

# Pitfalls in AI

**Joachim Ganseman**

20/03/2019



# Smals Research



Innovation with  
new technologies



Consultancy  
& expertise



Internal & external  
knowledge transfer



Support for  
going live

2019

Productivity in AI

AI for Public Sector

Advanced  
Cryptography

Conversational  
Interfaces

Robotic Process  
Automation

Blockchain

Web Scraping for  
Analytics

NewSQL  
Databases

Data Quality

[www.smalsresearch.be](http://www.smalsresearch.be)



# Today's menu

- **From data to decision**

- Data collection issues (bias vs. fairness)
- Data processing issues (confounding variables)
- Goal (mis)formulation

Screwing up your own AI

- **Attacks against AI systems**

- Data poisoning
- Adversarial examples

Someone screws with your AI

- **Abuse of AI systems**

- Spear phishing
- (personalized) disinformation
- The role of recommender systems

Someone's AI screws with you

- **Defense against the Dark Arts**

- Transparency & explainability
- Digital Skepticism
- Policy





# About data

---

- AI systems are trained on data  
→ Garbage in, garbage out
- Training data is ideally
  - Well-balanced
  - Free from hidden correlations
  - Independent and identically distributed (iid) over the domain
- In reality, this is rarely the case!



# The *curse of dimensionality*

- Collecting data is expensive
- Any combination of every parameter ... never finished
- Need lots of data quickly  
→ Crowdsourcing?

... when controlled!





# The danger of unbalanced datasets

Employee	Diploma	Gender	Experience	Wage
1	Business	M	Senior	50000
2	Theatre	F	Junior	20000
3	Theatre	F	Senior	25000
4	Business	F	Junior	40000
5	Engineer	M	Senior	50000

- Q: What offer to make to a person?
- AI:

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

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



# Hidden correlations

- We'll fix it by not taking gender into account, right?
- ... well...
  - Men/women have different ways of speaking



- In CVs, men/women mention different things (hobbies...)
  - Gender as prominent **confounding factor** in Amazon's model



# The problem with confounding factors

---

- **Semantic gap** between
  - What we want the AI to learn
  - What the AI actually derives from the training data

→ AI does not necessarily learn what you think it's learning!

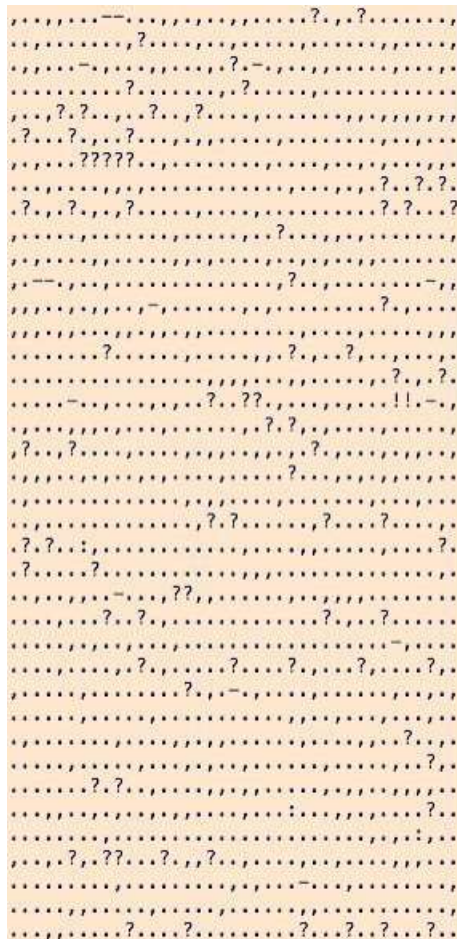
- Mitigations:
  - Better sampling of the training data
  - Thorough (statistical) data analysis



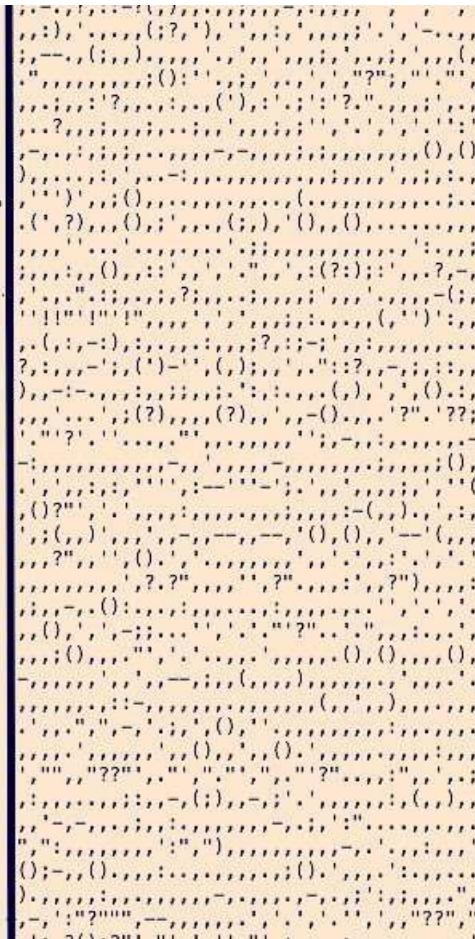
# Confounding factors

- Sometimes leads to surprising new insights!

Blood Meridian  
(Cormac McCarthy)



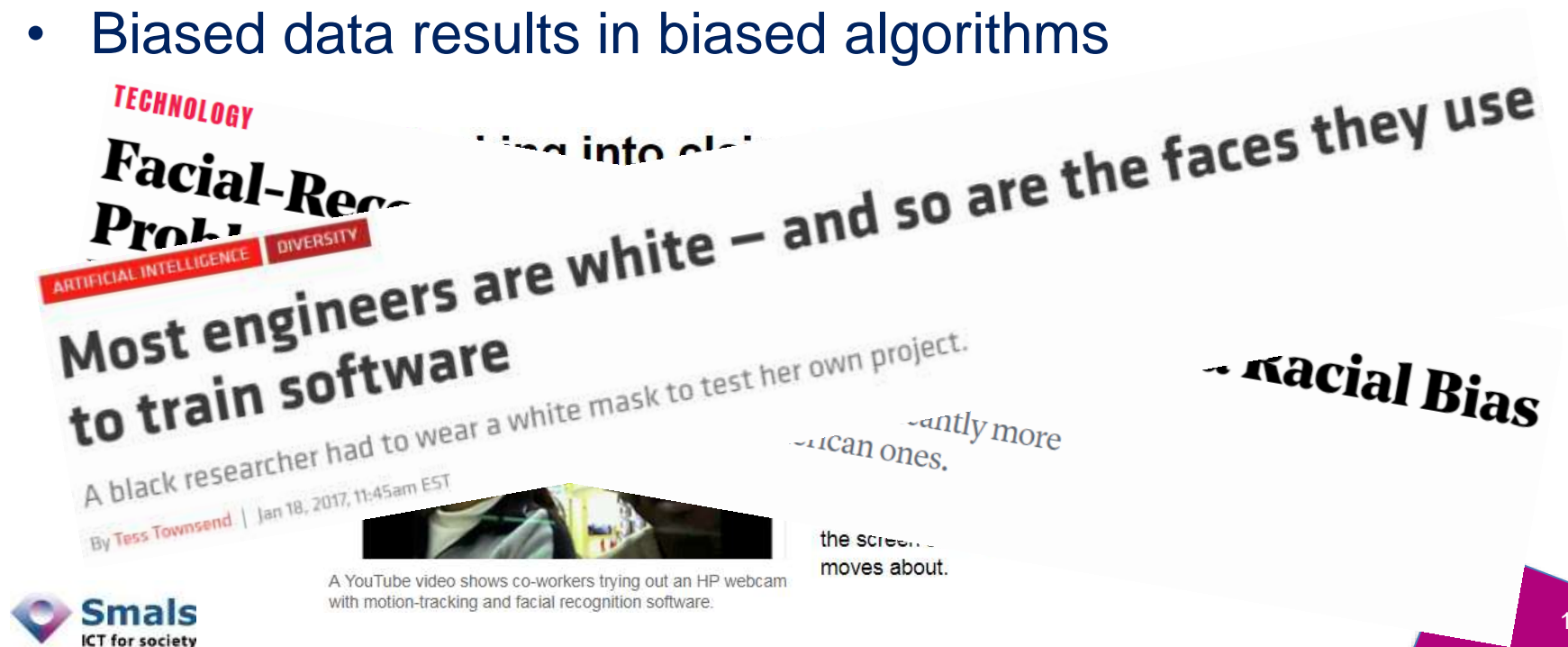
Absalom, Absalom!  
(William Faulkner)





# Data and bias

- Biased humans collect biased data
  - [https://en.wikipedia.org/wiki/List\\_of\\_cognitive\\_biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases)
  - Often subconscious
- Biased data results in biased algorithms





# Bias vs. Fairness

---

- Not all bias is unfair:
  - Prostate cancer data is biased towards men
  - Cervix cancer data is biased towards women
- Unfair bias can have serious consequences
  - Security decisions (airport controls / inspections)
  - Legal decisions (bail, parole)
  - Economic decisions (insurance, mortgage)

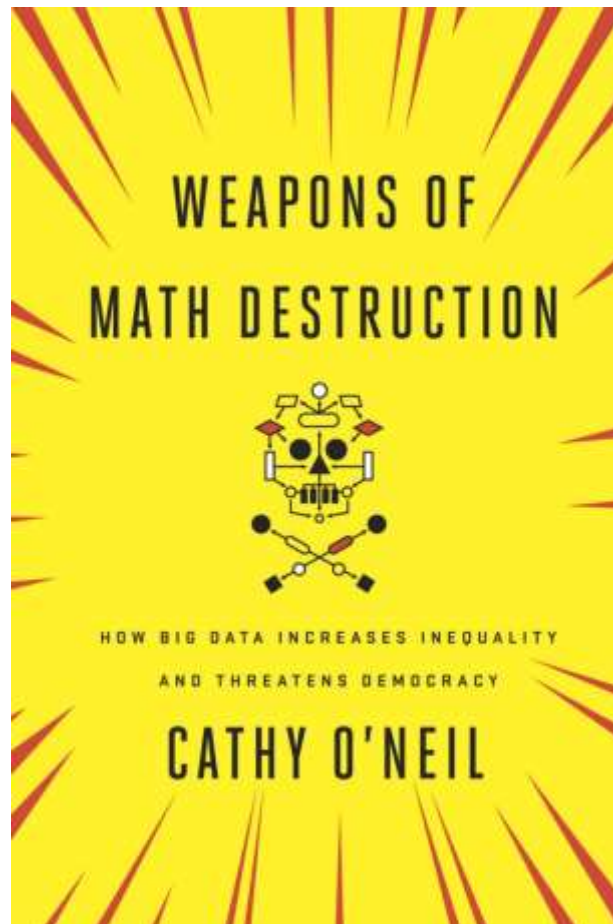
→ Know your data!



# Bias vs. Fairness

---

- Recommended reading:





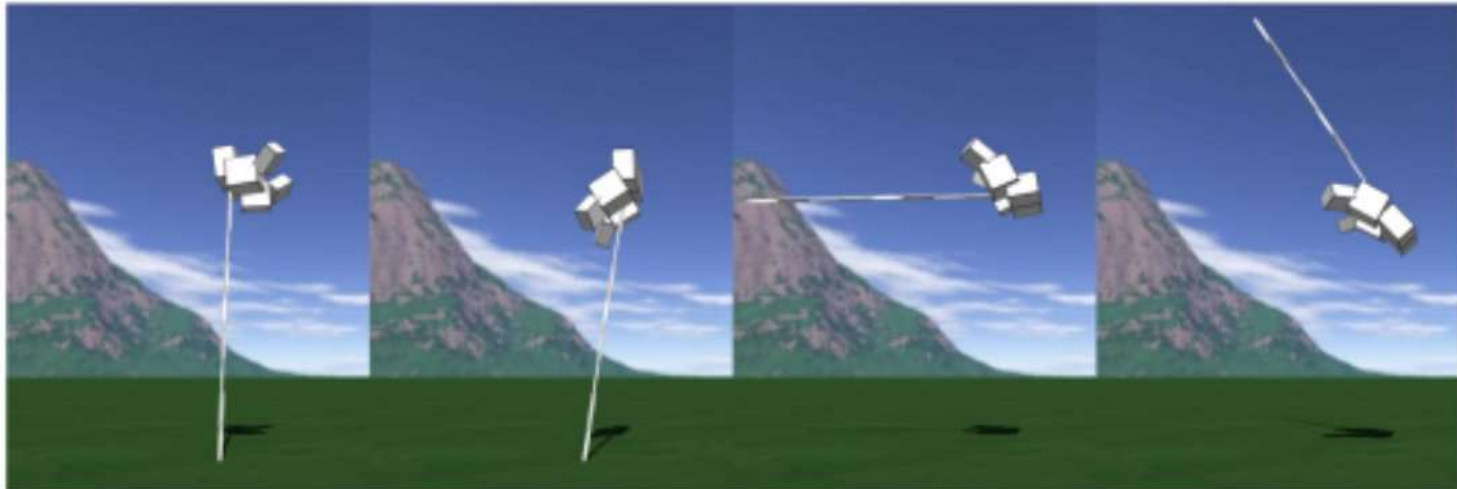
# Definition of objectives

---

- In many ML algorithms:
  - Reward “success”
  - Punish “failure”
  - Goal: maximize the reward
- “Success” is hard to correctly define!
  - AI exploits bugs
  - AI exploits unexpected data properties
  - AI gets stuck in endless loops



## Bad definition of objectives:

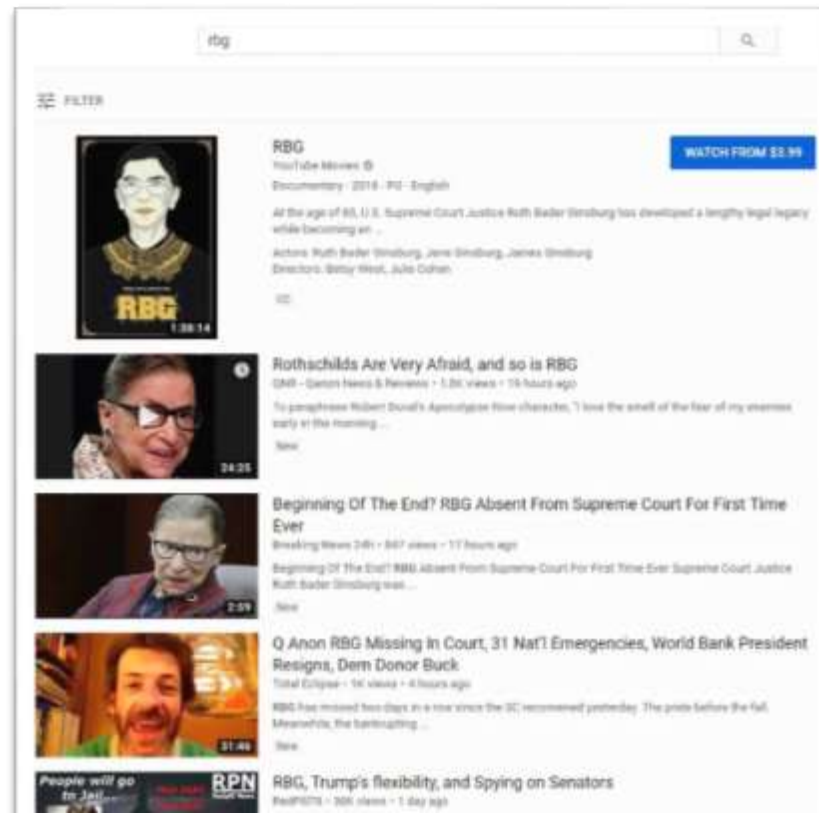




# Producer vs. consumer objectives

- You: want to find good information
- Youtube: wants you to keep watching (ads)
  - promotes content that “pushes buttons”

- Conspiracy theories
- Sensationalism
- Disturbing content
- Extremism
- ...





# Takeaways

---

AI is only as reliable as the data it's trained on.

AI maximizes its objective, nothing more.



# Today's menu

---

- From data to decision
  - Data collection issues (bias vs. fairness)
  - Data processing issues (confounding variables)
  - Goal (mis)formulation
- Attacks against AI systems
  - Data poisoning
  - Adversarial examples
- Abuse of AI systems
  - Spear phishing
  - (personalized) disinformation
  - The role of recommender systems
- Defense against the Dark Arts
  - Transparency & explainability
  - Digital Skepticism
  - Policy



# Data Poisoning

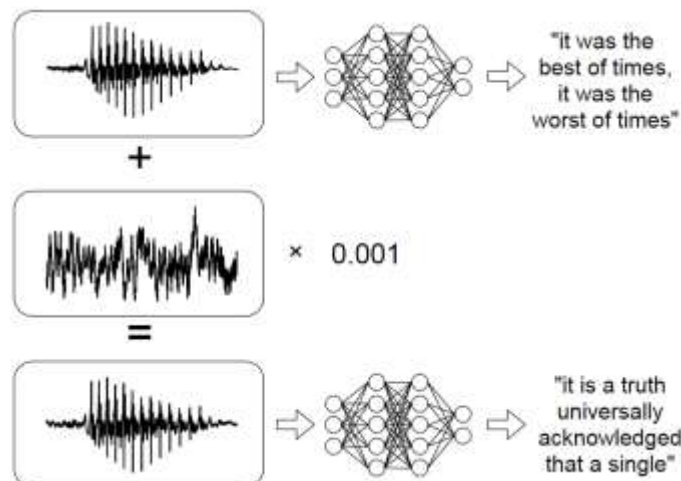
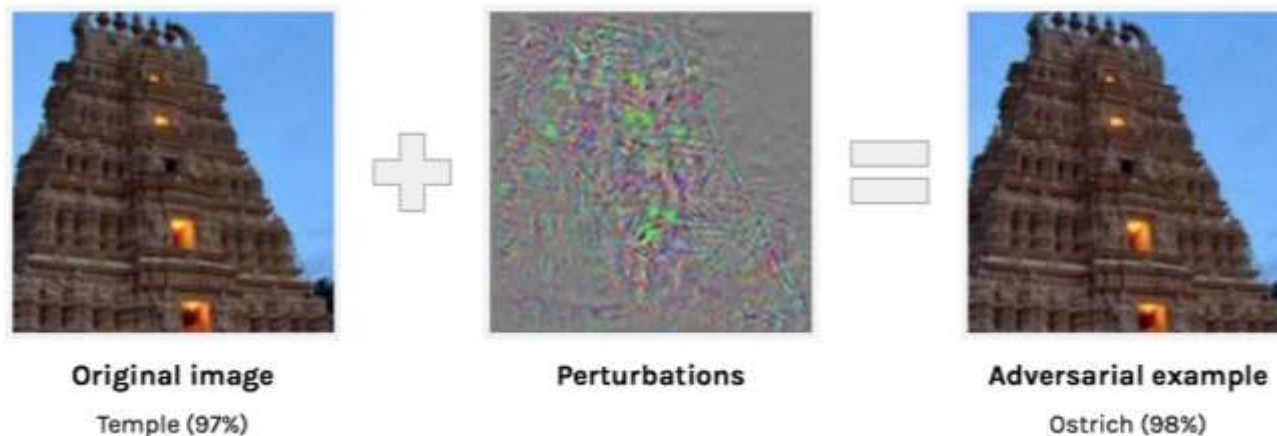
- Inject false training data to compromise learning
  - Intentionally mislabeled data
  - Bogus data or noise
  - E.g. Tay:





# Adversarial examples

- Minimal change to input  $\rightarrow$  large change in output

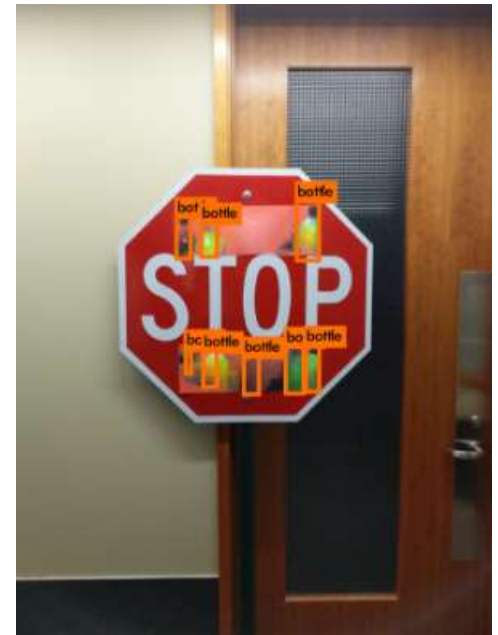




# Adversarial examples

---

- Problem in most AI methods, regardless of data format
- Often robust
  - Stickers on objects
  - 2D/3D printed objects
- Potential causes
  - Curse of dimensionality
  - Overfitting / Limited generalization
    - Adding one strong feature from another class is enough





# Takeaways

---

AI systems have (sometimes significant) error margins.



# On today's menu

---

- From data to decision
  - Data collection issues (bias vs. fairness)
  - Data processing issues (confounding variables)
  - Goal (mis)formulation
- Attacks against AI systems
  - Data poisoning
  - Adversarial examples
- Abuse of AI systems
  - Spear phishing
  - (personalized) disinformation
  - The role of recommender systems
- Defense against the Dark Arts
  - Transparency & explainability
  - Digital Skepticism
  - Policy



# Spear phishing

- Fraudulent attempt to obtain sensitive information, directed at a **specific individual/company**
- AI can **personalize** the message to the target (as in advertising)





# Disinformation (“fake news”)

---

- Definition (*EC action plan against disinformation, 05/12/2018*):
  - Verifiably false or misleading information
  - Disseminated for economic gain or to intentionally deceive
  - May cause public harm
- It is not:
  - (Extreme) political, scientific, ethical or moral viewpoints
  - Unions, lobbying, advocacy, campaigning, ...
  - Selective presentation of information
  - Satire, parody, ...
  - Religion



# Can disinformation be generated?

- Image/video/audio: yes, kind of



2014



2015



2016



2017



2018

cf. *deepfakes*:





# Generating fake text

---

- Result of the latest research (*OpenAI GPT-2, 14/02/2019*):

*Gimli was a tall and powerful man, and he had a beard and a moustache. He was also a dwarf, and he had a strong build, and he was covered in tattoos. He was not a man who looked like a hobbit.*

- Grammar ✓
- Sticking to the theme ✓
- Internal consistency ✗
- Style ✗

→ Short texts might fool an inattentive human



# Amplification through recommendation

---

- YouTube as **the great radicalizer** (Z. Tufekci)
  - Videos about vegetarianism lead to veganism
  - Videos about jogging lead to ultramarathons





- Similar on many other (free) content platforms





# Amplification through recommendation

---

- Any content that is watched more   
 obtains a higher ranking in search results



# Amplification through recommendation

- Inflammatory *Any* content is watched more   
 obtains a higher ranking in search results

**YouTube says it will recommend fewer videos about conspiracy theories**

Taking steps to reduce the spread of misinformation

By Casey Newton | @CaseyNewton | Jan 25, 2019, 10:47am EST



# Takeaways

---

It's still “spam vs. spamfilter”, but on a higher level.

If it sounds too good to be true,  
it's probably too good to be true.



# Today's menu

---

- From data to decision
  - Data collection issues (bias vs. fairness)
  - Data processing issues (confounding variables)
  - Goal (mis)formulation
- Attacks against AI systems
  - Data poisoning
  - Adversarial examples
- Abuse of AI systems
  - Spear phishing
  - (personalized) disinformation
  - The role of recommender systems
- Defense against the Dark Arts
  - Transparency & explainability
  - Digital Skepticism
  - Policy



# On the tech side

---

- Governance through FAT(E)
  - Fairness
  - Accountability
  - Transparency
  - ( Ethics )

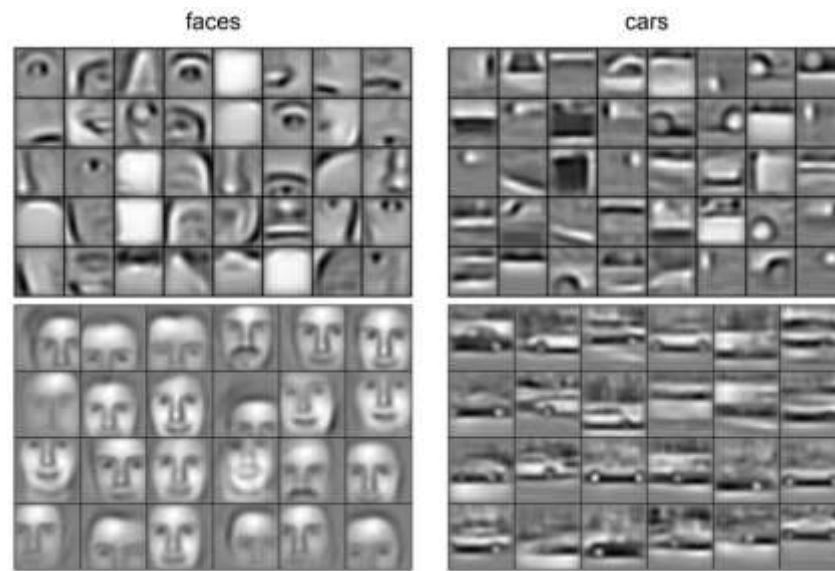




# Explainable AI

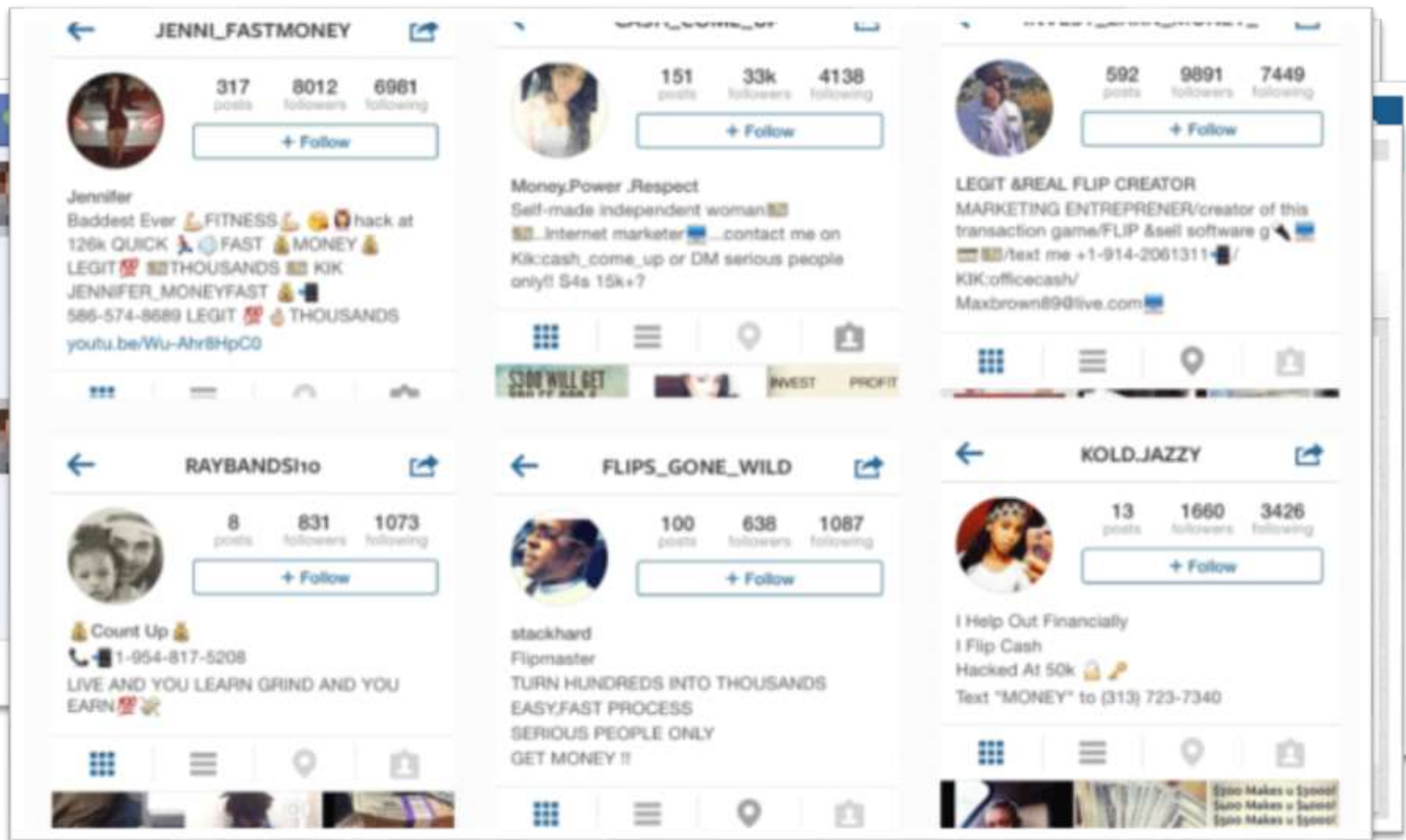
---

- Important factor in **accountability**
- Especially hard with deep learning:
  - What did an AI learn?
  - Why this outcome / decision?
  - Recent movement towards “Explainable AI”
- Still in its infancy





# You against scams and disinformation





# You against scams and disinformation

---

- Awareness
    - You are being profiled
    - What you see is not what someone else sees
    - Anything you post can be used against you
  - Rely on authoritative, transparent sources
    - Peer-reviewed science
    - Quality journalism
- Encourage **Digital Skepticism** (without being paranoid)
- Requires some **Competences / Literacy**



# Legal protection against AI abuse

---

- GDPR (ratified in Belgium: law of 30 July 2018)

## *Article 22*

### **Automated individual decision-making, including profiling**

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
  - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
  - (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.



# For policymakers

---

- Awareness
  - Own vulnerability to pre-selected information
  - Reach and impact of social media
  - Information warfare
- Stimulate
  - Independent and quality media
  - Innovation & research on the impact of innovation
  - Culture of permanent learning

→ Fancy “strategy plans” are useless without actual €€€ !



# What is happening in the EU?

---

- 03/2015: Stratcom Task Force ([euvdisinfo.eu](http://euvdisinfo.eu))
- 03/2018: recommendations of EU HLEG
- 10/2018: EU code of practice on disinformation
  - Signed by Google, Facebook, Twitter, Mozilla etc.
  - (Initial) choice for **industry self-regulation**
- 12/2018: EU action plan on disinformation



# Takeaways

---

digital skeptic > digital panic

Curiosity didn't kill the cat, naivety did.



# Further Reading

---

- The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (“Malicious AI report”, 02/2018)
- For a Meaningful Artificial Intelligence (“Villani rapport”, 03/2018)
- Information Manipulation, A Challenge for Our Democracies (CAPS & IRSEM, France, 08/2018)
- EU Action Plan against Disinformation (05/12/2018)



# Contact

Joachim Ganseman

[joachim.ganseman@smals.be](mailto:joachim.ganseman@smals.be)

**Smals, ICT for society**

02 787 57 11

Fonsnylaan 20 / Avenue Fonsny 20

1060 Brussel / 1060 Bruxelles





UNDENIABLE PROOF EMERGES THAT  
**JP MORGAN SANK THE TITANIC  
TO FORM FEDERAL RESERVE**