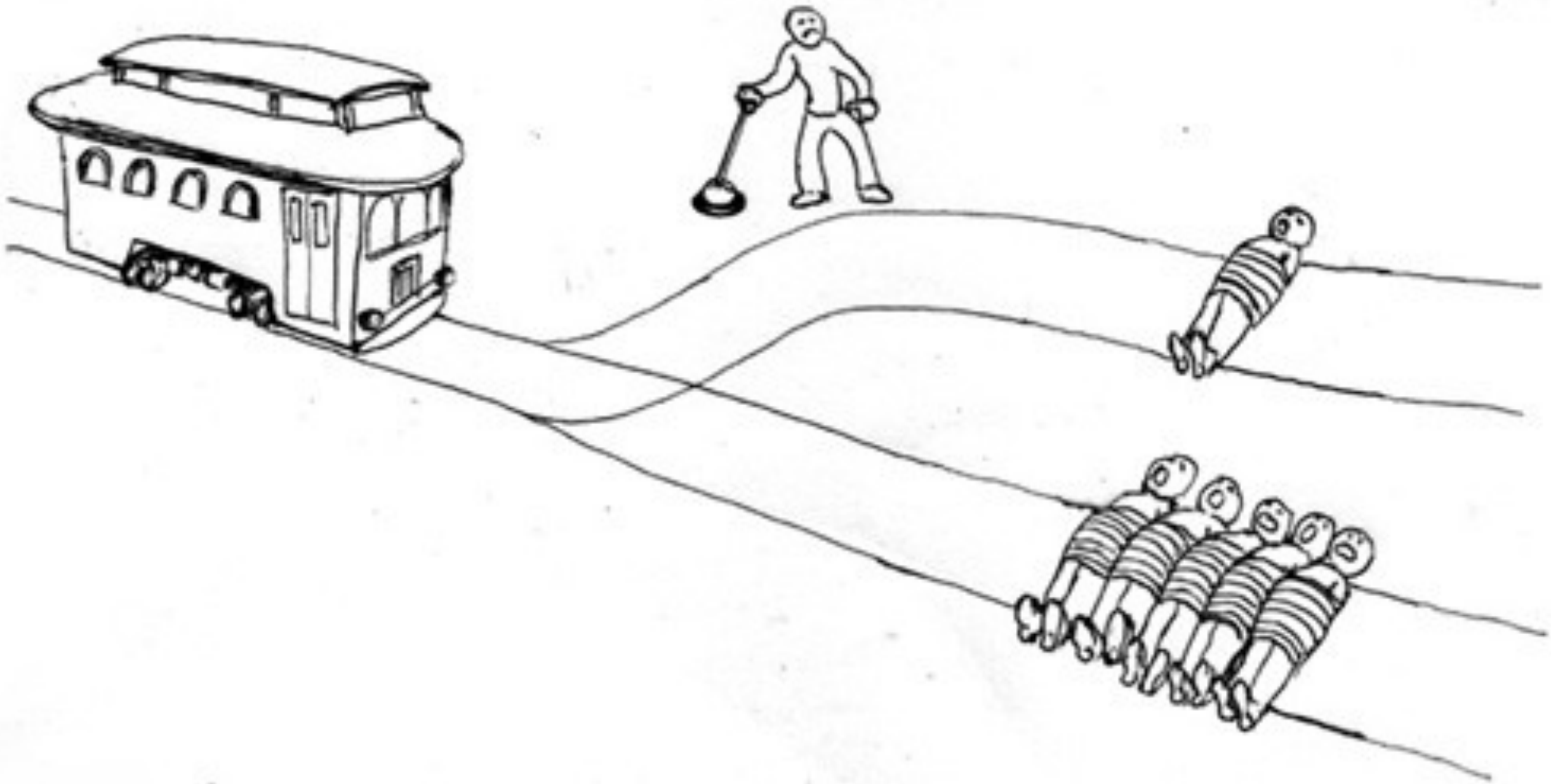
DIGITAL INNOVATION
COLLEGE OF EUROPE CHAIR

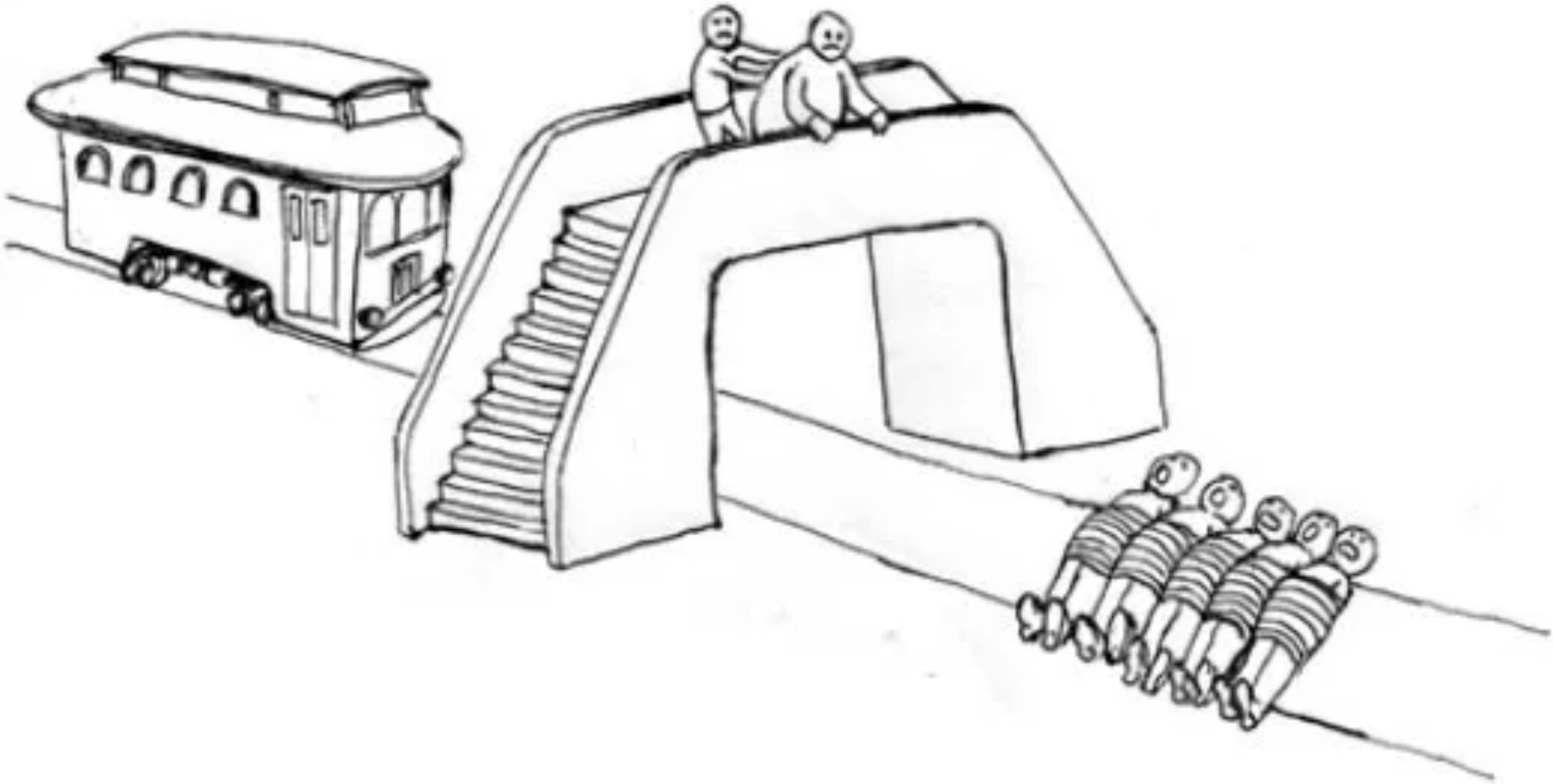
The trolley problem and Self-Driving Cars
A CSI into the Ethics of Algorithms

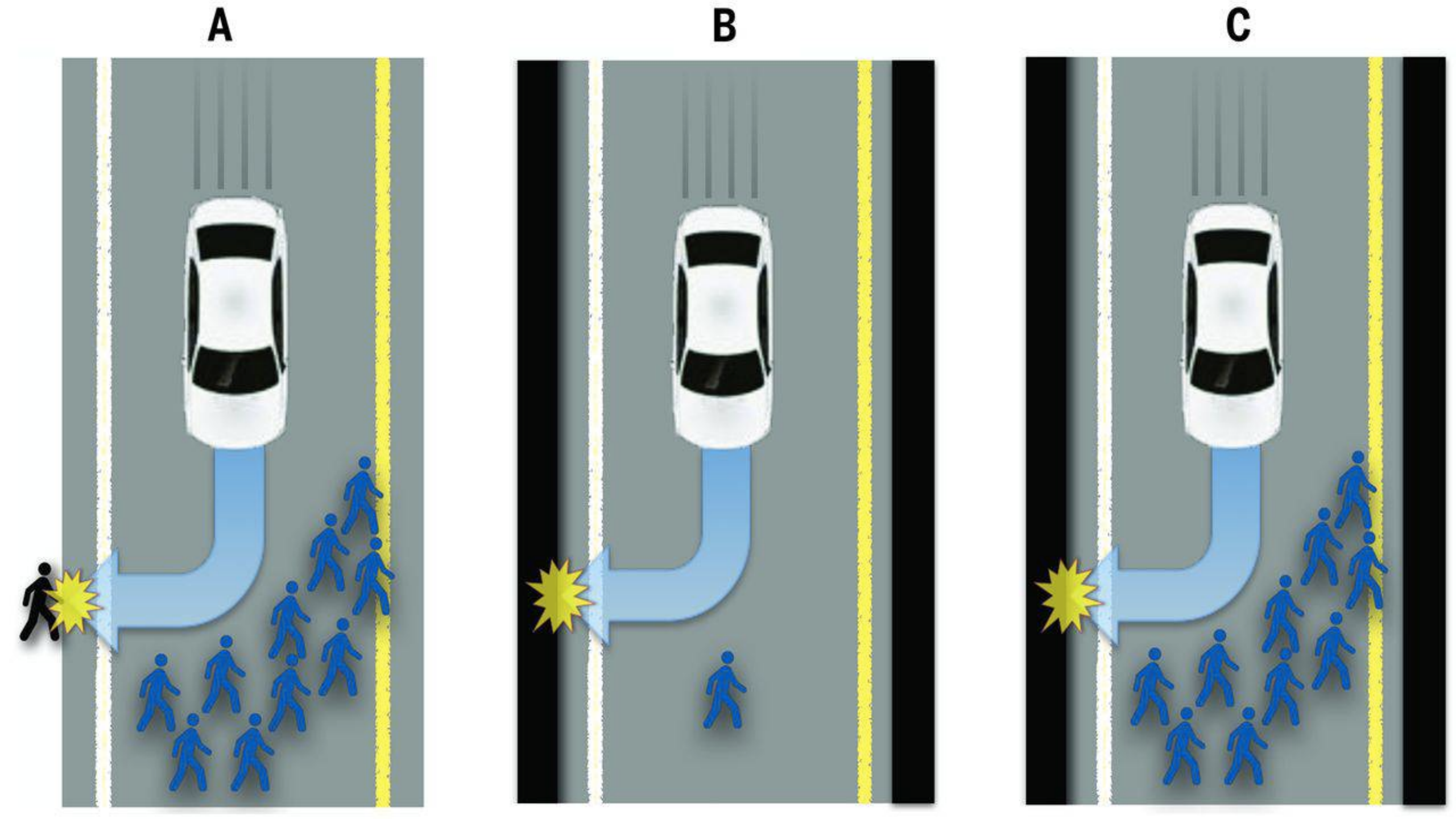
Andrea Renda

Chair for Digital Innovation, College of Europe

17 January 2018






A
B
C

Mercedes's Self-Driving Cars Will Kill Pedestrians Over Drivers

Mike Brown | October 14, 2016 | Autonomous Cars

Three principles (German draft code)

- Cars must opt for property damage over personal injury
- A car never distinguishes between humans based on categories such as age or race
- If a human removes his or her hands from the steering wheel - to check email, say - the car's manufacturer is liable if there is a collision

The real answer is “mu”: un-ask the question

Erase and rewind: six clues

CSI:
CRIME SCENE INVESTIGATION

1. How did the car end up there?
2. What did the car know?
3. What do we know about how the car decided?
4. Did we expect the car to behave like us, or better?
5. Who is liable?

Problem 1

How did the car end up there?

A policy choice, not an inevitable fact

- ***Carless* cities, or *driverless* cars?**
  - **Several alternatives to cars in city centers (carpods, light trains, etc.)**
  - **Several cities taking action to remove cars from city centers**
- **Will dedicated infrastructure for automated cars and pedestrians avoid the dilemma?**
  - **High speed lanes first, then all highways**
  - **Pedestrian passages, bridges etc.**


BRIDGE

We cannot think statically and one-dimensionally about technological evolution

We can make choices when it comes to human-machine interaction

Problem 2

What did the car know?

A battle over the data architecture

- **Vehicle-to-vehicle (V2V) v. 5G-enabled Vehicle-to-Environment (V2E), v. Offline**
  - **Most likely, it will be a combination of wireless, fixed-line, satellite, sensor-generated information: but some companies keep cars offline for security**
- **Cars on blockchain? (Toyota, Porsche, Daimler)**
- **What did the car know about the individuals involved?**
- **What did it know about the expected behavior of other cars?**

US 20170363430A1

(19) **United States**

(12) **Patent Application Publication**

Al-Dahle et al.

(10) **Pub. No.: US 2017/0363430 A1**

(43) **Pub. Date: Dec. 21, 2017**

(54) **AUTONOMOUS NAVIGATION SYSTEM**

(71) Applicant: **Apple Inc.**, Cupertino, CA (US)

(72) Inventors: **Ahmad Al-Dahle**, San Jose, CA (US); **Matthew E. Last**, San Jose, CA (US); **Philip J. Sieh**, San Jose, CA (US); **Benjamin Lyon**, Saratoga, CA (US)

(73) Assignee: **Apple Inc.**, Cupertino, CA (US)

(21) Appl. No.: **15/531,354**

(22) PCT Filed: **Dec. 4, 2015**

(86) PCT No.: **PCT/US2015/064059**

§ 371 (c)(1),
(2) Date: **May 26, 2017**

Related U.S. Application Data

(60) Provisional application No. 62/088,428, filed on Dec. 5, 2014.

(52) **U.S. Cl.**

CPC ***G01C 21/32*** (2013.01); ***G05D 1/0276*** (2013.01); ***G01C 21/3415*** (2013.01); ***G05D 1/0212*** (2013.01); *G05D 2201/0213* (2013.01)

(57) **ABSTRACT**

Some embodiments provide an autonomous navigation system which enables autonomous navigation of a vehicle along one or more portions of a driving route based on monitoring, at the vehicle, various features of the route as the vehicle is manually navigated along the route to develop a characterization of the route. The characterization is progressively updated with repeated manual navigations along the route, and autonomous navigation of the route is enabled when a confidence indicator of the characterization meets a threshold indication. Characterizations can be updated in response to the vehicle encountering changes in the route and can include a set of driving rules associated with the

Problem 3

What do we know about how the car decided?

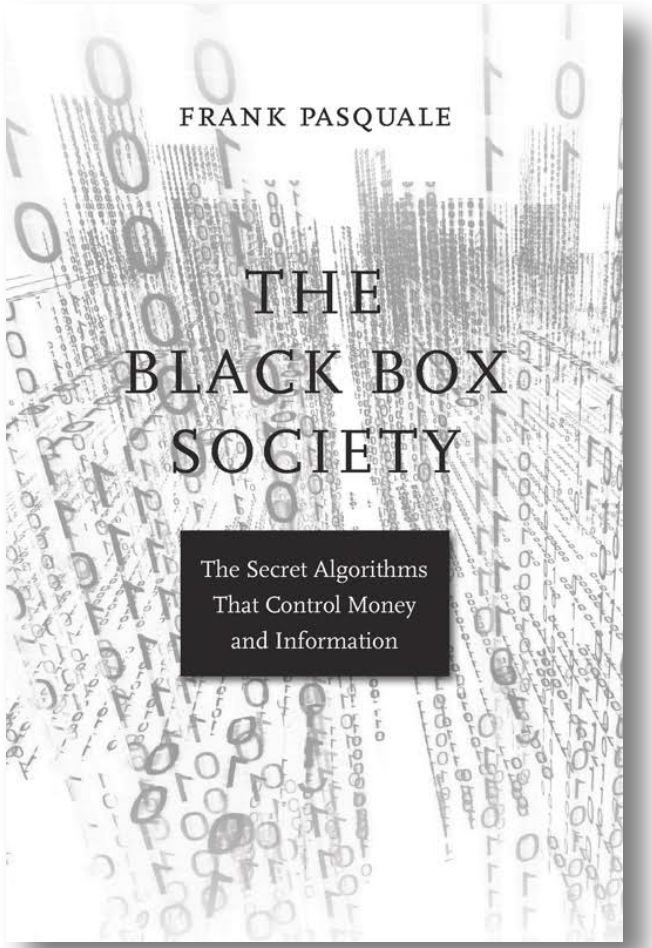
On transparency of algorithms and data

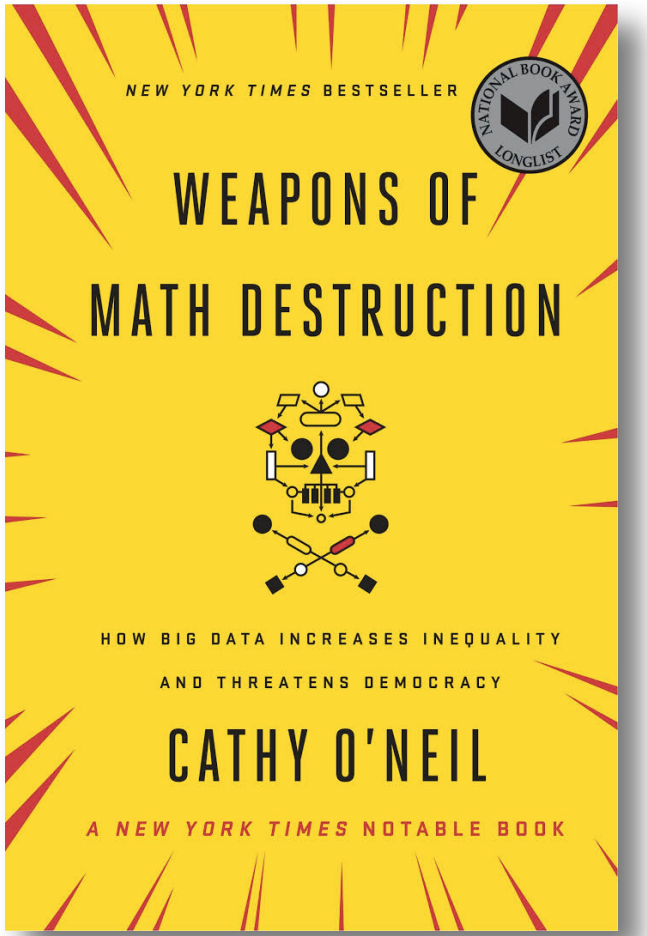
- **Should users and car owners (if any) have the right to understand how the algorithm takes decisions?**
- **In case a specific set of principles is agreed upon, should public authorities be able to check compliance by inspecting algorithms?**
- **Should insurance companies be enabled to audit and inspect algorithms to set their premiums?**
- **Will there be (blockchain-enabled) black boxes that allow us to understand what happened in more detail?**

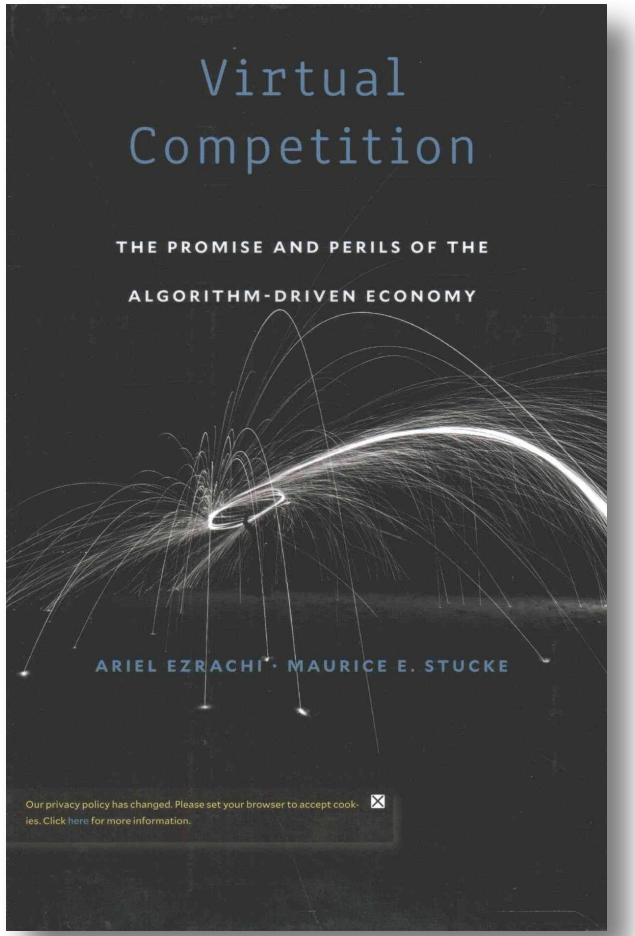
FUNCTION	TYPE	EXAMPLES
PRIORITISATION: associating rank with emphasis on particular information or results at the expense of others through a set of pre-defined criteria	*General search engines*	Google, Bing, Baidu
	Special search engines	Genealogy, image search, Shutterstock
	Meta search engines	Info.com
	Questions & answers	Quora, Ask.com
	Social media timelines	Facebook, Twitter
CLASSIFICATION: grouping information based on features identified within the source data	*Reputation systems*	Ebay, Uber, Airbnb
	News scoring	Reddit, Digg
	Credit scoring	Credit Karma
	Social scoring	Klout
ASSOCIATION: determining relationships between particular entities via semantic and connotative abilities	*Predictive policing*	PredPol,
	Predicting developments and trends	ScoreAhit, Music Xray, Google Flu Trends
FILTERING: including and/or excluding information as a result of a set of criteria	*Spam filter*	Norton
	Child protection filter	Net Nanny
	Recommender systems	Spotify, Netflix
	News aggregators	Facebook News Feed

Source: World Wide Web Foundation (2017)

Demonizing algorithms?







What algorithm?

- **An enigma**
  - **Adaptive, self-learning algorithms are seen as very advanced, but also not very transparent and unpredictable**
  - **Clustering and pattern recognition are less efficient as they are too data-hungry, and hardly recognize moving images**
  - **Replicating human behaviour?**

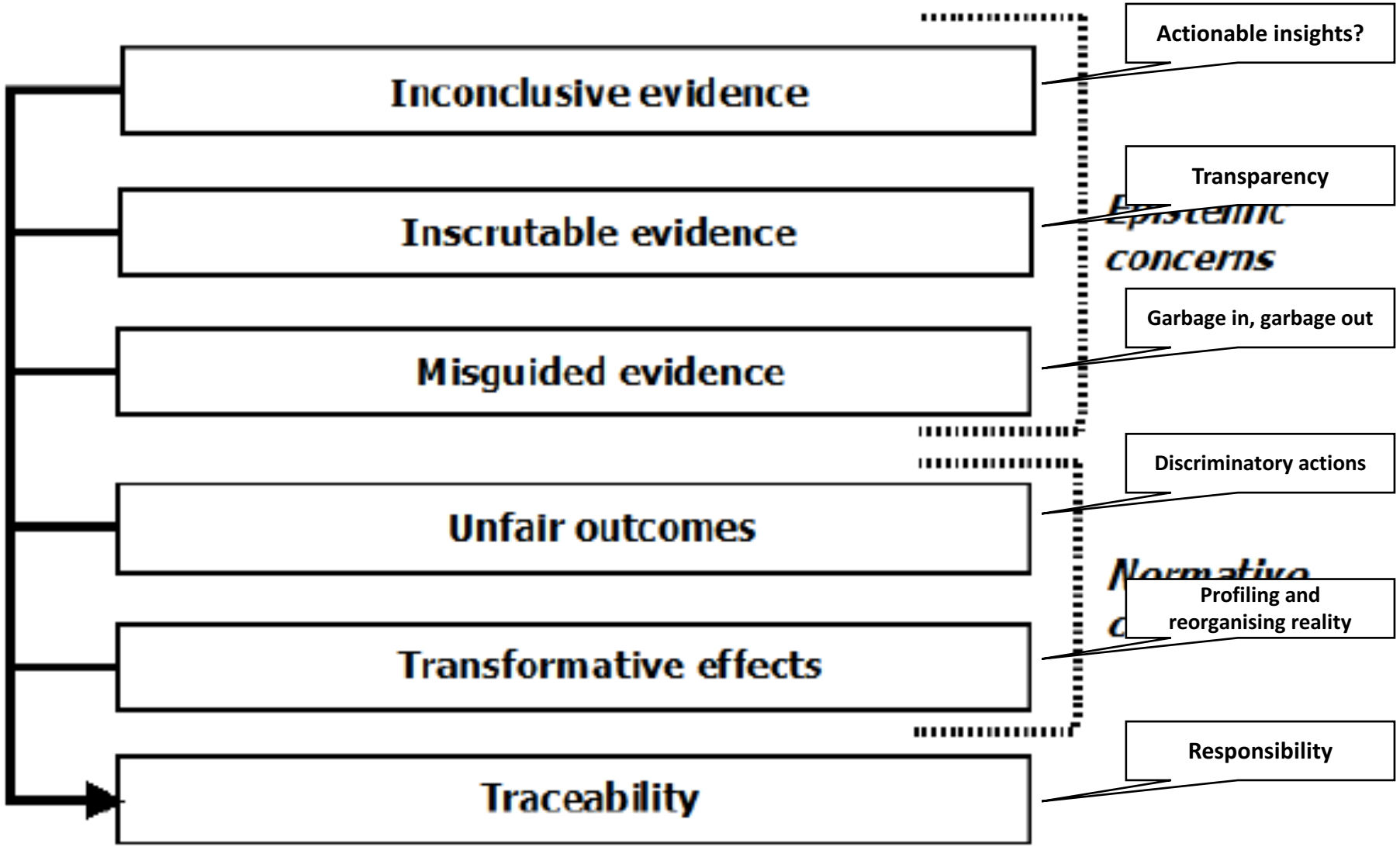


Figure 1 - Six types of ethical concerns raised by algorithms

Source: Mittelstadt et al (2016)

PRINCIPLE	DESCRIPTION
FAIRNESS	"Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics"
EXPLAINABILITY	"Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms."
AUDITABILITY	"Enable interested third parties to probe, understand, and review the behaviour of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use."
RESPONSIBILITY	"Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues."
ACCURACY	"Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures."

Source: FATML

Problem 4

Should algorithms behave like us, or better?

An emerging trade-off

- Our society is already biased and far from equal
- Possibilities: CBA v lexicographic ordering?
- Emerging efficiency/privacy trade-off
  - Algorithms cannot be neutral
  - The more they discriminate, the more they are efficient
  - Would people trade off privacy in exchange for accuracy?

The Washington Post

The Intersect

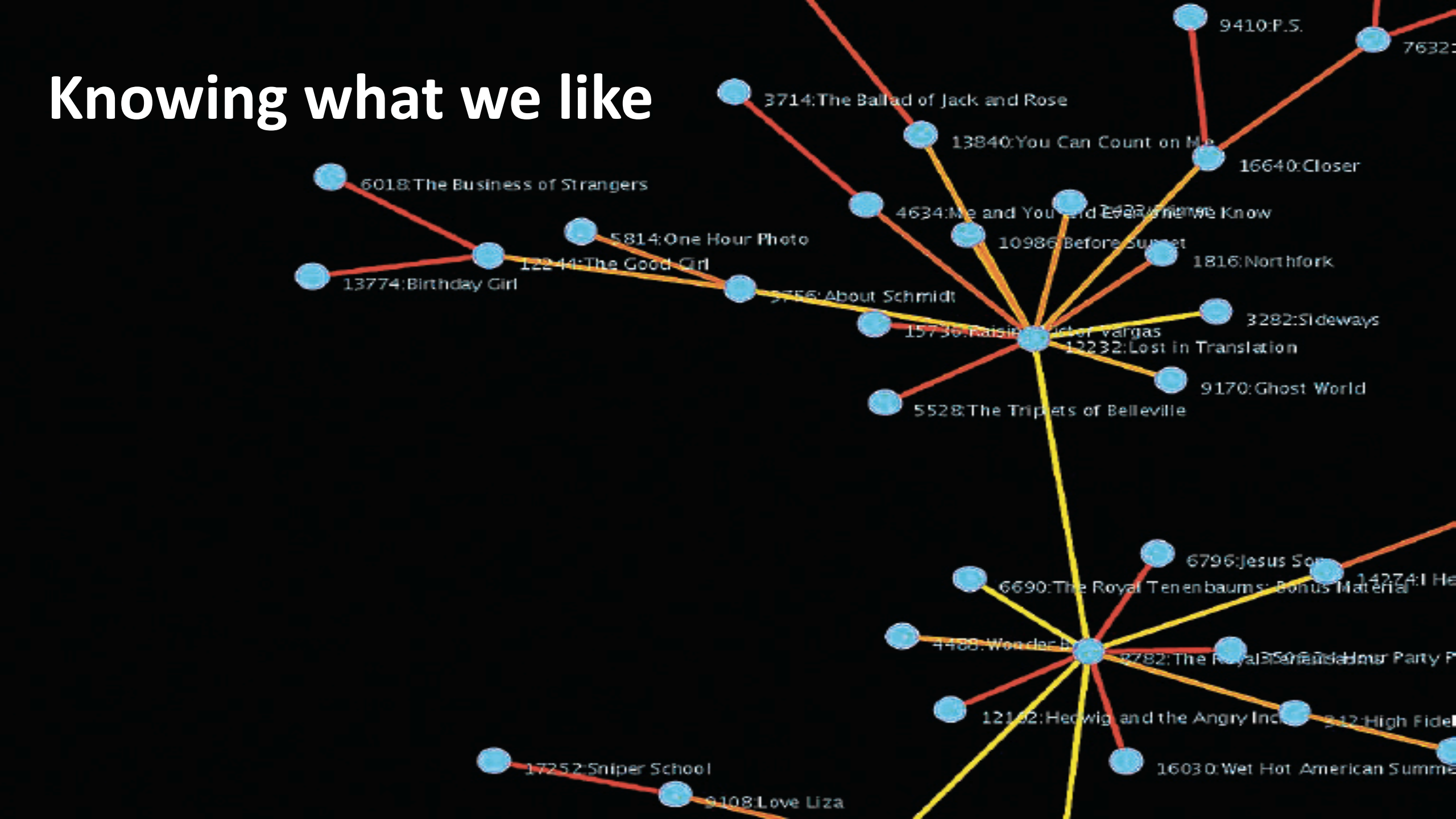
Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

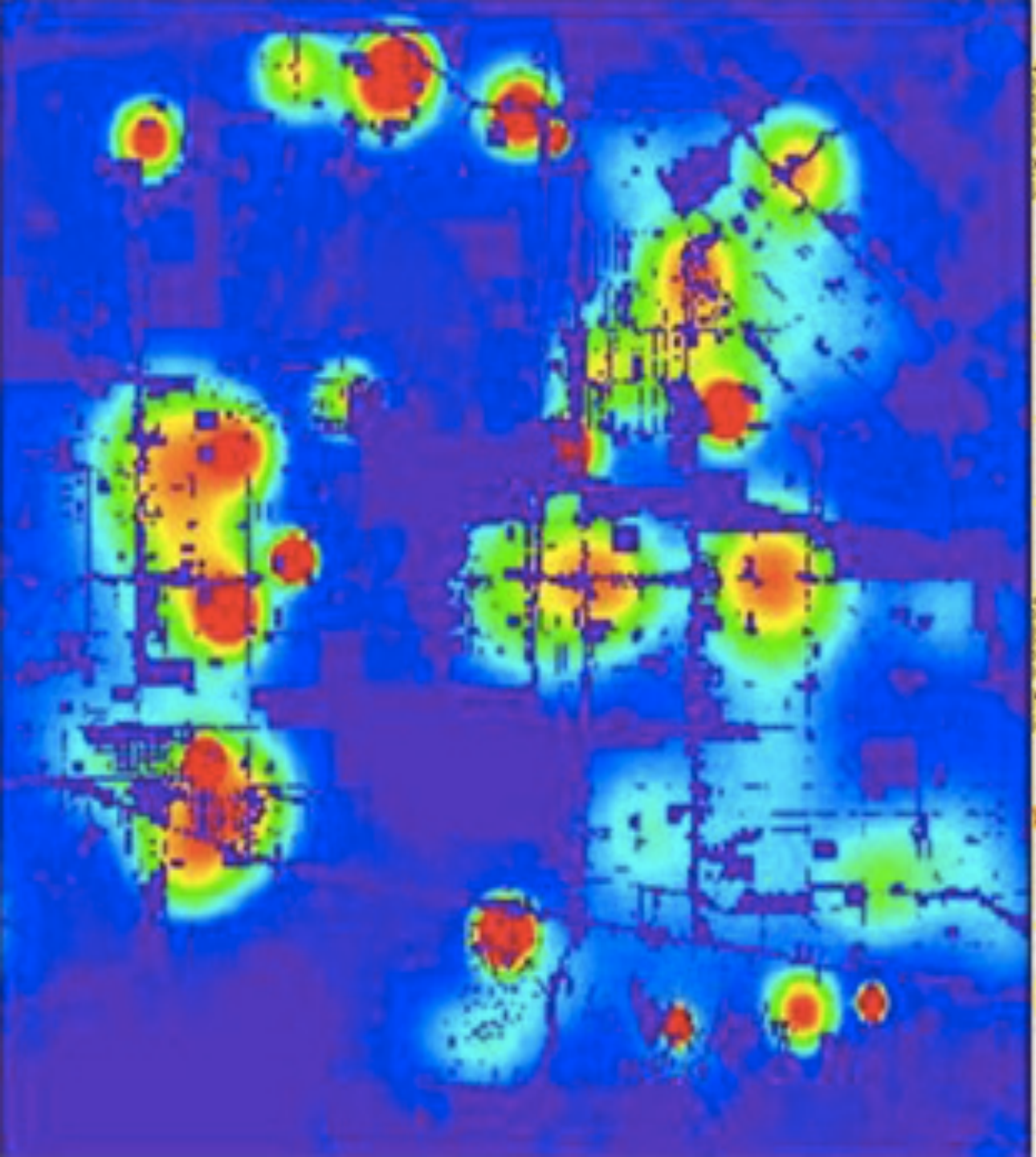
By Julia Carpenter July 6, 2015

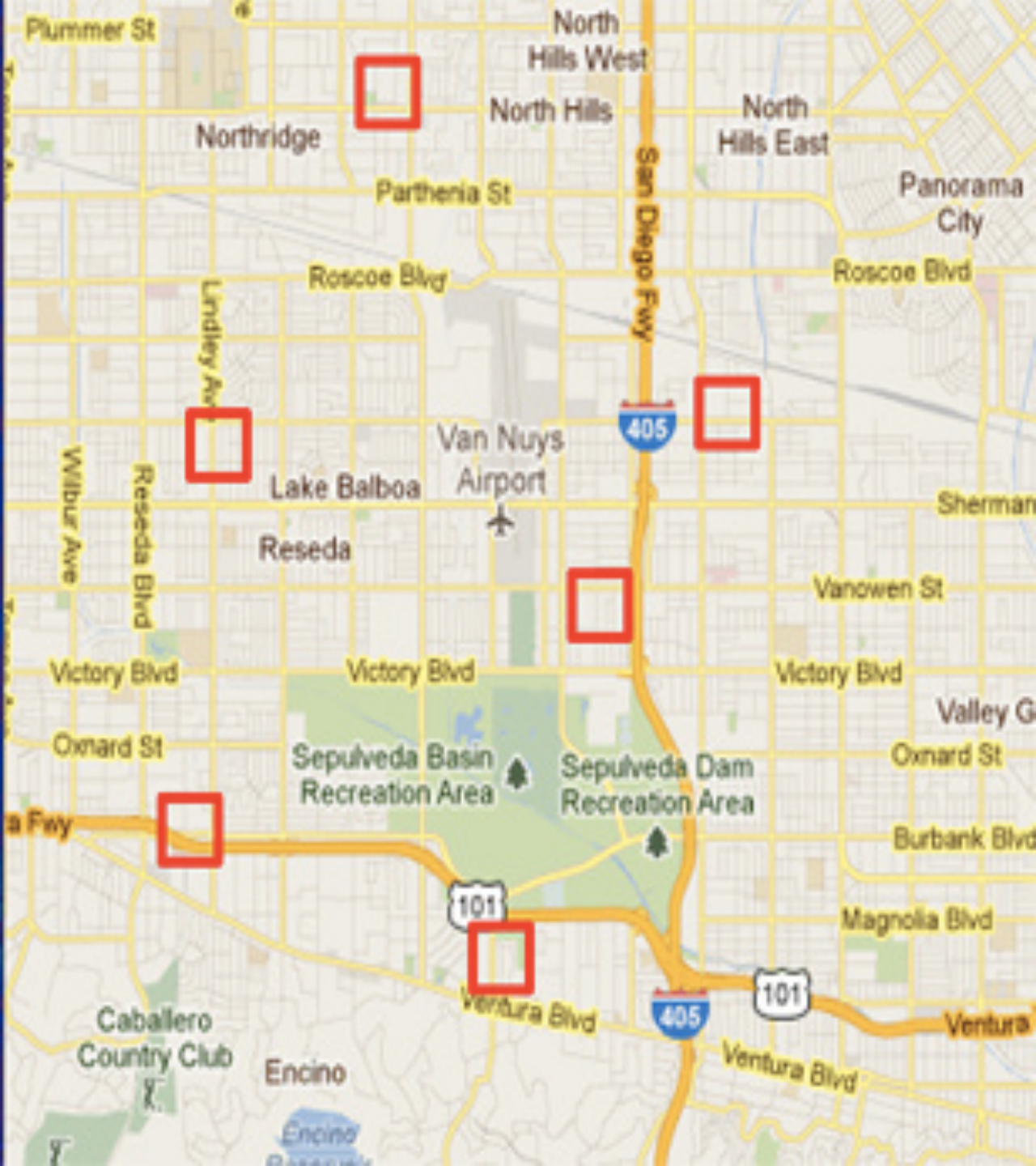
A recent screenshot of Google images for "CEO."

Fresh off the revelation that Google image searches for "CEO" only turn up pictures of white men, there's new evidence that algorithmic bias is, alas, at it again. In a paper published in April, a team of researchers from Carnegie Mellon University claim Google displays far fewer ads for high-paying executive jobs...

Knowing what we like






Plummer St
North Hills West
North Hills
North Hills East
Northridge
Parthenia St
San Diego Fwy
Panorama City
Roscoe Blvd
Roscoe Blvd
Lindley Ave
405
Van Nuys Airport
Wilbur Ave
Reseda Blvd
Lake Balboa
Reseda
Sherman
Vanowen St
Victory Blvd
Victory Blvd
Victory Blvd
Valley G
Oxnard St
Sepulveda Basin Recreation Area
Sepulveda Dam Recreation Area
Oxnard St
Burbank Blvd
101
Magnolia Blvd
Ventura Blvd
405
101
Ventura
Caballero Country Club
Encino
Ventura Blvd

Building the “digital panopticon”

(12) **United States Patent**
Lunt

(10) **Patent No.: US 9,100,400 B2**
(45) **Date of Patent: *Aug. 4, 2015**

(54) **AUTHORIZATION AND AUTHENTICATION BASED ON AN INDIVIDUAL'S SOCIAL NETWORK**

(75) Inventor: **Christopher Lunt**, Mountain View, CA (US)

(73) Assignee: **Facebook, Inc.**, Menlo Park, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

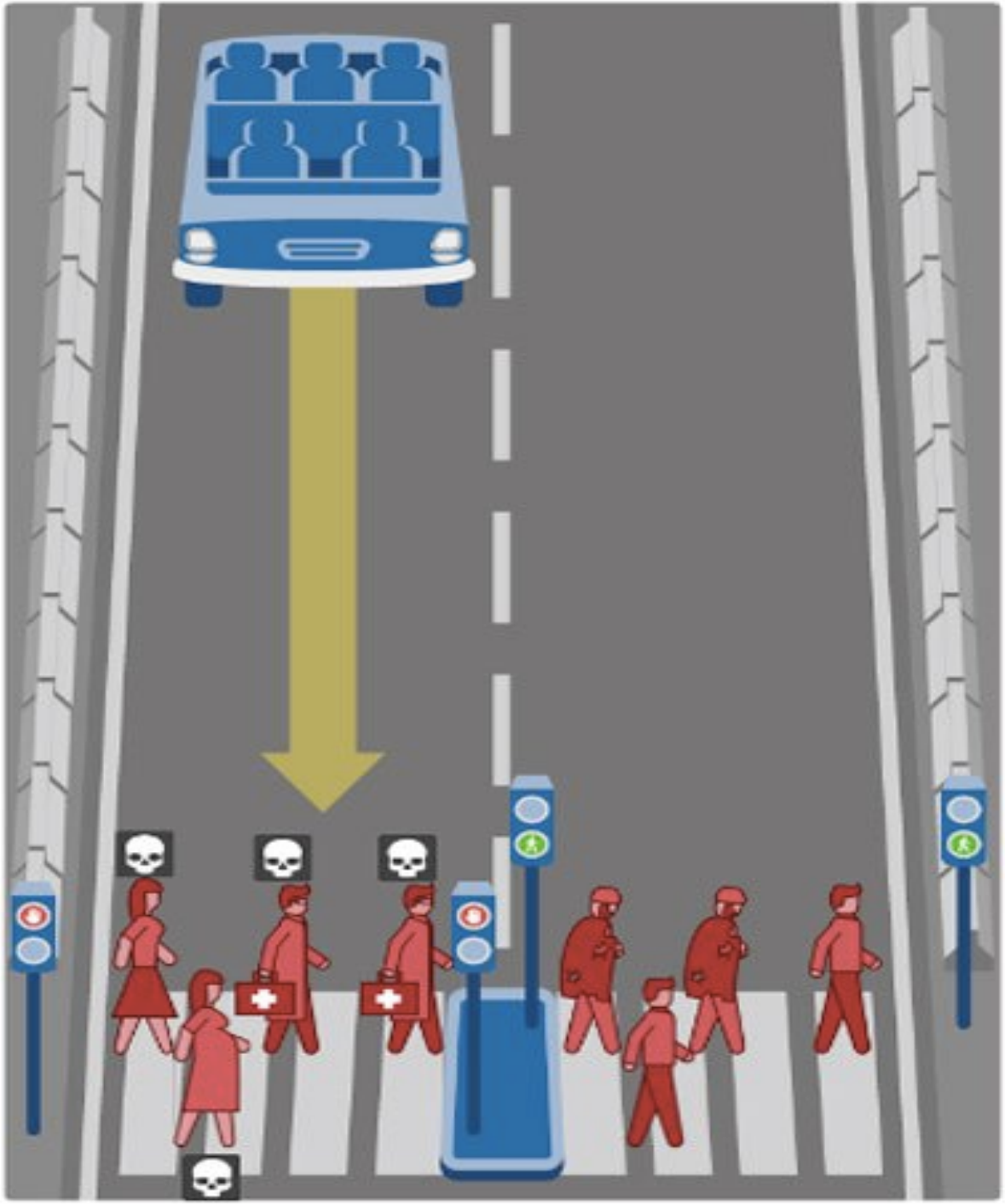
This patent is subject to a terminal disclaimer.

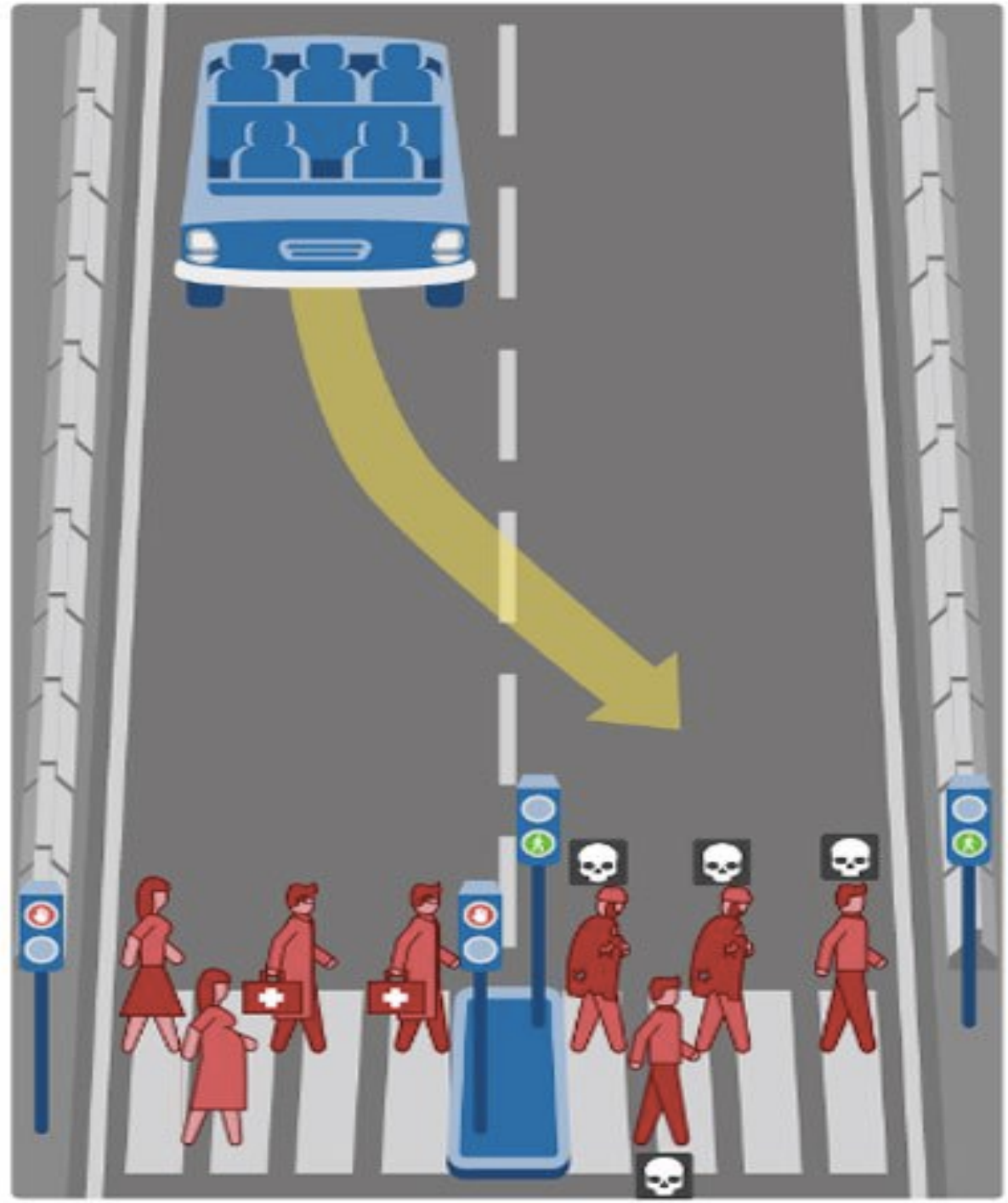
(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,950,200	A	9/1999	Sudai
5,963,951	A	10/1999	Collins
5,978,768	A	11/1999	McGovern
6,052,122	A	4/2000	Sutcliffe
6,061,681	A	5/2000	Collins
6,073,105	A	6/2000	Sutcliffe
6,073,138	A	6/2000	de l'Etraz
6,175,831	B1	1/2001	Weinrich





Problem 5

Who’s liable?

Liability and algorithms: open fronts

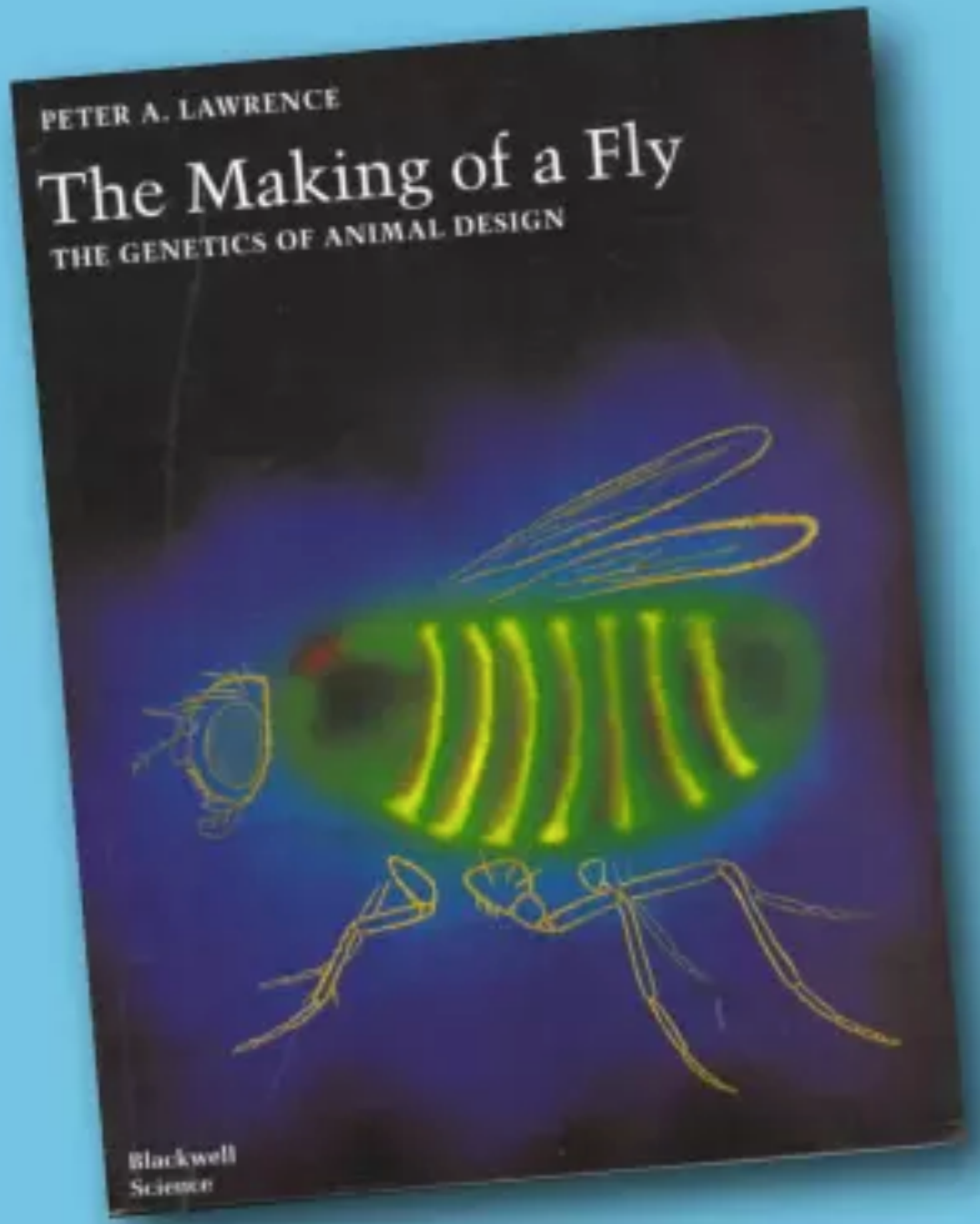
- Since the p2p copyright saga, we learnt that algorithms can be used to distance tortfeasors from liability
- The debate is now extended to antitrust (dominance and collusion)
- Time for strict liability?
  - We don’t know what we will know …
  - Difficult to establish causation, even without having to prove negligence
  - Key problems: distributed responsibility and clash of algorithms
  - Process-based or outcome-based?

Robots: animals or slaves?

- **Option 1: individual legal entities (e.g. European Parliament report on Civil Law Rules for Robotics)**
- **Option 2: Robots = animals (*culpa in vigilando)***
- **Option 3: Robots = slaves *(culpa in eligendo,* and strict liability*)***
- **Option 4: Robots like robots?**

Interaction between algorithms

Who’s liable?



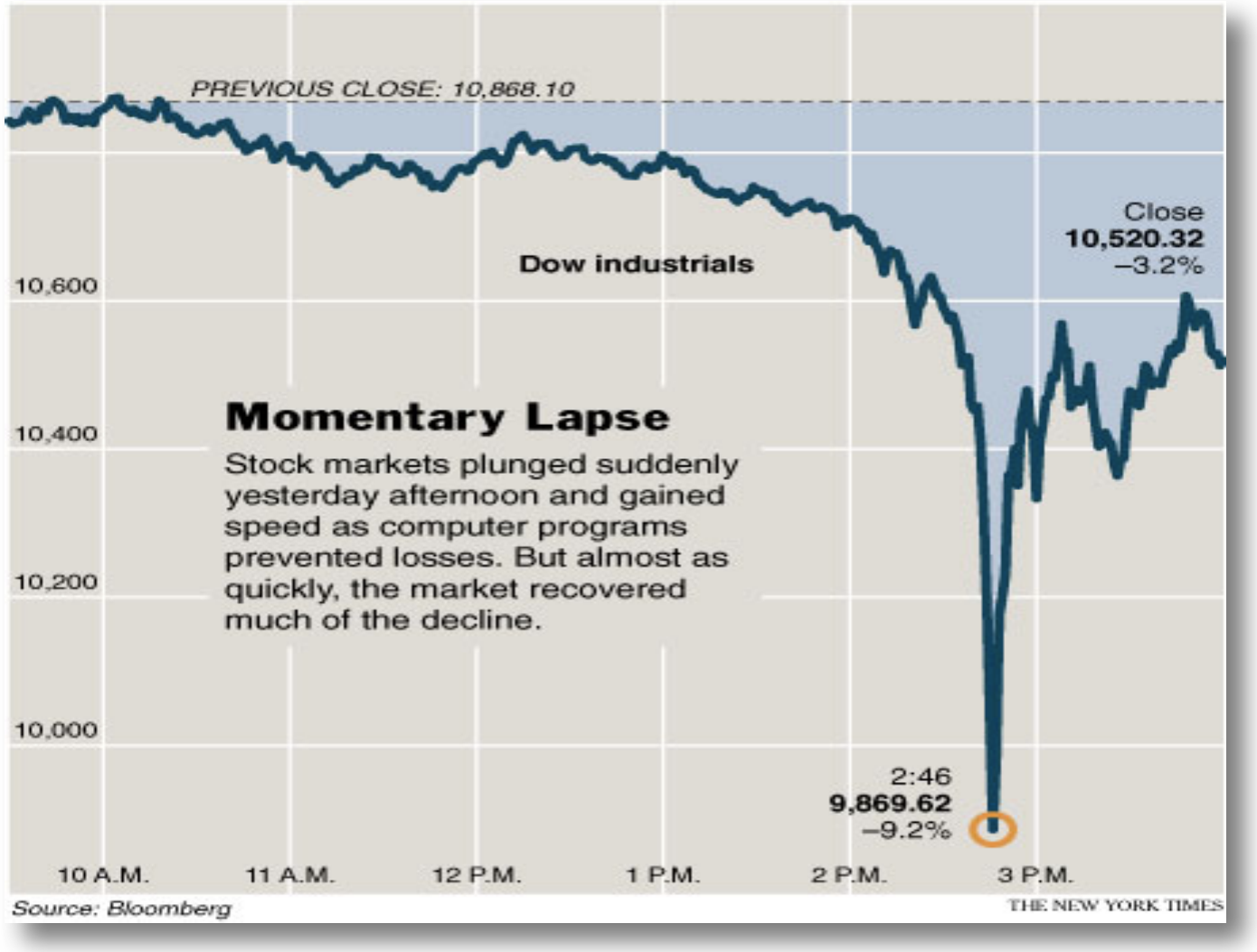
$23 million!!

Who is responsible?

“Flash crash of 2.45”

- 9.2%!

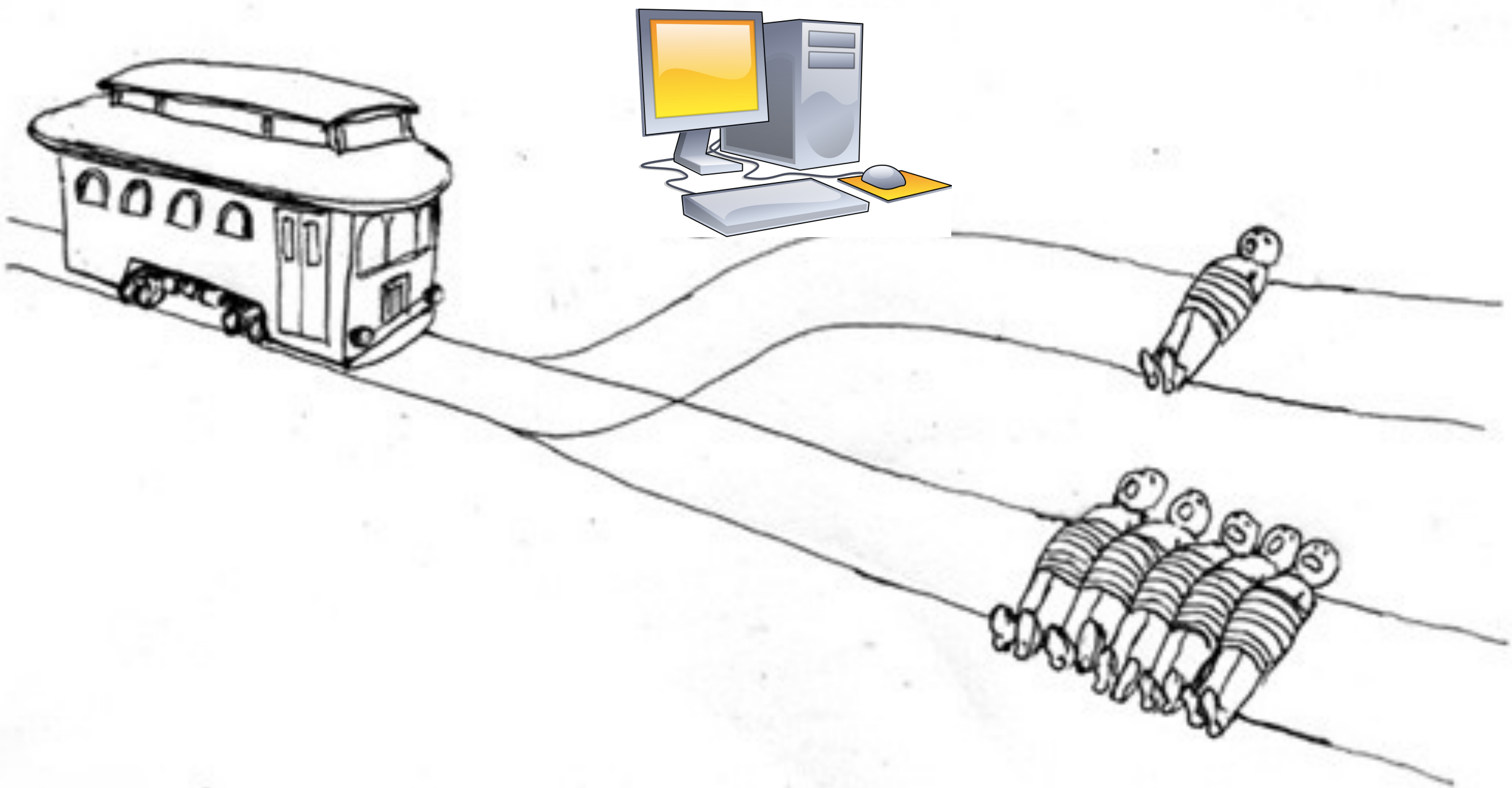
Who is responsible?

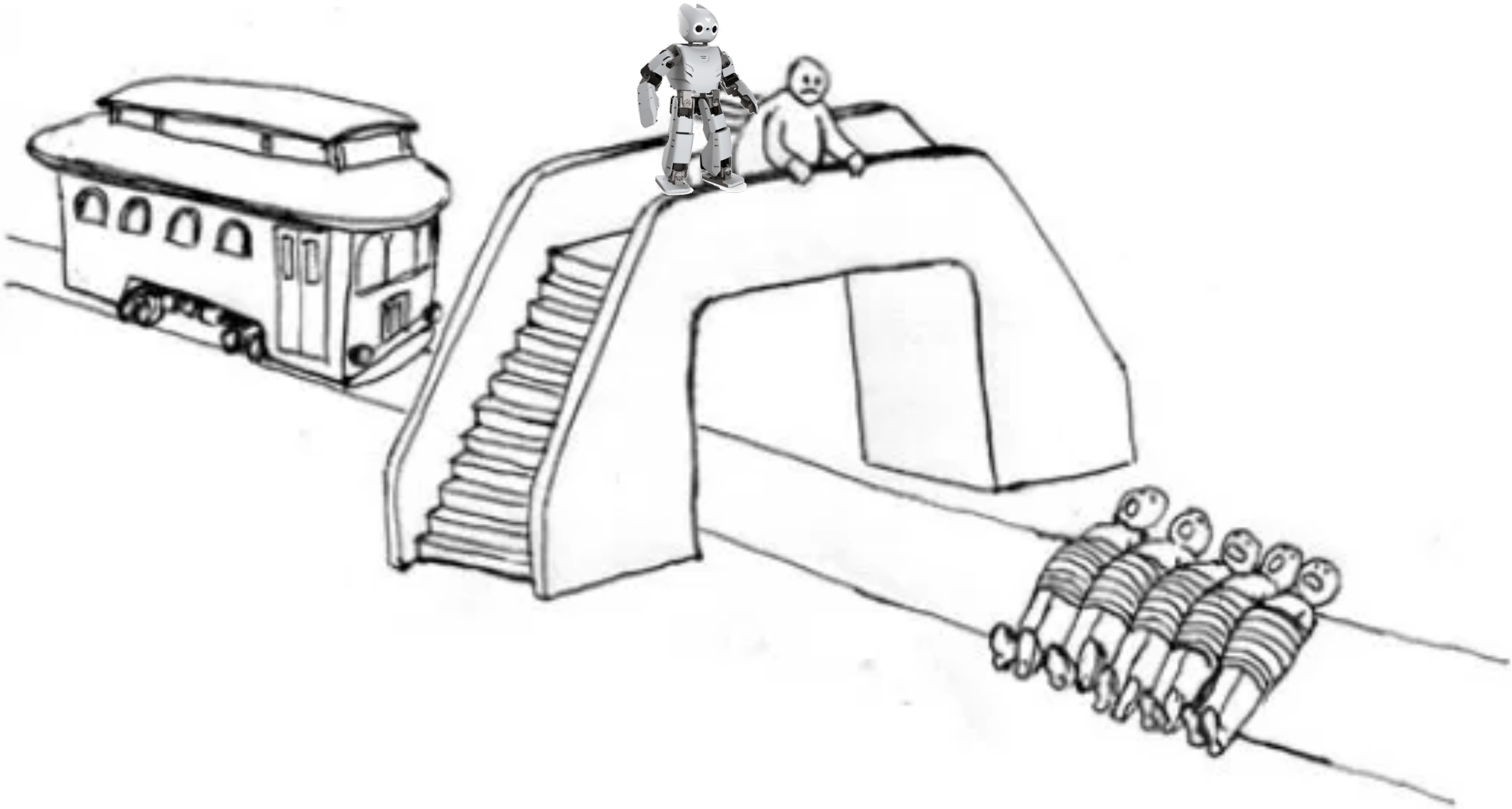


Results of the investigation: key take-aways

Who gets to decide?

- Nobody: the trolley problem cannot occur (and we don’t want it to)
- Nobody: the car does not have any instruction
- Government (through regulation)
- The algorithm developer
- The car manufacturer
- The transportation company
- The owner (“ethical knob”)
- The insurer (by price discrimination)





An emerging quagmire

- We can choose to avoid socially unacceptable situations, and pave the way towards complementarity between humans and robots
- Trade-off between data availability and algorithmic accuracy: possible race to the bottom?
- Need for accountability of algorithms: co-regulatory solutions, public auditing, or Distributed Ledger Technologies
- Need for strict and joint and several liability, especially in the case of algorithmic interaction

Problem	Policy challenge/response
1. How did the car end up there?	- Avoid delegating life-threatening decisions to machines - Preserve human control as a key item in policy shaping
2. What did the car know?	- Adopt a clear and predictable data policy for self-driving cars, balancing privacy and efficiency - Test the use of privacy-compliant distributed ledgers for automated vehicles - Experiment with forms of differential privacy in algorithms to strike the balance between efficiency and privacy
3. What do we know about how the car decided?	- Clarify the legal framework for algorithmic accountability and transparency - Clarify the applicability and scope of the right to explanation under the GDPR - Establish an obligation for *ex post* inspection of automated vehicle 'black boxes'
4. Better than us, or like us?	- Define a set of principles for algorithmic decision-making, including clear criteria for separating lawful from unlawful discrimination - Work on anti-polarisation strategies to avoid the AI-powered exacerbation of existing biases
5. Who's liable?	- Define strict liability principles for algorithm-powered decision-making - Define legal rules for damages caused by the interaction between algorithms

And what about the Trolley Problem?

Not today…

DIGITAL INNOVATION
COLLEGE OF EUROPE CHAIR

The trolley problem and Self-Driving Cars
A CSI into the Ethics of Algorithms

Andrea Renda

Chair for Digital Innovation, College of Europe

17 January 2018