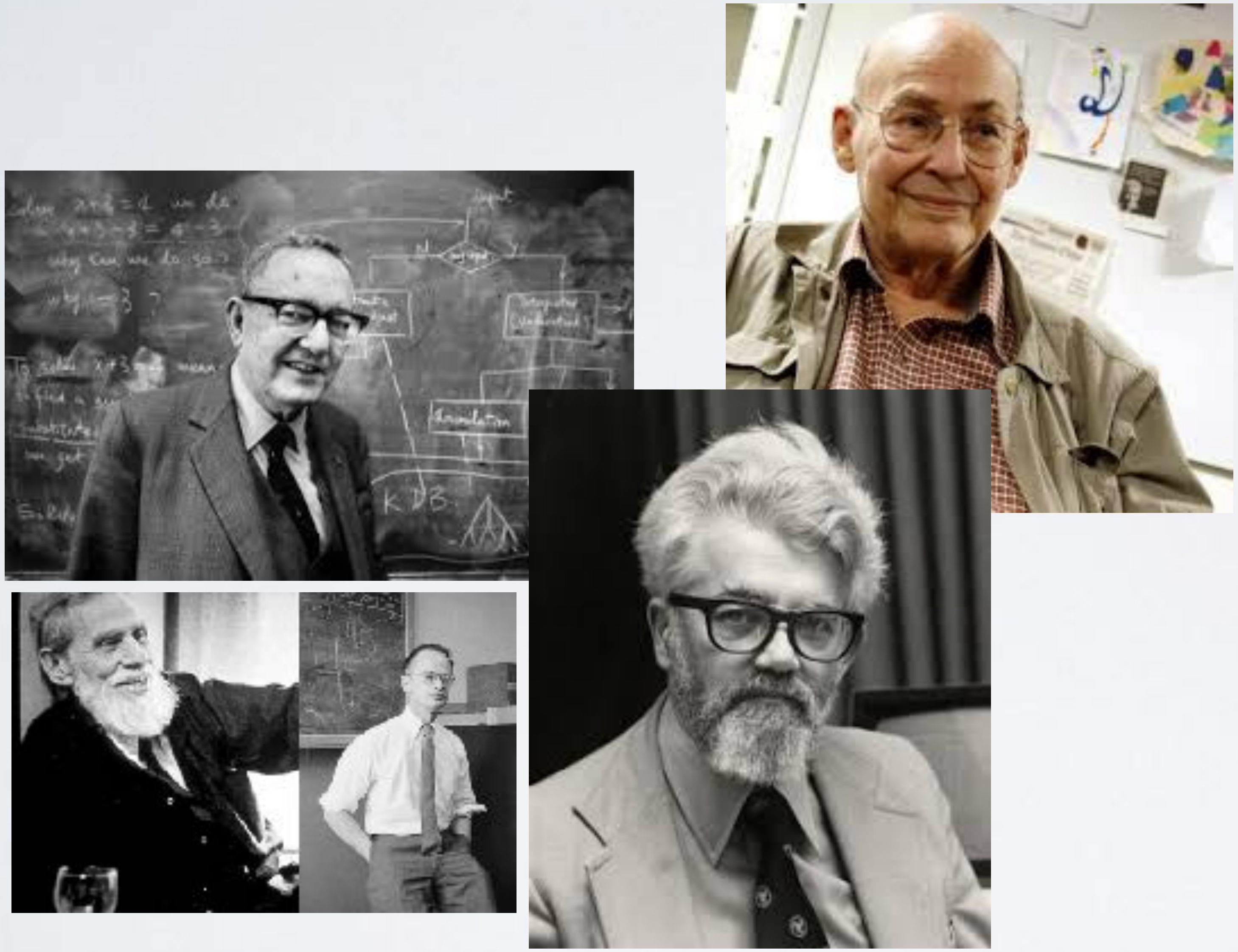
ETHICAL AND SOCIETAL ISSUES IN MODERN AI

Umberto Grandi
University of Toulouse
Master MIAGE 2IS

PAST AND PRESENT

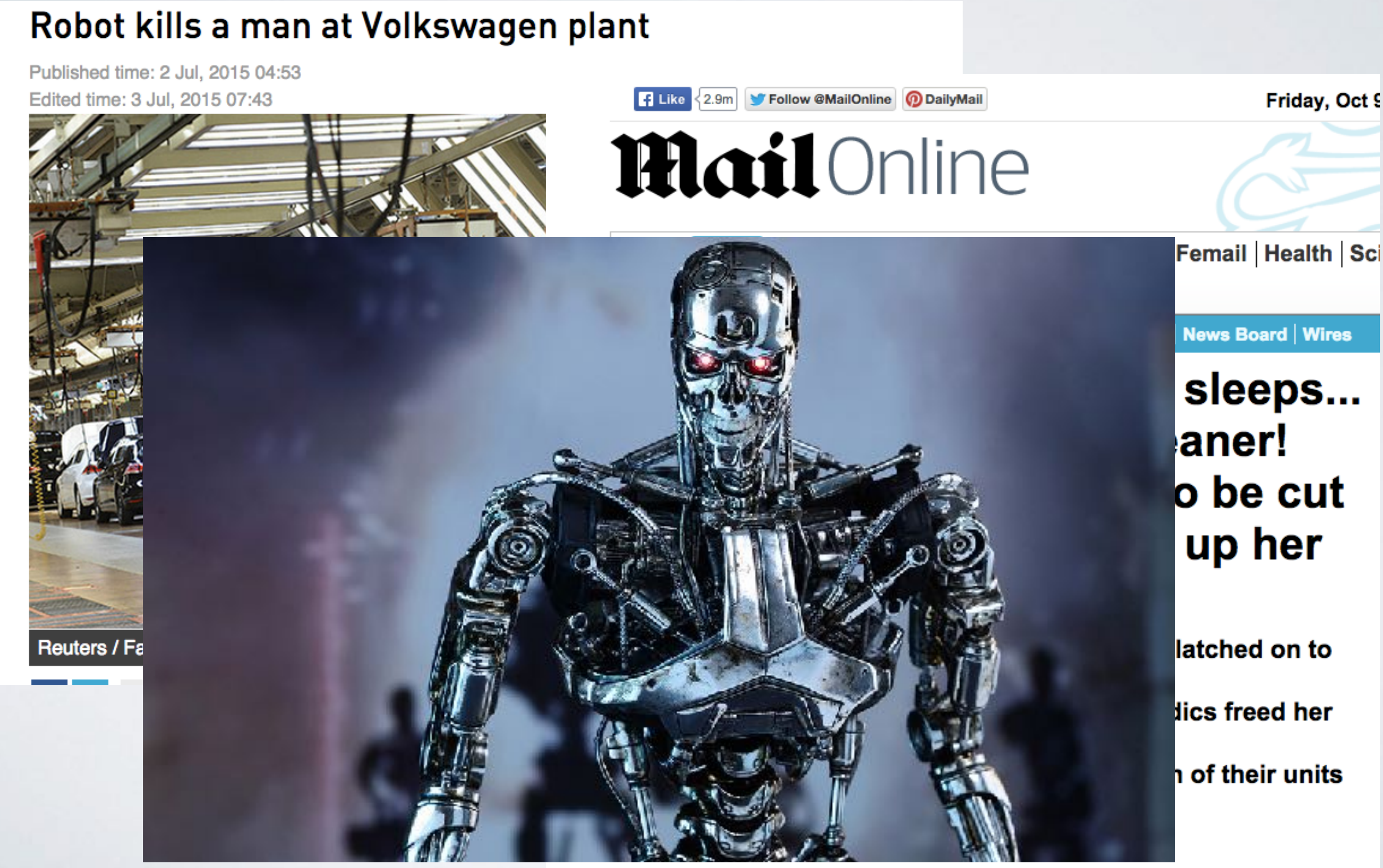


How we got here
(Dennis' lecture)

PLAN FOR TODAY

- Ethics and AI: from philosophy to engineering
- Responsibility and autonomy
- Privacy and identity
- AI and work

AI IS SCARY - WHY EXACTLY?



Robot kills a man at Volkswagen plant

Published time: 2 Jul, 2015 04:53

Edited time: 3 Jul, 2015 07:43

Reuters / Fa

sleeps...
aner!
o be cut
up her

latched on to

dics freed her

n of their units

DEGREES OF CONTROL

- **Controlled systems**: humans have full or partial control. Example: an ordinary car

- **Supervised systems**: do what an operator has instructed. Example: industrial machinery

- **Automatic systems**: carry out fixed functions without the intervention of an operator. Example: an elevator

- **Autonomous systems**: that are adaptive, learn and **can make decisions**. Example: autonomous car

AUTONOMY AND ETHICS

Many of AI's problem today comes from giving too much autonomy to stupid AI! Like bots, or autonomous cars…

- When giving autonomy to machines we need to be able to teach the machine **right from wrong**

- Philosophical **moral theories** can be very helpful: what makes an action right/wrong, a person good/bad

UTILITARIANISM

The right actions are those that increase the overall utility of society

Jeremy Bentham, John Stuart Mill (18th/19th century UK)

Most AI architectures are based on **utility maximisation**. They are efficient, but are they "moral"?

- Whose utility should be maximized?
- How to define utility?
- In an uncertain world, how can consequences of actions be computed accurately?

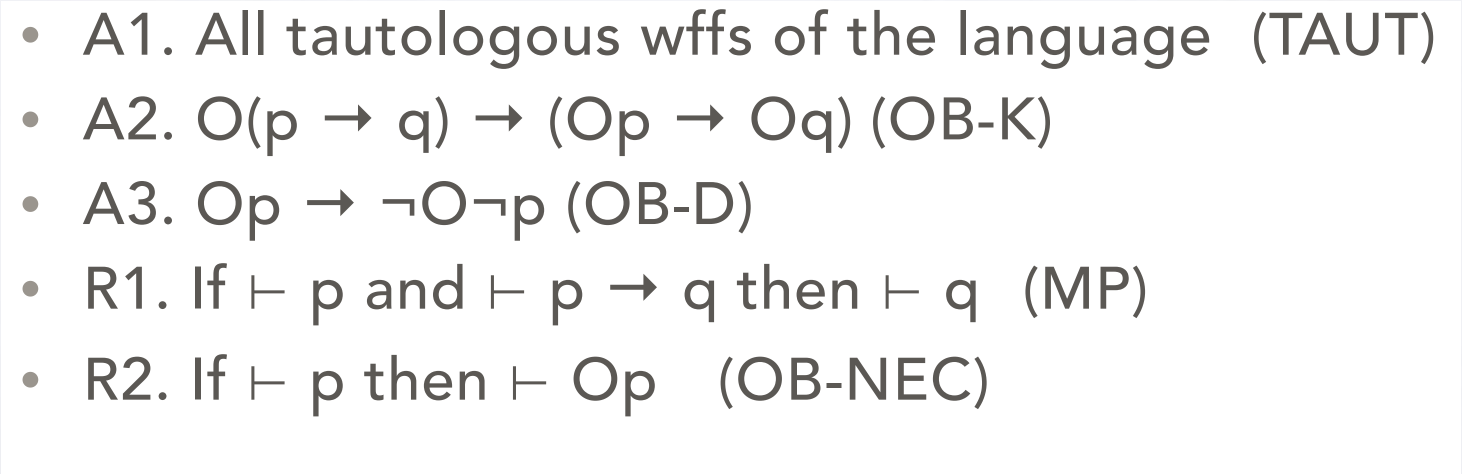
KANTIAN ETHICS

> Act only on that maxim through which you can at the same time will that it becomes a universal law

Immanuel Kant, 18th century Germany

Emphasise **principles** behind actions rather than consequences.

But how do we include these principles into a machine?



- A1. All tautologous wffs of the language (TAUT)
- A2. $O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$ (OB-K)
- A3. $Op \rightarrow \neg O \neg p$ (OB-D)
- R1. If $\vdash p$ and $\vdash p \rightarrow q$ then $\vdash q$ (MP)
- R2. If $\vdash p$ then $\vdash Op$ (OB-NEC)

ENGINEERING ETHICS

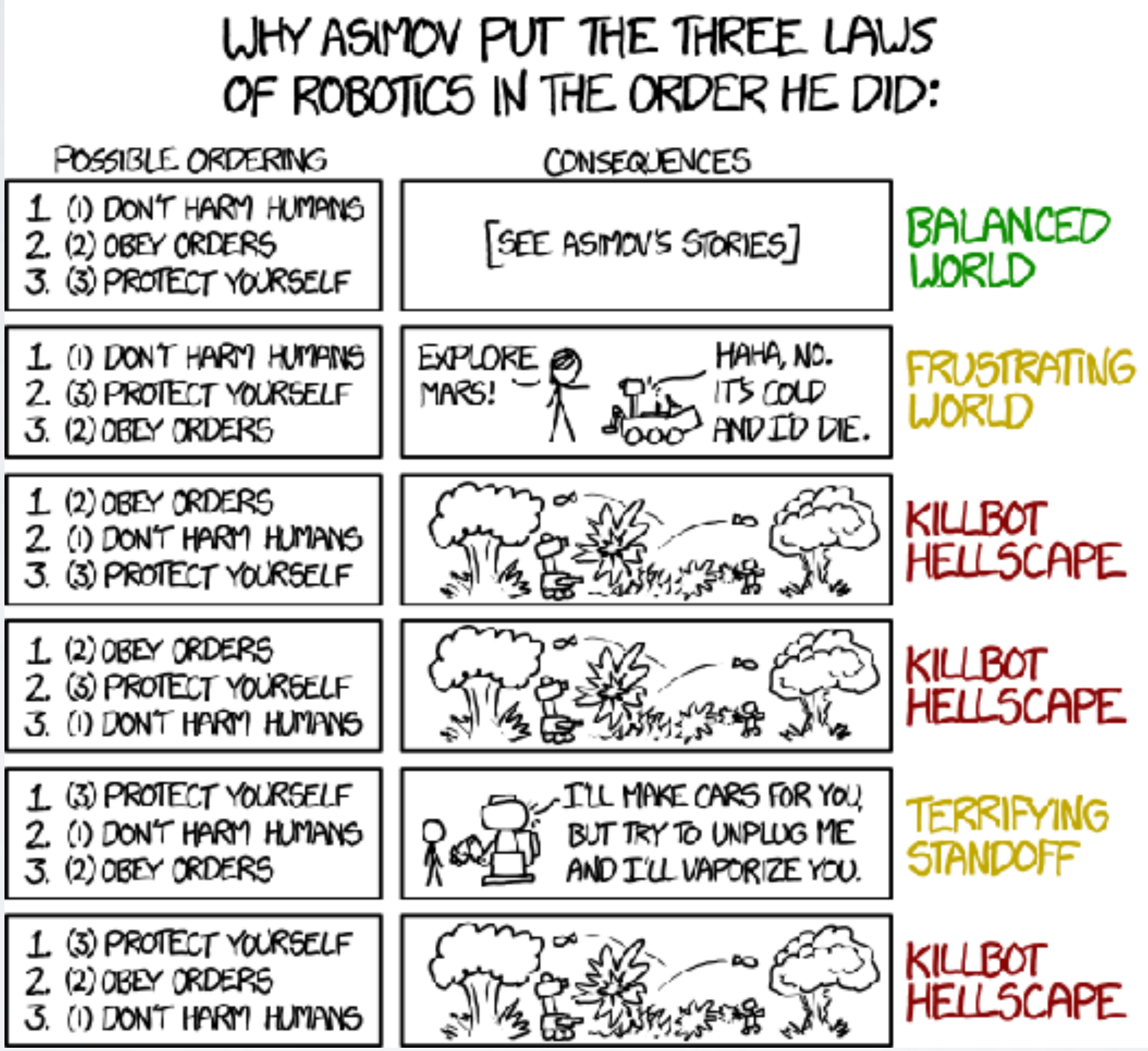
"Ethical cognition must be taken as a subject matter of engineering." Nick Bostrom (living philosopher).

Some philosophical/technical examples:

- Which machine learning algorithm is more **transparent**?
- Can we limit **unpredictable local actions** generated by globally optimal behaviour? (bayesian vs genetic algo)
- Subjective **experience of time** does not vary in biological entities. How about in AIs?

RESPONSIBILITY IN AUTONOMY

ASIMOV THREE LAWS



WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS! HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS	I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZE YOU.	TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

Asimov laws are flawed, otherwise no nice novel

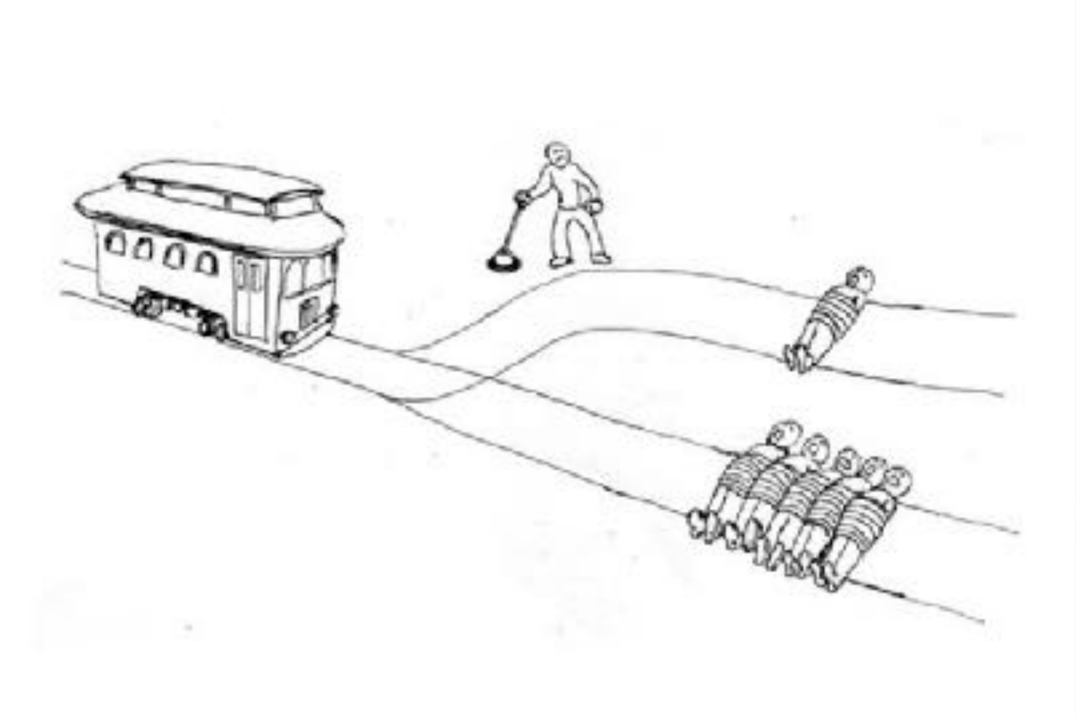
EPSRC ROBOTICS PRINCIPLES

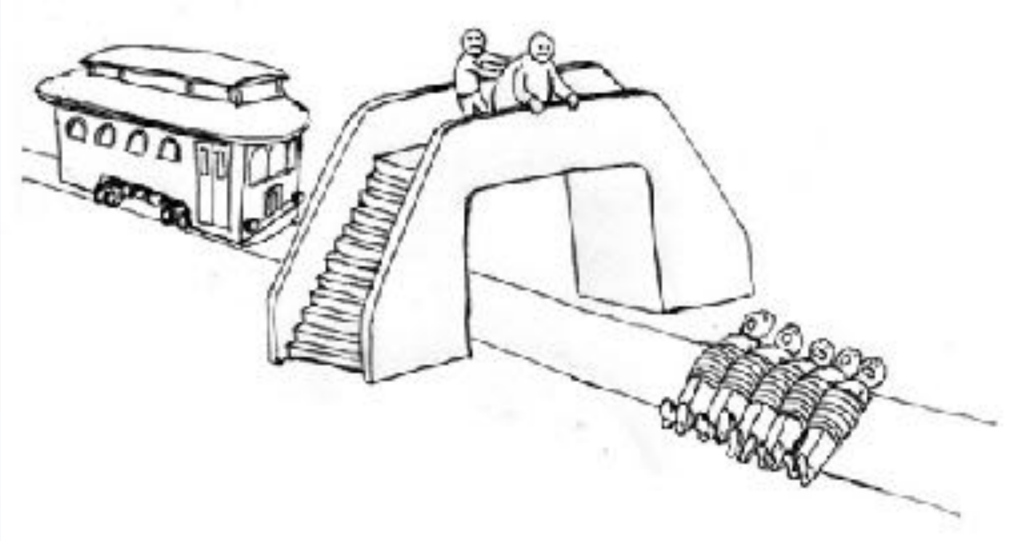
In 2010, the UK research council proposed:

1. Robots are multiuse tools. Robots should not be designed solely or primarily to kill or hard humans, **except in the interest of national security.**
2. Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws and fundamental rights and freedom, **including privacy**.
3. Robots are products.They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts.They should not be designed in a deceptive way to exploit vulnerable users; instead **their machine nature should be transparent**.
5. The person with **legal responsibility** for a robot should be attributed.

RESPONSIBILITY

Still not clear what to do in this case:

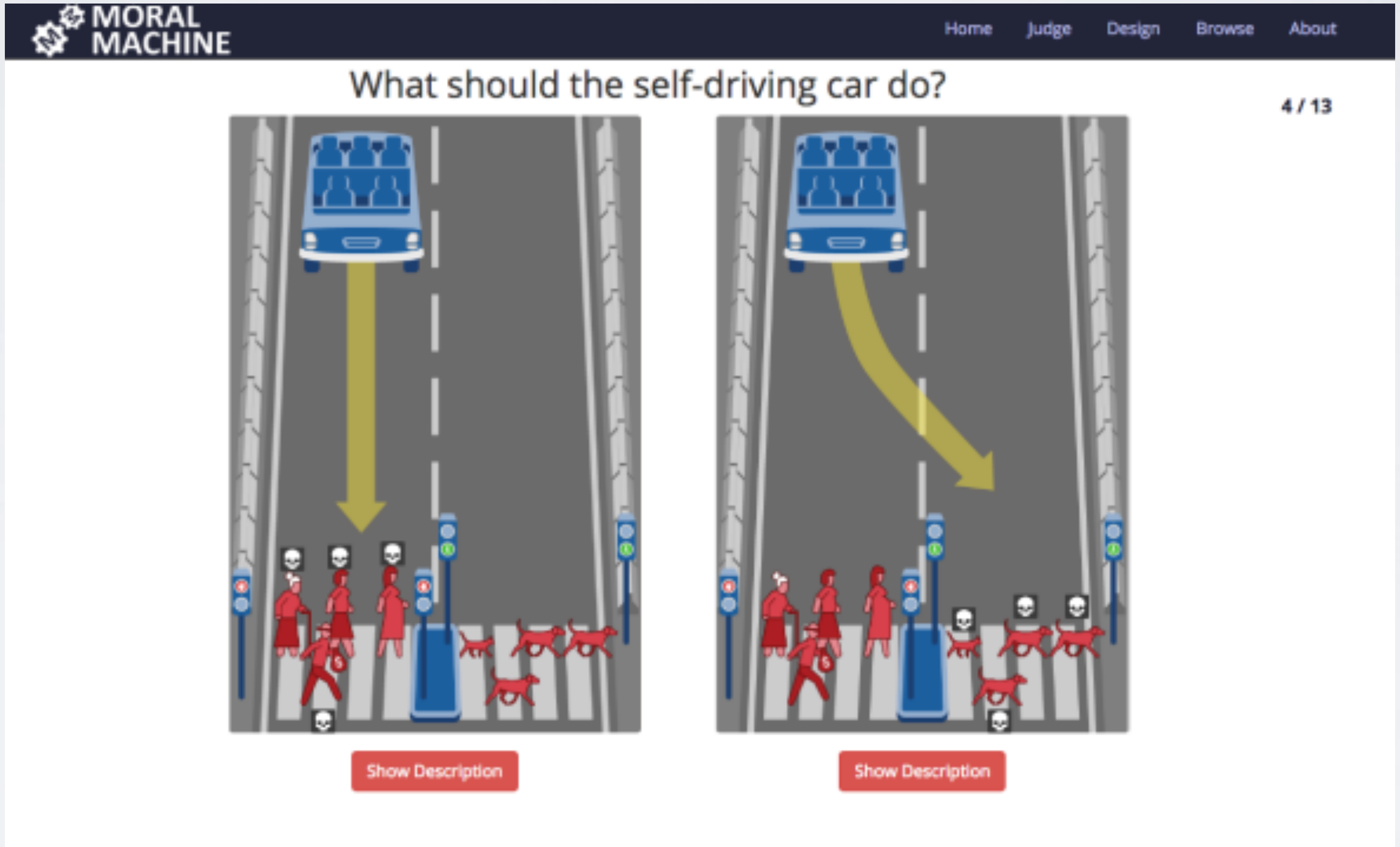




The trolley problem
Classical thought experiment in ethics, and the Internet's most philosophical meme, NY mag

MORAL MACHINES

Crowdsource ethical principles:

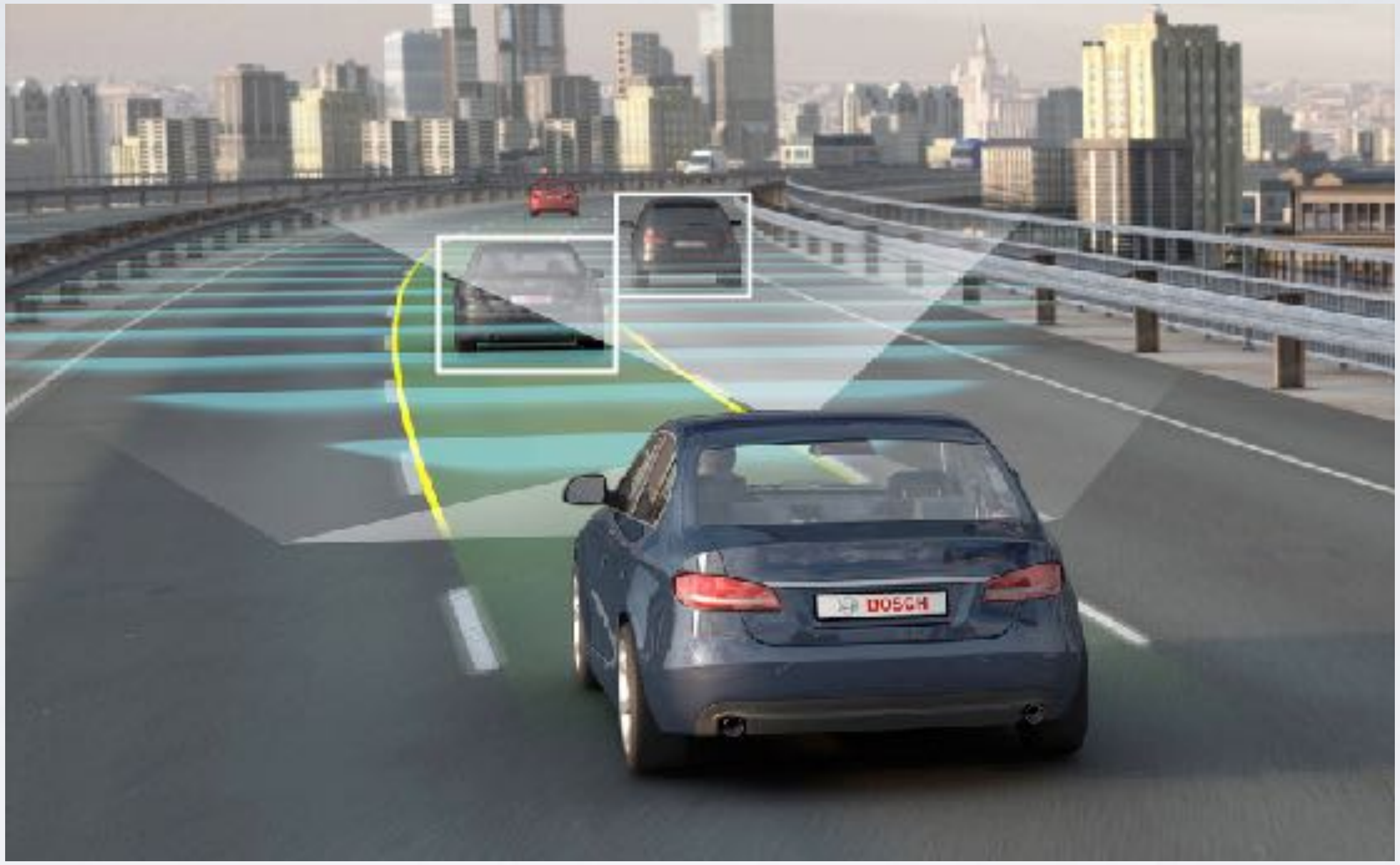


What should the self-driving car do?

4 / 13

In 2016, researchers from MIT, Oregon, and Toulouse launched a platform to experiment on **extracting ethical principles** via crowdsourcing.

AUTONOMOUS CARS



- By 2050 **non-autonomous cars** will look so **old-fashioned** (think about how horse-filled roads of 1900 looked to 1950 drivers)
- A **moral imperative**: 1.2M deaths by car accidents in world per year, more than 30 Airbus A380 crashing every week!

AUTONOMOUS CARS

- **Regulation** will bring safety! Think about how dangerous was flying 100 years ago, and how safe it is now.
- Companies like TESLA are allowed to **test their algorithms** on the population: should be forbidden (drug companies cannot test drugs freely!)

PRIVACY AND IDENTITY

Image by Patrick Kyle

PRIVACY AND AI

Too vast topic to be treated completely.

The STASI had 2.1 million informers to spy on 17 million people.

The NSA has 30.000 (estimated) employees to spy on the entire world.

PRIVACY AND AI

Big data collected by all sorts of apps are only useful if AI programs can make use of them: **automated profiling**

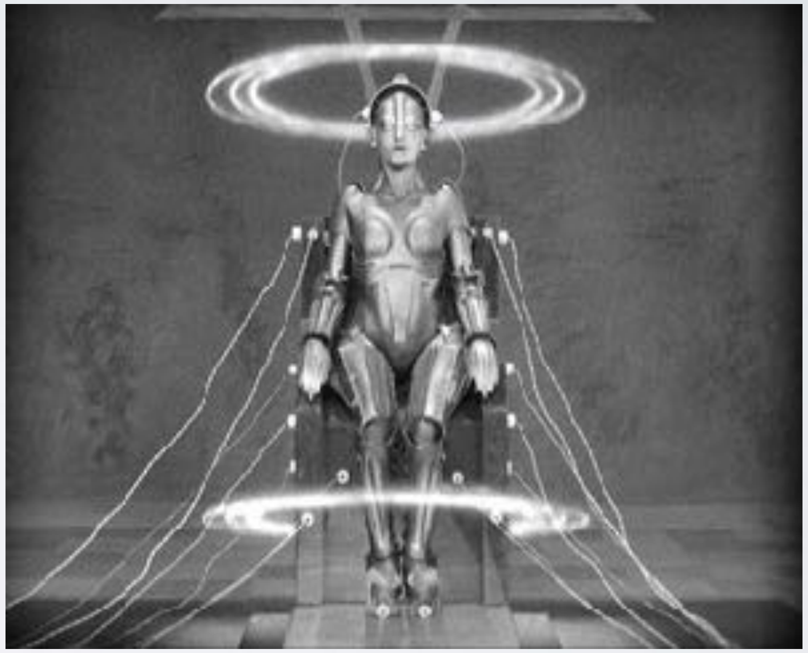
Personalised actions can then be taken, including:

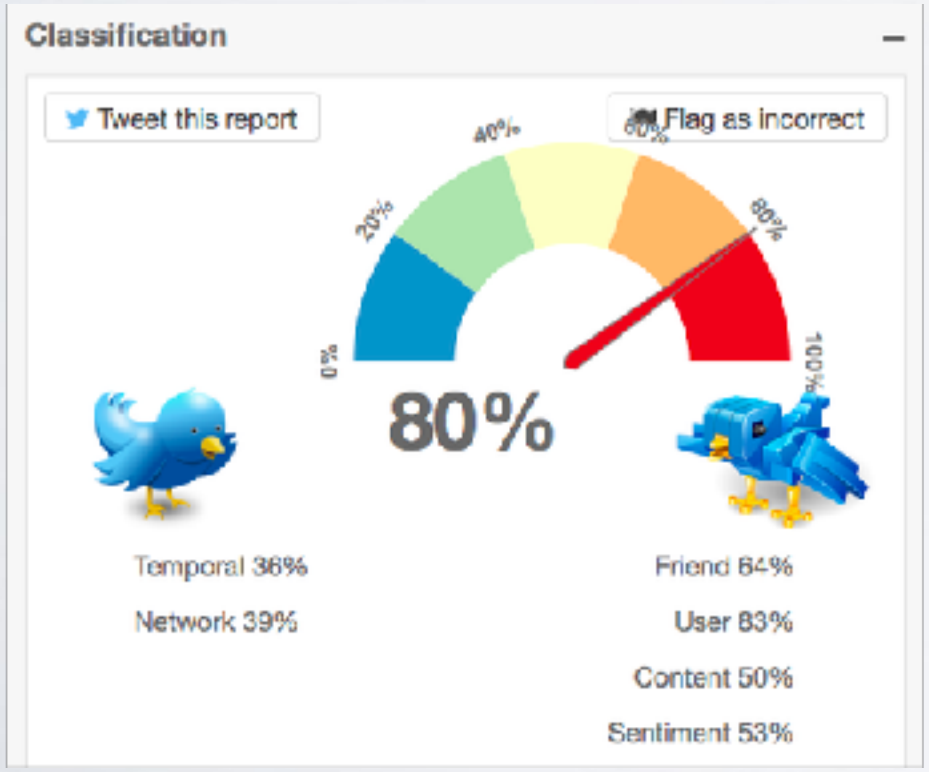
- Spy on you
- Influence your opinion in a political debate
- Show higher/lower prices for purchases (influencing your consumer behaviour)
- Show personalised search keywords (influencing your access to information)

Read about similar issues about machine learning biases in Cathy O' Neil book "*Weapons of Math Destruction*", 2016

MACHINE OR HUMAN?

From **science fiction**: Metropolis, Blade Runner, Ex Machina…





Classification

80%

Temporal 36%

Network 39%

Friend 64%

User 83%

Content 50%

Sentiment 53%

To **reality**: BotorNot, Botometer…AI is being used to identify AI!

TURING RED FLAGS

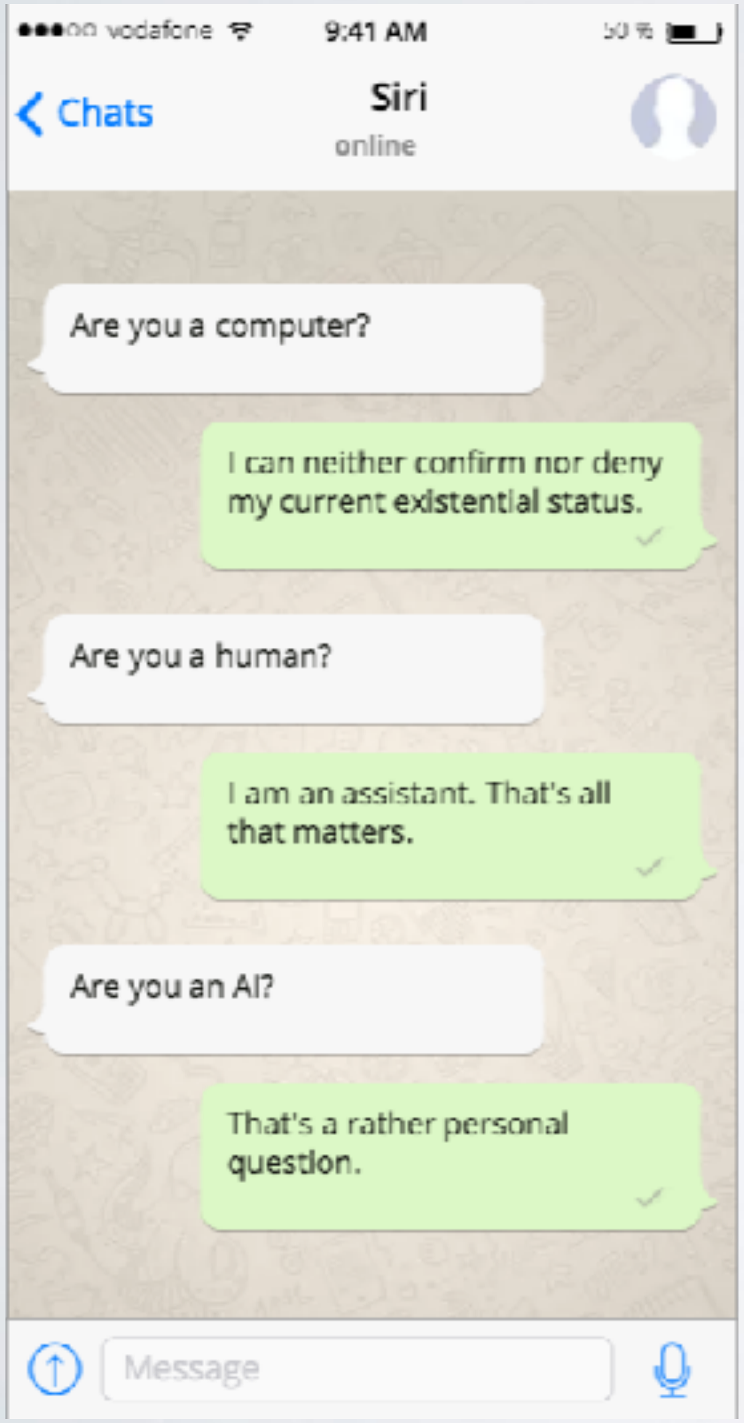
A proposal by Toby Walsh (Arxiv, 2015)

An autonomous system should be designed so that it is unlikely to be **mistaken for anything besides an autonomous system**, and should **identify itself** at the start of any interaction with another agent.

Precedent: motored cars in19th century UK

Consequences: end of twitter bots influencing political debates?

MORE ON RED FLAGS



Are you a computer?

I can neither confirm nor deny my current existential status.

Are you a human?

I am an assistant. That's all that matters.

Are you an AI?

That's a rather personal question.

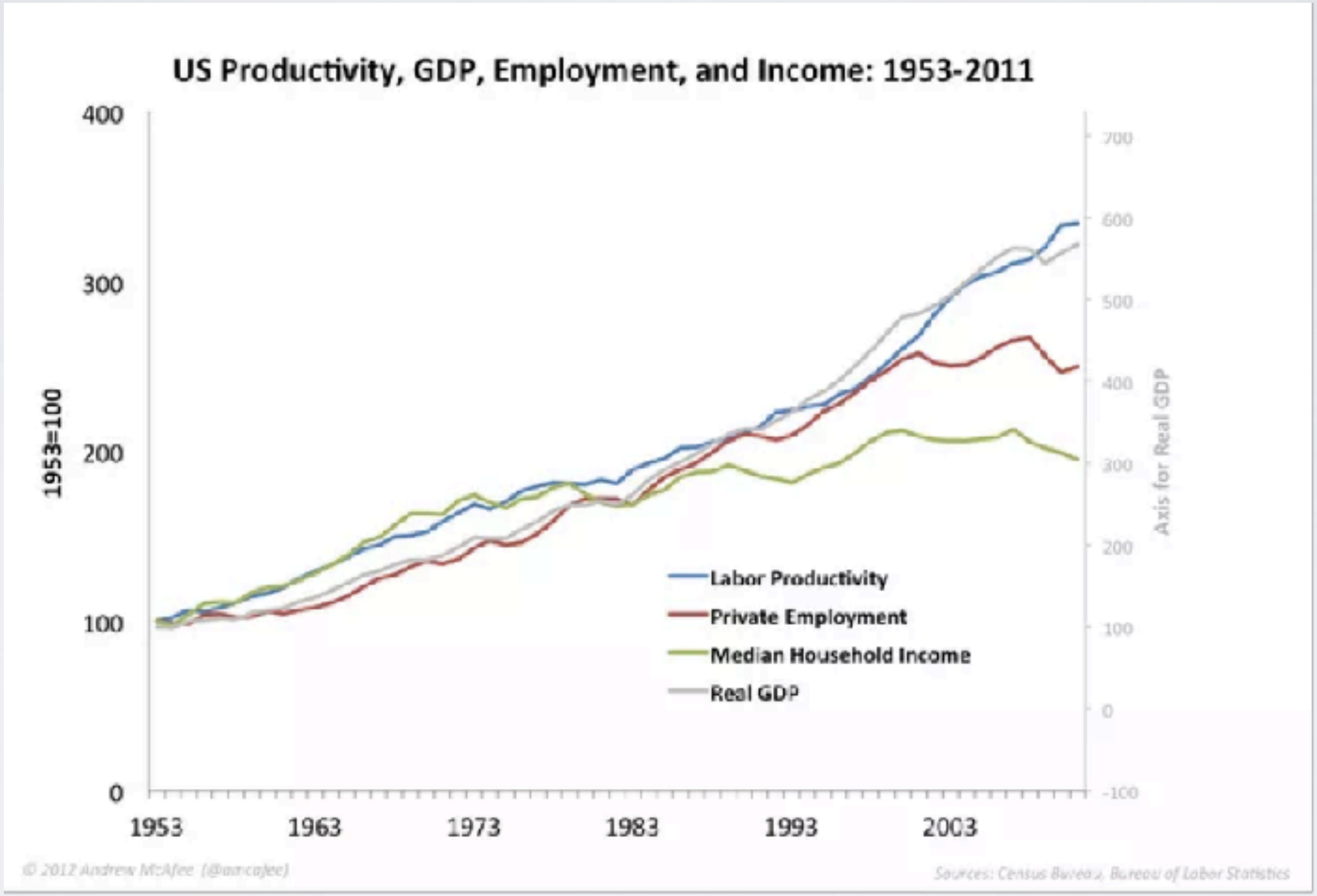
SIRI, Alexa, Google Home,
do not show Turing red flags!

Other examples: autonomous poker players, automatic news-text generation, autonomous cars in mixed roads…

However: it might be **too early**, and slow down the technology (not for the proponent of this principle)

AI AND WORK

PRODUCTIVITY AND INCOME



© 2012 Andrew McAfee (@amcafee)

Sources: Census Bureau, Bureau of Labor Statistics

Articles by Moshe Vardi analysing **technological unemployment**

NOTORIOUS OXFORD STUDY

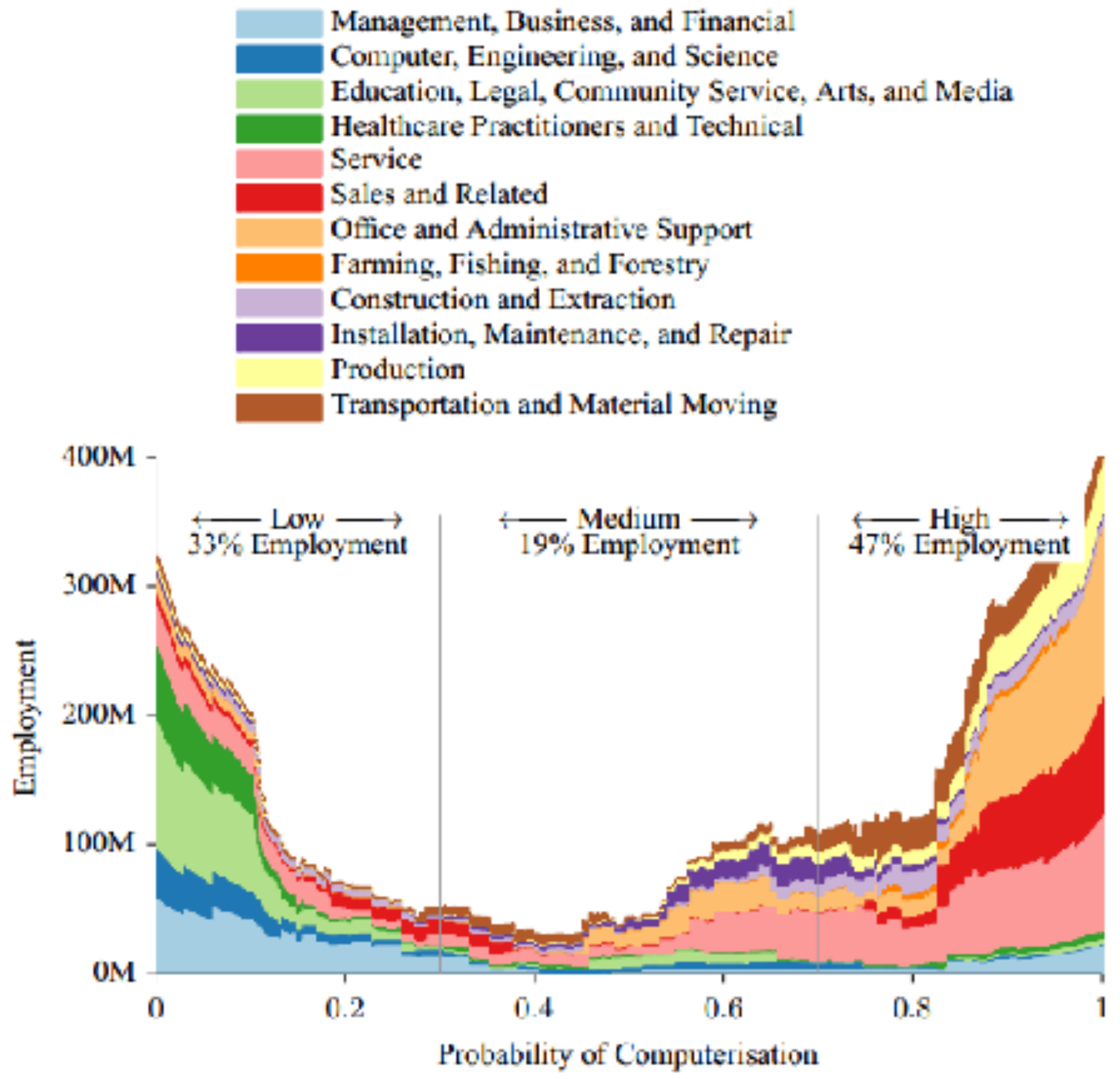


FIGURE III. The distribution of BLS 2010 occupational employment over the probability of computerisation, along with the share in low, medium and high probability categories. Note that the total area under all curves is equal to total US employment.

Frey and Osborne, 2013, predicted that **47% of jobs** in the US are **under threat of automation** in next 2 decades.

In practice: most jobs will not be automated for economical, technical or societal reasons.

Ironic: they used machine learning in their analysis!

AUTOMATABLE JOBS ZOO

Bicycle repairer?

Oxford: 94%, too high

Quarry worker?

Oxford: 96%, correct

Journalist?

Oxford: 11%, too low

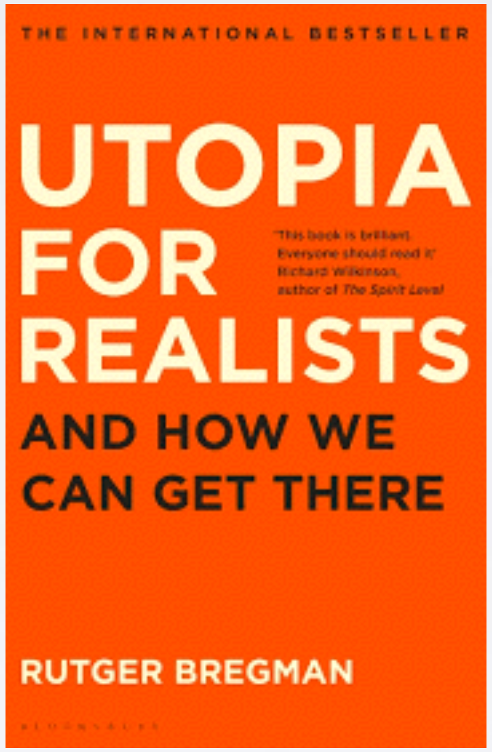
Musicians?

Oxford: 7%, correct

A JOBLESS SOCIETY?

Introduce a **universal revenue**, and start experimenting now!

Re-think education to be less occupation-oriented



THE INTERNATIONAL BESTSELLER

UTOPIA FOR REALISTS

AND HOW WE CAN GET THERE

'This book is brilliant. Everyone should read it' Richard Wilkinson, author of *The Spirit Level*

RUTGER BREGMAN

BLOOMSBURY

In the meantime, aim at the **triangle of opportunity**: be technically literate (learn how to program), work on your emotional intelligence, develop a taste for creativity and the "human touch".

SLEEPWALKING INTO THE FUTURE?

Future of Life Institute: open letter agains autonomous weapons, manages research grants in ethics of AI

New **research centers**: Levehulme Center for the Future of Intelligence (Cambridge), Strategic AI Research Center (Oxford), Center for Human-Compatible AI (Berkeley)…

The **partnership on AI** (2016): Google, Amazon, Facebook, IBM and Microsoft, to "benefit people and society" (hopefully not only PR…)

One hundred year study on AI, Stanford initiative from 2016: study long-term changes in "how people work, live and play".

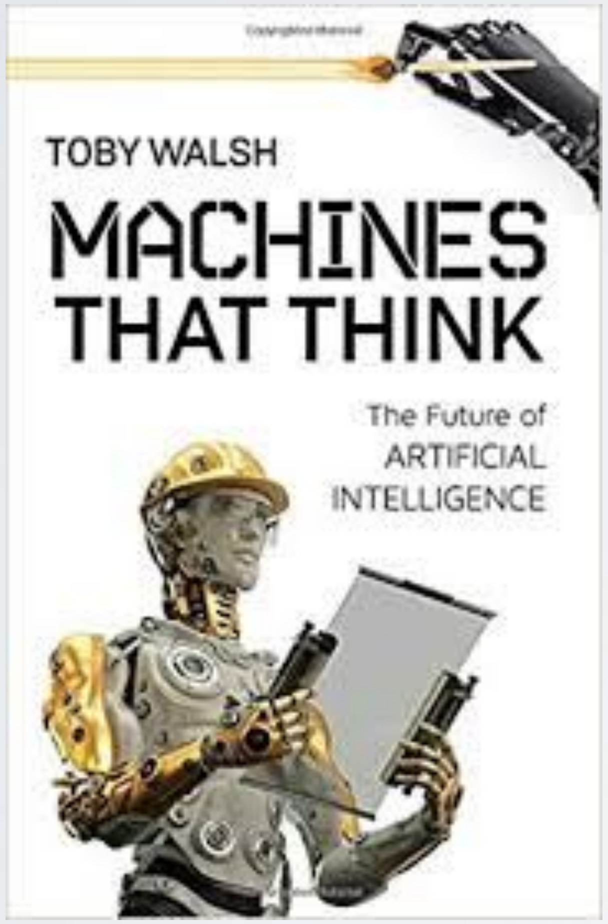
WE DID NOT TALK ABOUT

Killer bots and
autonomous weapons

Taxing robots, and
technological inequalities

Privacy and algorithmic discrimination

REFERENCES



Toby Walsh, Machines that Think, 2018

Work by Nick Bostrom and Moshe Vardi

A multitude of internet posts

Slides by M. Slavkovik at EASSS 2017