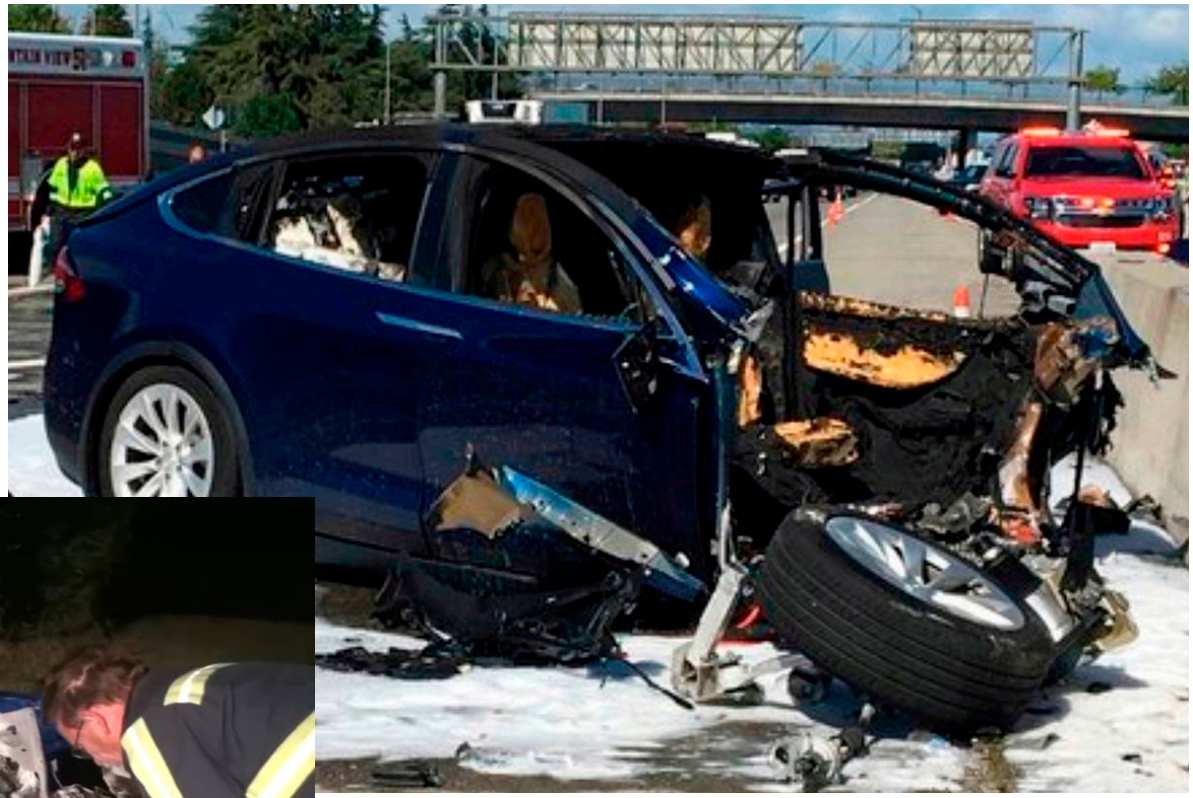
Bias and Trust in AI

Joshua A. Kroll
jkroll@nps.edu
CS 4000: Harnessing AI
8 September 2021

NAVAL
POSTGRADUATE
SCHOOL



Why are the self-driving cars crashing?

- Autonomous vehicle control systems aren't good enough for all real-world use cases yet.
  - Corner cases: the world is variable. Humans can handle this in ways the computers haven't learned to yet.
    - Humans can ***Generalize***, acting well in scenarios they haven't seen yet. Algorithms can't.
- To solve this, the autopilot "disengages" when it's not sure what to do
  - But is the human paying attention? Maybe not.
- Can we make the automation (AI) better?
  - Sure, but how do we know we're getting better? When is the automation ready?

Safety, accidents, & *automation bias*

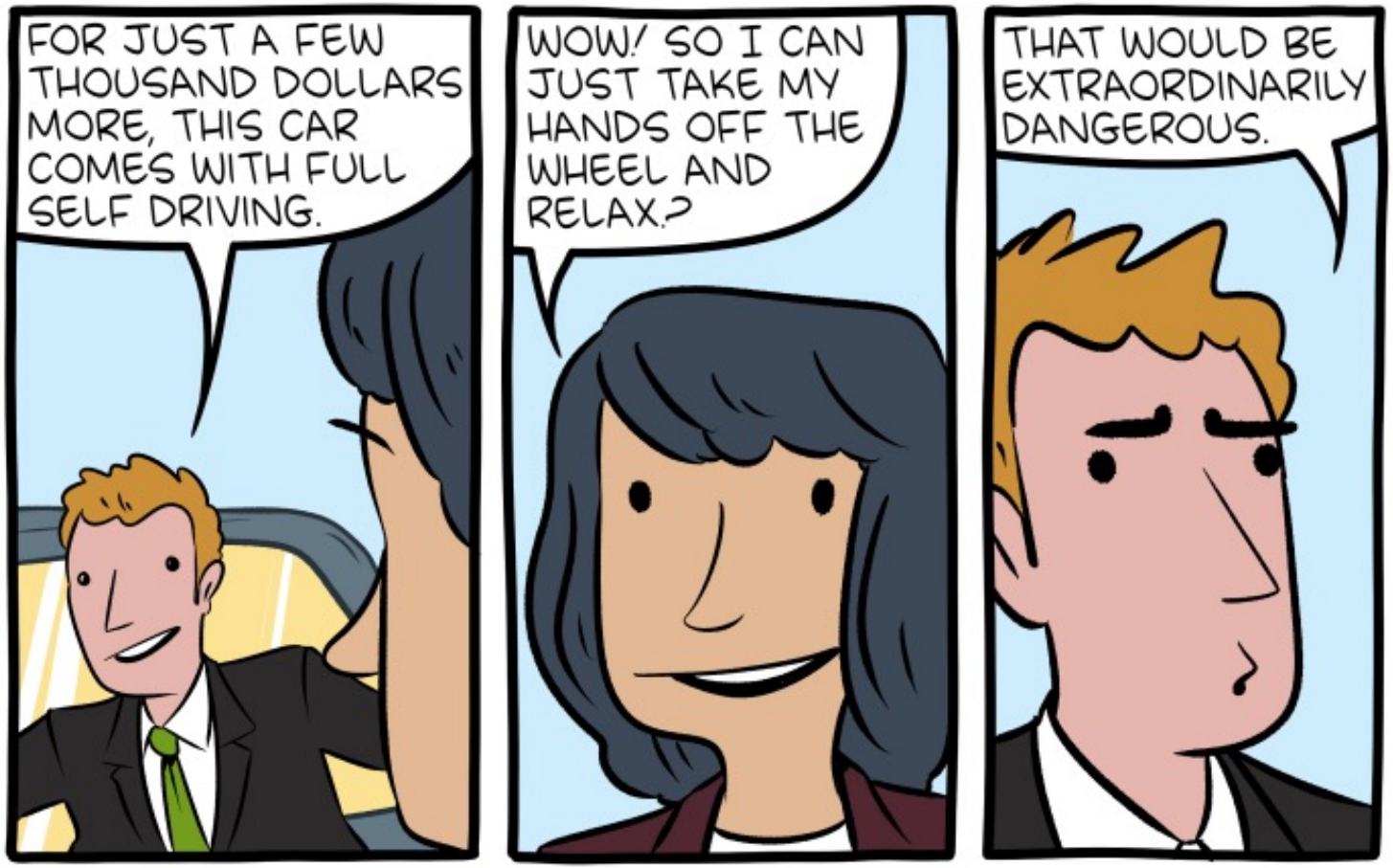
- Automation complacency and passive vigilance
- Mode confusion
- Automation ironies



Is this an AI-specific problem?

- Lots of automated systems suffer these accidents (e.g., your spelchkr)
- These accidents are deadly and affect even trained crews!
- Automation Complacency
  - Tendency to favor the output of machines/software over contradictory observations or intuitions, even when the machine is wrong.
- Mode Confusion
  - When a system responds differently to control based on its state, operators may put in “correct” inputs for the state they believe the system to be in, which are “incorrect” in the state the machine is actually in. That is, the human is confused about which state or “mode” the system is in.
  - Sometimes called “automation surprises” in the literature

Automation Complacency



Mode Confusion

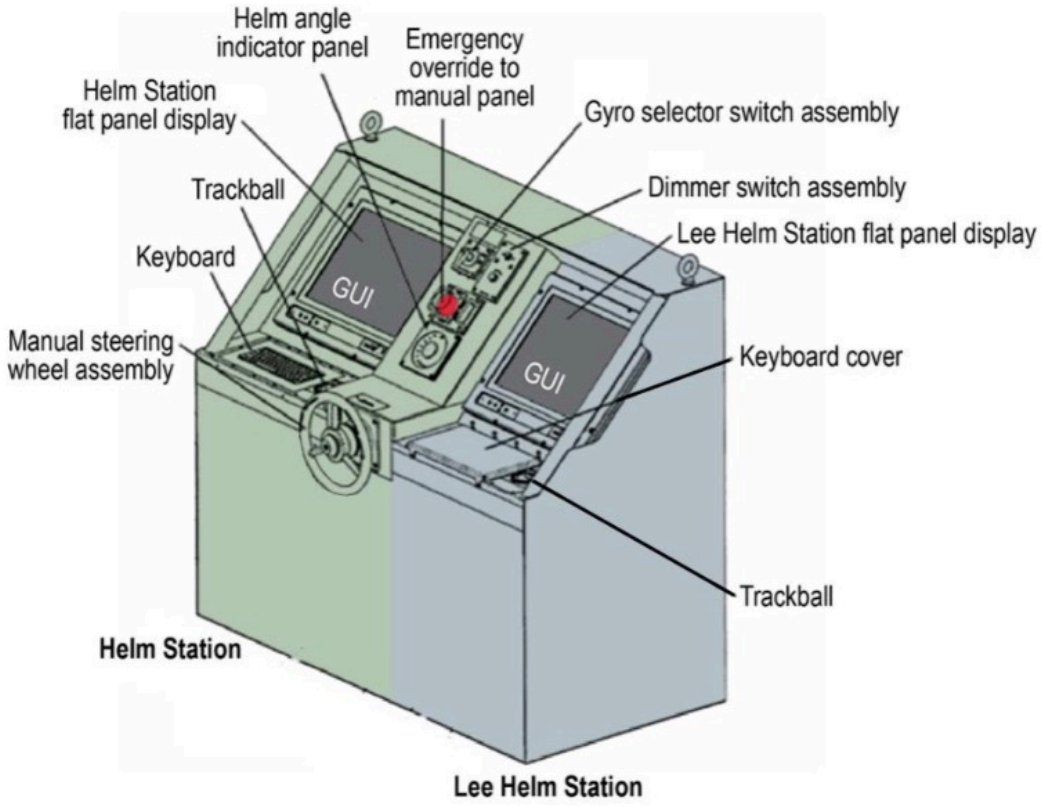


Figure 4. *John S McCain* SCC. (Drawing from IBNS technical manual; color added by NTSB)

Unsafe at any AUC

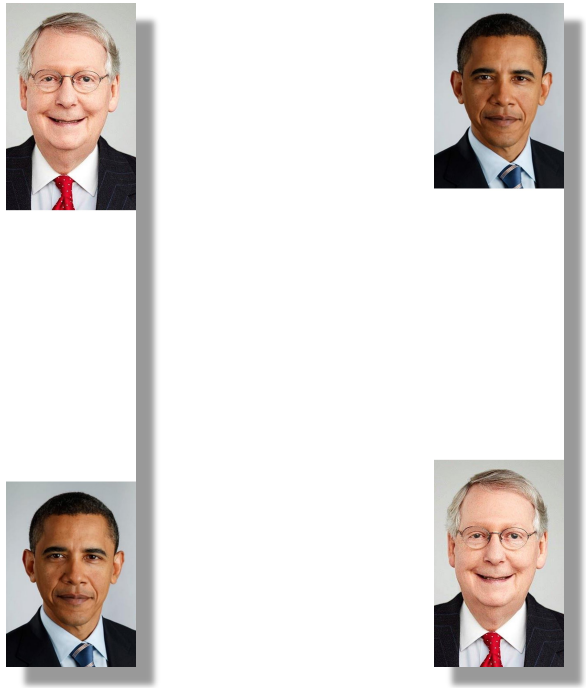
- The most common failure mode is misrepresenting the world
  - **Confounding and Spurious Correlation:**
    Identifying relationships in data which aren't real or which are controlled by something outside the data
  - **Failures of Measurement:**
    Misrepresenting the world through choices about gathering data or by making bad modeling decisions
- This is often described as "AI bias"
  - But is bias the right problem to address?
- Instead: focus on *harms* – allocative, representational, or dignitary

AI Harms

- **Allocative:** An algorithm mis-allocates a resource or punishment, reflecting existing social biases
- **Representational:** An algorithm fails to represent populations in an equitable way, particularly marginalized populations
- **Dignitary:** By using an algorithm to make a decision, human agency and recourse for bad/wrong decisions is lost

Discrimination and Allocative Harm

Bias and Discrimination: Representation

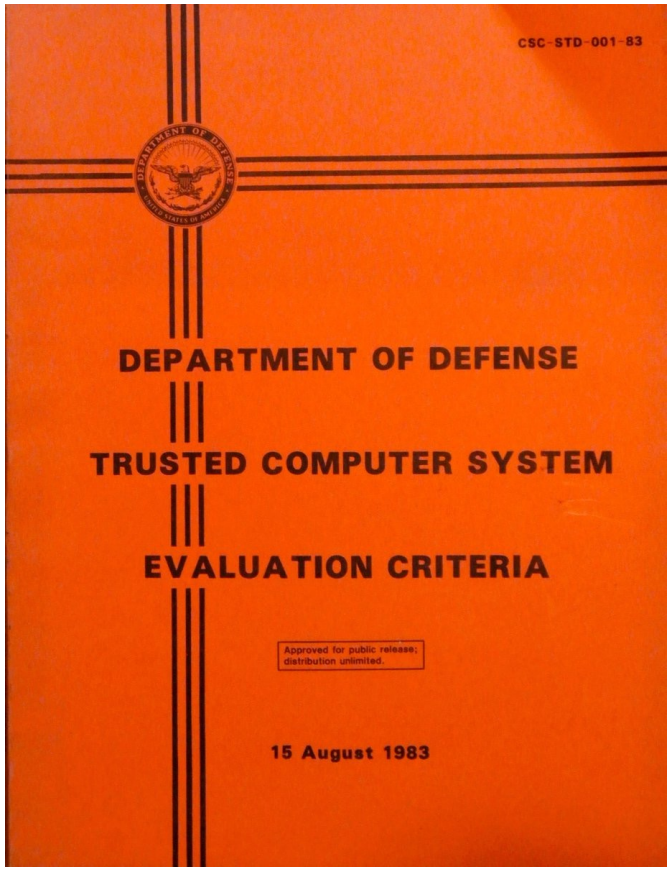


'A white mask worked better': why algorithms are not colour blind

When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing

Loss of Dignity/Autonomy

Trusted Computing Base

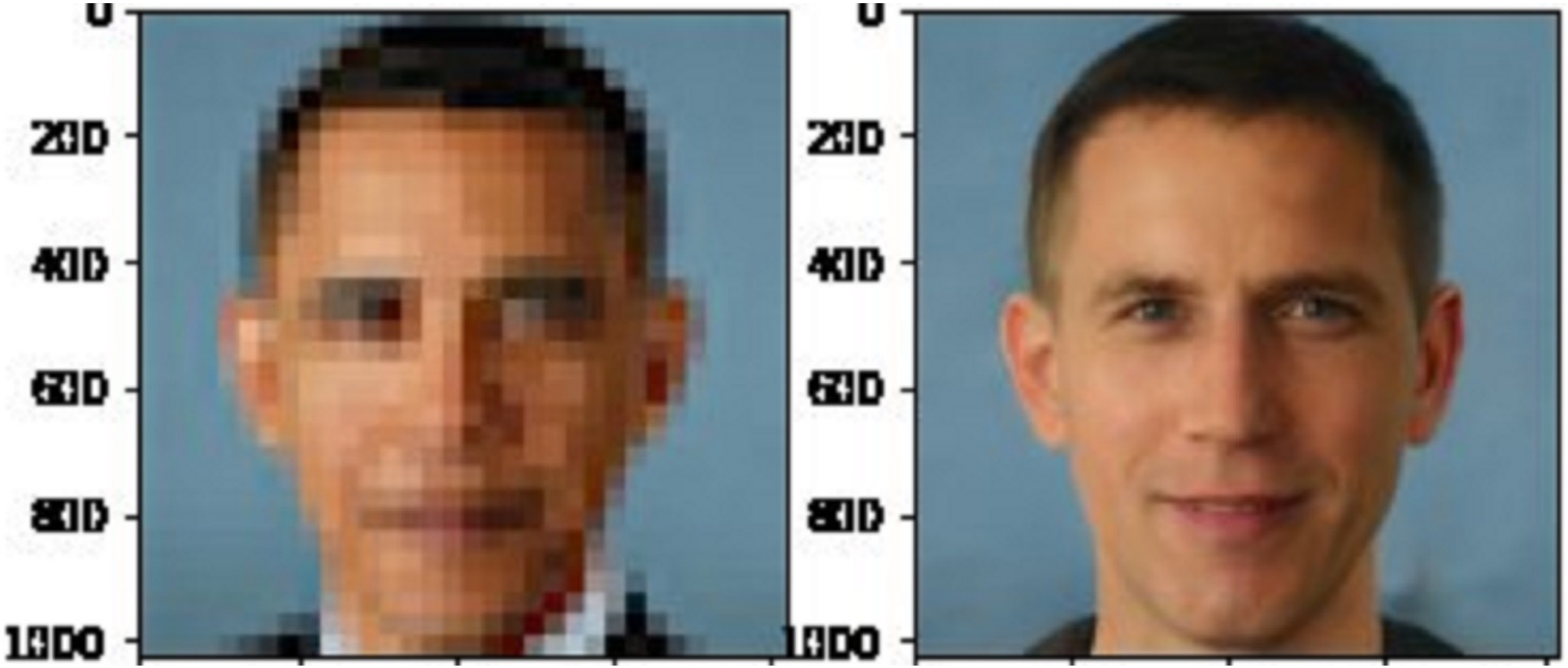


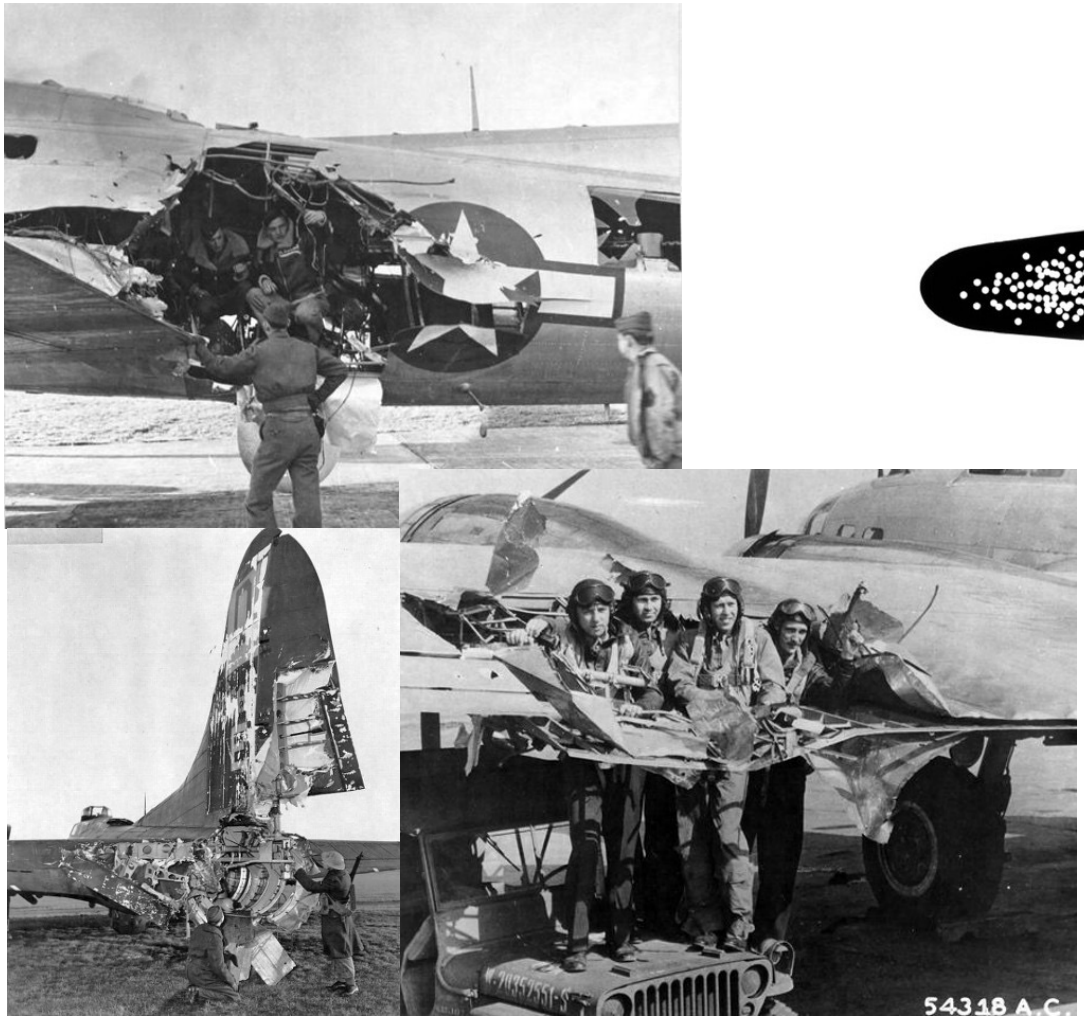
Trust and Trustworthiness

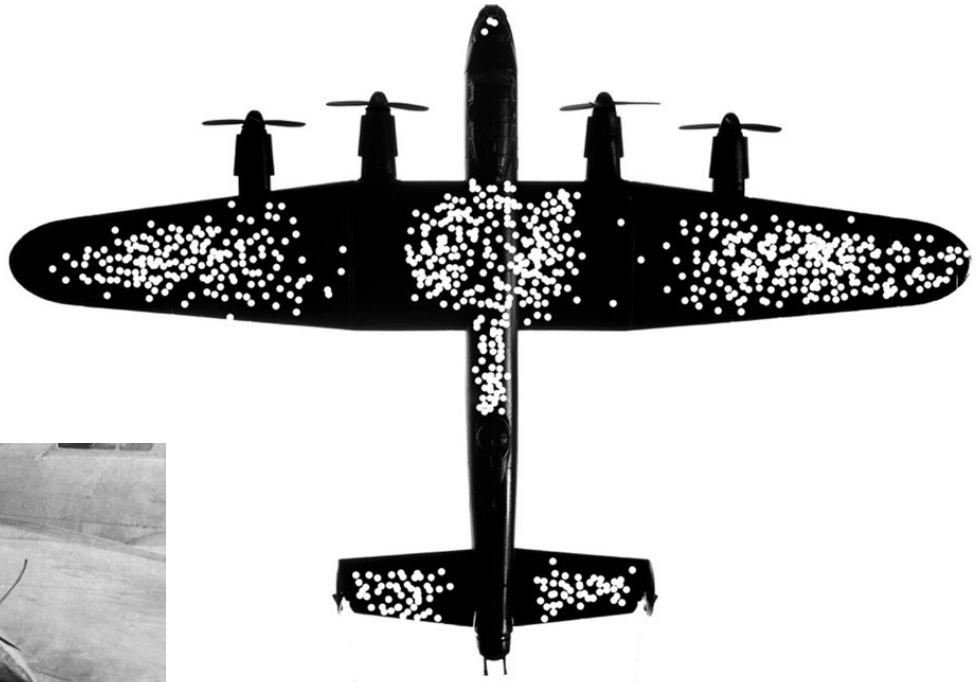
- In information security, **trusted** refers to components that can break key system invariants (i.e., can break the security policy)
- **Trustworthy** systems and components are those that *deserve to be trusted* (i.e., aren't likely to break, so we can rely on them safely)
  - Possibly, these are the systems we "have to" rely on: GCHQ defines trusted components as those "whose integrity cannot be assured by external observation of its behavior whilst in operation".

Statistical Biases

- **Statistical bias:** A *systematic error* or *deviation* from the "true" value
  - But what is the true value, and how do we know?
  - Independent from the notion of bias as a prejudice
- 2 important classes:
  - **Sampling Bias:**
    Collecting data (a sample) such that parts of the intended population have different probabilities of being included (harming generalization)
  - **Selection Bias:**
    Error introduced by selecting data in a non-representative way (harming the ability to make conclusions from the data)


0
200
400
600
800
1000
0
200
400
600
800
1000


54318 A.C.



Proxies

No Fairness Through Blindness

- Simply eliminating a sensitive attribute rarely works.
- Each feature may have a small correlation with sensitive status, even if it seems non-sensitive.
  - E.g, 59% of US users of `pinterest.com` are female.
  - With ~7.7 billion people, only need 33 bits of information to uniquely identify each person (less for groups).
- Often, we need the value of a sensitive attribute
  - E.g., medical diagnosis. Few women with testicular cancer. Drugs/treatments/diseases may affect different groups differently.
  - Forcing outcomes to be independent of sensitive status harms model validity.

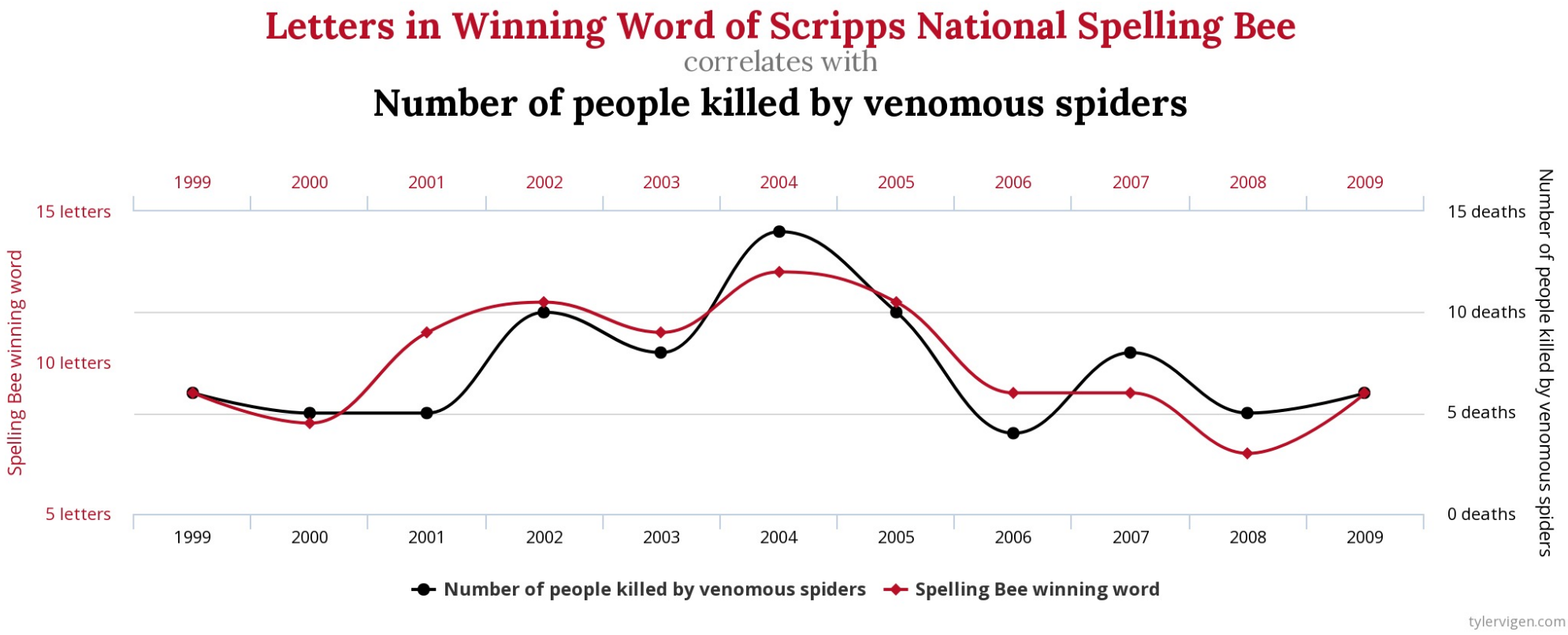
Does Accuracy Eliminate Bias?

- Suppose there's a rare cancer. It's always fatal, and it affects one out of every million people.
- I have a new blood test for this disease and I claim it's 99.9% accurate.
- You get the test and it says you don't have the cancer.
- *Should you believe me? How confident are you?*

Accuracy and Bias: Some math

- Suppose my test is completely bogus, and always comes back negative
  - If the cancer affects 1/1,000,000 people, it will only be wrong for those people
  - If I test 1,000,000 people and I get 1 wrong answer, I'm actually 99.9999% accurate!
- Thus, a test which gives no information about the disease is highly accurate! This is a problem for detecting rare phenomena

Model Evaluation



Confounding and the limits of data

- Data analysis can reveal relationships in data
  - This is amazing, and very useful
  - This is amazing, but very dangerous
- Relationships in data aren't always real
  - We need to understand the relationships outside the data and why they might or might not exist
  - Design experiments to differentiate real relationships from fake ones
  - Relationships can change over time → ***Concept Drift***

Can we recognize untrustworthy ML?

Article | Open Access | Published: 11 January 2021

Facial recognition technology can expose political orientation from naturalistic facial images

Michal Kosinski

Scientific Reports **11**, Article number: 100 (2021)

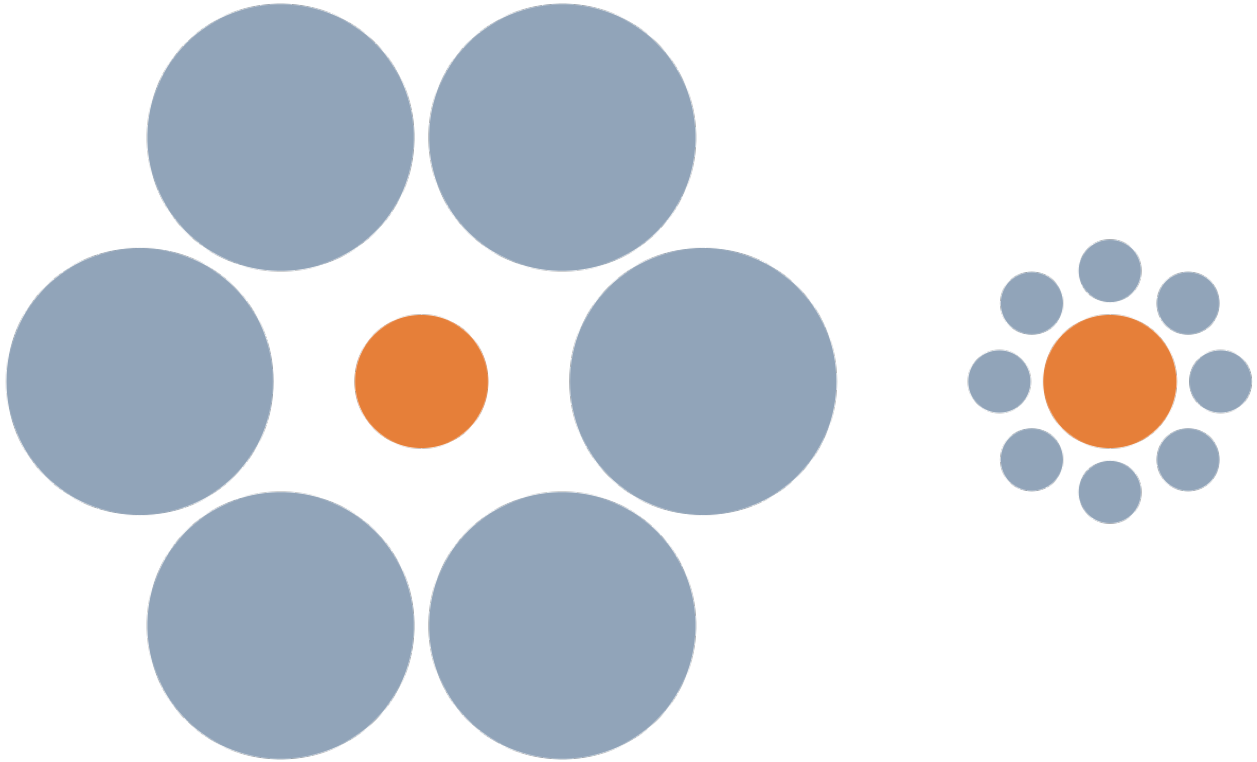
Is bias always bad?

Should we be de-biasing our data?

There’s no de-biasing AI

- Trustworthy AI is a *governance* problem
- We must interpret the the scope of a problem, the relationships/biases in data, the bounds within which a system works
  - Is a system safe always? When used by trained operators? For what purposes?
- *De-biasing* invents new relationships rather than representing the world as it is and acting within that framework

Human Cognitive Bias



The system includes people!

Bias and Trust are Human Problems

- Humans fail in predictable ways (e.g., *anchoring* is a bias that makes the two orange circles in the picture above appear different sizes when they're the same size)
  - Humans also have prejudices, and these are often reflected in data or in modeling decisions that support AI systems
  - That prejudice is expectable/predictable doesn't excuse it from an ethics/values perspective – we must act to counteract it
- Humans are always a part of our systems
  - We must account for ways humans behave sub-optimally and design our systems to manage this
  - AI systems must also convince the humans who interact with them (or who are affected by them without any control) that the whole system (including the humans) acts in a trustworthy way, upholding the system's goals

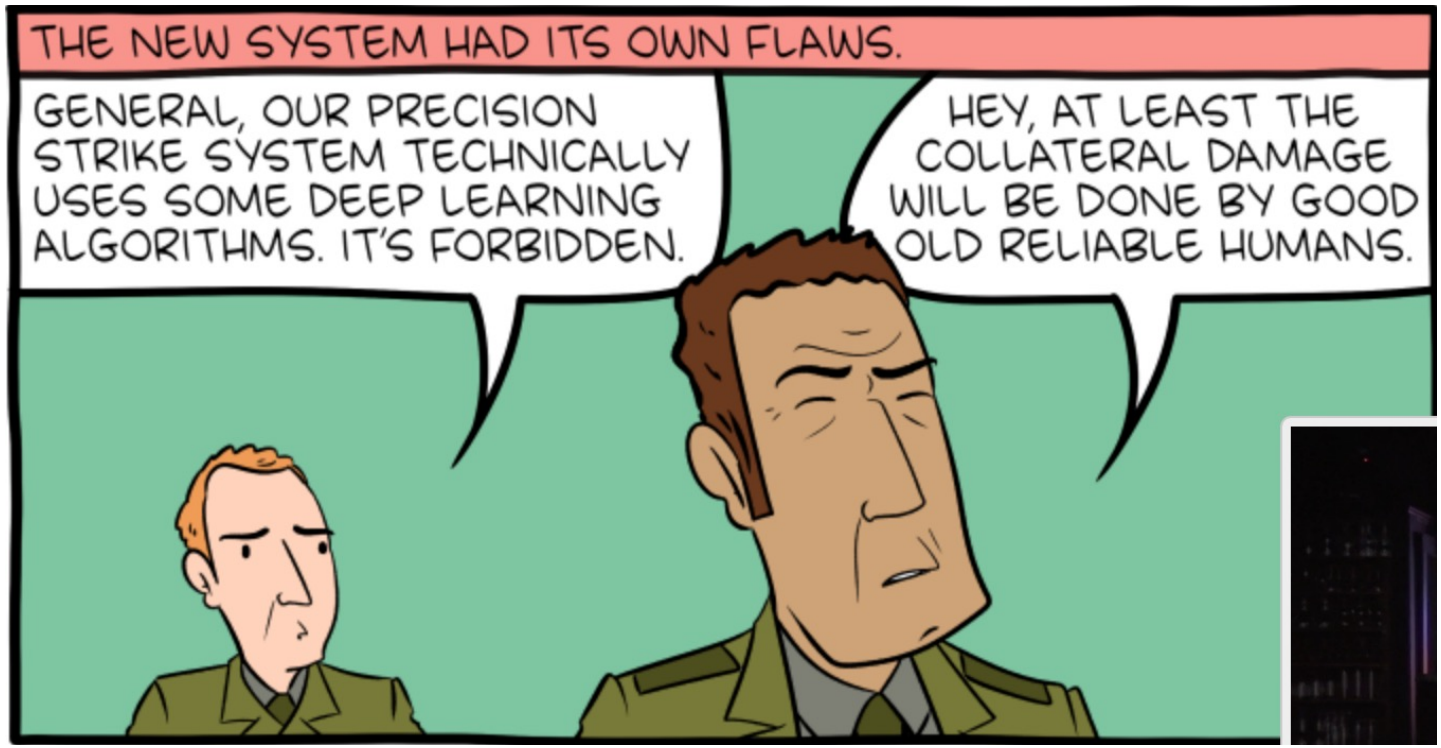

FALCONVIEW
DELL

AI Bias Issues in the DoD Enterprise

- Are we recruiting/hiring/retaining/promoting the right people?
  - Diversity makes us stronger and more capable
- Medical Applications
- Acquisitions
- Policy Fusion

DoD AI Ethics Principles (2020)

1. **Responsible**. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
2. **Equitable**. The Department will take deliberate steps to minimize unintended bias in AI capabilities.
3. **Traceable**. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.
4. **Reliable**. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
5. **Governable**. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.


THE NEW SYSTEM HAD ITS OWN FLAWS.
GENERAL, OUR PRECISION STRIKE SYSTEM TECHNICALLY USES SOME DEEP LEARNING ALGORITHMS. IT'S FORBIDDEN.
HEY, AT LEAST THE COLLATERAL DAMAGE WILL BE DONE BY GOOD OLD RELIABLE HUMANS.

Questions? Reach out!

Take CS4340 in Spring 2022!

jkroll@nps.edu
https://jkroll.com