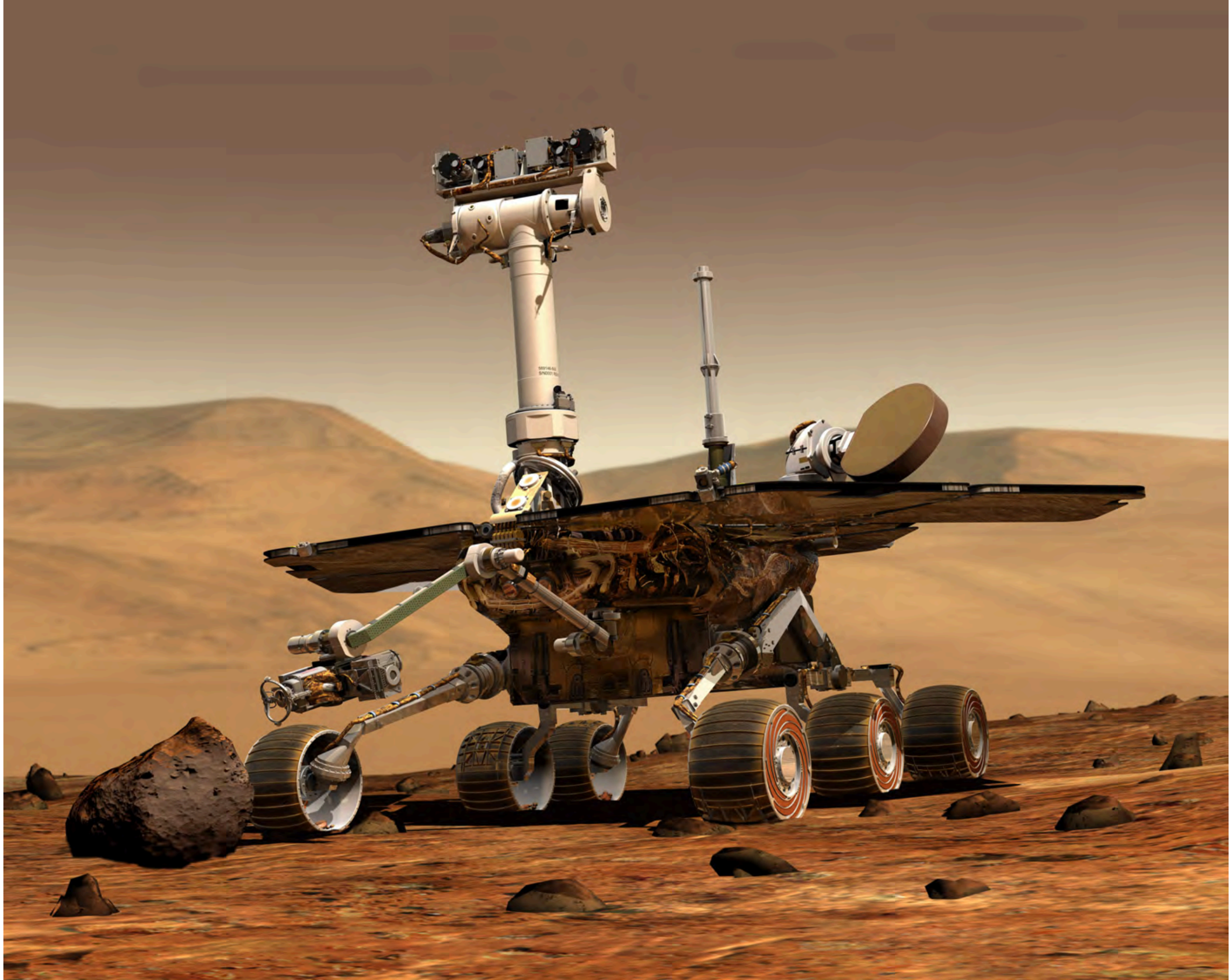
A Path towards Trustworthy and Responsible Artificial Intelligence

Joshua A. Kroll
Naval Postgraduate School

Online Seminar of the Fondazione Ugo Bordoni
24 September 2020

NAVAL
POSTGRADUATE
SCHOOL

©2009 ESA - CNES - ARIANESPACE / Photo Optique Vidéo CSG

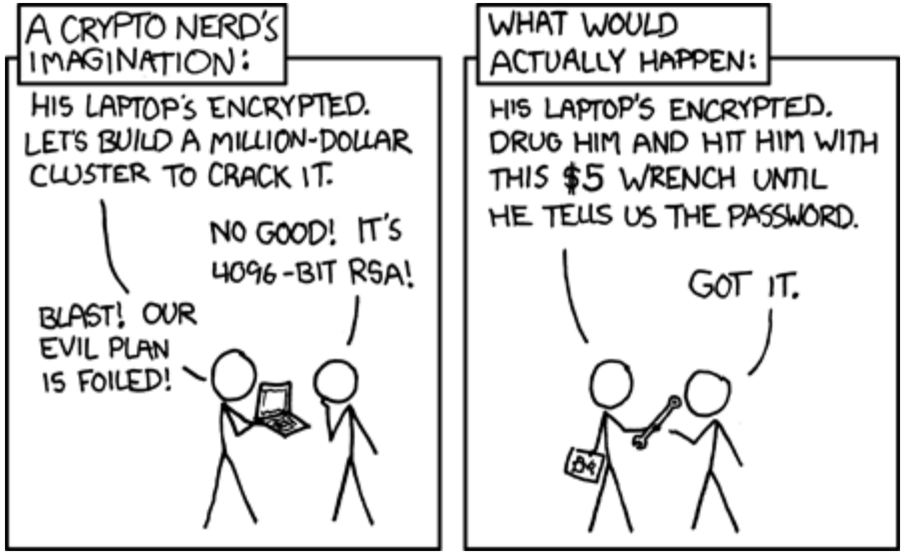


Reformulating the Problem

- Perhaps we modeled the wrong problem, or the wrong way, to achieve our goal and another solution would serve the goal better
  - For example, in criminal justice, studies have shown that sending people reminders or paying their transit fare both improve appearance rates in court

- Or perhaps we should solve a different problem or use no computer system at all.

Samir Passi and Solon Barocas, (2019). “Problem Formulation and Fairness.” ACM Conference on Fairness, Accountability, and Transparency.


A CRYPTO NERD'S IMAGINATION:
HIS LAPTOP'S ENCRYPTED. LET'S BUILD A MILLION-DOLLAR CLUSTER TO CRACK IT.
NO GOOD! IT'S 4096-BIT RSA!
BLAST! OUR EVIL PLAN IS FOILED!
WHAT WOULD ACTUALLY HAPPEN:
HIS LAPTOP'S ENCRYPTED. DRUG HIM AND HIT HIM WITH THIS $5 WRENCH UNTIL HE TELLS US THE PASSWORD.
GOT IT.

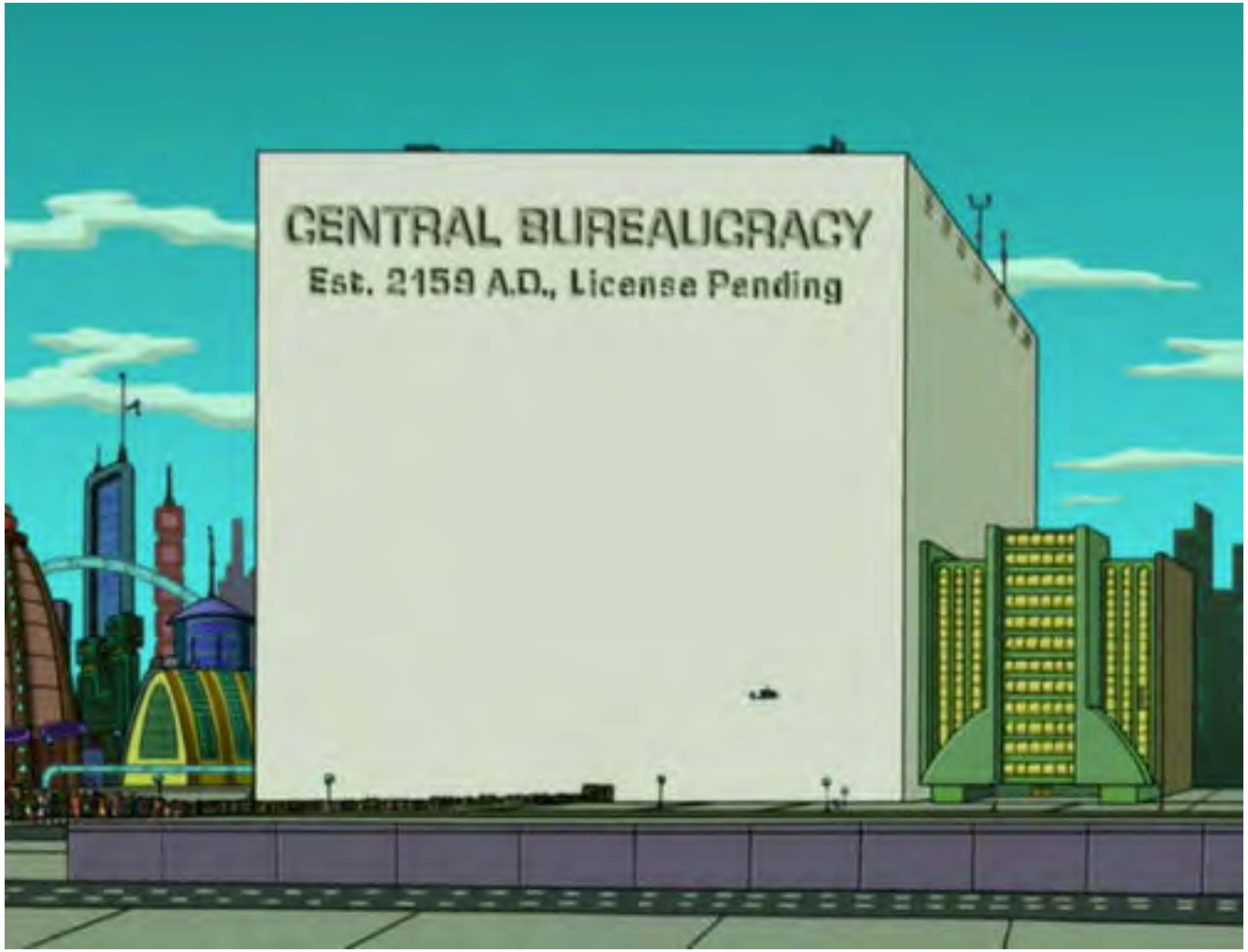
Audit and Assessment

- Goal is to establish system *behavior* and understand *outcomes*
- An **audit** is an examination of evidence to ensure it is materially correct
- An **assessment** measures something or calculates a value for it
- Structured review of evidence about a system
    - Black box vs. white box
    - Static vs. Dynamic

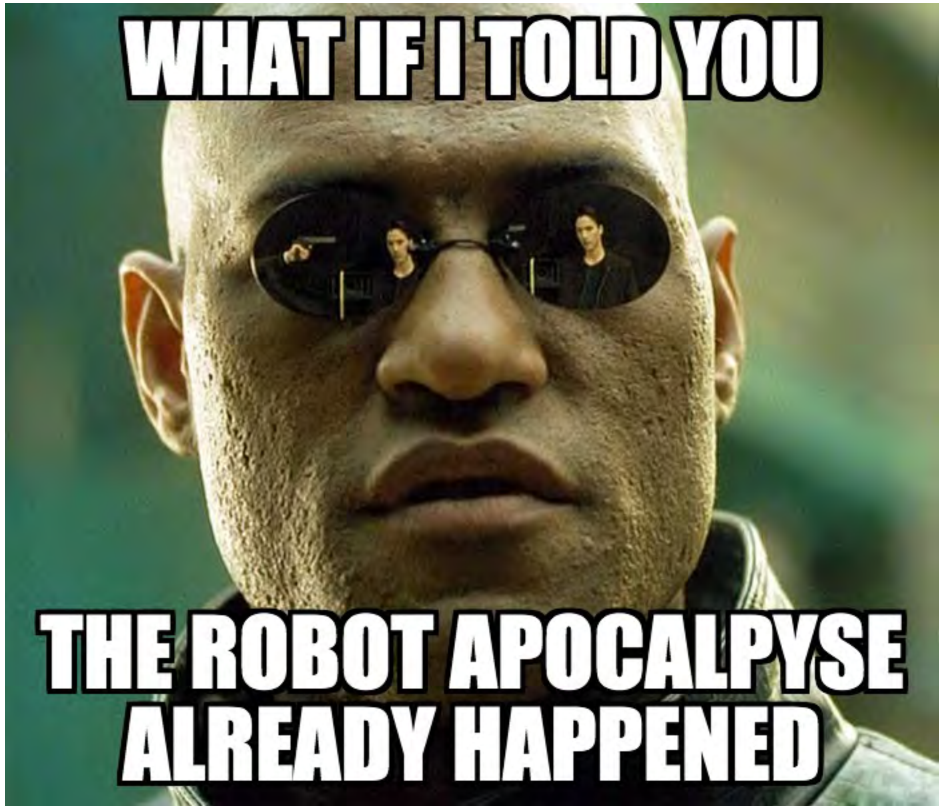
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

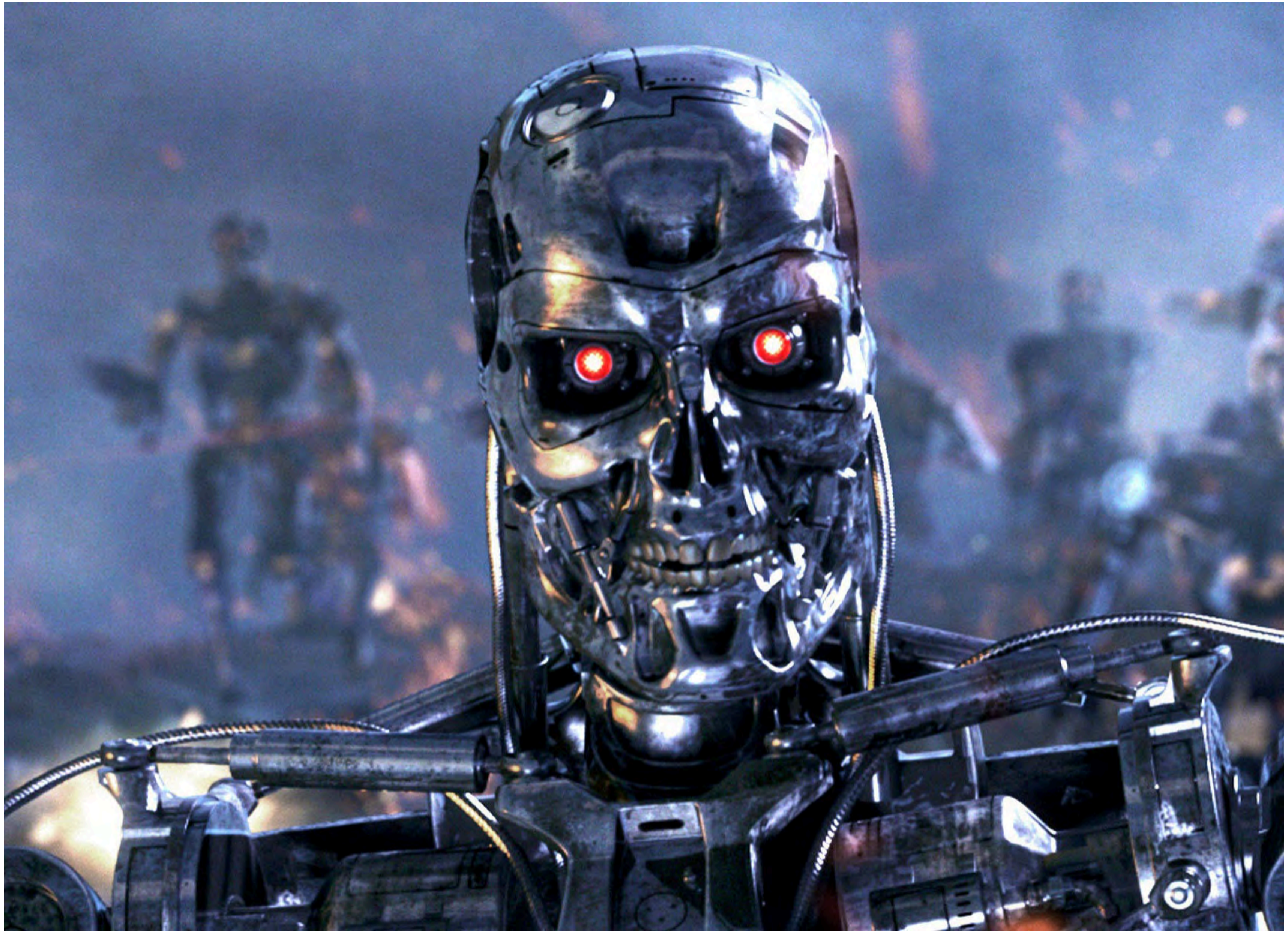
Impact Assessments

- A special kind of assessment, developed according to a formal, structured process, for assessing the economic, social, and environmental impacts of public policy or the deployment of new technologies before they are implemented
    - Impact assessments for automation
- Process
    - Impact analysis
    - Consultation of stakeholders
- Publication?
    - Requirements to disclose vs. incentives


CENTRAL BUREAUCRACY
Est. 2159 A.D., License Pending

WOPR
War Operation Plan Response


WHAT IF I TOLD YOU
THE ROBOT APOCALPYSE ALREADY HAPPENED

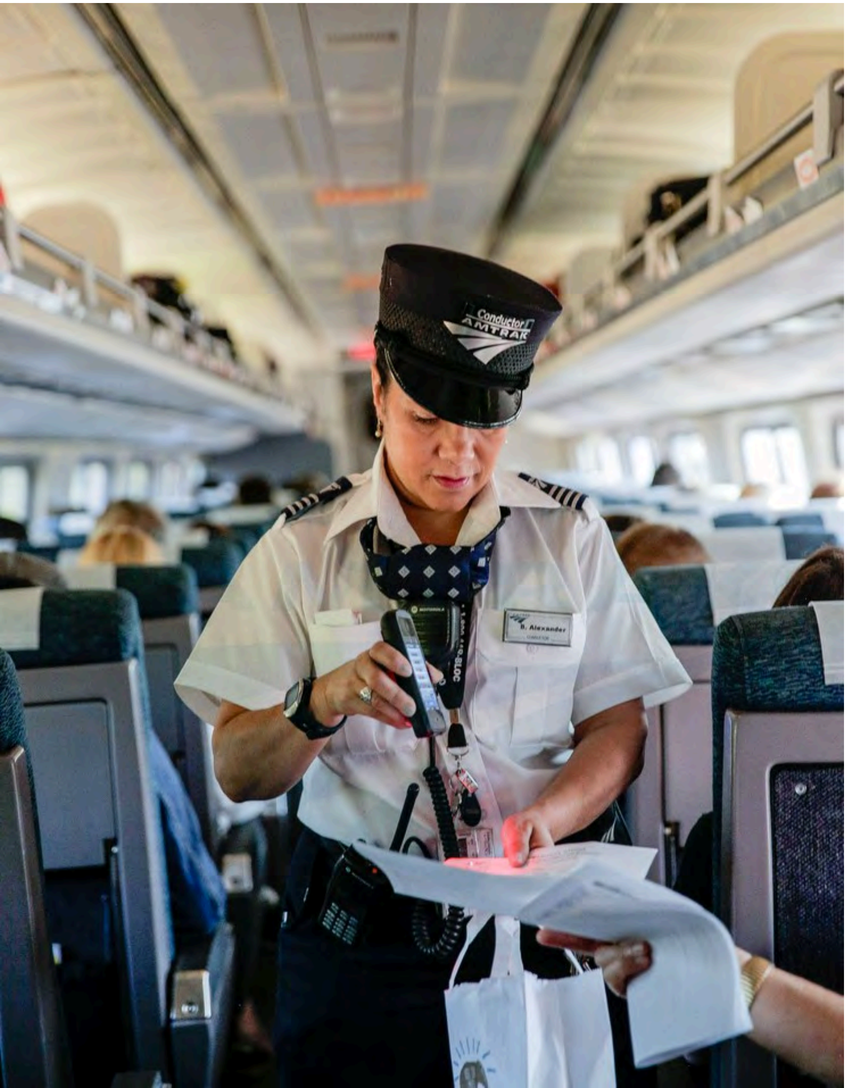


PSA
Hapag-Lloyd
HYUNDAI
CMA CGM
EVERGREEN
MOL
K LINE

What is automation?

- Operation of a process according to a ***set of established rules***
  - Generally, the rules are referred to as a **specification**
    - Explicit specifications
    - Implicit specifications
    - Can be simple or highly complex
  - Can be implemented in a variety of ways, both human and machine
    - In software (software controllers, AI)
    - Mechanically (looms, autopilots)
    - Via groups of people (bureaucracies, “computers”)
- Often, automation works in tandem with humans
  - Very few scenarios are truly “fully automated”

Problem Concepts and Automation



→

Humans are also a factor

- AI will, in general, look more like autopilot than Commander Data
- **Automation Paradox**: Automation makes humans more important to the supervision of systems (because fewer humans are responsible for more output) yet leaves them less experienced and prepared to manage those systems
  - A risk is that humans become **moral crumple zones**, present only to be blamed for the system’s failures.

Elish, Madeleine Clare. "Moral crumple zones: Cautionary tales in human-robot interaction." *Engaging Science, Technology, and Society* 5 (2019): 40-60.

Accountability

- Broadly desirable, hard to pin down
- Unit of Analysis
- Accountability is a *relationship*: some entity is accountable to another entity
- Allows problem re-formulation:
    - Entity held accountable can act freely subject to guess about oversight
        - Still allows for rules
    - Entity holding accountable can do so based on facts of situation, without need for general rule

Joshua Kroll (2020). "Accountability in Computer Systems" In *Oxford Handbook of the Ethics of AI*, Frank Pasquale, Markus Dubber, and Sunit Das, Eds. Oxford University Press.

Accountability in Computer Systems

- Levels of Abstraction
  - System Components: Accountability as ***Accounting***

Joshua Kroll (2020). “Accountability in Computer Systems” In *Oxford Handbook of the Ethics of AI*, Frank Pasquale, Markus Dubber, and Sunit Das, Eds. Oxford University Press.

Accountability in Computer Systems

- Levels of Abstraction
  - System Components: Accountability as ***Accounting***
  - System Control and Ownership: Accountability as ***Responsibility***

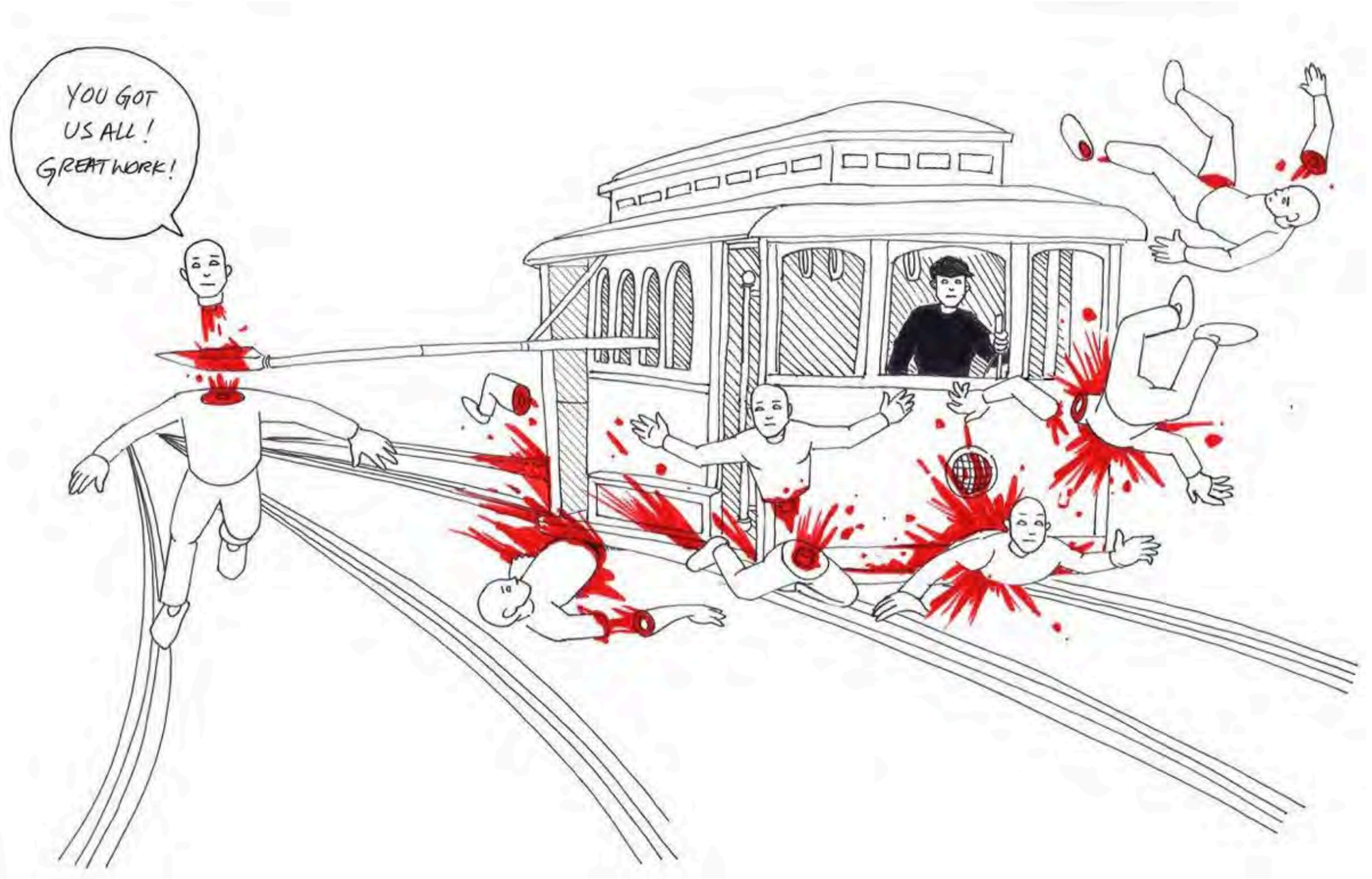
Joshua A. Kroll "Accountability in Computer Systems" In *Oxford Handbook of the Ethics of AI*. Markus D. Dubber, Frank Pasquale, and Sunit Das, Eds. Oxford University Press. 2020.

Accountability in Computer Systems

- Levels of Abstraction
  - System Components: Accountability as ***Accounting***
  - System Control and Ownership: Accountability as ***Responsibility***
  - The System in Context: Accountability as ***Fidelity to Norms and Values***

Joshua A. Kroll "Accountability in Computer Systems" In *Oxford Handbook of the Ethics of AI*. Markus D. Dubber, Frank Pasquale, and Sunit Das, Eds. Oxford University Press. 2020.

But What of Ethics?



Accountability

- *Accounting* for a system: Keeping records to support outcomes
  - Supports other requirements, like reproducibility
- *Responsibility*: Tying actions to consequences
  - People, not machines
- *Respect for Values & Norms*: Matches political, social, legal expectations
  - Possibly counterproductive to specify! Desirably unspecified/underspecified.
  - Often managed through *oversight* and *review* rather than specification

- Values are ideals. Accountability is achievable.

What does that mean in practice?

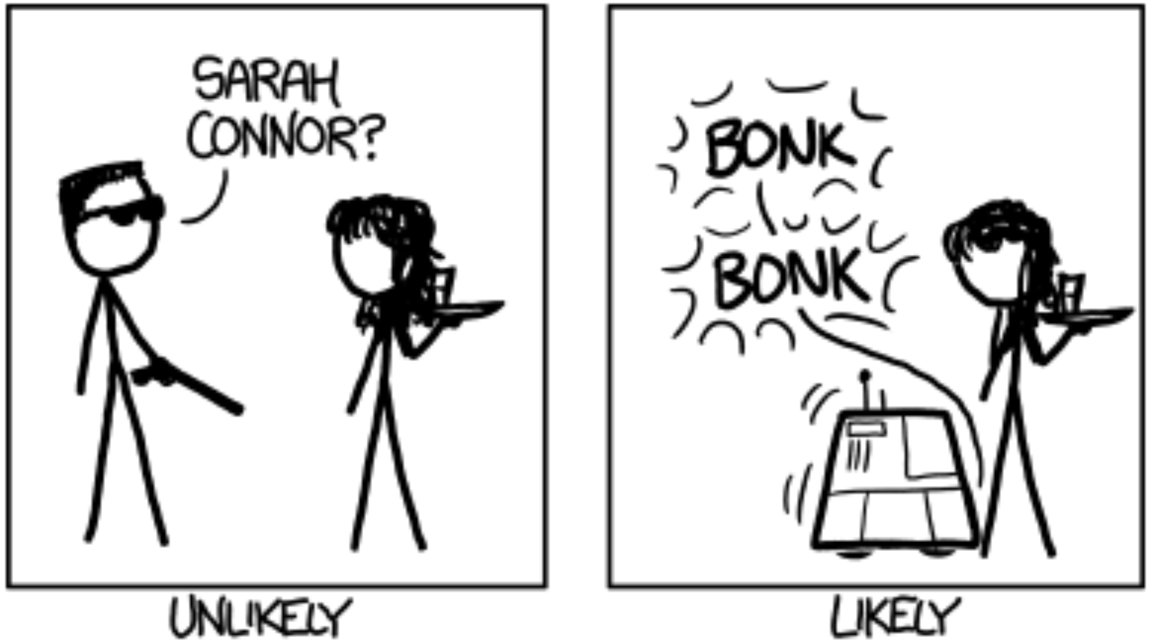
- Solve the right problem in the right way
  - Embrace interpretive space: Don’t specify what can’t be specified
  - Leave for humans that which is human
  - Measure twice, predict once
- Produce evidence where it is needed and to whom it is needed
  - Development: specs, requirements, tests, acceptance criteria, UX, …
  - Operation: inputs/outputs, decisions, logs, …
  - Post-deployment: performance measures, monitoring, abuse/debugging, …
- Use advanced tools when they help you, avoid them when they don’t
  - Cryptography, verification, abstraction, …

How do we build trustworthy *systems*?

- Assess entire systems, not isolated components

- Human factors and the automation paradox
  - Build systems so they can be operated effectively and integrated into organizations well
  - Design systems so they support human decision-makers. Verify that this is so.

- Governance, Oversight, Assessment, and Audit
  - Bureaucracies don’t work without oversight and governance. Why should AI?
  - Accountability as a design goal
    - Bridge from “accounting for” a system to the values the system embodies

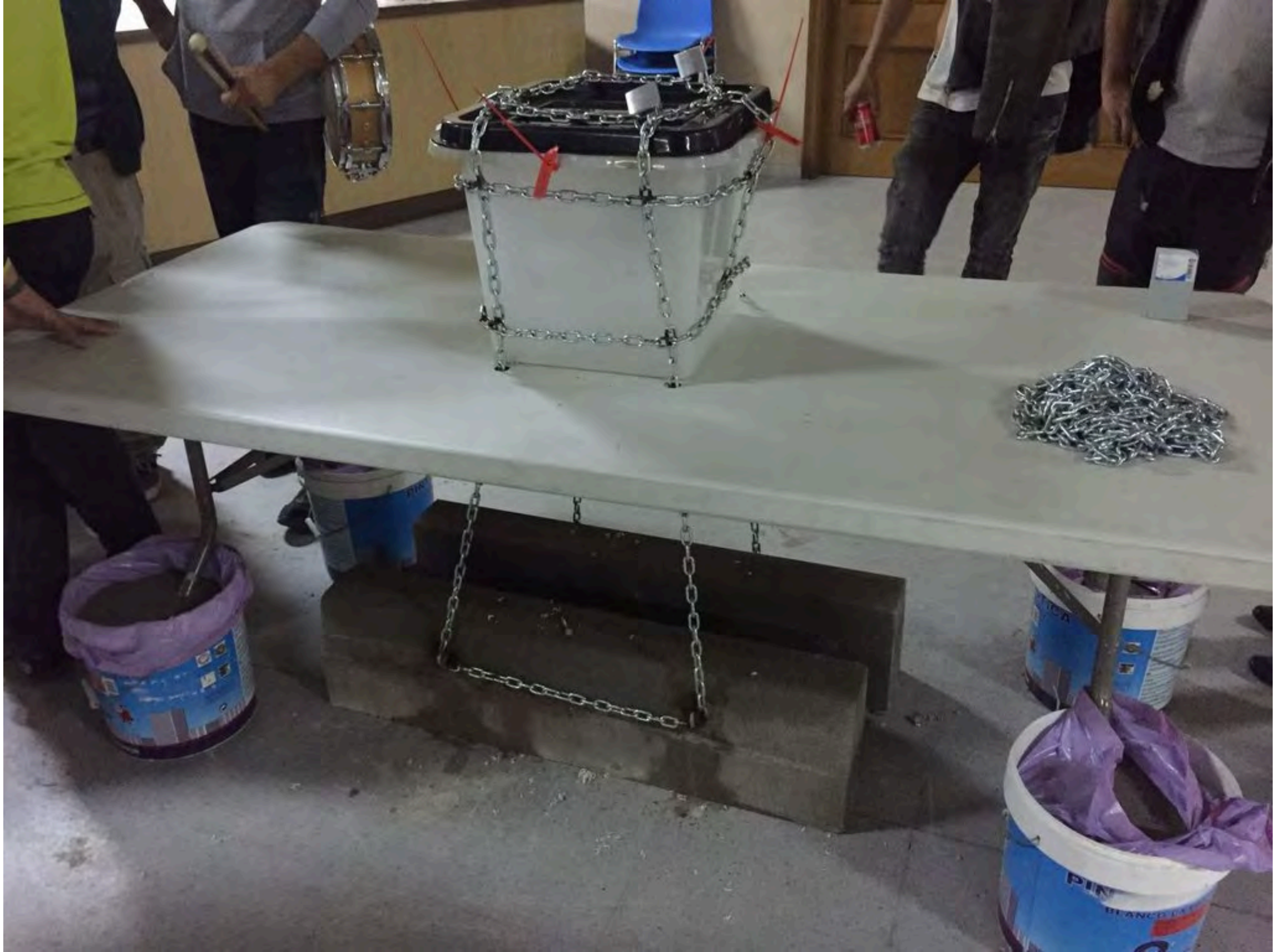


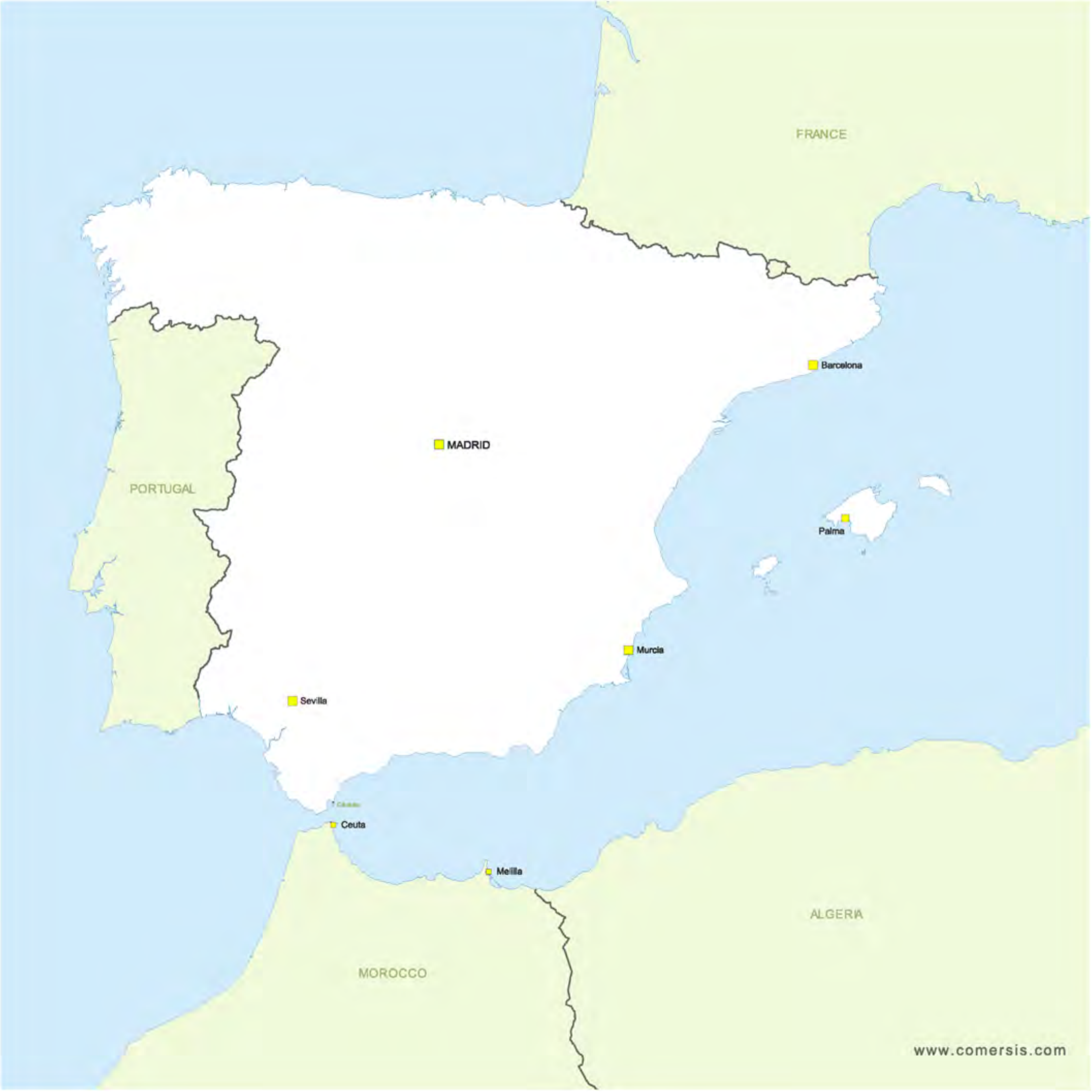
Questions? Reach out!
jkroll@nps.edu
https://jkroll.com

32
02
44
48
47

Rules vs. Standards

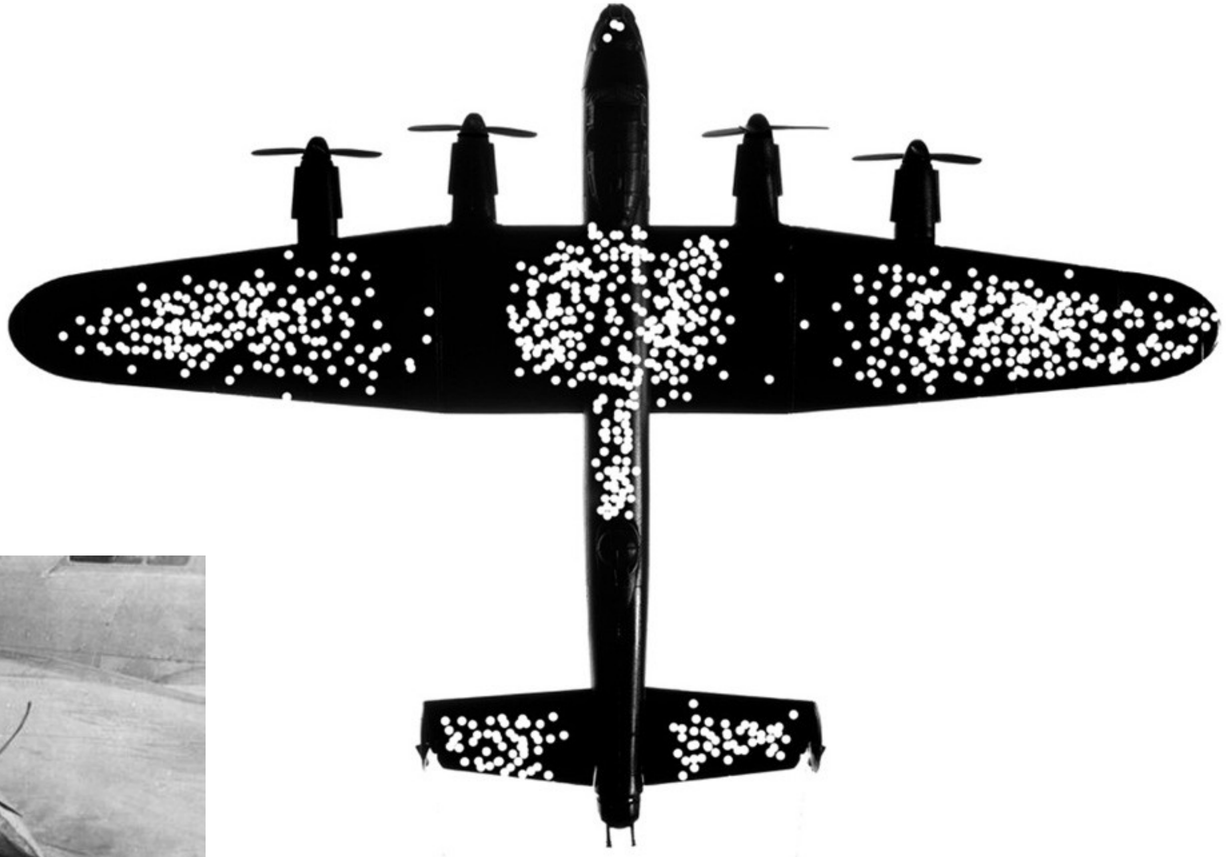
- **Rules** require no interpretation, so can be objectively applied both quickly and in a large number of cases
  - Consistency/Reproducibility/Predictability of outcomes
  - Once rule is set and facts determined, outcome is also determined
  - E.g., “You must be at least 35 years old to become president.”
- **Standards** are more flexible, allowing for some discretion in application
  - Flexibility to balance countervailing concerns or mitigating circumstances
  - Require a decision-maker to apply
  - E.g., “You must be sufficiently mature to become president.”
- **Principles** are open-ended but mandatory considerations
  - E.g., “No one should profit from their own misdeeds.”
  - E.g., “No citizen is above the law.”





www.comersis.com

W-2035551-S
54318 A.C.



“On Exactitude in Science”

Jorge Luis Borges, *Collected Fictions*, translated by Andrew Hurley.

…In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.

© Harry Marmot 2014
withinstrikingdistance.wordpress.com
타는 곳
Tracks 乘車
인천국제공항
Incheon Int'l Airport
仁川國際空港
고객안내센터
A'REX