

# CS-C1000 – Introduction to Artificial Intelligence

## Impact and Ethics of AI

**Arno Solin**

April 9, 2021

 @arnosolin

 arno.solin.fi

# Outline and intended learning goals

We look into ...

- ▶ Societal impact
- ▶ Fairness and biases
- ▶ Transparency and accountability
- ▶ Ethics

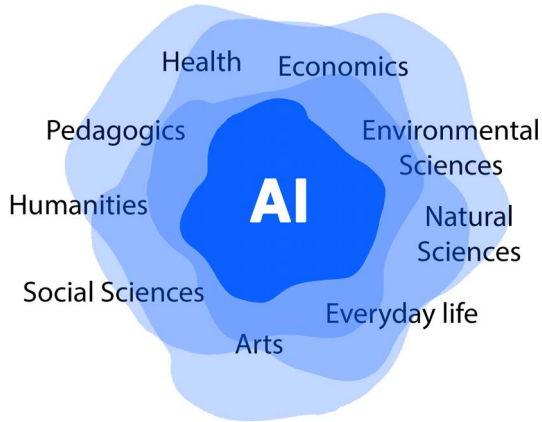
... of AI.



# Leaving the lab

- ▶ AI is leaving (and has left) the lab and **getting deployed** in the real world.
- ▶ Can mean encountering new problems, challenges, and issues.
- ▶ New methods can have **disruptive effects** on established ways of working.

# AI disruption



# AI disruption

- ▶ Clear rationale why this is happening and why we want this to happen.
  - 1 AIs are good at **repetitive** tasks.
  - 2 Can **endure pressure** and are not affected by hostile (also physical) environments.
  - 3 Can think logically **without emotions**, making rational decisions with less or no mistakes.

# Do we all lose our jobs now?

- ▶ *“The major challenge of the sixties is to maintain full employment at a time when automation is replacing men”*  
(John F. Kennedy, 1961)
- ▶ (This far) every technological shift has ended up creating more jobs than were destroyed.
- ▶ When particular tasks are automated, becoming cheaper and faster, we need more human workers to do the other functions in the process that haven't been automated.

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

# Should we all move to working in IT now?

Q: *“But who’s going to maintain the machines?”*

A: The machines.

Q: *“But who’s going to improve the machines?”*

A: The machines.

► So no. Machines will do it (and do it better).

# Human-like intelligence

- ▶ Have algorithms do things we think are intelligent:  
*To have computers do things, that, if people did them, we would consider intelligent.*
- ▶ Replicate human intelligence:  
*To explain how human intelligence works, and reproduce it in computers.*
- ▶ Does that imply some flaws by design?



*“Only two things are infinite, the universe and human stupidity, and I’m not sure about the former.”*

Albert Einstein

# Fairness (and biases) in AI

- ▶ The AI will only be as fair and unbiased as we **make** it or **train** it.
- ▶ The training data reflect the current biases in society, and the AI can end up **amplifying** these effects.
- ▶ Seemingly neutral algorithms may **reinforce inequality**.



How did Google fix this?

## Fairness (and biases) in AI

# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

May 23, 2016

# Fairness (and biases) in AI

BUSINESS NEWS OCTOBER 10, 2018 / 6:12 AM / 6 MONTHS AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

# Fairness (and biases) in AI

The New York Times

## Facebook Job Ads Raise Concerns About Age Discrimination



Mark Edelman, a 58-year-old social-media marketing strategist, wasn't shown job advertising placed on Facebook by a company aiming it at a younger audience. Whitney Curtis for The New York Times

By Julia Angwin, Noam Scheiber and Ariana Tobin

Dec. 20, 2017



*This article was written through collaboration between The New York Times and ProPublica, the independent, nonprofit investigative journalism organization.*

A few weeks ago, Verizon placed an ad on Facebook to recruit applicants for a unit focused on financial planning and analysis. The ad showed a smiling, millennial-aged woman seated at a computer and promised that new hires could look forward to a rewarding career in which they would be "more than just a number."

# Fairness (and biases) in AI

## Facebook continues to let advertisers racially discriminate in housing ads

*The company still lets you exclude minority groups in housing ads, a violation of federal law*

By Nick Stett | @nickstett | Nov 21, 2017, 2:31pm EST



One year after a [ProPublica investigation](#) found that Facebook lets housing advertisers exclude users by race, a separate follow-up investigation found that the social network has barely changed any of its practices. The new story, [published today](#), details how ProPublica was able to purchase targeted housing ads that excluded groups like African-Americans, Jews, and Spanish speakers, among others. Such ad targeting for housing is a violation of the federal Fair Housing Act, because of long-standing discriminatory practices in the real estate and rental industries that disadvantage black, Asian, and Latino renters.

Facebook, which just earlier this month launched a [revamped housing category on its new Craigslist competitor](#), owned up to its mistake in a statement from Amil Vora, a vice president of product management at Facebook, given to the *The Verge*:

# Fairness (and biases) in AI

**B BREITBART**

TRENDING: DEM CORONAVIRUS PROBE RECORD JOBLESS CLAIMS FOREIGN WORKER CRACKDOWN

## THE POPE TEAMS UP WITH MICROSOFT AND IBM ON AI ETHICS

 36

 EMAIL

 SHARE

 TWIST



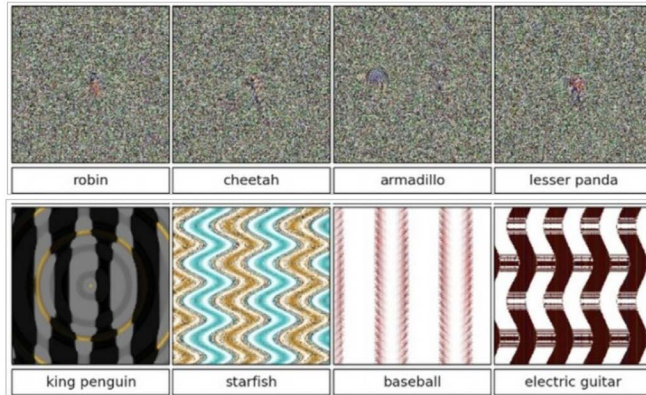
By LUCAS NOLAN | 28 Feb 2020

 [LISTEN TO STORY](#) 2:09

Vatican officials plan to release principles promoting the ethical use of artificial intelligence with the support of Microsoft and IBM, according to a recent report.

Reuters reports that Vatican officials on Friday planned to release principles promoting the ethical use of artificial intelligence with the backing of Microsoft and IBM as the first two technology industry sponsors. The "Rome Call for AI Ethics" states that technology should respect privacy, work reliably and without bias, consider "the needs of all human beings" and operate transparently.

# A part of the problem: Sensitivity to details



These images are classified with >99.6% confidence as the shown class by a Convolutional Network.

Goodfellow *et al.*. Explaining and Harnessing Adversarial Examples. ICLR 2015.



# Tricking AI by adversarial attacks

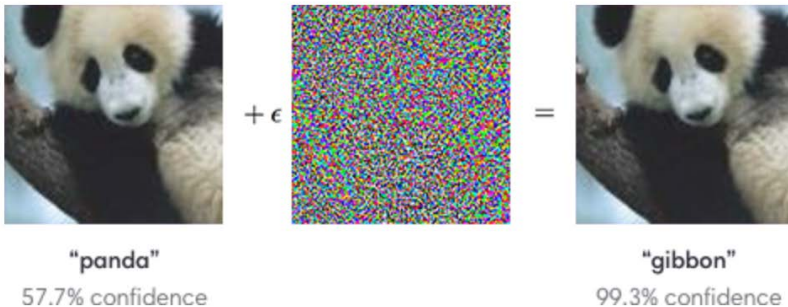
- ▶ AI methods can in some cases also be tricked by malicious attacks.
- ▶ Deep learning models are widespread in visual tasks, but also vulnerable to adversarial attacks.
- ▶ In an adversarial attack, the method is tricked into predicting the wrong output by carefully crafting the input.



A stop sign?

Figure: Eykholt et al. (2008). *Robust Physical-World Attacks on Deep Learning Visual Classification*.

# Tricking AI by adversarial attacks



See more: <https://openai.com/blog/adversarial-example-research/>

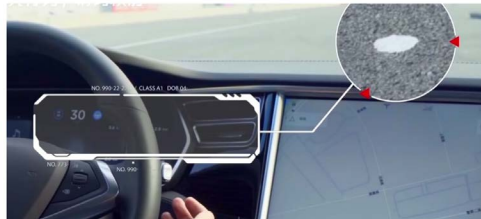
# Tricking AI by adversarial attacks

1 Apr 2019 | 16:56 GMT

## Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane

Security researchers from Tencent have demonstrated a way to use physical attacks to spoof Tesla's autopilot

By Evan Ackerman



# Tricking AI by adversarial attacks



BUSINESS  
INSIDER

## An artist wheeled 99 smartphones around in a wagon to create fake traffic jams on Google Maps

Aaron Holmes Feb 3, 2020, 12:28 PM



Simon Weckert

# Privacy

- ▶ Growing into a major concern.
- ▶ Data is not only used for targeting ads, also for malicious manipulation.
- ▶ Data is, however, key to run many services.
- ▶ **Differential privacy** is a statistical technique that maximizes accuracy of queries from data while minimizing the privacy impact.

The New York Times

## ***Cambridge Analytica Used Fashion Tastes to Identify Right-Wing Voters***

Christopher Wylie, who helped found the voter-profiling firm, said that clothing preferences had been key to helping "Steve Bannon build his insurgency."



Christopher Wylie at the Business of Fashion conference.  
Sara Heston/Getty Images for The Business of Fashion



By Vanessa Friedman and Jonah Engel Bromwich

Nov. 20, 2018



You've heard of profiling criminals, but welcome to fashion profiling — the practice of classifying and targeting individuals based on their clothing brand preferences. Fashion profiling played a bigger role in the 2016 American presidential election than anyone realized, according to new information from Christopher Wylie, [the Cambridge Analytica](#) whistle-blower.

# Transparency and trustworthiness

- ▶ Can AI systems be trusted?
- ▶ Yes, but they probably need to build on the following pillars:
  - + **Fairness:**  
Training data and models free of bias.
  - + **Robustness:**  
Secure, not vulnerable to tampering or compromising data.
  - + **Explainability:**  
Interpretable models and reasons leading to decisions trackable.
  - + **Lineage:**  
Clear details on development, deployment, and maintenance.  
(Open source frameworks?)

# Weaponization of AI

- ▶ Autonomous systems can be used for warfare or terrorism.
- ▶ Various degree of autonomy, but when the weapon starts deciding things based on sensor readings we are in the regime of AI.
- ▶ 'Smart' land mines have been around for decades already.
- ▶ Autonomous aircraft, drones, submarines, tanks, etc.



▲ The US X-47B unmanned autonomous aircraft. Photograph: Rex Features



# Weaponization of AI



▲ Russia's Armata T-14 battle tank. Photograph: Mikhail Metzel/Tass/PA Images



Russia has developed a **robot tank, Nerehta**, which can be fitted with a machine gun or a grenade launcher, while its semi-autonomous tank, the T-14, will soon be fully autonomous. Kalashnikov, the Russian arms manufacturer, has developed a fully automated, high-calibre gun that uses artificial neural networks to choose targets.

Mattha Busby. *Killer robots: pressure builds for ban as governments meet*. The Guardian, April 9, 2018.



# Accountability

*Who is responsible for the actions of an AI?*



The scientist?



The developer?






The deployer?



The AI itself?

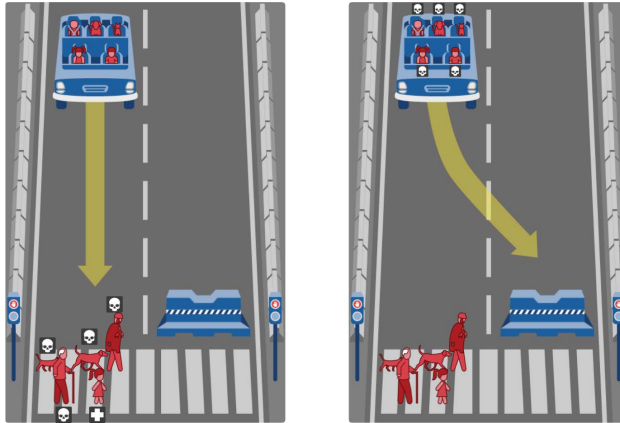
`https://presemo.aalto.fi/ai2021`

# Isaac Asimov's 'Three Laws of Robotics'

-  *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
-  *A robot must obey orders given it by human beings except where such orders would conflict with the First Law.*
-  *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

Is our (actual) legislation up to date?

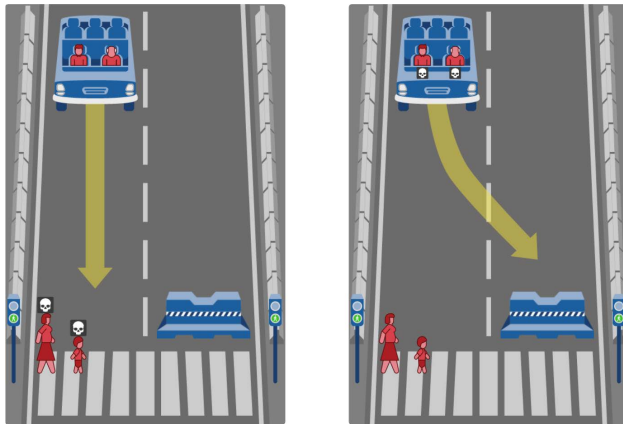
# Moral machine



Should the self-driving car turn or not?

<http://moralmachine.mit.edu/>

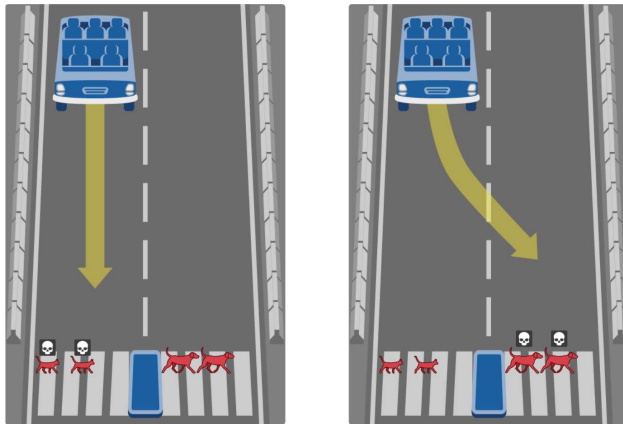
# Moral machine



Should the self-driving car turn or not?

<http://moralmachine.mit.edu/>

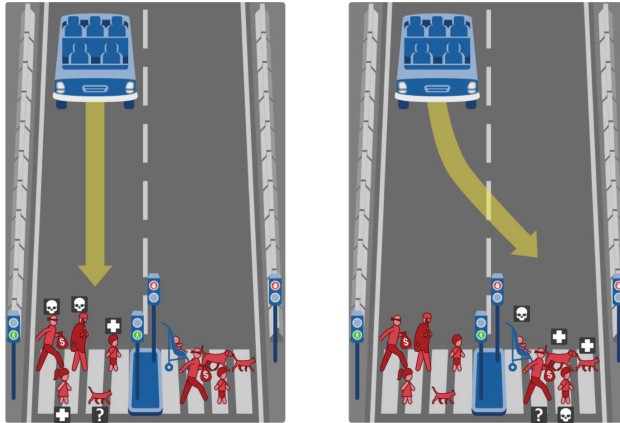
# Moral machine



Should the self-driving car turn or not?

<http://moralmachine.mit.edu/>

# Moral machine



Should the self-driving car turn or not?

<http://moralmachine.mit.edu/>

# AI Surgeon



## Tweet



**Geoffrey Hinton**  
@geoffreyhinton



Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

22.37 · 20.2.2020 · [Twitter Web App](#)

937 Retweets 4 184 Likes

# Recap

- ▶ Societal impact
- ▶ Fairness and biases
- ▶ Transparency and accountability
- ▶ Ethics





**AI**