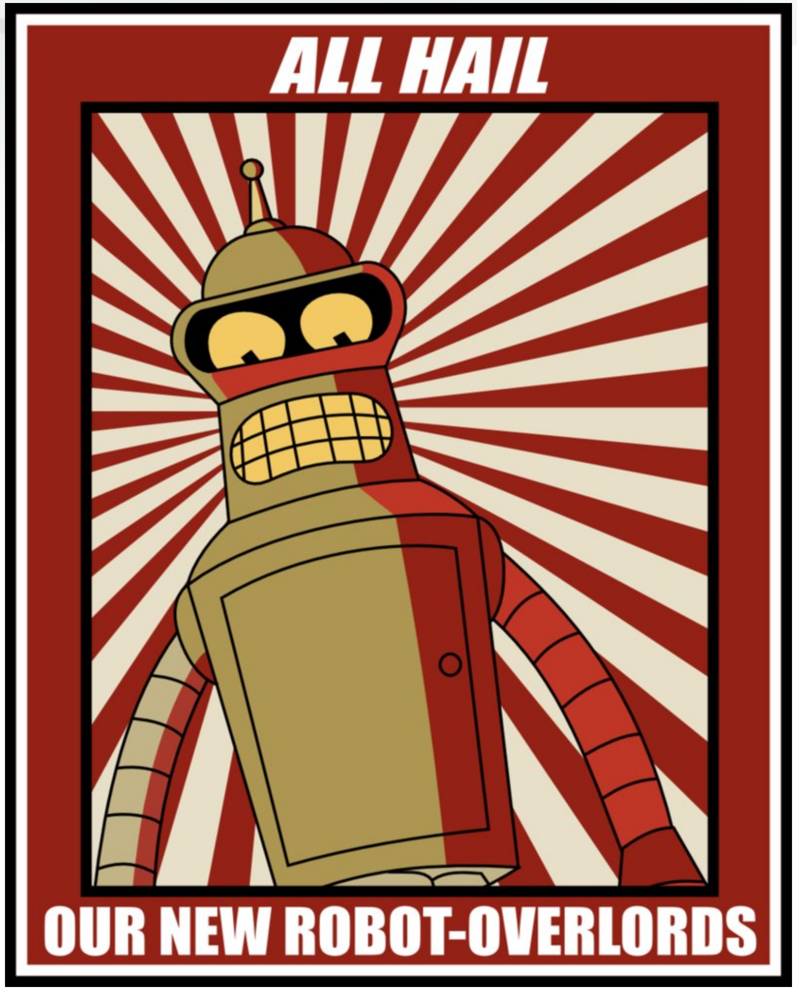
- Model is built with training data
  - Model learns to perform a task from that data
    - Ex: how to play Go, how to detect cancer in a radiograph, catching a phish
  - Possibility to improve ("learn") over time without human intervention
  - Unintended lessons learned
    - Tay
    - Generative Adversarial Network

    https://openai.com/blog/emergent-tool-use/



Narrow vs General

- General (AGI)
  - What scares most people
  - The rise of the sentient machine
  - Robots replacing humans
  - Technically possible, but not right now
- Narrow (ANI)
  - Limited in scope/use case
  - Machines assisting humans
  - Wide variety of deployments today and growing



Machine Learning

Supervised, Unsupervised

- Supervised
  - Training data is labeled, input/output pairs
  - Model accuracy highly dependent on labels/representation
  - Bias-variance balance
- Unsupervised
  - Training data – input only
  - Pattern detection

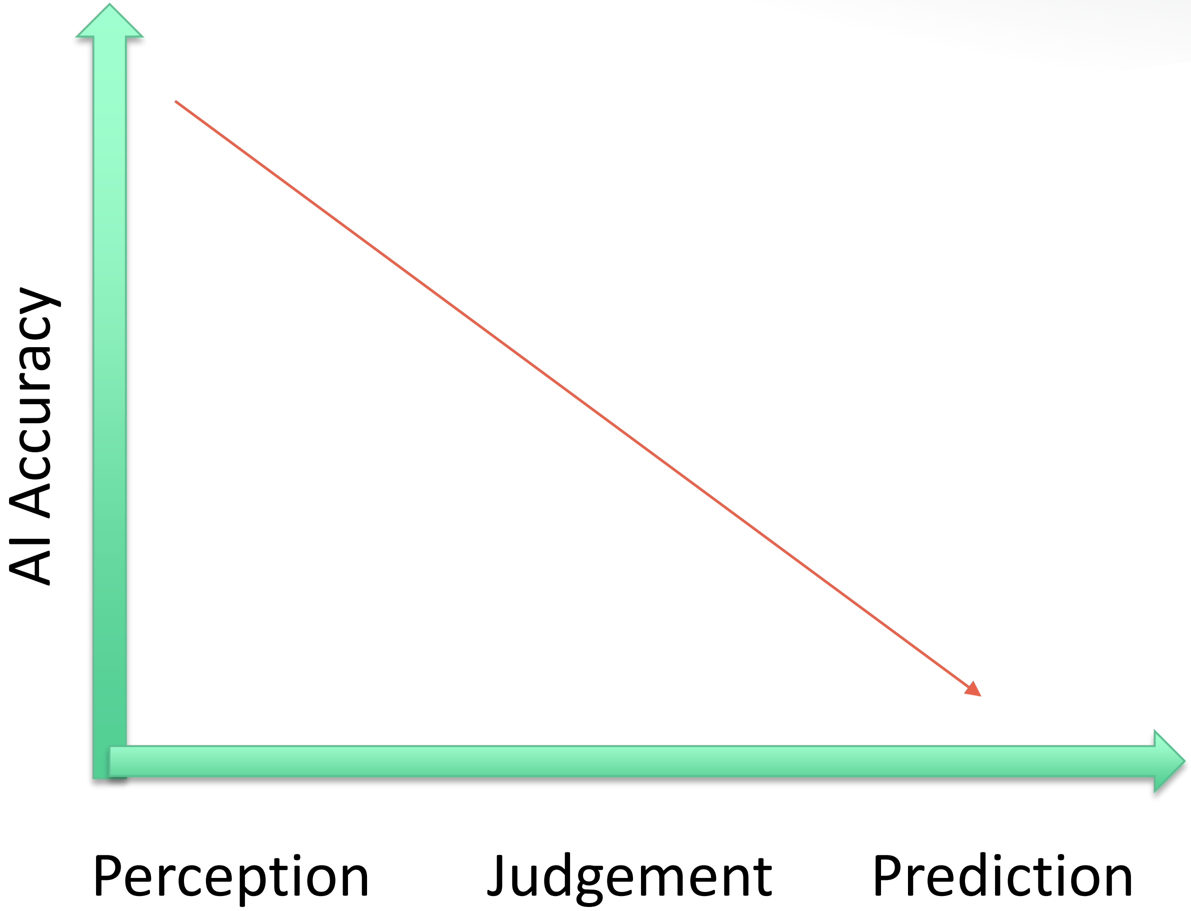
Reniforcement, Overfitting

- Reinforcement
  - Concerned with actions by agent or node
  - Observation – Reward – Action – Repeat
- Overfitting
  - What is it?
  - Why do we care?

Ingest vs Retrieval Speeds

- How the data is organized matters
  - Data structures (computationally speaking, searches are expensive)
    - Trees
    - Neural networks
  - Language chosen (high level languages use more processing power and resources, normally this does not matter, *but* remember Moore's law?)
    - Java
    - C

AI and Behaviors Balancing Act



(Arvind Narayanan 2019)

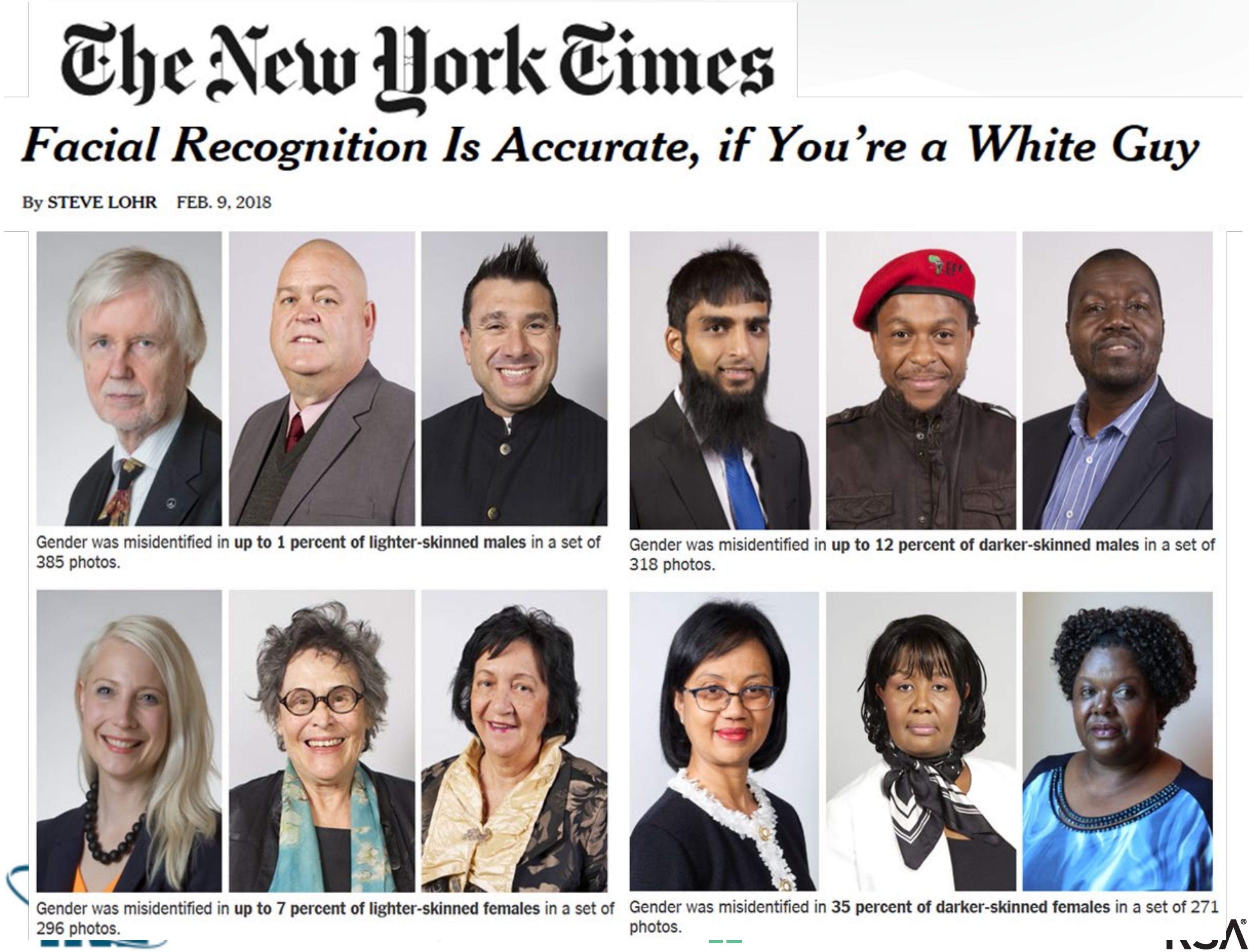
Training and Test/Production Sets

- Training – Data used to train the model
  - Neural Net
  - Naïve Bayes
- Validation – Data used to confirm model fit
  - Correct for overfitting
- Test – Data used to test the model
  - Usually “fresh” (holdout), not used in training
- What about data poisoning?

The Importance of Classification

- Classifiers are key to answering questions with ML
  - Process of organizing elements into classes
  - Example: Keanu Reeves pictures (class one)
  - All other “non-Keanu” pictures (class two)
- How does a classifier separate data elements?
- Takes information as input
  - Series of data points
    - Health vitals, financial info
  - Block of pixels
    - Picture, radiogram
  - Text/Audio
- Outputs a prediction
  - Probability that something is found in a given set (e.g., “is Keanu in this picture?”)
  - Answer to a yes/no question (e.g., “is this a human face?”)

The Problem with Bias



The New York Times

Facial Recognition Is Accurate, if You're a White Guy

By **STEVE LOHR** FEB. 9, 2018

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

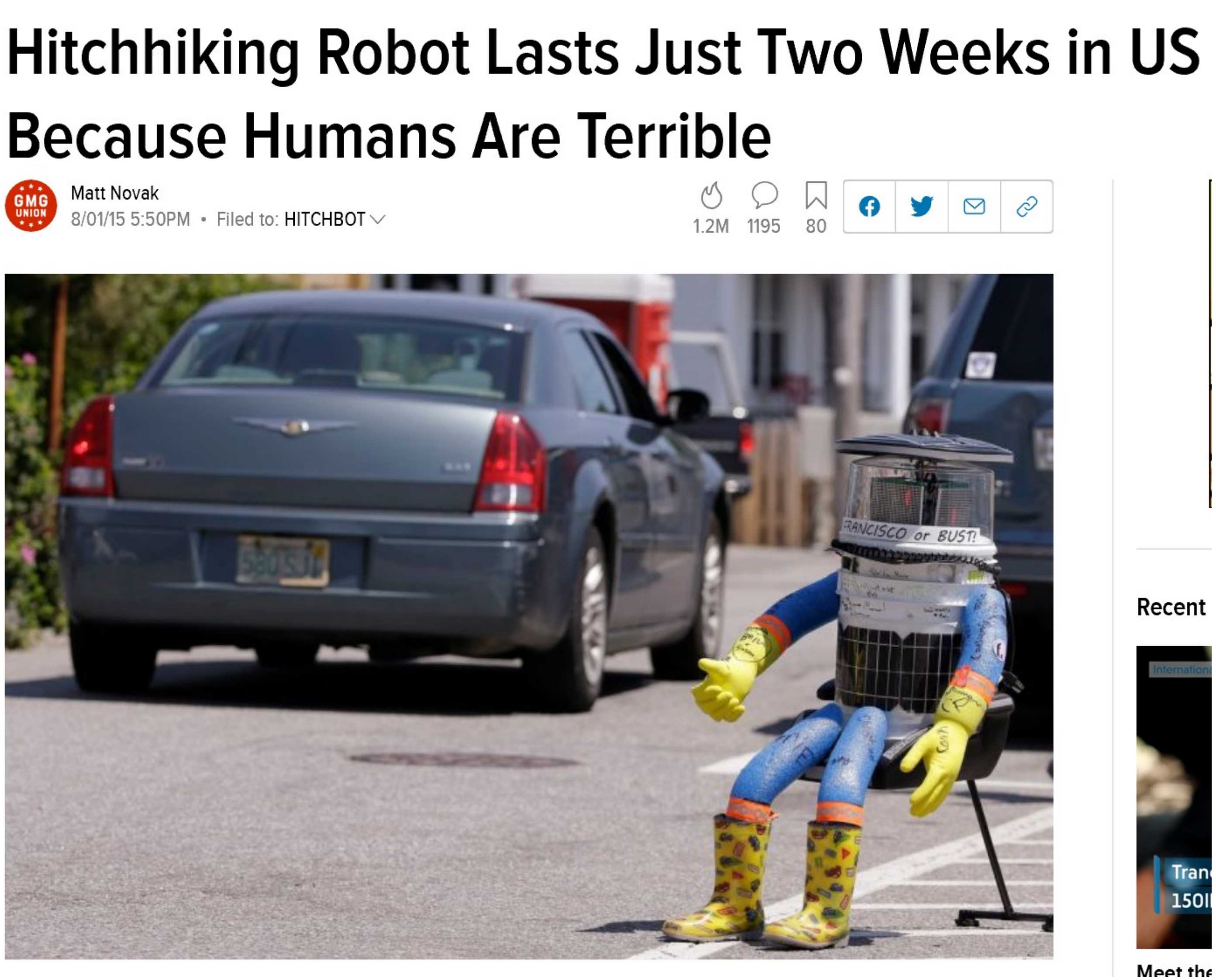
Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Solving for Bias

- Understand how bias can be introduced and affect recommendations
  - Programmers work off varying cognitive models
  - Programmers are biased, and biases can be good or bad
- Attract diverse pool of AI talent
- Develop analytical techniques to detect and eliminate bias
- Human review and domain expertise

The Accountability Problem

- Who is responsible when things go wrong?
  - Legal rights?
  - Human rights?
  - AI/Robot rights?



Hitchhiking Robot Lasts Just Two Weeks in US Because Humans Are Terrible

Matt Novak
8/01/15 5:50PM • Filed to: HITCHBOT

1.2M 1195 80

Recent

Tran
150l

Meet the

AML/MUAI

- Adversarial Machine Learning (AML)
  - Data poisoning is one technique to counter AML
  - Denial or disruption
  - Data manipulation (not the same as poisoning)
- Malicious Use of Artificial Intelligence (MUAI)
  - Differs from AML, MUAI exploits features of behavior for unanticipated outcomes.
    - Bots used to inflame political discourse
    - Denial of service attacks
    - Taking advantage of a set of features developed in one domain and applying to an unrelated domain.

RSAConference2020

Sifting Out the Fool's Gold

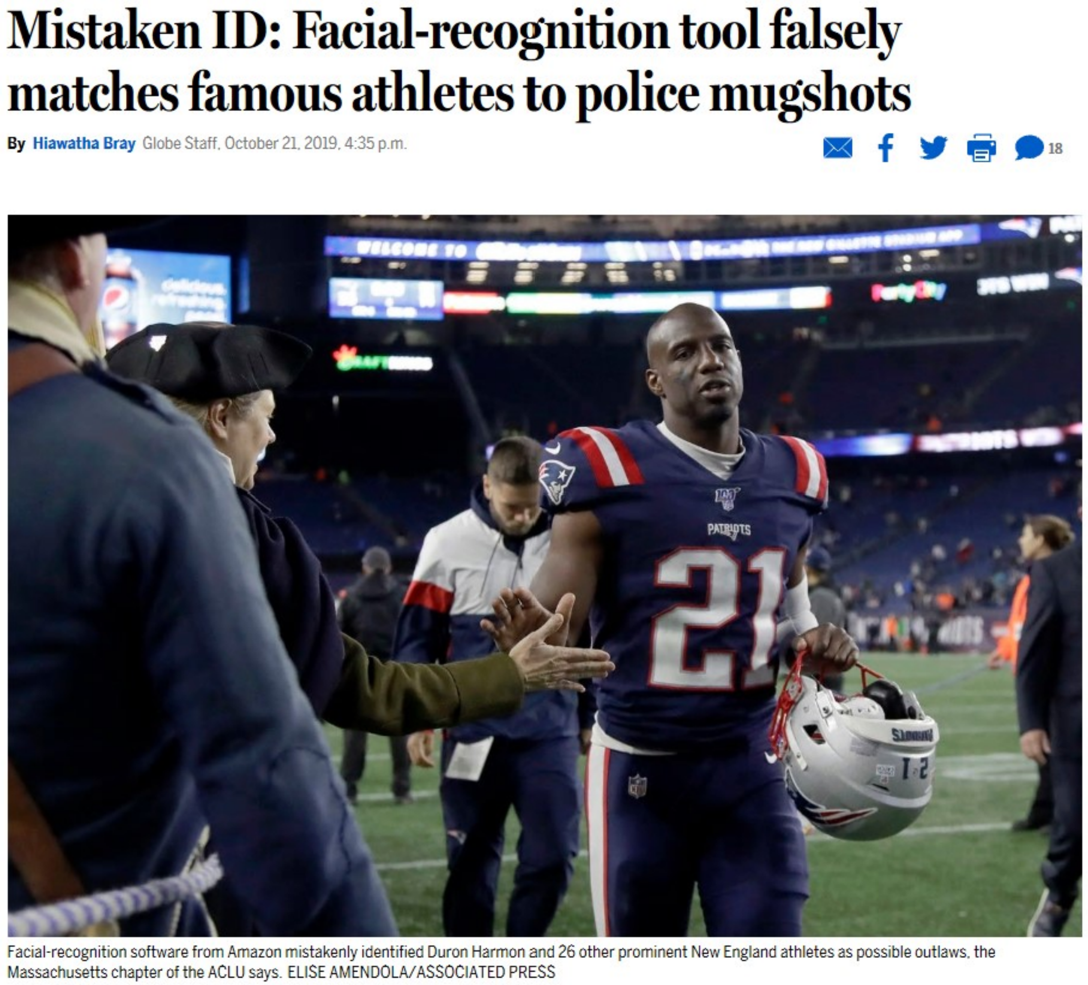
5 Questions to ask your Vendors

Question 1 - AI types and ML algorithms?

- If they've got a "super model"
  - They may not understand how the tech works
- AI Types revisited
  - AGI, ANI, and ASI (Super Intelligence)
- What algorithms are in use?
  - Why did the vendor select for those algorithms?
  - Naïve Bayes/Lasso Regression (Supervised)
  - Temporal Difference/Q-Learning (Reinforcement)

Question 2 – Data Sets for Training?

- What data sets are used for training and how are they labeled?
  - Bias data sets lead to biased outcomes from ML and AI
  - Improper labels and training – inaccuracies and potential failure
- Is there a human in the loop?
  - Process to correct?



Mistaken ID: Facial-recognition tool falsely matches famous athletes to police mugshots

By Hiawatha Bray Globe Staff, October 21, 2019, 4:35 p.m.

💬 18

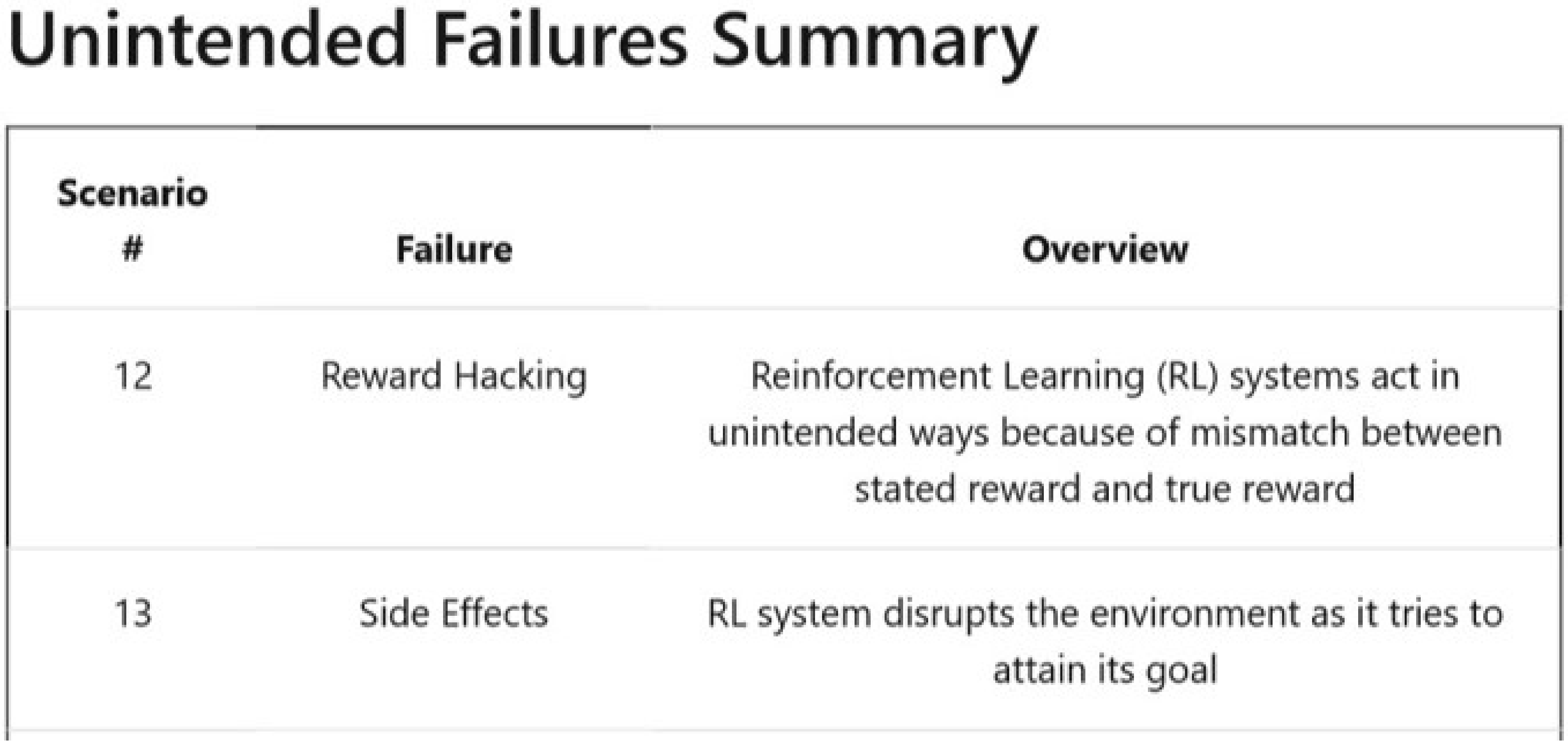
Facial-recognition software from Amazon mistakenly identified Duron Harmon and 26 other prominent New England athletes as possible outlaws, the Massachusetts chapter of the ACLU says. ELISE AMENDOLA/ASSOCIATED PRESS

Microsoft

INL Idaho National Laboratory

Question 3 – Is the AI Resilient to Attack?

- How is the AI/ML resilient to attack?



Unintended Failures Summary

Scenario #	Failure	Overview
12	Reward Hacking	Reinforcement Learning (RL) systems act in unintended ways because of mismatch between stated reward and true reward
13	Side Effects	RL system disrupts the environment as it tries to attain its goal

Images Source: https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning

Question 4 – What's the Real ROI?

- Is the vendor making specific claims?
  - Ex: Solution will reduce analyst hunt time!
    - Can they quantify by how much?
  - Better catch rate of malware!
    - How much faster? What kinds of malware?

- Vendor should have trial and customer data to back up those claims

Question 5 – References/POC Support?

- The best data science in the world isn't useful to a business if it's not solving for a business need
  - Your problem may not be a nail
  - Discuss existing user experience
  - Assess suitability of outcomes
- Will the vendor support a POC or Bake-off?
  - *Our ML finds fileless malware others can’t!*
  - Test that assertion in practice before signing the agreement.

Applying What You've Learned

- Next Week
  - Review these slides and do additional reading as needed
  - Share the 5 questions with your team and partners
- Three Months
  - Incorporate concepts from the 5 questions into RFPs
- Six Months
  - Institutionalize the 5 questions as part of the AI/ML procurement process
  - Or part of the AI/ML dev and build

Recommended Reading

- Homo Deus: A Brief History of Tomorrow, Yuval Noah Harari
- Applied Artificial Intelligence: A Handbook For Business Leaders, Mariya Yao, Adelyn Zhou, and Marlene Jia
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Aurélien Géron
- Introduction to Machine Learning with Python: A Guide for Data Scientists, Andreas C. Müller

RSA®Conference2020

Questions?

Thank you!