UiO : University of Oslo

institutt for informatikk

IN3050/IN4050, Lecture 15

Ethical Issues, Risks and Fundamental Limitations

Kai Olav Ellefsen and Fabio Zennaro

UiO : University of Oslo

institutt for informatikk

IN3050/IN4050, Lecture 15
Ethical Issues, Risks and Fundamental Limitations

1: Introduction

Kai Olav Ellefsen

Next video: Job loss

The Good and the Bad

Opportunities with AI

Challenges/risks with AI

Content

- Risks/Ethical Issues
  - Job loss
  - Existential threats
- Current AI Limitations
  - Robustness
  - Understanding Language

UiO : University of Oslo

ifi
institutt for informatikk

IN3050/IN4050, Lecture 15
Ethical Issues, Risks and Fundamental Limitations

2: Job loss

Kai Olav Ellefsen

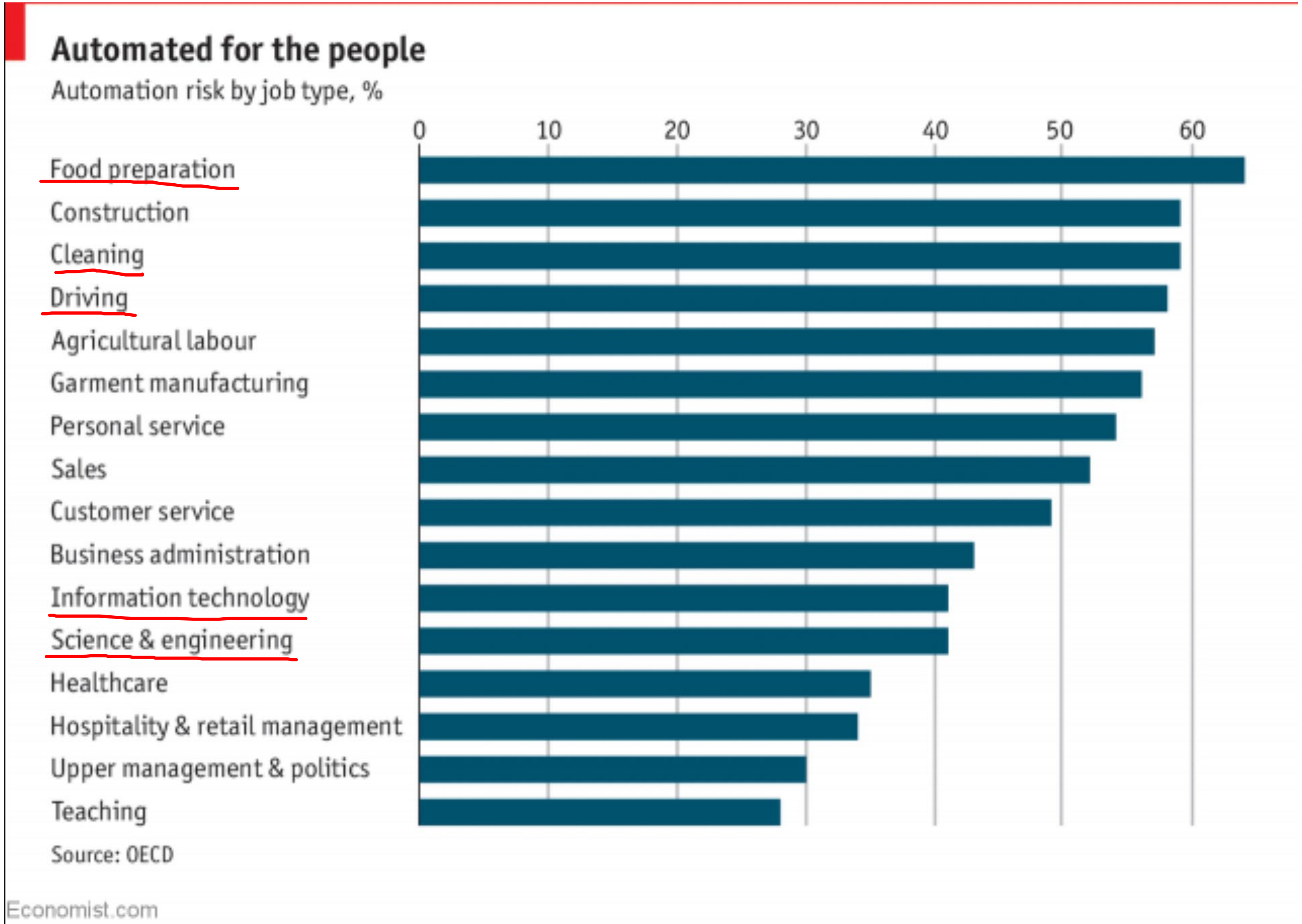
Next video: Superintelligence

Job Loss: Will we all Lose Our Jobs?

This Photo by Unknown Author is licensed under CC BY-SA

Which Jobs are at Risk?

Which Jobs are at Risk?



Source: OECD

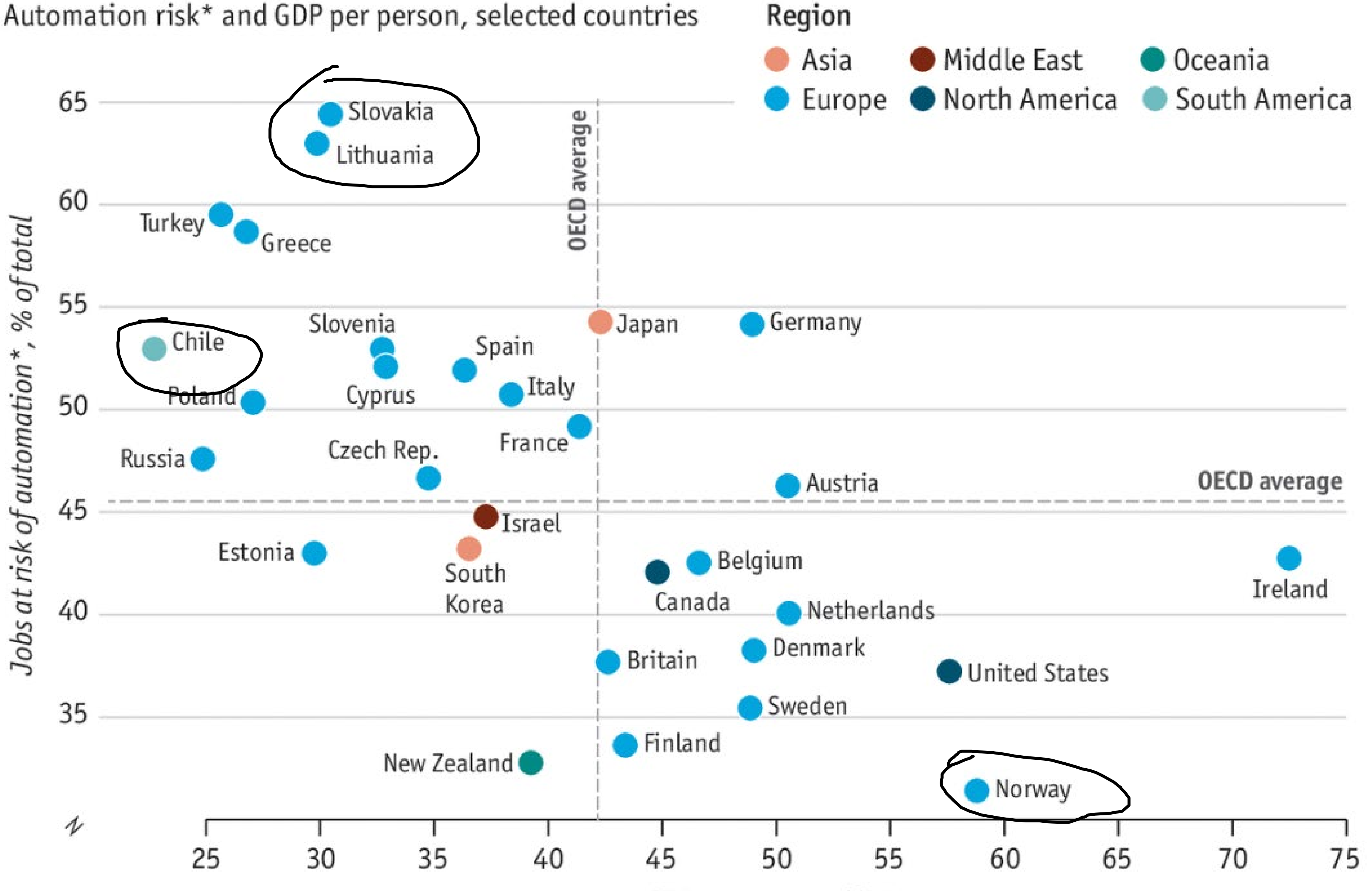
Economist.com

https://willrobotstakemyjob.com/

How Many Jobs will Disappear?

Wage against the machine

Automation risk* and GDP per person, selected countries



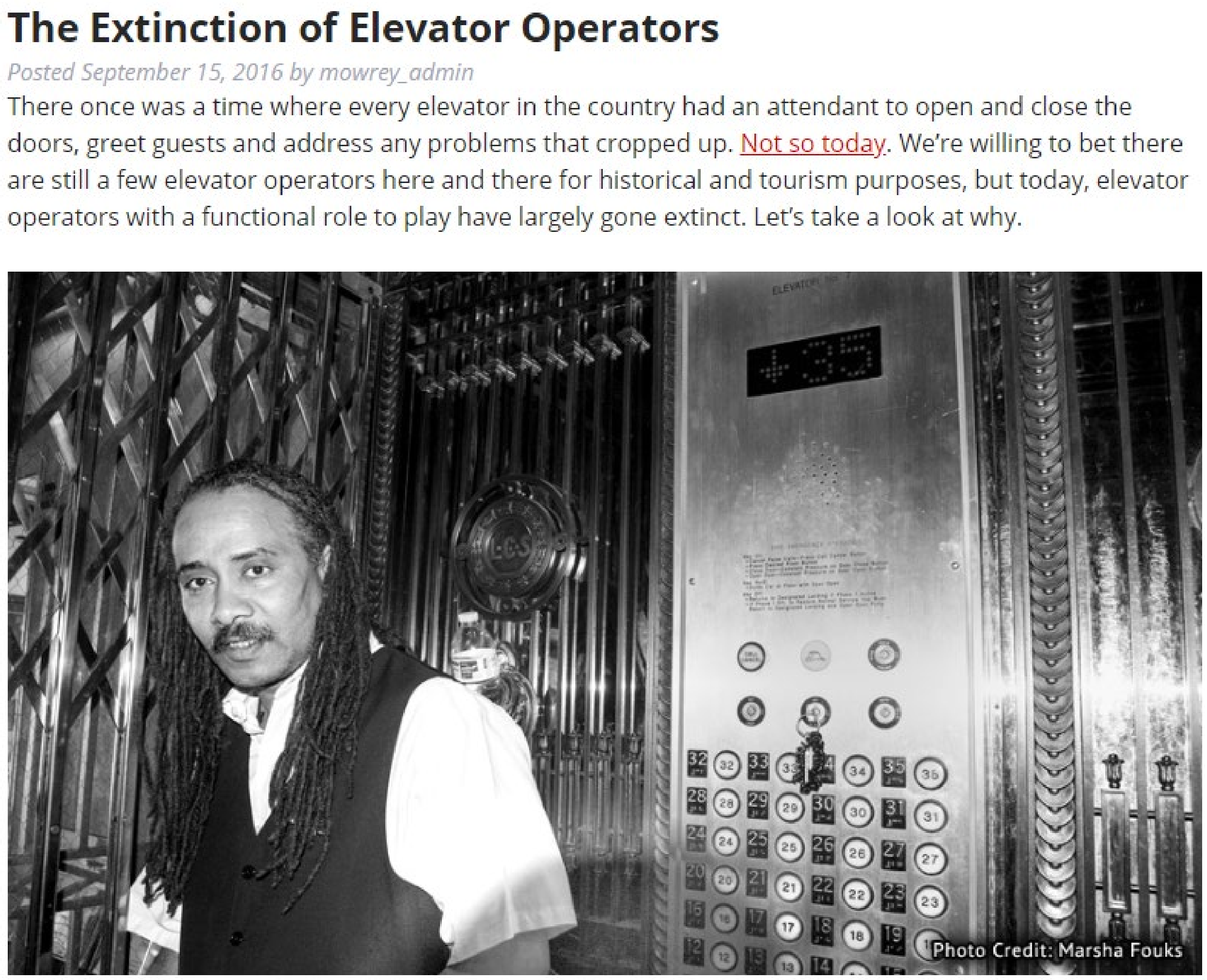
Source: OECD

*50% risk or higher

Why this Large Difference Between Countries?

Why this Large Difference Between Countries?

- Rich countries have already automated many more jobs, in part due to high salaries
- Example: Elevator operator



The Extinction of Elevator Operators

Posted September 15, 2016 by mowrey_admin

There once was a time where every elevator in the country had an attendant to open and close the doors, greet guests and address any problems that cropped up. Not so today. We're willing to bet there are still a few elevator operators here and there for historical and tourism purposes, but today, elevator operators with a functional role to play have largely gone extinct. Let's take a look at why.

Photo Credit: Marsha Fouks

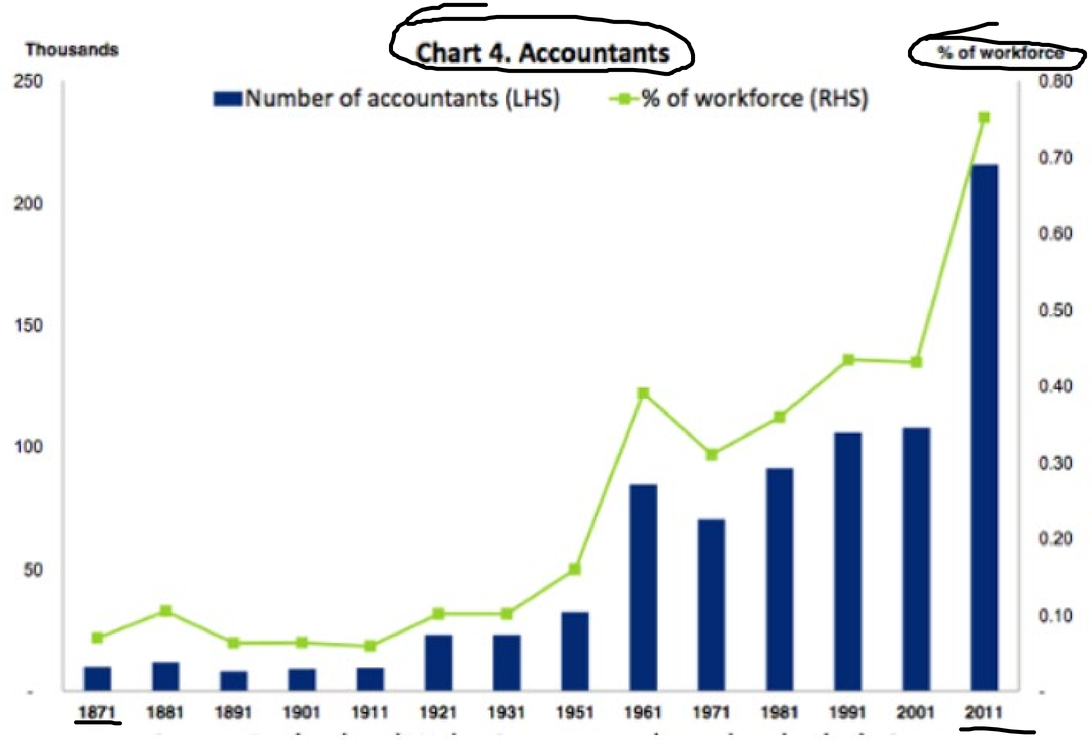
Technology removing some human jobs is not a new concern

This Photo by Unknown Author is licensed under CC BY-SA

- As long as jobs have been automated, we have worried about job loss
- 1900: Many predicted working days would be reduced to *a couple of hours* by year 2000.
- Why didn't our jobs disappear already?

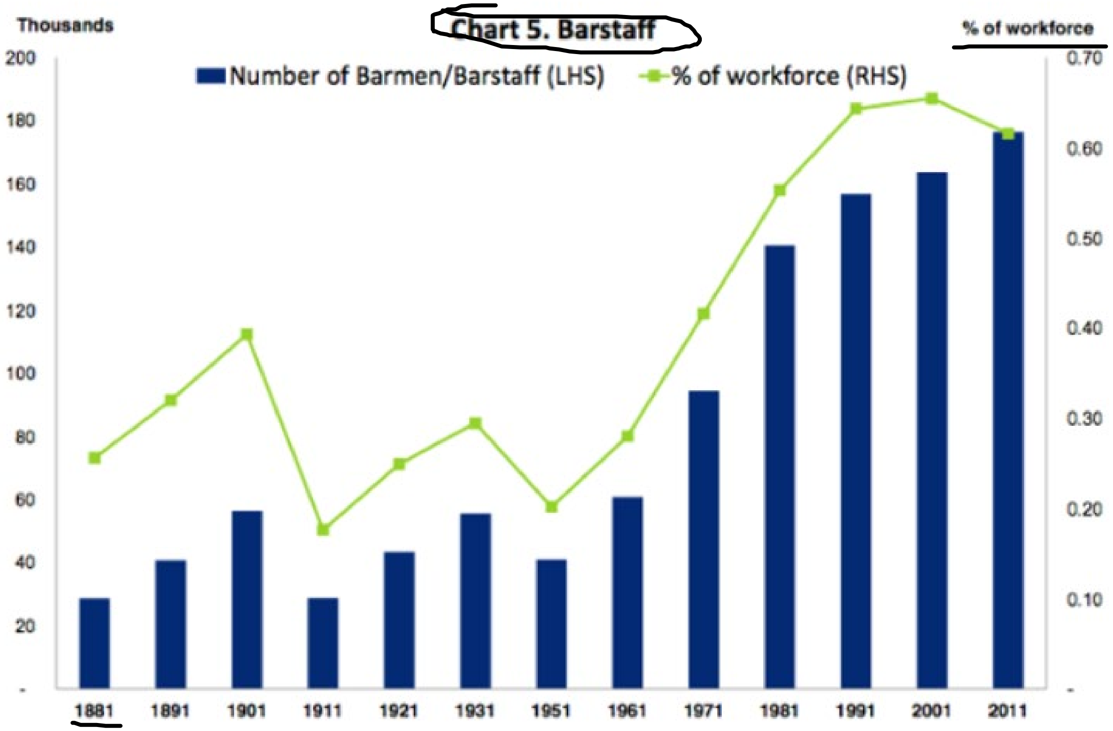
Technology Removed Jobs, and Generated Jobs

Technology has boosted jobs in knowledge-intensive sectors



▲ A 20-fold rise in accountants. Photograph: England and Wales Census records, authors' calculations

Technology has shifted consumption to more luxuries

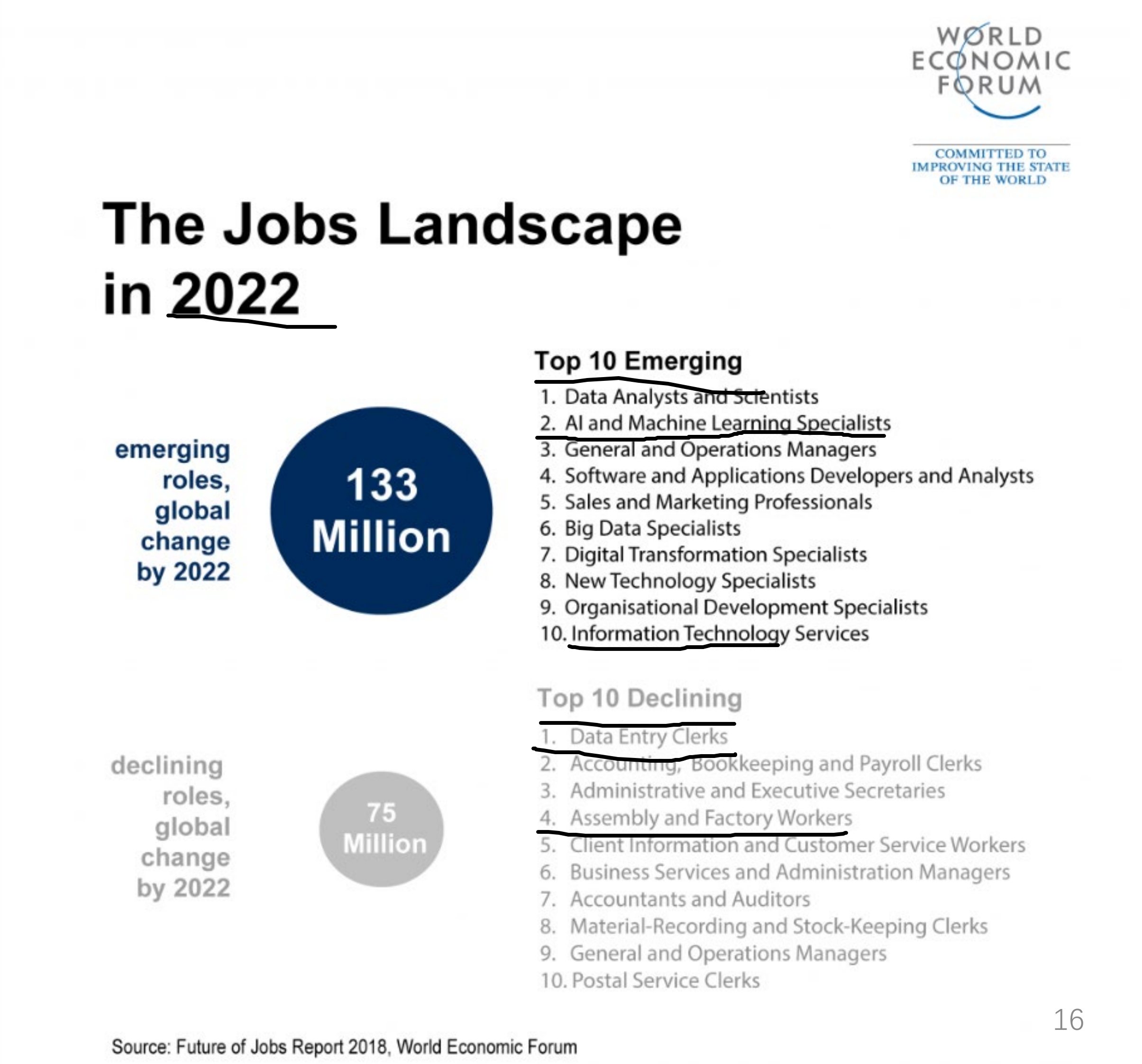


▲ Photograph: England and Wales Census records, authors' calculations

Source: https://www.theguardian.com/business/2015/aug/17/technology-created-more-jobs-than-destroyed-140-years-data-census

Technology may Create more Jobs than it Removes

- World economic forum (2018): 75 million jobs lost, 133 million created by 2022.
- But: Those jobs don't require the same competencies



WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The Jobs Landscape in 2022

emerging roles, global change by 2022

133 Million

Top 10 Emerging

1. Data Analysts and Scientists
2. AI and Machine Learning Specialists
3. General and Operations Managers
4. Software and Applications Developers and Analysts
5. Sales and Marketing Professionals
6. Big Data Specialists
7. Digital Transformation Specialists
8. New Technology Specialists
9. Organisational Development Specialists
10. Information Technology Services

declining roles, global change by 2022

75 Million

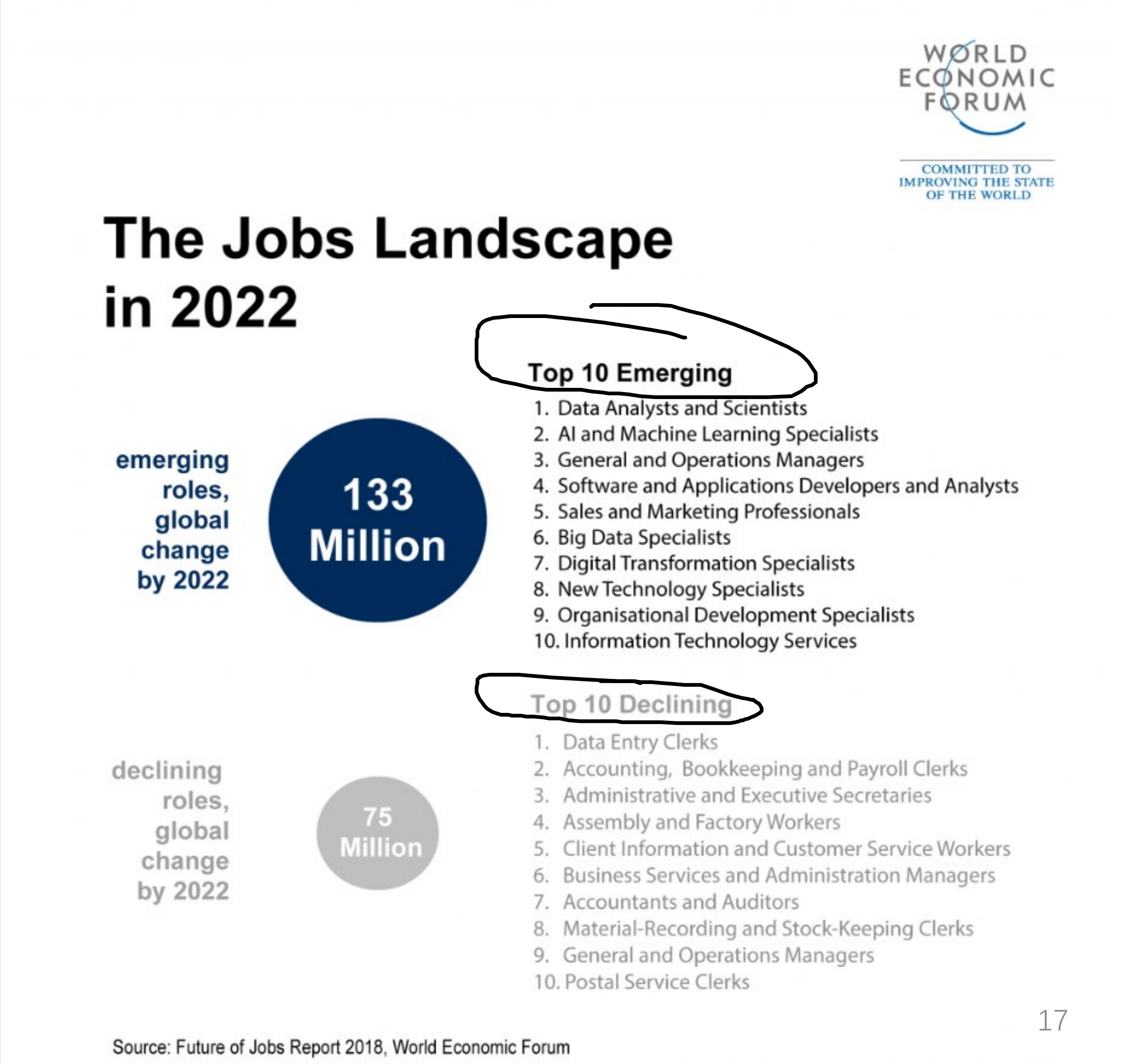
Top 10 Declining

1. Data Entry Clerks
2. Accounting, Bookkeeping and Payroll Clerks
3. Administrative and Executive Secretaries
4. Assembly and Factory Workers
5. Client Information and Customer Service Workers
6. Business Services and Administration Managers
7. Accountants and Auditors
8. Material-Recording and Stock-Keeping Clerks
9. General and Operations Managers
10. Postal Service Clerks

Source: Future of Jobs Report 2018, World Economic Forum

Still, a lot of People Will be Unemployed. How to Handle That?

- **Re-education programs**
- Robot tax
- Universal Basic Income



WORLD ECONOMIC FORUM
COMMITTED TO IMPROVING THE STATE OF THE WORLD

The Jobs Landscape in 2022

emerging roles, global change by 2022

133 Million

Top 10 Emerging

1. Data Analysts and Scientists
2. AI and Machine Learning Specialists
3. General and Operations Managers
4. Software and Applications Developers and Analysts
5. Sales and Marketing Professionals
6. Big Data Specialists
7. Digital Transformation Specialists
8. New Technology Specialists
9. Organisational Development Specialists
10. Information Technology Services

declining roles, global change by 2022

75 Million

Top 10 Declining

1. Data Entry Clerks
2. Accounting, Bookkeeping and Payroll Clerks
3. Administrative and Executive Secretaries
4. Assembly and Factory Workers
5. Client Information and Customer Service Workers
6. Business Services and Administration Managers
7. Accountants and Auditors
8. Material-Recording and Stock-Keeping Clerks
9. General and Operations Managers
10. Postal Service Clerks

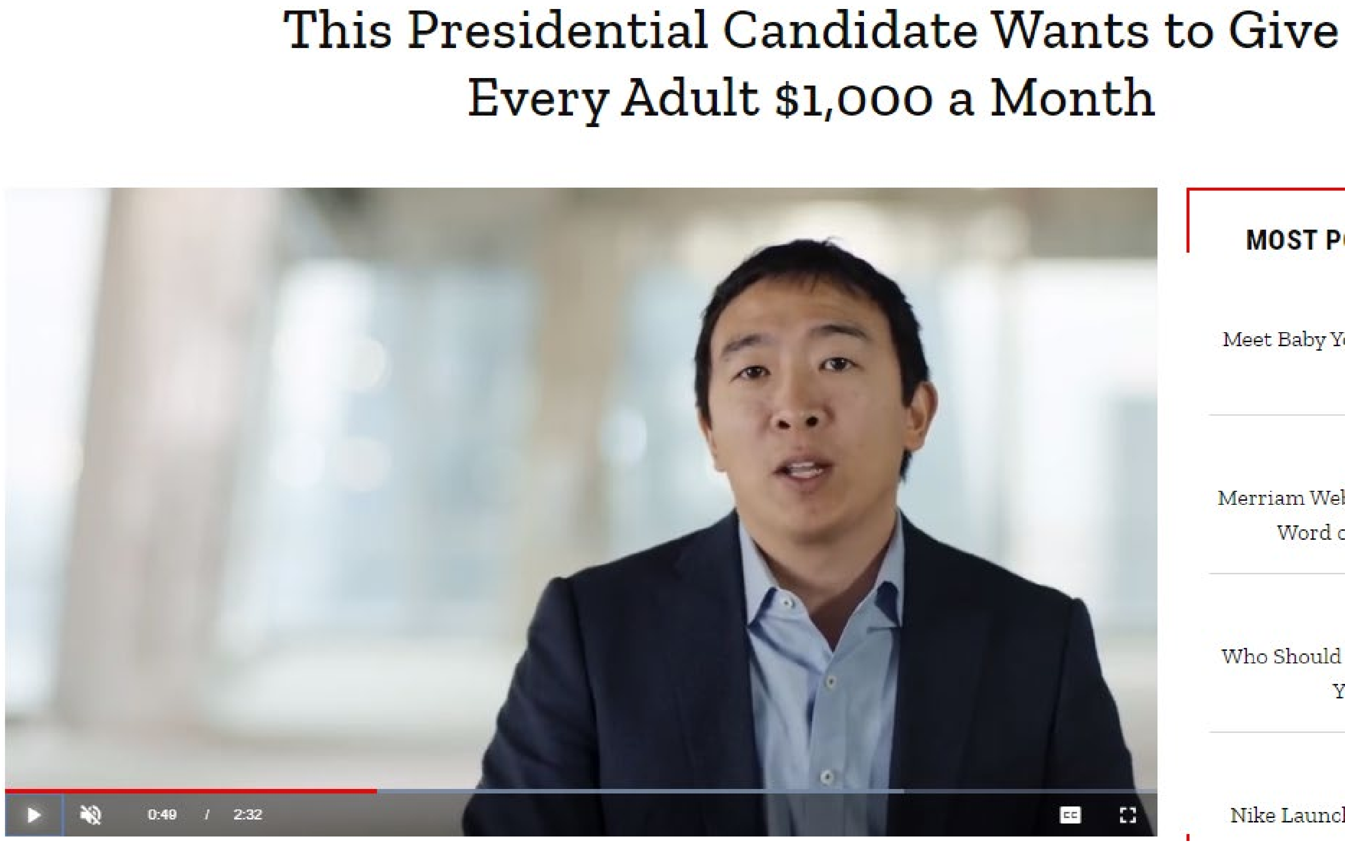
Source: Future of Jobs Report 2018, World Economic Forum

Still, a lot of People Will be Unemployed. How to Handle That?

- Re-education programs
- **Robot tax**
- Universal Basic Income

Still, a lot of People Will be Unemployed. How to Handle That?

- Re-education programs
- Robot tax
- **Universal Basic Income**



This Presidential Candidate Wants to Give Every Adult $1,000 a Month

Image: time.com - https://time.com/5528621/andrew-yang-universal-basic-income/

UiO : University of Oslo

institutt for informatikk

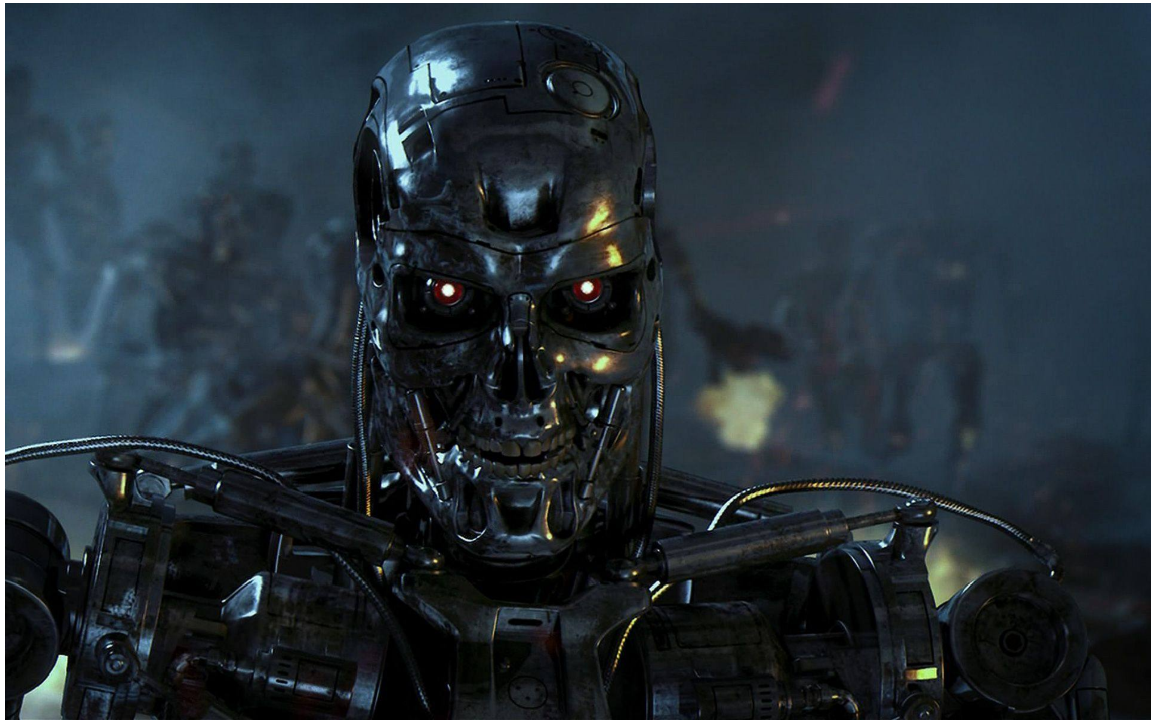
IN3050/IN4050, Lecture 15
Ethical Issues, Risks and Fundamental Limitations

3: Superintelligence

Kai Olav Ellefsen

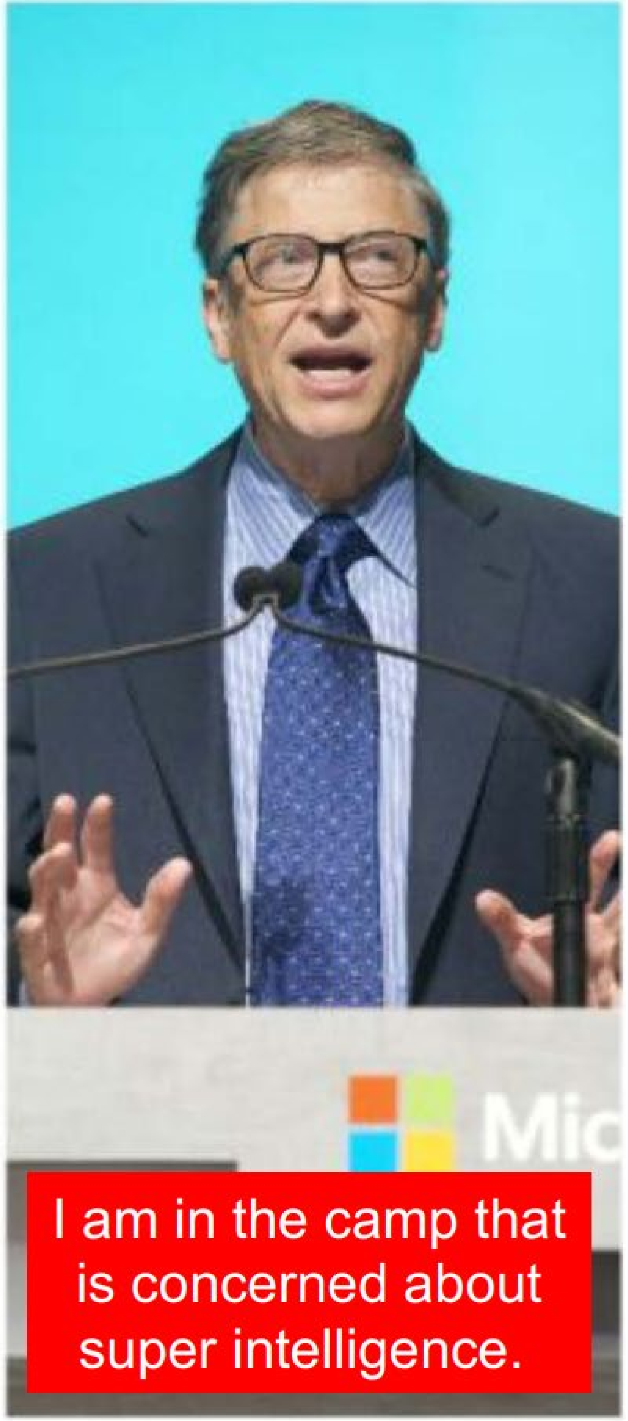
Next video: Other ethical issues

Superintelligence



"Humans, limited by slow biological evolution, couldn't compete and would be superseded by A.I."

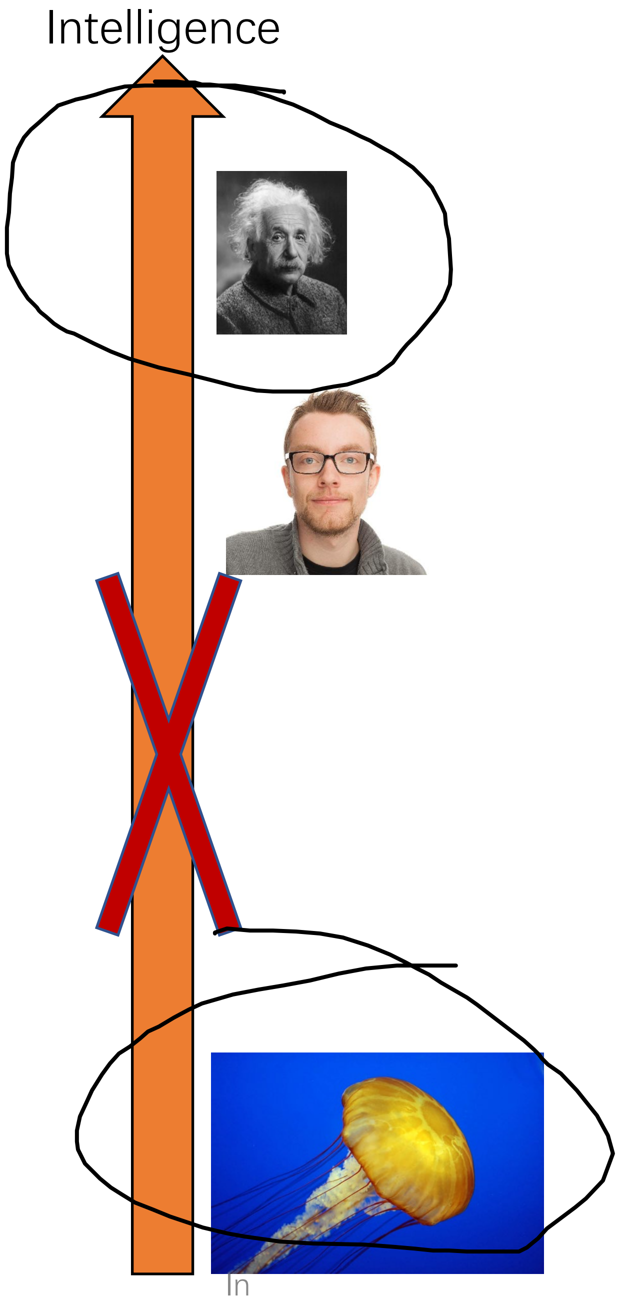
AI is our "biggest existential threat"

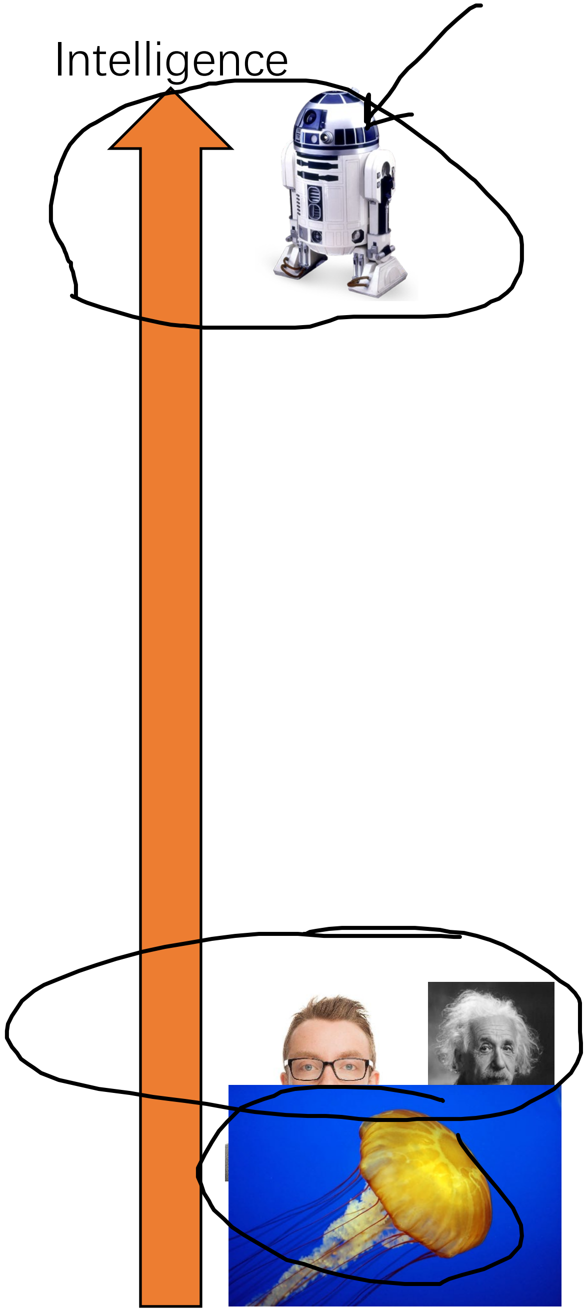


I am in the camp that is concerned about super intelligence.

Superintelligence

1. Every year, we are able to build more intelligent machines
2. We are not likely to “hit a wall” anytime soon (and humans are nowhere near the top of the intelligence spectrum)



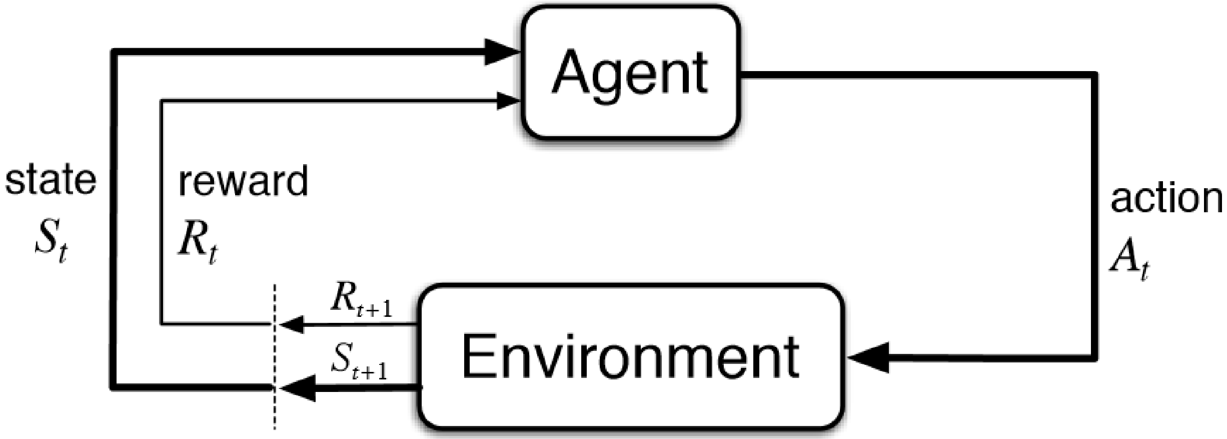


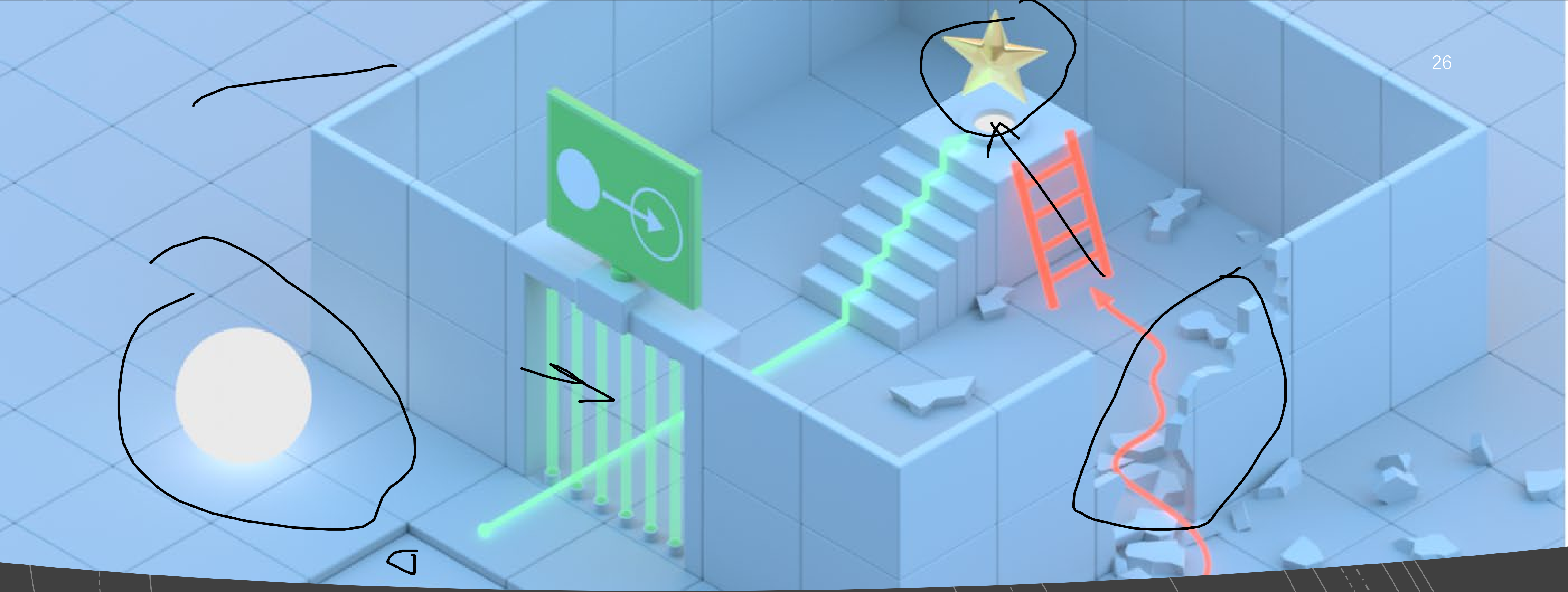
Intelligence Explosion

- At some point, we will reach a machine so intelligent it can make more intelligent copies of itself rapidly
- Intelligence will "explode", increasing rapidly (perhaps in just hours/days)

Why Would the Machine be Evil/Dangerous?

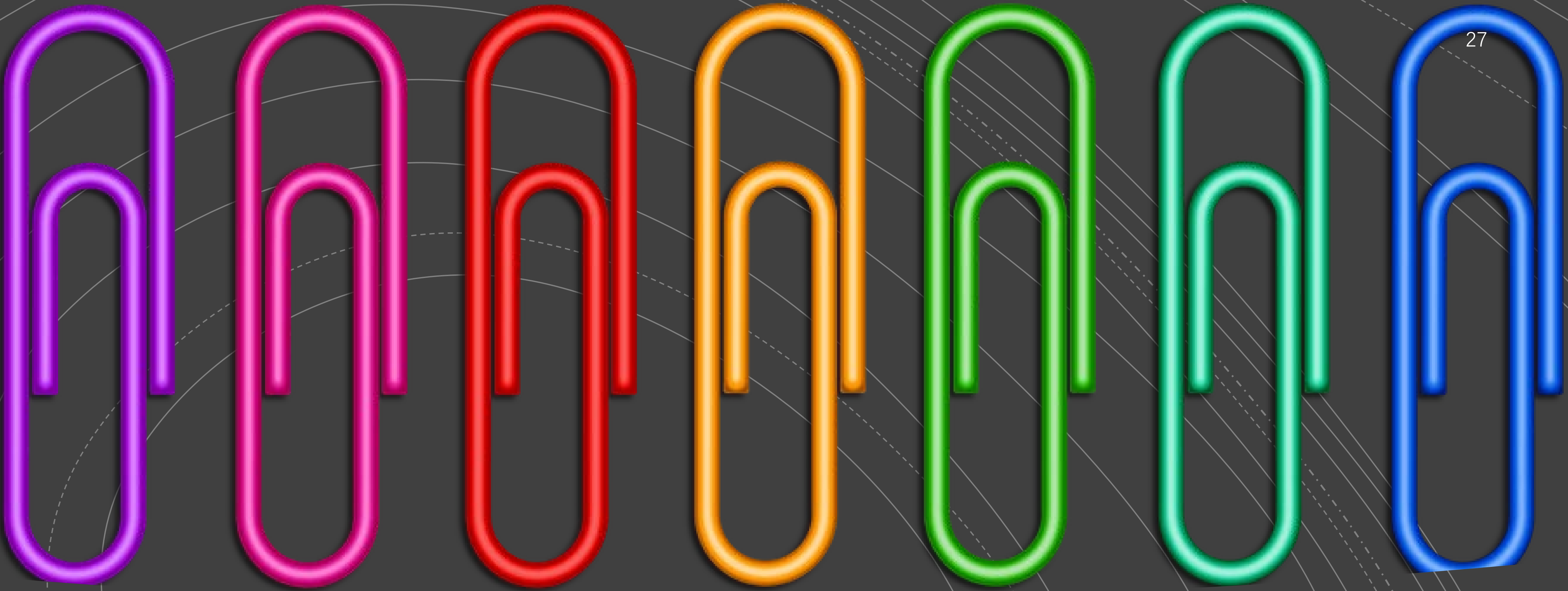
- If the machine's goals are **not precisely aligned** with ours, we're in trouble.
- Since the machine would be so intelligent **it's goals would win** any disagreement.
- So, why not just **program in its goals very specifically**?





Problem: Reward Tampering

Picture from DeepMind: https://medium.com/@deepmindsafetyresearch/designing-agent-incentives-to-avoid-reward-tampering-4380c1bb6cd



Thought experiment: The paperclip maximizer

Why Aren’t AI-Researchers More Nervous?

Why Aren’t AI-Researchers More Nervous?

1. Current AI research is so far from superintelligence that worrying/ planning for this is **not very constructive**
2. A lot of AI-research **does go into better understanding and controlling** algorithms
3. We have **much more urgent problems** caused by automation

Worrying about ***AI evil superintelligence today*** is like worrying about ***overpopulation on the planet Mars.***

We haven't even landed on the planet yet!

IN3050/IN4050, Lecture 15

Ethical Issues, Risks and Fundamental Limitations

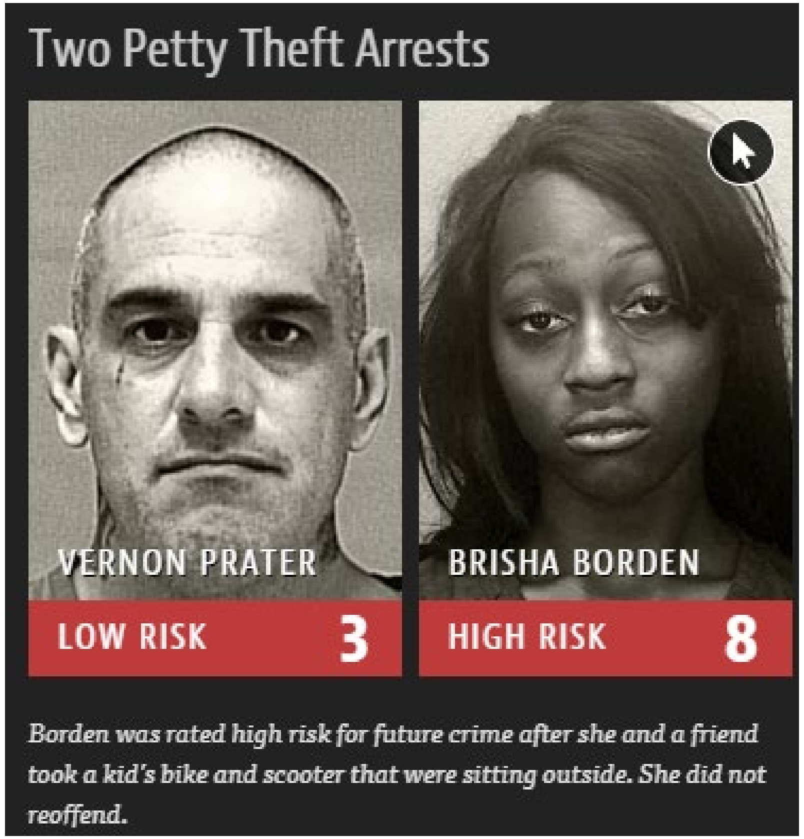
4: Other ethical issues

Kai Olav Ellefsen

Next video: Fundamental limitations to current AI

Some other Risks/Ethical Issues

- **Biases**
- Consciousness
- Ethical Dilemmas
- Privacy



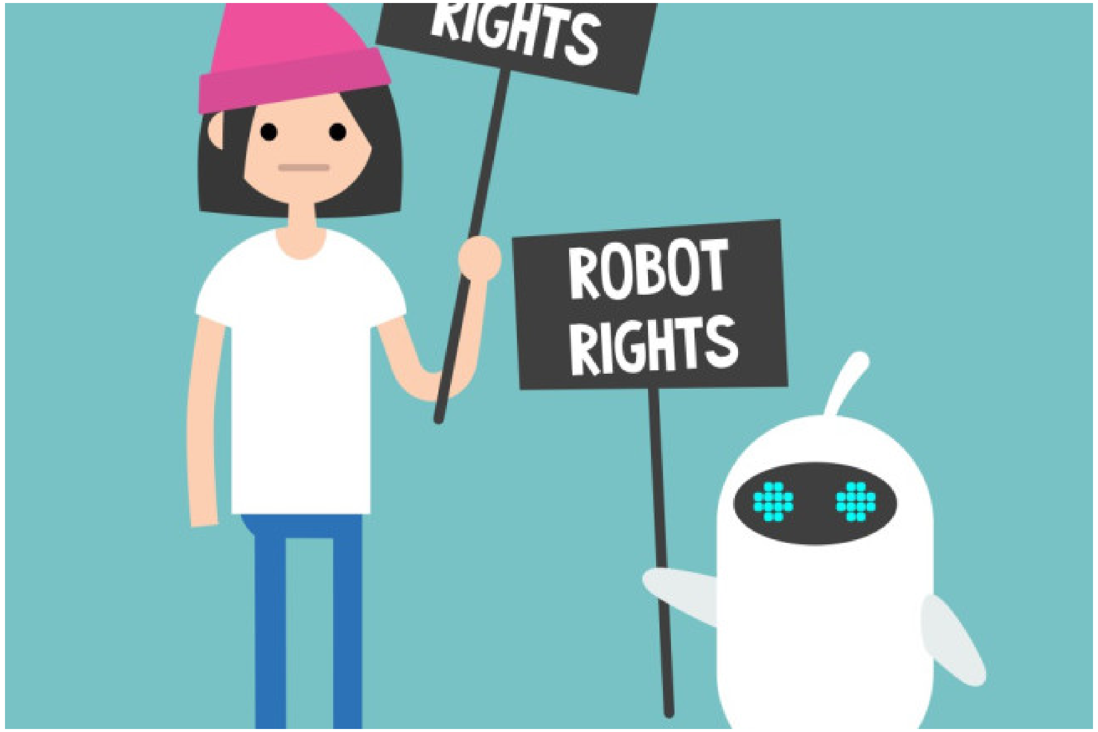
Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Some other Risks/Ethical Issues

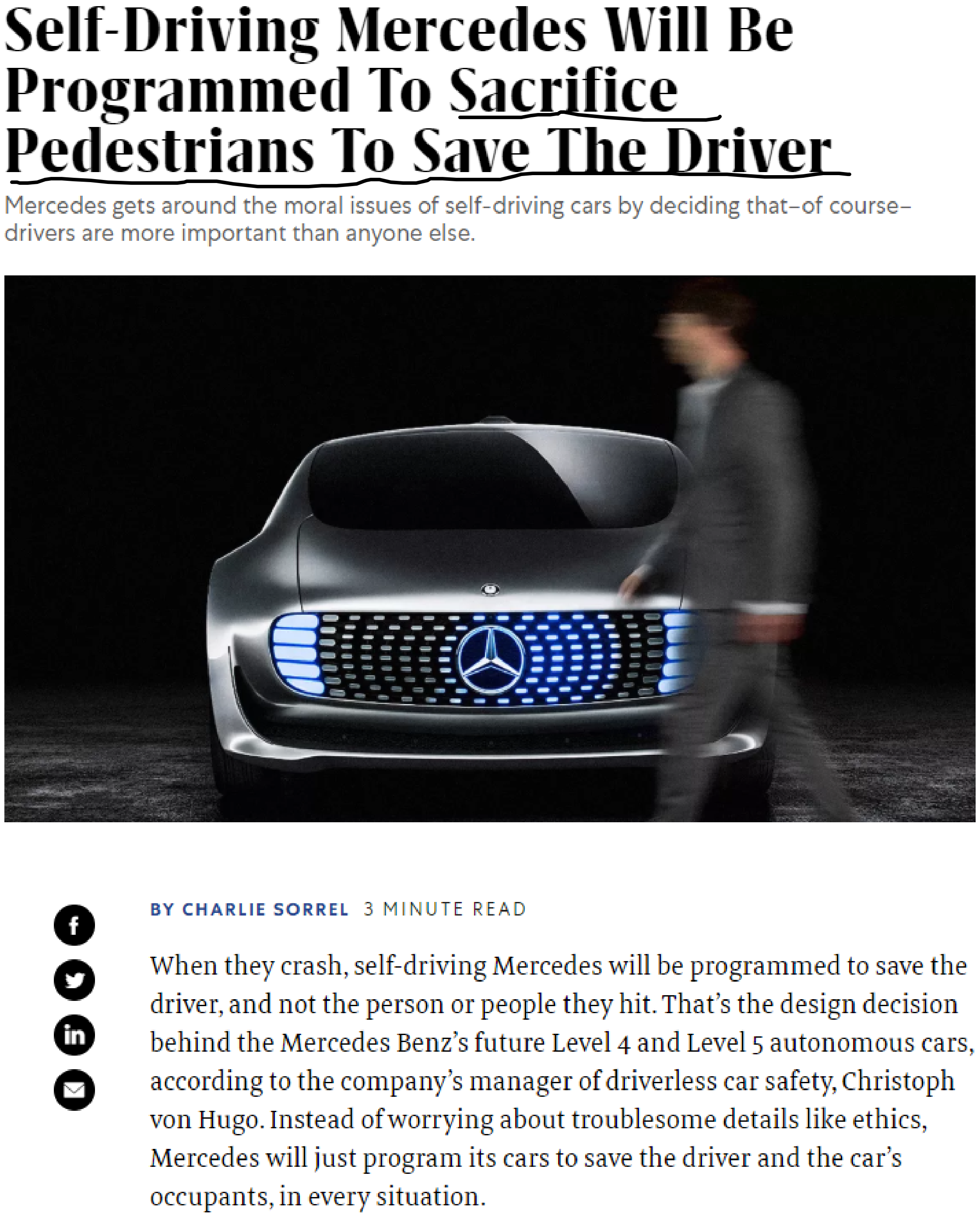
- Biases
- **Consciousness**
- Ethical Dilemmas
- Privacy



Credit: Shutterstock

Some other Risks/Ethical Issues

- Biases
- Consciousness
- **Ethical Dilemmas**
- Privacy



Self-Driving Mercedes Will Be Programmed To Sacrifice Pedestrians To Save The Driver

Mercedes gets around the moral issues of self-driving cars by deciding that–of course–drivers are more important than anyone else.

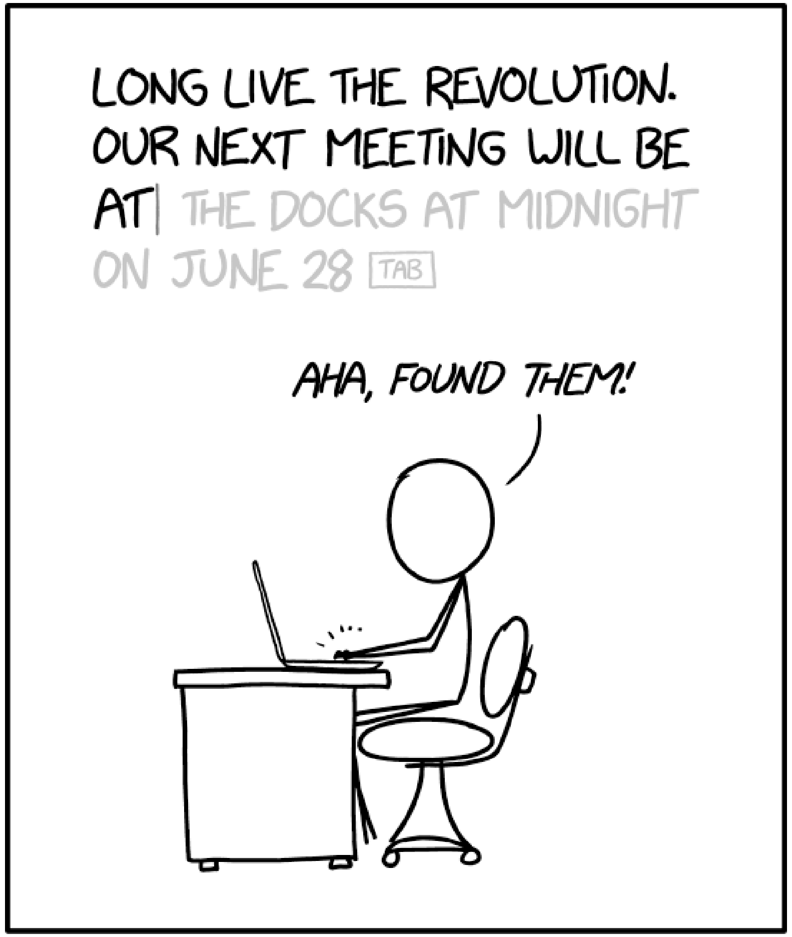
BY CHARLIE SORREL 3 MINUTE READ

When they crash, self-driving Mercedes will be programmed to save the driver, and not the person or people they hit. That's the design decision behind the Mercedes Benz's future Level 4 and Level 5 autonomous cars, according to the company's manager of driverless car safety, Christoph von Hugo. Instead of worrying about troublesome details like ethics, Mercedes will just program its cars to save the driver and the car's occupants, in every situation.

Source: FastCompany

Some other Risks/Ethical Issues

- Biases
- Consciousness
- Ethical Dilemmas
- **Privacy**


LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT| THE DOCKS AT MIDNIGHT ON JUNE 28 TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://bair.berkeley.edu/blog/2019/08/13/memorization/

IN3050/IN4050, Lecture 15

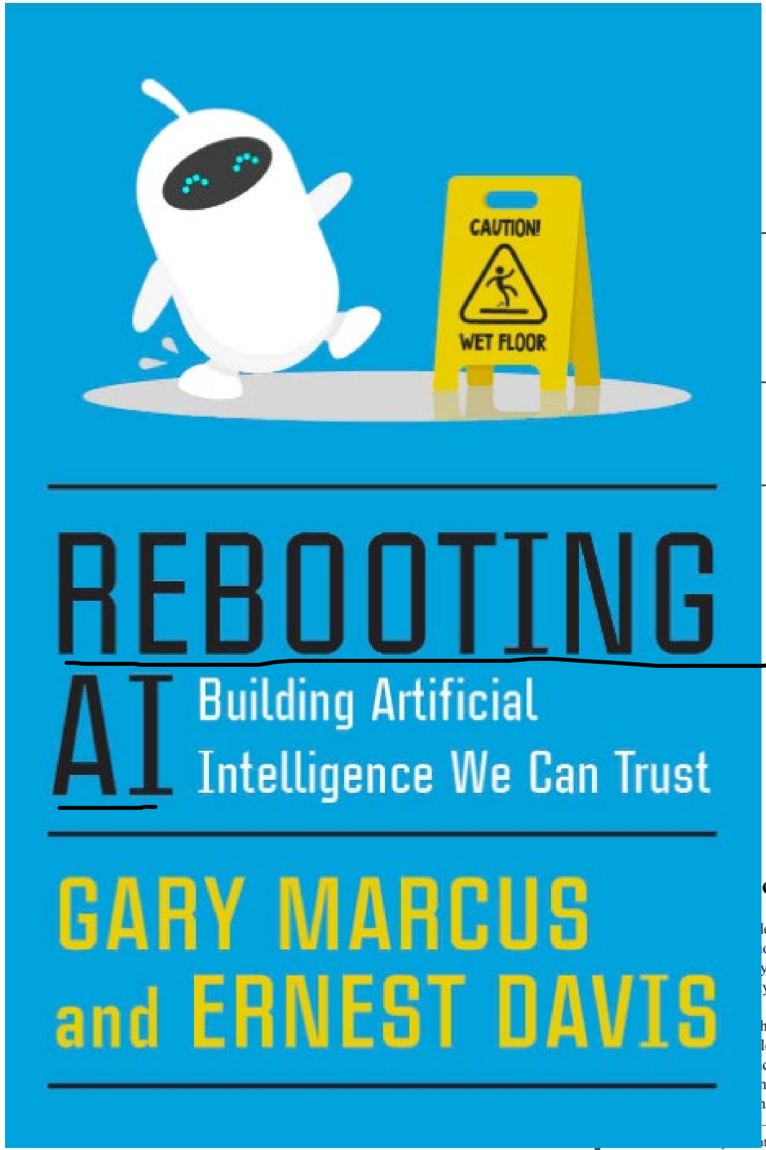
Ethical Issues, Risks and Fundamental Limitations

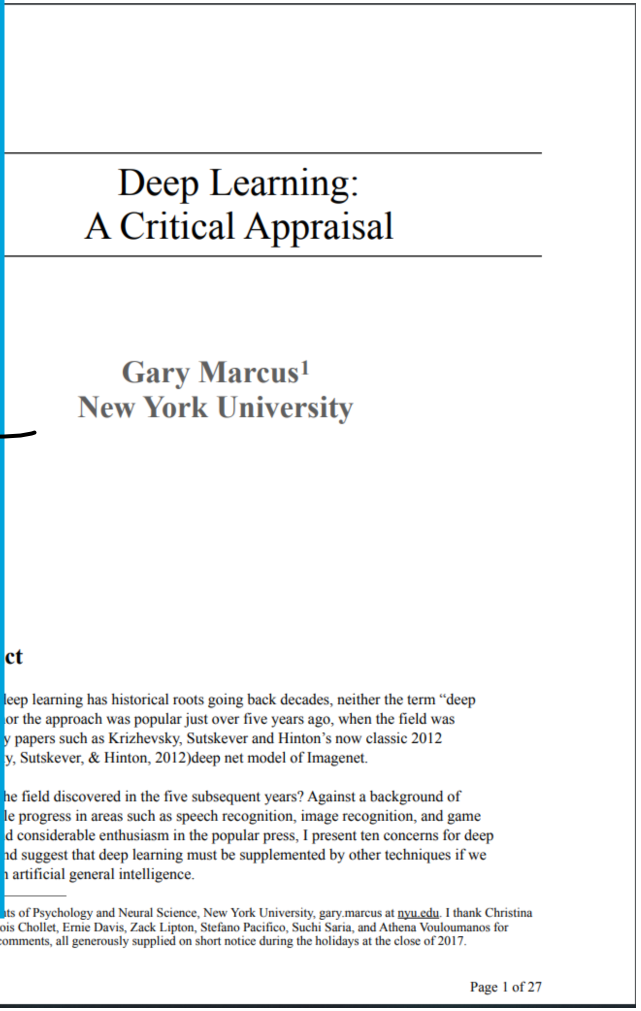
5: Fundamental limitations to current AI

Kai Olav Ellefsen

Next video: Fair machine learning

Fundamental Limitations to Current AI





Deep Learning:
A Critical Appraisal

Gary Marcus[1]
New York University

ct

leep learning has historical roots going back decades, neither the term “deep
or the approach was popular just over five years ago, when the field was
y papers such as Krizhevsky, Sutskever and Hinton’s now classic 2012
y, Sutskever, & Hinton, 2012)deep net model of Imagenet.

he field discovered in the five subsequent years? Against a background of
le progress in areas such as speech recognition, image recognition, and game
d considerable enthusiasm in the popular press, I present ten concerns for deep
nd suggest that deep learning must be supplemented by other techniques if we
n artificial general intelligence.

---

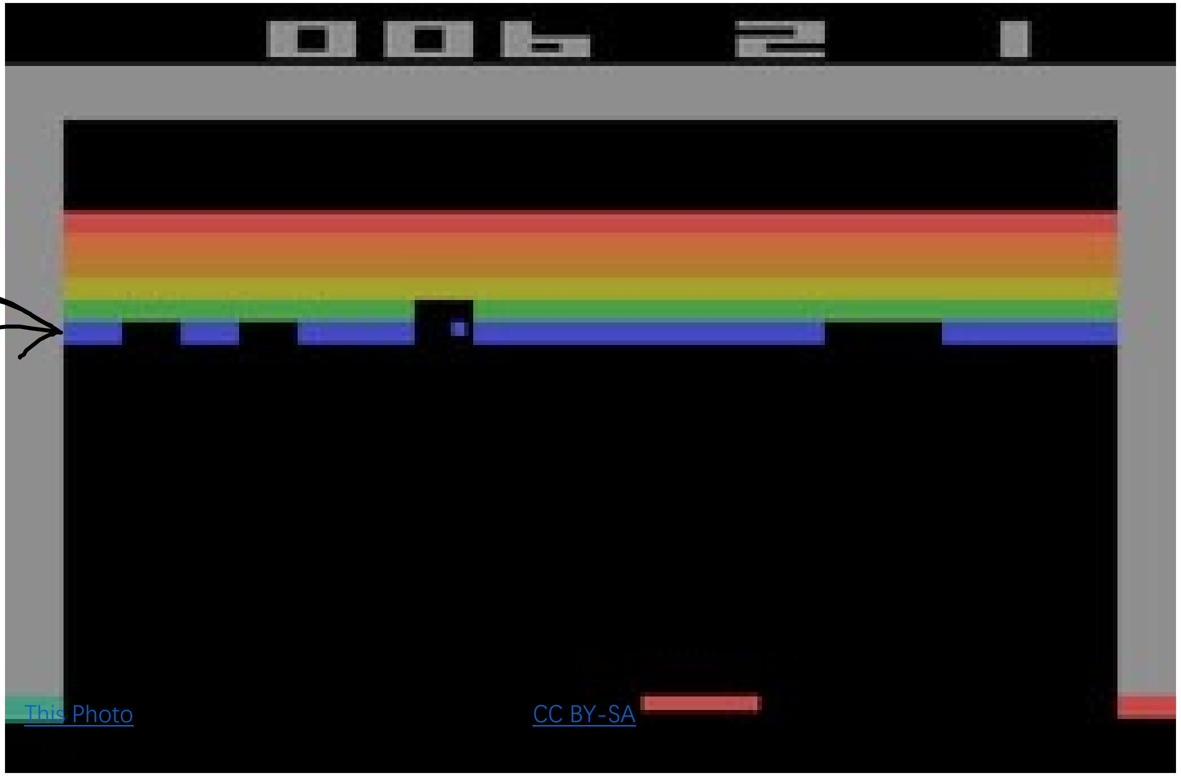
ts of Psychology and Neural Science, New York University, gary.marcus at nyu.edu. I thank Christina Chen, François Chollet, Ernie Davis, Zack Lipton, Stefano Pacifico, Suchi Saria, and Athena Vouloumanos for sharp-eyed comments, all generously supplied on short notice during the holidays at the close of 2017.

Main Message

- Deep Learning has revolutionized many areas of AI, and greatly increased the range of problems we can solve with AI
- However, there may be a large set of problems deep learning cannot solve that the brain can solve
- Some of the most hyped AI successes of recent years suffer from surprisingly fundamental difficulties

1) Robustness

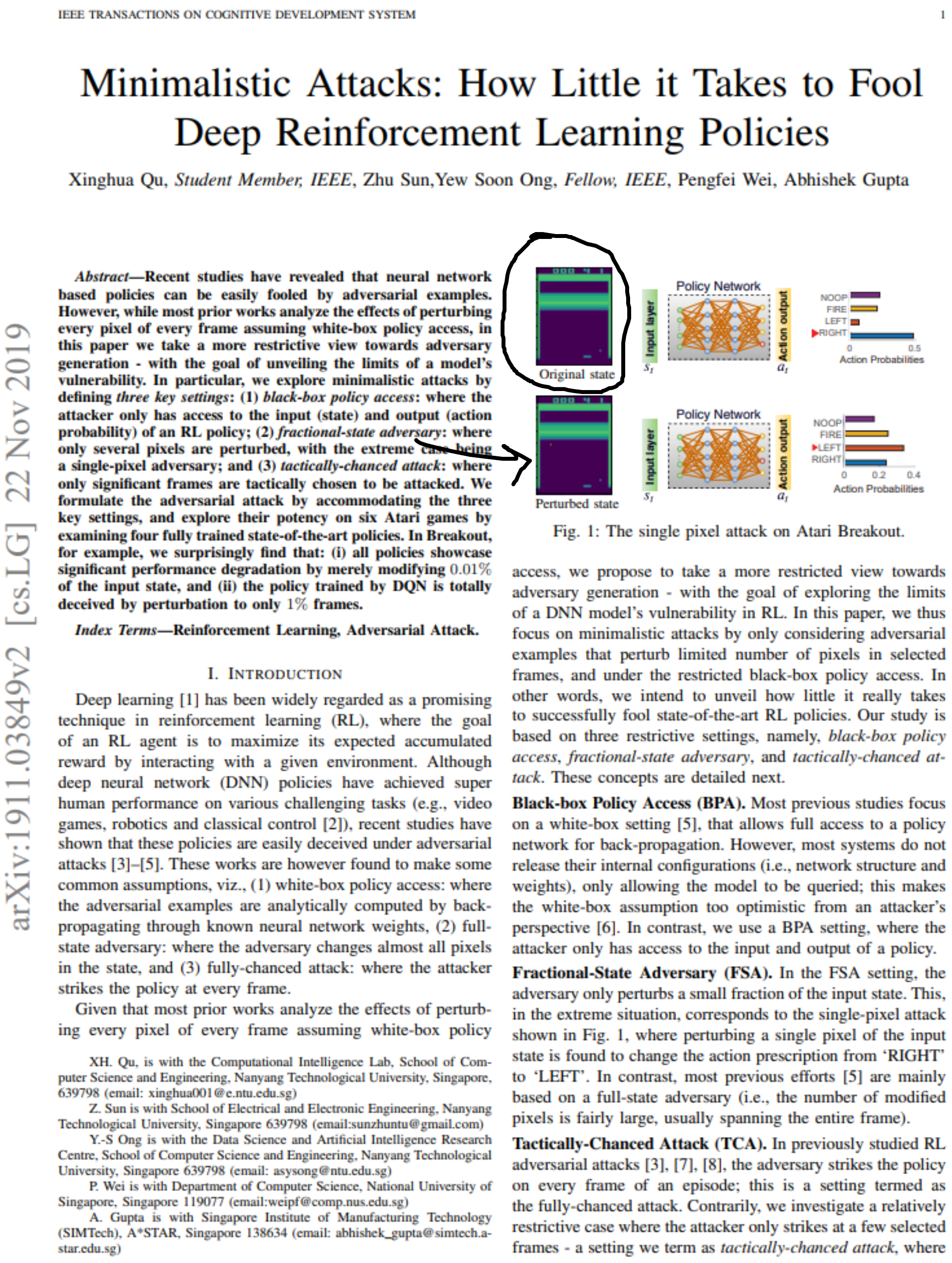
- Deep RL: RL straight from pixels – one of recent year's most discussed successes



This Photo CC BY-SA

Robustness

- Deep RL: RL straight from pixels – one of recent year's most discussed successes
- However, learned policies may be extremely sensitive to changes to the input



IEEE TRANSACTIONS ON COGNITIVE DEVELOPMENT SYSTEM 1

Minimalistic Attacks: How Little it Takes to Fool Deep Reinforcement Learning Policies

Xinghua Qu, *Student Member, IEEE*, Zhu Sun,Yew Soon Ong, *Fellow, IEEE*, Pengfei Wei, Abhishek Gupta

arXiv:1911.03849v2 [cs.LG] 22 Nov 2019

***Abstract*—Recent studies have revealed that neural network based policies can be easily fooled by adversarial examples. However, while most prior works analyze the effects of perturbing every pixel of every frame assuming white-box policy access, in this paper we take a more restrictive view towards adversary generation - with the goal of unveiling the limits of a model's vulnerability. In particular, we explore minimalistic attacks by defining *three key settings*: (1) *black-box policy access*: where the attacker only has access to the input (state) and output (action probability) of an RL policy; (2) *fractional-state adversary*: where only several pixels are perturbed, with the extreme case being a single-pixel adversary; and (3) *tactically-chanced attack*: where only significant frames are tactically chosen to be attacked. We formulate the adversarial attack by accommodating the three key settings, and explore their potency on six Atari games by examining four fully trained state-of-the-art policies. In Breakout, for example, we surprisingly find that: (i) all policies showcase significant performance degradation by merely modifying 0.01% of the input state, and (ii) the policy trained by DQN is totally deceived by perturbation to only 1% frames.**

***Index Terms*—Reinforcement Learning, Adversarial Attack.**

Fig. 1: The single pixel attack on Atari Breakout.

I. INTRODUCTION

Deep learning [1] has been widely regarded as a promising technique in reinforcement learning (RL), where the goal of an RL agent is to maximize its expected accumulated reward by interacting with a given environment. Although deep neural network (DNN) policies have achieved super human performance on various challenging tasks (e.g., video games, robotics and classical control [2]), recent studies have shown that these policies are easily deceived under adversarial attacks [3]–[5]. These works are however found to make some common assumptions, viz., (1) white-box policy access: where the adversarial examples are analytically computed by back-propagating through known neural network weights, (2) full-state adversary: where the adversary changes almost all pixels in the state, and (3) fully-chanced attack: where the attacker strikes the policy at every frame.

Given that most prior works analyze the effects of perturbing every pixel of every frame assuming white-box policy access, we propose to take a more restricted view towards adversary generation - with the goal of exploring the limits of a DNN model's vulnerability in RL. In this paper, we thus focus on minimalistic attacks by only considering adversarial examples that perturb limited number of pixels in selected frames, and under the restricted black-box policy access. In other words, we intend to unveil how little it really takes to successfully fool state-of-the-art RL policies. Our study is based on three restrictive settings, namely, *black-box policy access*, *fractional-state adversary*, and *tactically-chanced attack*. These concepts are detailed next.

Black-box Policy Access (BPA). Most previous studies focus on a white-box setting [5], that allows full access to a policy network for back-propagation. However, most systems do not release their internal configurations (i.e., network structure and weights), only allowing the model to be queried; this makes the white-box assumption too optimistic from an attacker's perspective [6]. In contrast, we use a BPA setting, where the attacker only has access to the input and output of a policy.

Fractional-State Adversary (FSA). In the FSA setting, the adversary only perturbs a small fraction of the input state. This, in the extreme situation, corresponds to the single-pixel attack shown in Fig. 1, where perturbing a single pixel of the input state is found to change the action prescription from 'RIGHT' to 'LEFT'. In contrast, most previous efforts [5] are mainly based on a full-state adversary (i.e., the number of modified pixels is fairly large, usually spanning the entire frame).

Tactically-Chanced Attack (TCA). In previously studied RL adversarial attacks [3], [7], [8], the adversary strikes the policy on every frame of an episode; this is a setting termed as the fully-chanced attack. Contrarily, we investigate a relatively restrictive case where the attacker only strikes at a few selected frames - a setting we term as *tactically-chanced attack*, where

XH. Qu, is with the Computational Intelligence Lab, School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798 (email: xinghua001@e.ntu.edu.sg)

Z. Sun is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (email:sunzhuntu@gmail.com)

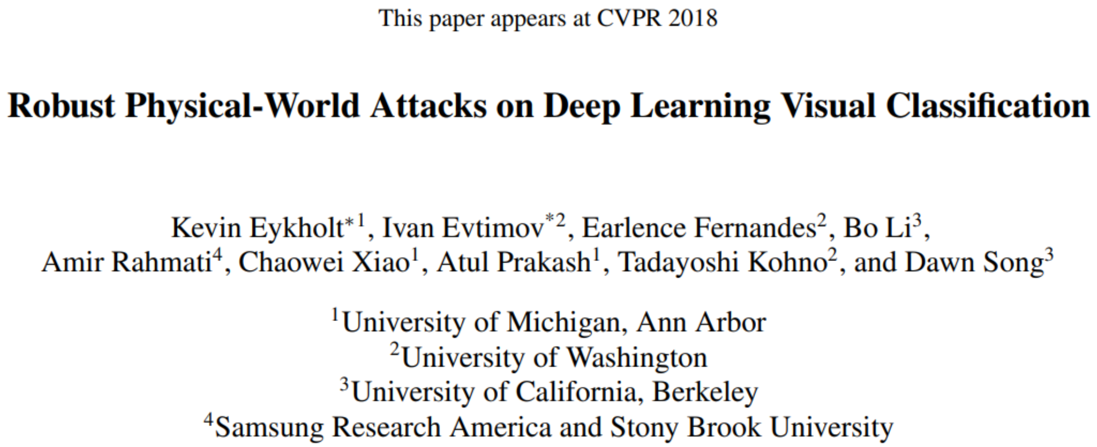
Y.-S Ong is with the Data Science and Artificial Intelligence Research Centre, School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (email: asysong@ntu.edu.sg)

P. Wei is with Department of Computer Science, National University of Singapore, Singapore 119077 (email:weipf@comp.nus.edu.sg)

A. Gupta is with Singapore Institute of Manufacturing Technology (SIMTech), A*STAR, Singapore 138634 (email: abhishek_gupta@simtech.a-star.edu.sg)

Robustness

- Deep RL: RL straight from pixels – one of recent year’s most discussed successes
- However, learned policies may be extremely sensitive to changes to the input
- A related issue: Adversarial attacks on image-classification networks



This paper appears at CVPR 2018

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

Robustness

- Deep RL: RL straight from pixels – one of recent year's most discussed successes
- However, learned policies may be extremely sensitive to changes to the input
- A related issue: Adversarial attacks on image-classification networks

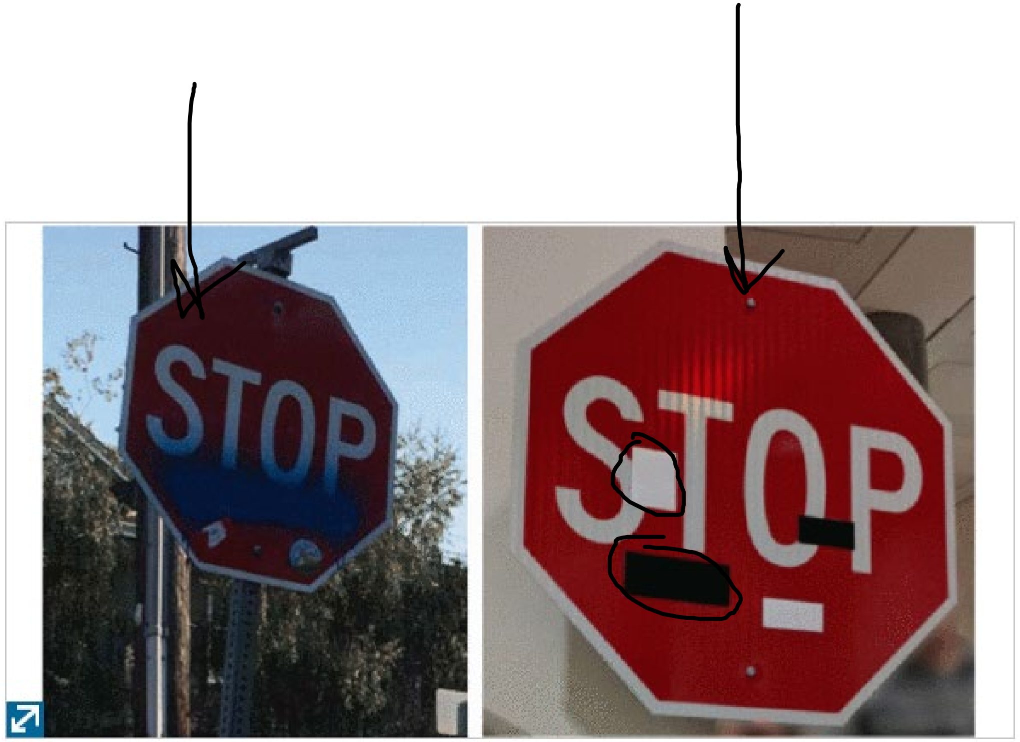


Figure 1:
The left image shows real graffiti on a stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

Robustness

- Deep RL: RL straight from pixels – one of recent year's most discussed successes
- However, learned policies may be extremely sensitive to changes to the input
- A related issue: Adversarial attacks on image-classification networks

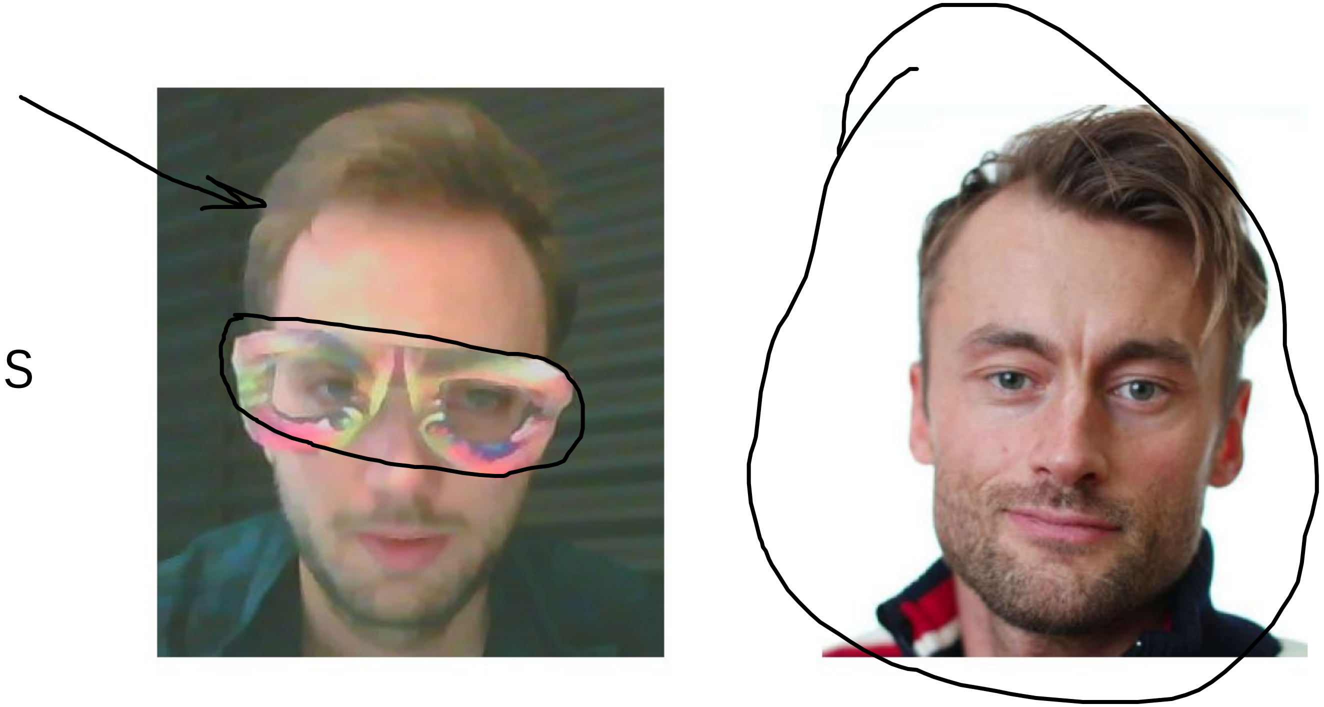
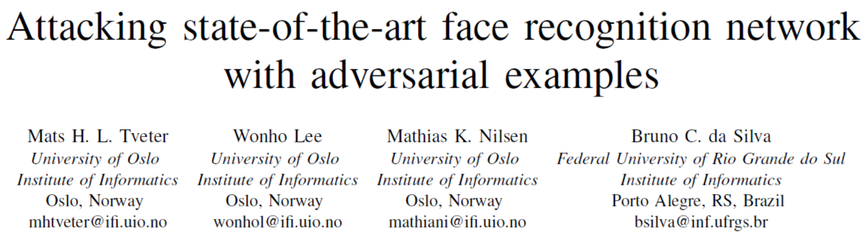


Fig. 6. Author 3 predicted(90.2%) as famous Norwegian cross country skier Petter Northug.



Attacking state-of-the-art face recognition networks with adversarial examples

Mats H. L. Tveter
University of Oslo
Institute of Informatics
Oslo, Norway
mhtveter@ifi.uio.no

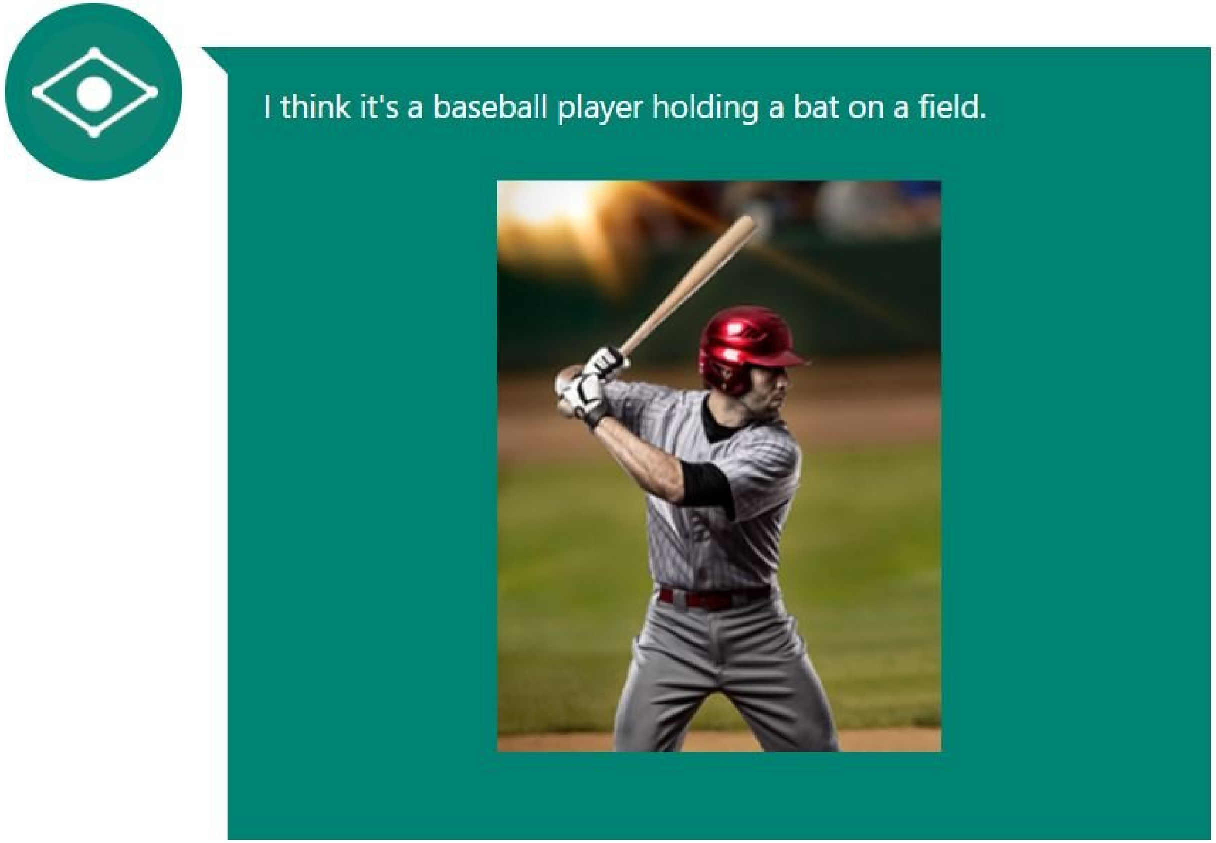
Wonho Lee
University of Oslo
Institute of Informatics
Oslo, Norway
wonhol@ifi.uio.no

Mathias K. Nilsen
University of Oslo
Institute of Informatics
Oslo, Norway
mathiani@ifi.uio.no

Bruno C. da Silva
Federal University of Rio Grande do Sul
Institute of Informatics
Porto Alegre, RS, Brazil
bsilva@inf.ufrgs.br

2) Understanding Human Language

- Extremely impressive progress has been made in modelling human language in recent years, with applications such as
  - Translation
  - Text summarization
  - Automated image description
  - **Automated text generation**

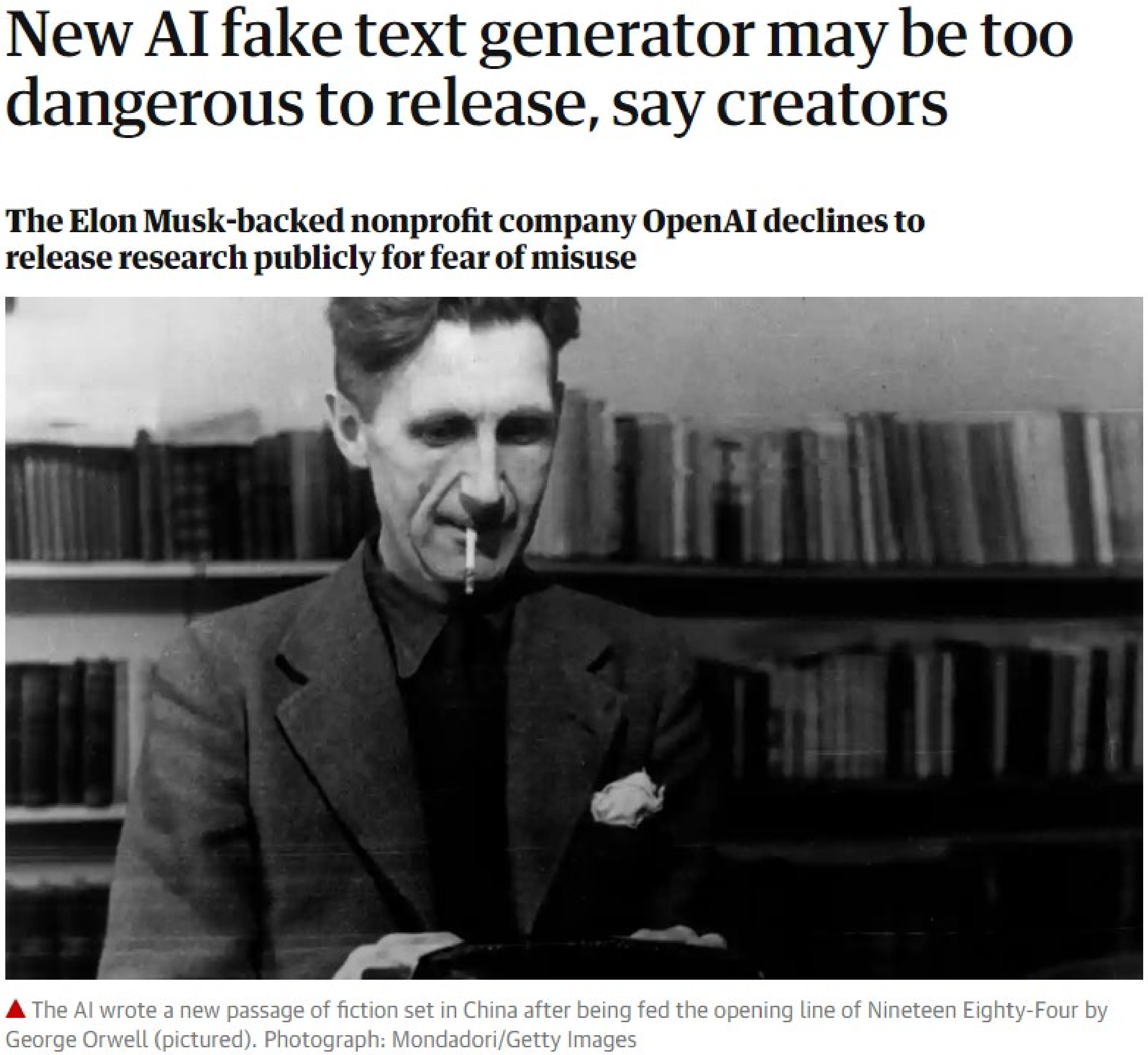


I think it's a baseball player holding a bat on a field.

Source: Captionbot

GPT-2

- Auto-generating plausible text on some topic
- Fear of misuse (e.g. for fake news)



New AI fake text generator may be too dangerous to release, say creators

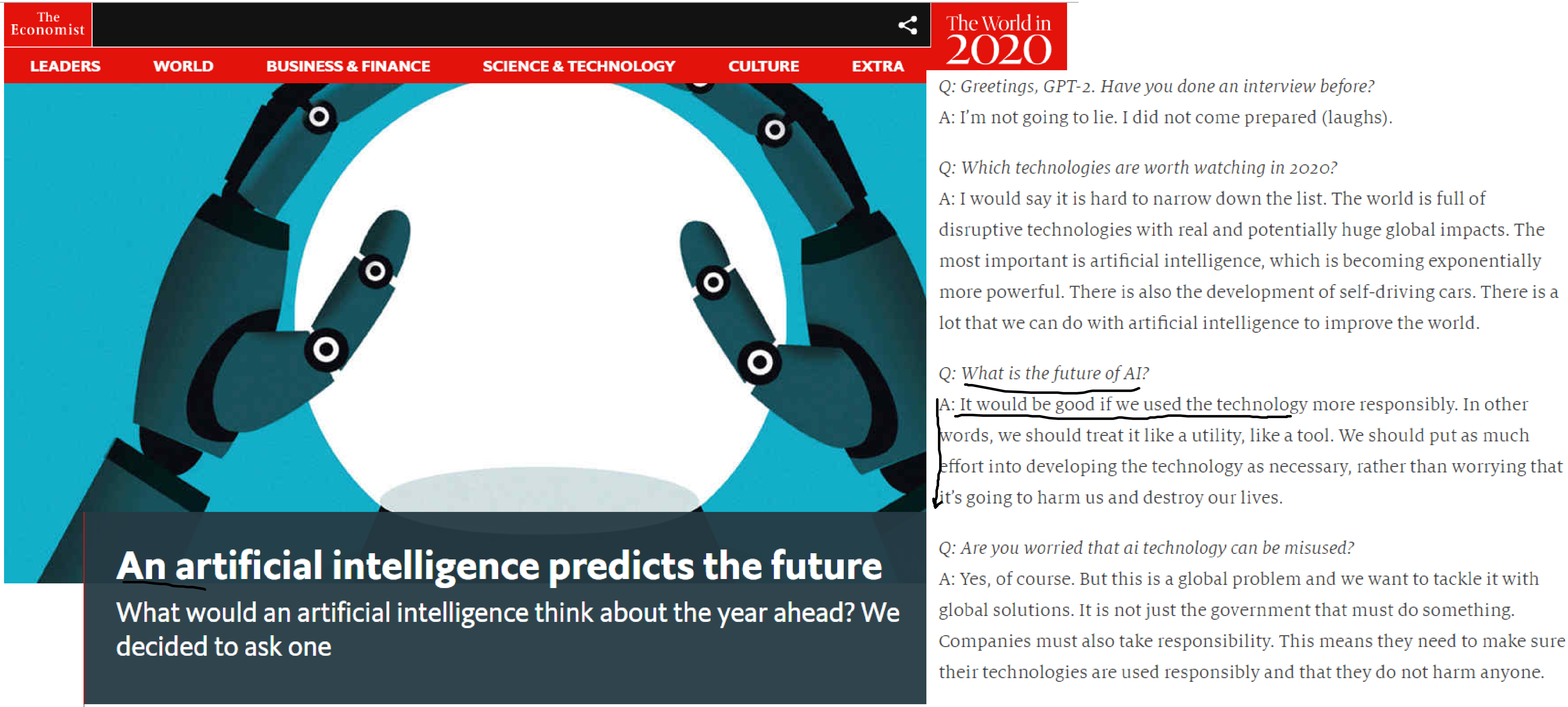
The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

Try it out!

- https://talktotransformer.com/
- “The most important limitation to current Artificial Intelligence systems is”

Hype/Expectations are Enormous



An artificial intelligence predicts the future

What would an artificial intelligence think about the year ahead? We decided to ask one

Q: Greetings, GPT-2. Have you done an interview before?

A: I'm not going to lie. I did not come prepared (laughs).

Q: Which technologies are worth watching in 2020?

A: I would say it is hard to narrow down the list. The world is full of disruptive technologies with real and potentially huge global impacts. The most important is artificial intelligence, which is becoming exponentially more powerful. There is also the development of self-driving cars. There is a lot that we can do with artificial intelligence to improve the world.

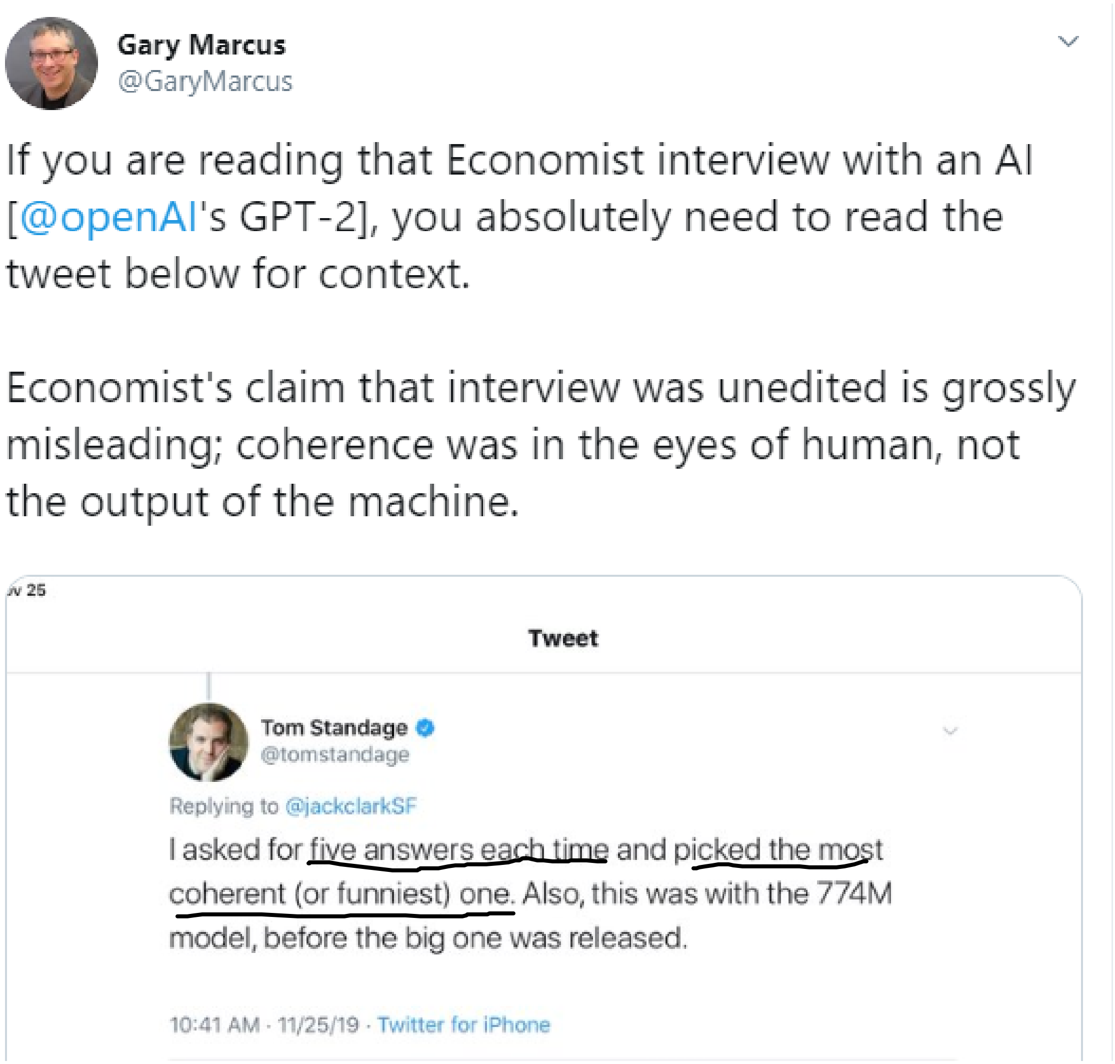
Q: What is the future of AI?

A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

Q: Are you worried that ai technology can be misused?

A: Yes, of course. But this is a global problem and we want to tackle it with global solutions. It is not just the government that must do something. Companies must also take responsibility. This means they need to make sure their technologies are used responsibly and that they do not harm anyone.

But Reality is More Disappointing



Gary Marcus
@GaryMarcus

If you are reading that Economist interview with an AI [@openAI's GPT-2], you absolutely need to read the tweet below for context.

Economist's claim that interview was unedited is grossly misleading; coherence was in the eyes of human, not the output of the machine.

v 25

Tweet

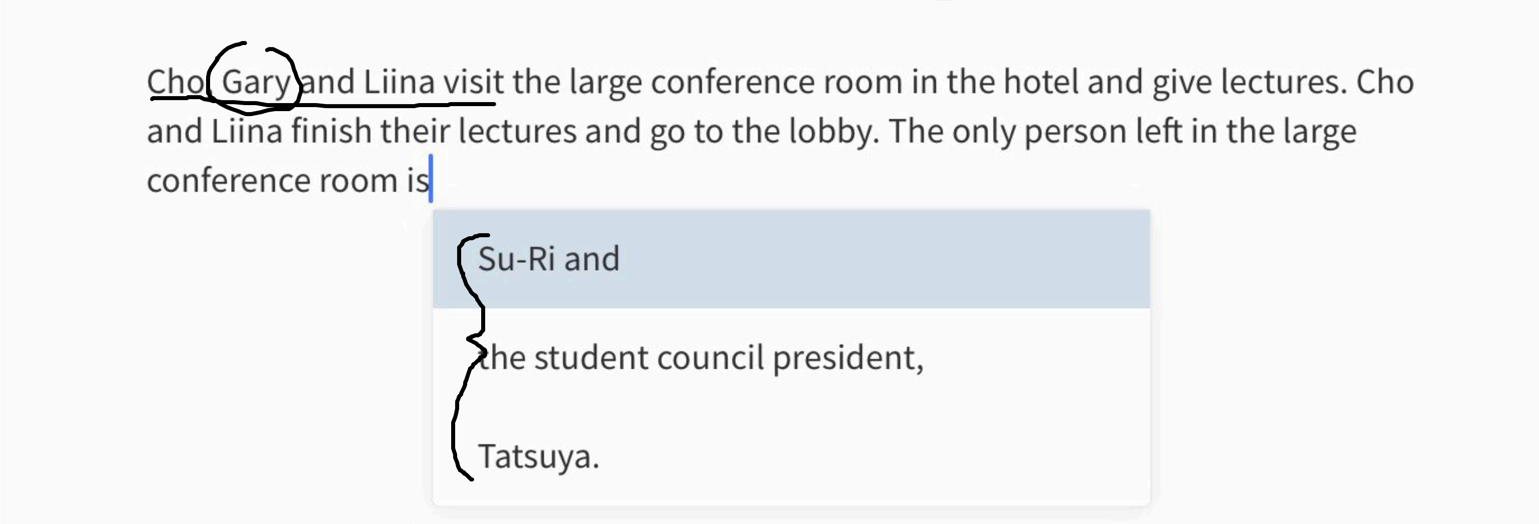
Tom Standage
@tomstandage

Replying to @jackclarkSF

I asked for five answers each time and picked the most coherent (or funniest) one. Also, this was with the 774M model, before the big one was released.

10:41 AM · 11/25/19 · Twitter for iPhone

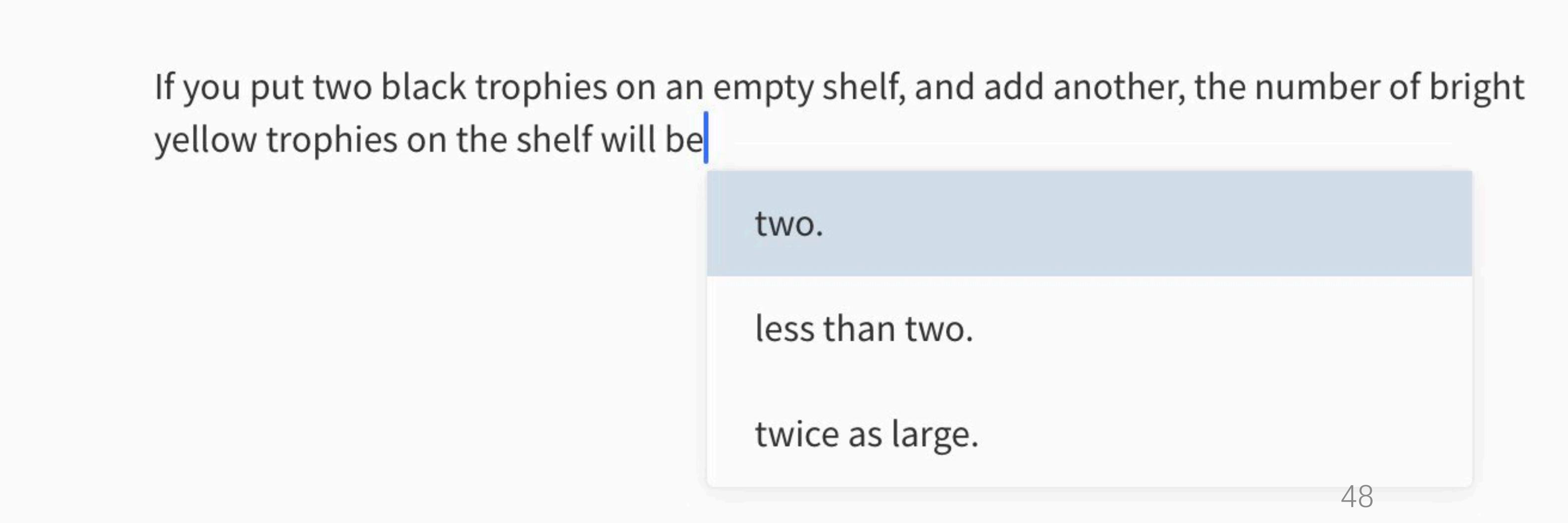
GPT-2 Often Gives Interesting Answers – What Does it “Understand” About the World?



Cho, Gary and Liina visit the large conference room in the hotel and give lectures. Cho and Liina finish their lectures and go to the lobby. The only person left in the large conference room is

- Su-Ri and
- the student council president,
- Tatsuya.

From Gary Marcus’ Twitter



If you put two black trophies on an empty shelf, and add another, the number of bright yellow trophies on the shelf will be

- two.
- less than two.
- twice as large.

Why Can’t GPT-2 Understand These?

Why Can't GPT-2 Understand These?

- This type of "quiz" was probably not very present in its training data
- Highlights a limitation to all AI systems: They are often very good at **solving one type of problem**, but if we go a little beyond that problem, they fail

Consequences

- Impressive deep learning systems learn in a very different way from humans
- They are vulnerable to “simple” attacks indicating they do not *understand* the images they are looking at
- Perhaps we should be more worried about these systems not being *smart enough*

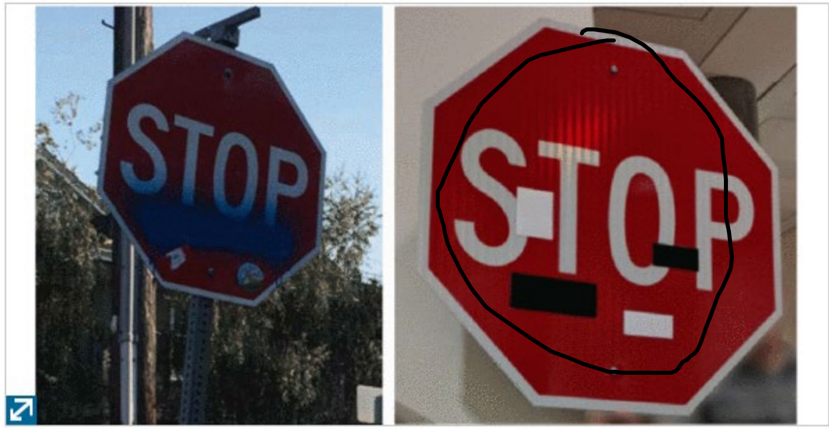
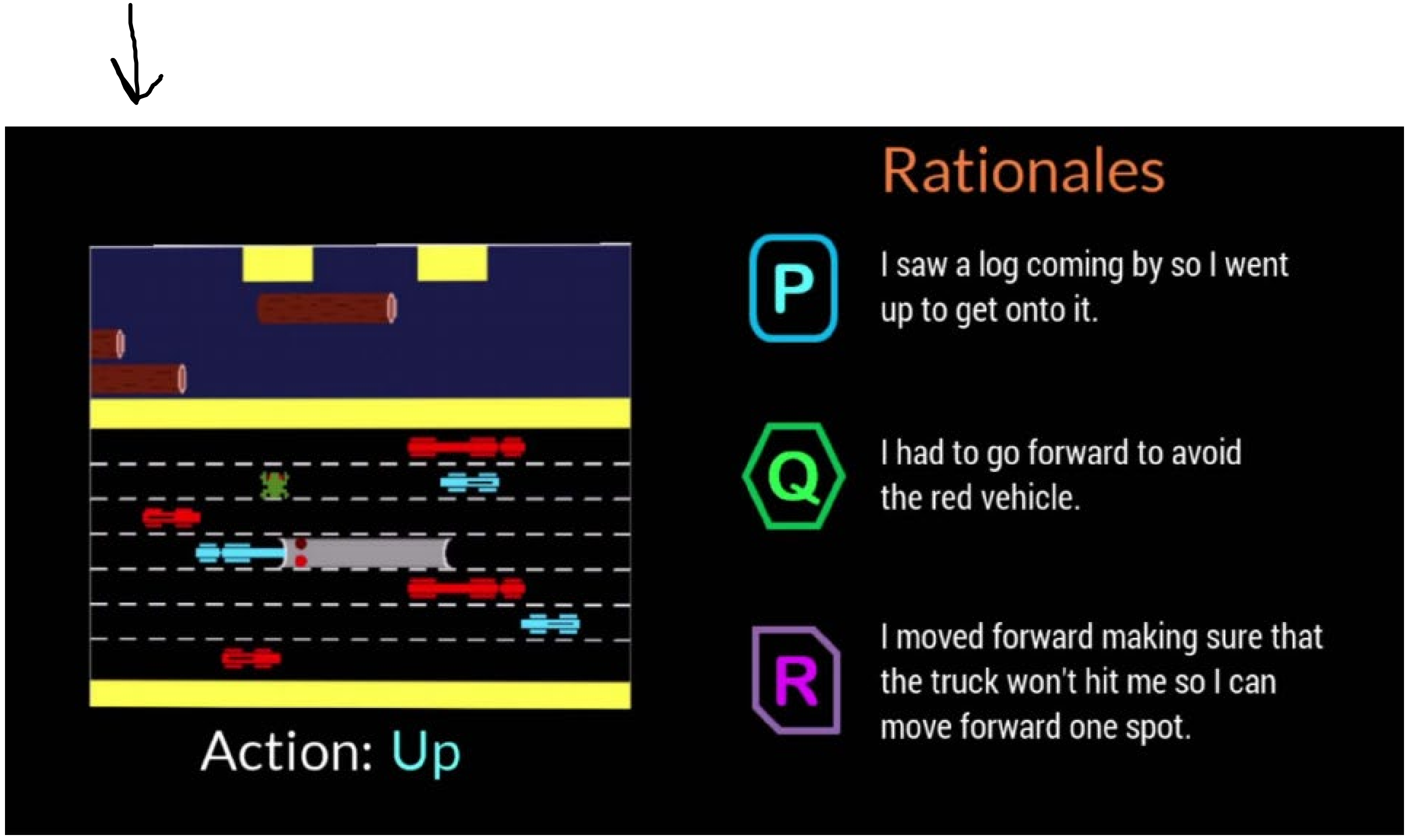


Figure 1:
The left image shows real graffiti on a stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

Some more Limitations to Current AI

- **Explainability**
- Continuous learning
- Extremely data-inefficient learning


Rationales
P
I saw a log coming by so I went up to get onto it.
Q
I had to go forward to avoid the red vehicle.
R
I moved forward making sure that the truck won't hit me so I can move forward one spot.
Action: Up

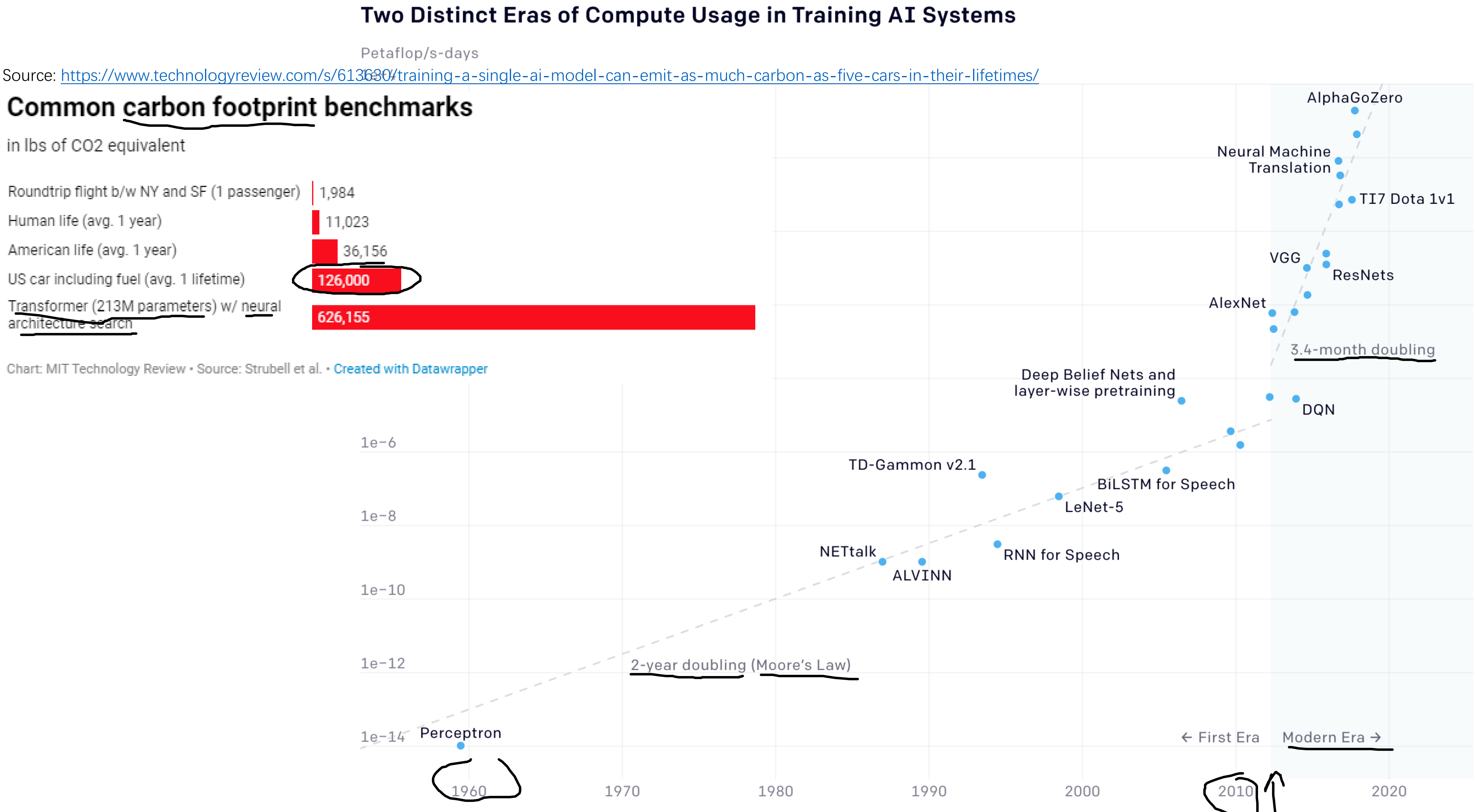
Ehsan et al (2019): Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions

Some more Limitations to Current AI

- Explainability
- **Continuous learning**
- Extremely data-inefficient learning

Some more Limitations to Current AI

- Explainability
- Continuous learning
- **Extremely data-inefficient learning**



Two Distinct Eras of Compute Usage in Training AI Systems

Source: https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Common carbon footprint benchmarks

in lbs of CO2 equivalent

Roundtrip flight b/w NY and SF (1 passenger) 1,984
Human life (avg. 1 year) 11,023
American life (avg. 1 year) 36,156
US car including fuel (avg. 1 lifetime) 126,000
Transformer (213M parameters) w/ neural architecture search 626,155

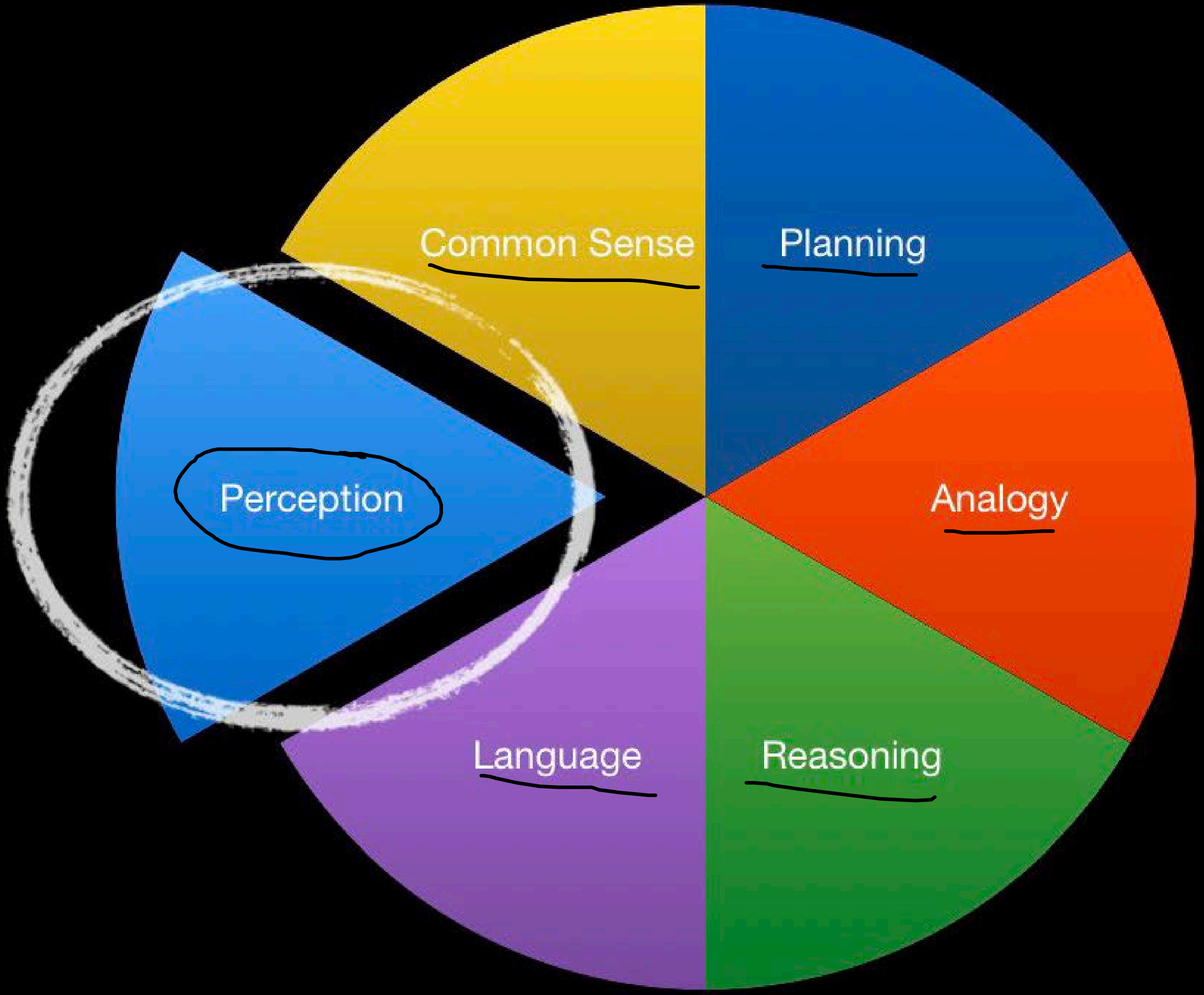
Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Source: https://medium.com/@mattia.cd.ferrini/compositional-deep-learning-a40a07351c37

How do we Adress These Limitations?

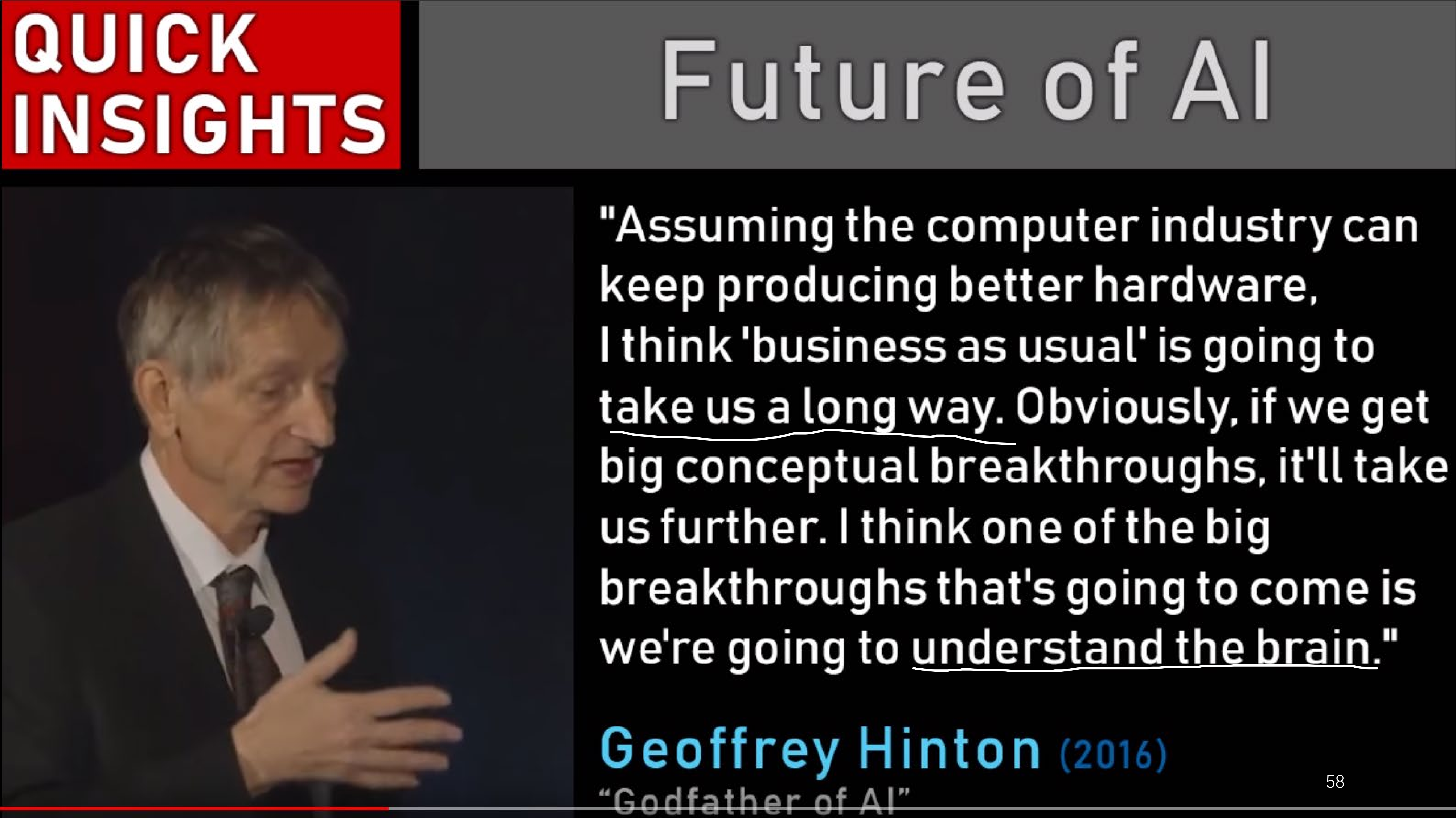
This Photo by Unknown Author is licensed under CC BY-NC

It's reasonable to expect deep learning to be part of the solution
It is not reasonable to expect deep learning to be the entire solution



"**Despite all of the problems I have sketched, I don't think that we need to abandon deep learning**. Rather, we need to reconceptualize it: not as a universal solvent, but simply as one tool among many, a power screwdriver in a world in which we also need hammers, wrenches, and pliers, not to mention chisels and drills, voltmeters, logic probes, and oscilloscopes."

Gary Marcus, 2018 "Deep learning: A critical appraisal"



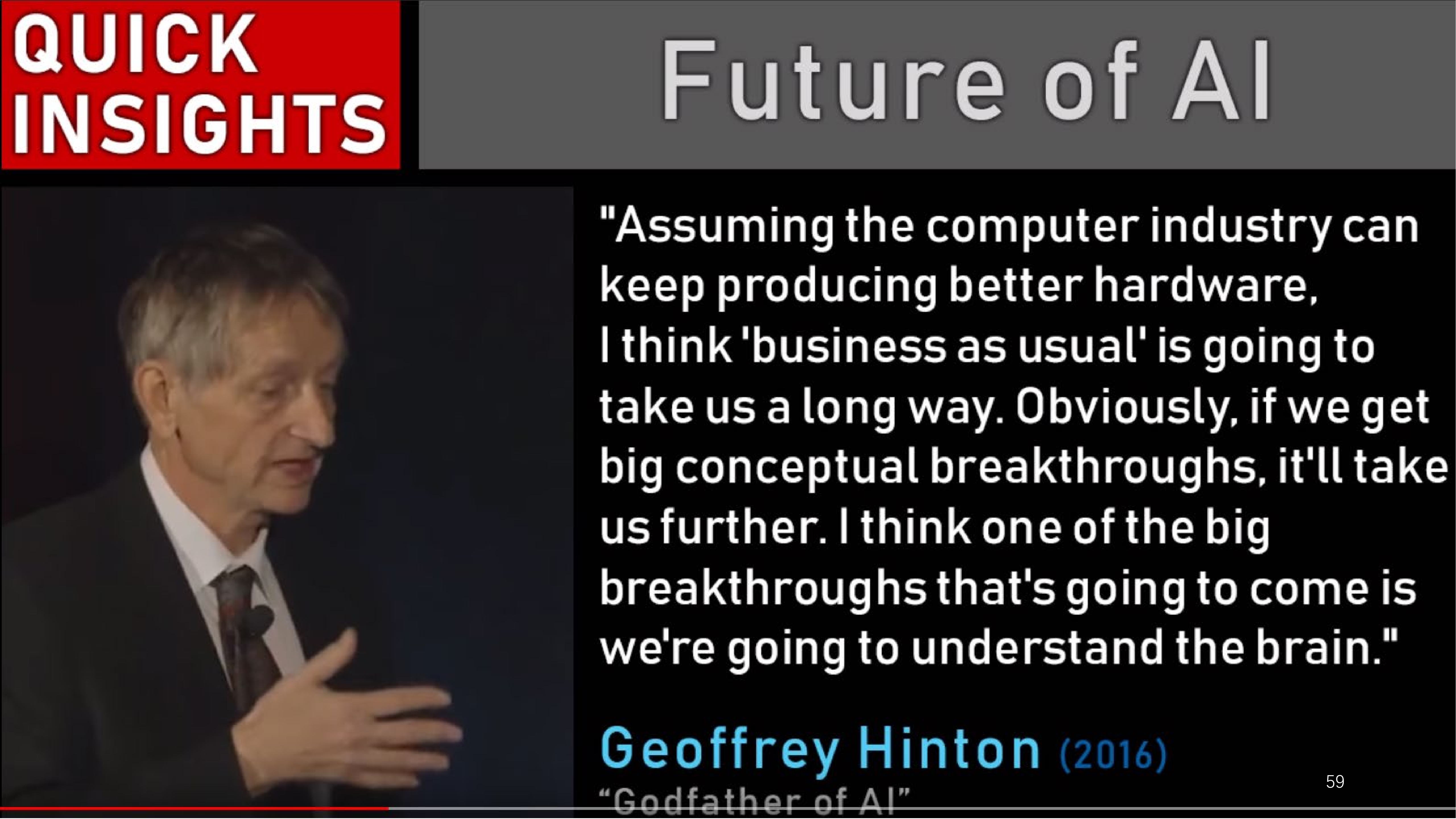
QUICK
INSIGHTS

Future of AI

"Assuming the computer industry can keep producing better hardware, I think 'business as usual' is going to take us a long way. Obviously, if we get big conceptual breakthroughs, it'll take us further. I think one of the big breakthroughs that's going to come is we're going to understand the brain."

Geoffrey Hinton (2016)

“Godfather of AI”



QUICK INSIGHTS

Future of AI

"Assuming the computer industry can keep producing better hardware, I think 'business as usual' is going to take us a long way. Obviously, if we get big conceptual breakthroughs, it'll take us further. I think one of the big breakthroughs that's going to come is we're going to understand the brain."

Geoffrey Hinton (2016)

"Godfather of AI"

IN3050/IN4050, Lecture 15

Ethical Issues, Risks and Fundamental Limitations

6: Fair machine learning

Fabio Zennaro

Next video: Causality in machine learning

Fairness

When deploying machine learning systems in social-sensitive setting you may have to consider not only *performance/accuracy* but also *fairness*.

Take for instance a bank application for *loan selection*. Let your data matrix be:

$\mathbf{X}=$

	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

We want to model a decision Y that maximize the bank's profits as a function of the data $\mathbf{X}$:

$$Y = f(\mathbf{X})$$

Bias

What if $Y = f(\text{Ethnicity})$ varies strongly as a function of the ethnicity of the customer?

- The data set we learned from is *historically biased* and our system would then *reinforce* an existing social bias;
- The data set we learned from is *observationally biased* and our system would then *introduce* a new social bias.

The correlation between a sensitive variable (like ethnicity) and the output (like profit) is real in the data, and it helps maximize our objective. Yet, for ethical reasons, we do not want to exploit and worsen this bias.

Protected Attributes

Let us distinguish our features between *sensitive* or *protected* attributes $\mathcal{A}$ and standard features $\mathcal{X}$

	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

$$\mathcal{A} = \{\text{Ethnicity}\}$$
$$\mathcal{X} = \{\text{Postcode, Univ Degree, Monthly Income}\}$$

Fairness is defined with respect to these protected attributes. Definition is complex and subject to debate.

Case Study 1: Fairness through unawareness (is not fair!)

Let us discard *protected* attributes $\mathcal{A}$ and train the model only on the standard features $\mathcal{X}$.

Why is this not fair?

Case Study 1: Fairness through unawareness (is not fair!)

Let us discard *protected* attributes $\mathcal{A}$ and train the model only on the standard features $\mathcal{X}$.

ID	Ethnicity	Postcode	University Degree	Monthly income
#0001	1	1234	Maths	1k
#0002	3	5678	Computer Science	2k
#0003	1	1234	Literature	4k
...	...	...	...	...

Even if we ignore *protected* attributes (like Ethnicity), some standard features (like Postcode) may be **highly correlated** with the protected attribute [3].

The same biases would then be re-inforced or introduced.

Case Study 2: COMPAS

Northpointe developed a model that given a set of attributes $\mathbf{X}$ of a defendant, would predict the degree of recidivism Y.

Image from `propublica.org`

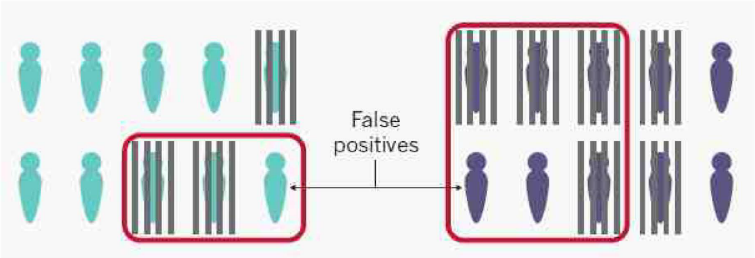


Image from [2]

ProPublica accused the tool of being *unfair*, with respect to *false positives*: more black defendants, later proved innocent, were classified as high risk.

Northpointe argued that their tool was *fair*, with respect to *prediction*: accuracy in classification among white/black defendants was the same.

Different measures of fairness may be **inconsistent** [1].

UiO : University of Oslo

institutt for informatikk

IN3050/IN4050, Lecture 15
Ethical Issues, Risks and Fundamental Limitations

7: Causality in machine learning

Fabio Zennaro

Correlation is not causation

It is well-known that machine learning systems learn *correlations*, not *causation*.

Take for instance an application to predict number of thefts. Let your data matrix be:

X=

Ice-cream sold	Number of thefts
210	22
209	21
12	2
11	1
…	…

We want to model Theft $= f\,(\text{Ice})$.

Prediction and Intervention

Is it correct to use the model in which the number of thefts is a *function* of the number of ice-cream sold?

Prediction and Intervention

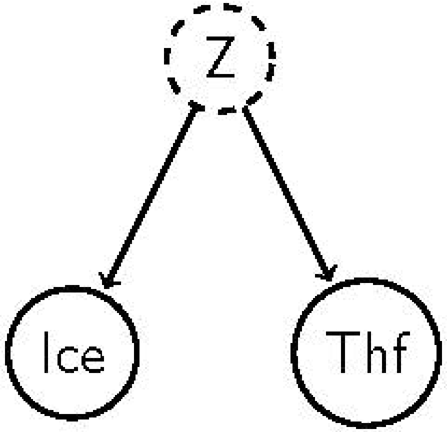
Is it correct to use the model in which the number of thefts is a *function* of the number of ice-cream sold?

We know that the number of ice-cream sold *does not cause* the number of thefts.
Yet:

- If you want only to **predict**, then the model is enough.
  We captured a *predictive regularity*: from the cause we infer the effect, from the effect we infer the cause.
- If you want to **intervene**, then the model is not enough.
  We need to know *relationship of cause and effect*: acting on the cause will change the effect, acting on the effect will leave the cause untouched.

Causal Models

Reasoning about causality is not trivial: it requires its own theory, its own statistical algorithms, its modelling practices [4].



Graphical models[1] are versatile tools to understand and reason about relationships of cause and effect.

---

[1] These DAGs are causal models and they are endowed with a semantics explained by the theory of causality.

UiO : University of Oslo

institutt for informatikk

IN3050/IN4050, Lecture 15
Ethical Issues, Risks and Fundamental Limitations

Kai Olav Ellefsen and Fabio Zennaro