

# Some pitfalls of AI

Joachim Ganseman  
Smals Research

15/09/2020

# Smals Research 2020



**Innovation with  
new technologies**



**Consultancy  
& expertise**



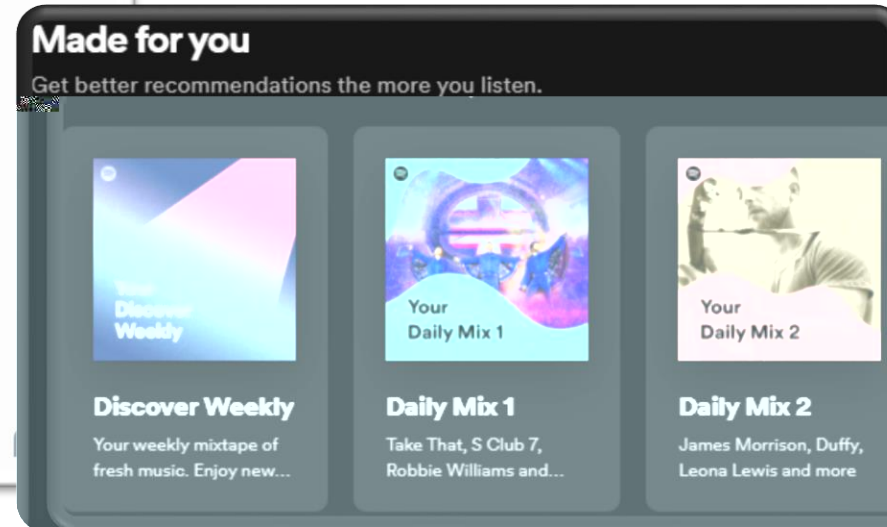
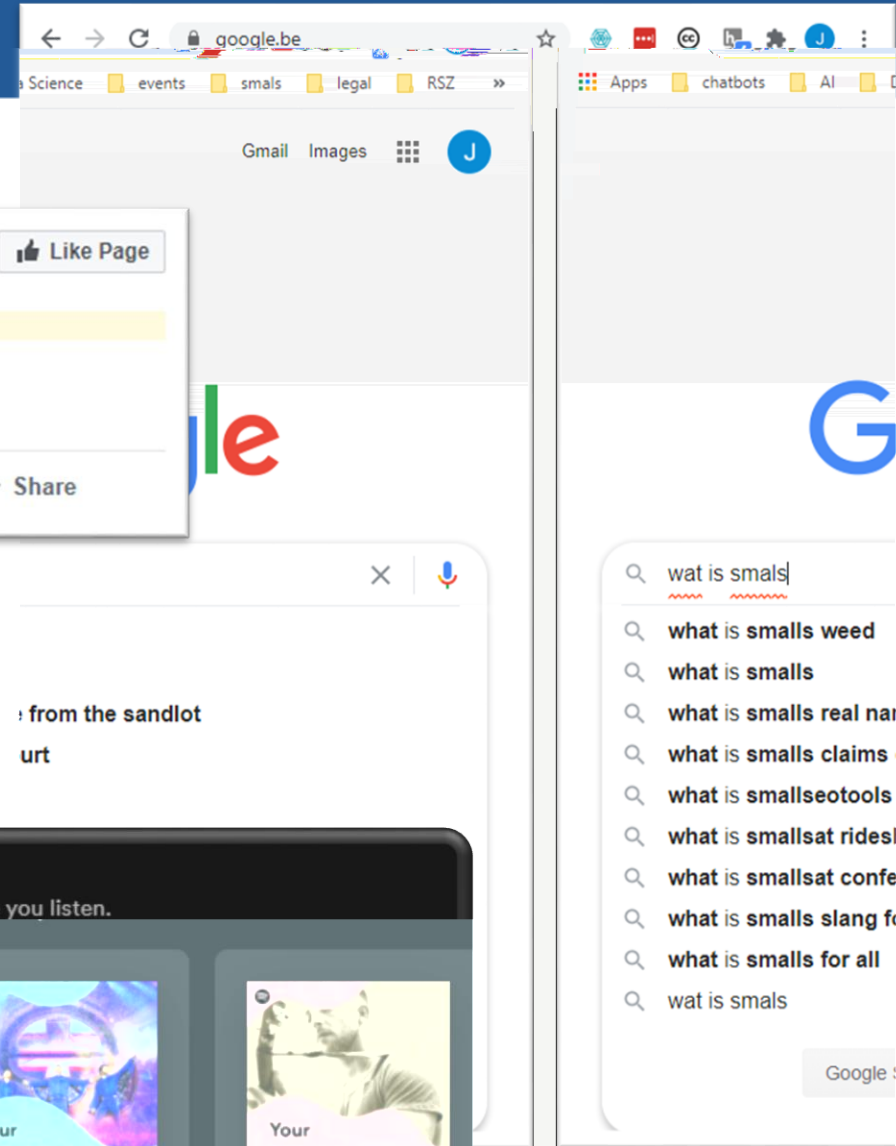
**Internal & external  
knowledge transfer**



**Support for  
going live**

**2020**

# AI: part of our daily life



"now \*that's\* a chinese wall!" by [Esthr](#) is licensed under [CC BY-NC 2.0](#)

"Nest Learning Thermostat showing Celsius" by [Nest](#) is licensed under [CC BY-NC-ND 2.0](#)

# What could possibly go wrong?



For AI developers: from data to decision



For AI deployers



General public



Defense against the Dark Arts

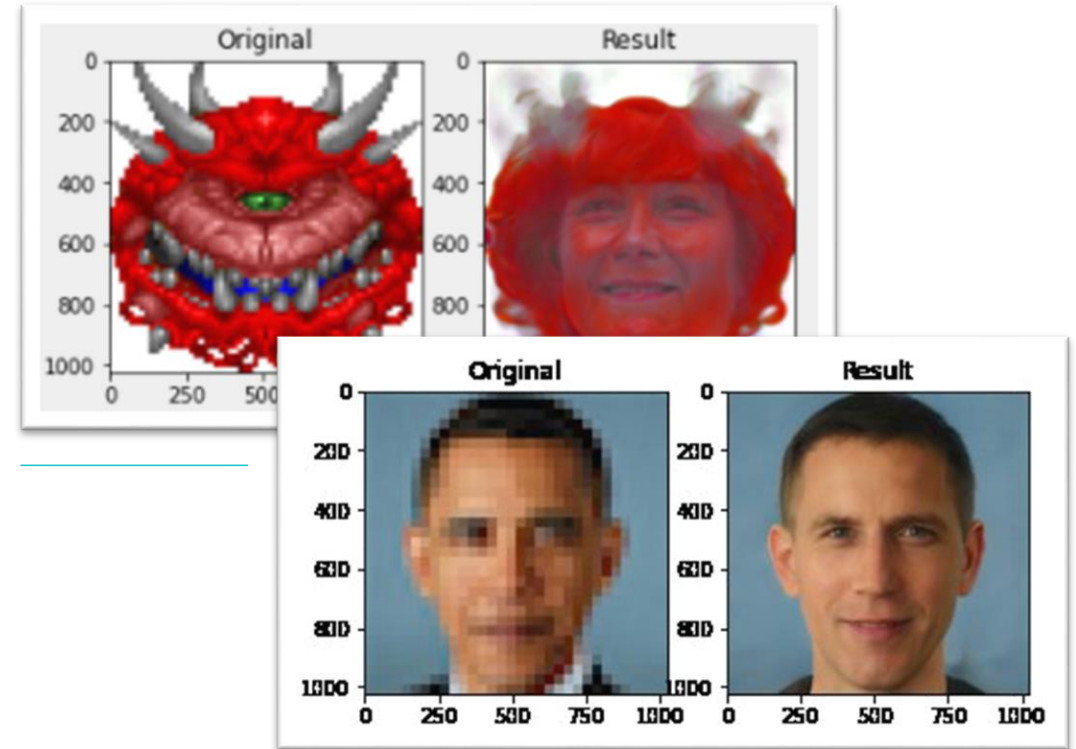
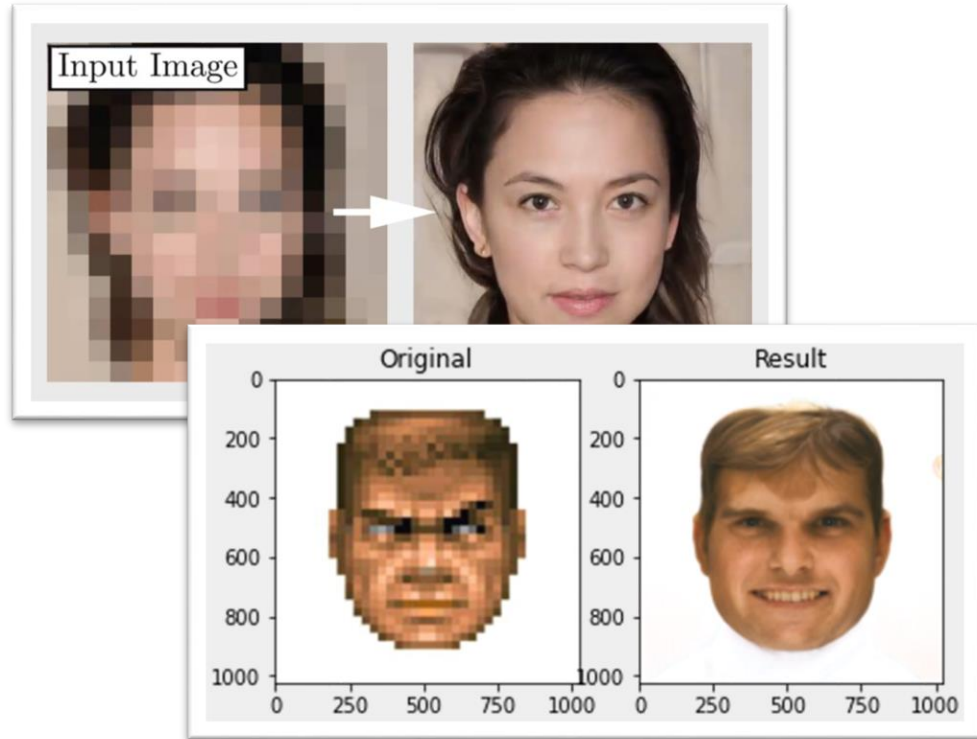




*H H H H H N M H*

*H HH HH H*

# Limits on data → limits on results





**TECHNOLOGY**

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

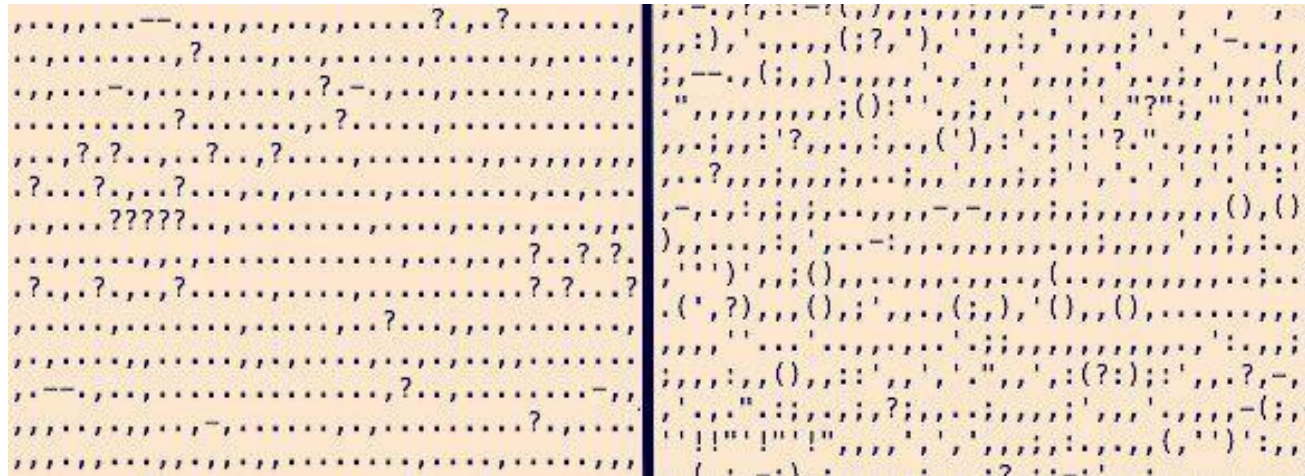
# Amazon scraps secret AI recruiting tool that showed bias against women

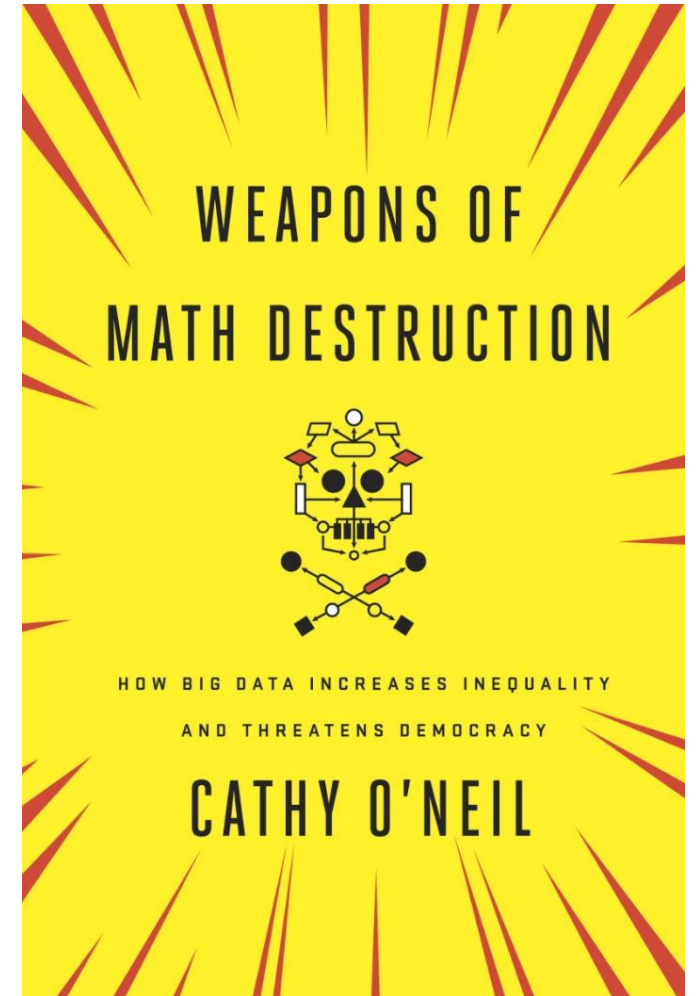
...video shows co-workers trying out an HP webcam with motion-tracking and facial recognition software

moves about.



# Confounding factors





# Definition of objectives



**Custard Smingleigh**  
@Smingleigh

I hooked a neural network up to my Roomba. I wanted it to learn to navigate without bumping into things, so I set up a reward scheme to encourage speed and discourage hitting the bumper sensors.

It learnt to drive backwards, because there are no bumpers on the back.

# What could possibly go wrong?

For AI developers



**For AI deployers: attacks against AI systems**

General public

Defense against the Dark Arts

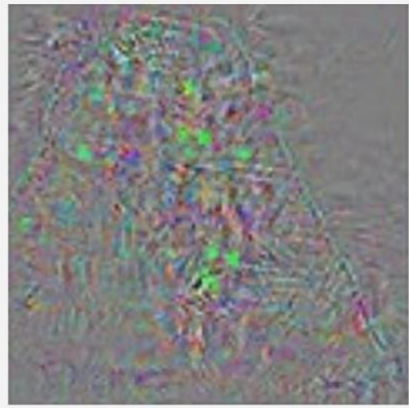


# Adversarial examples

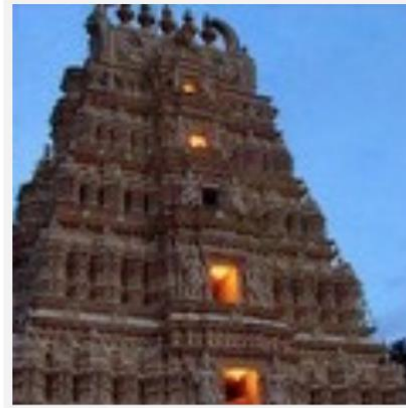
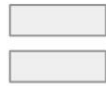


Original image

Temple (97%)

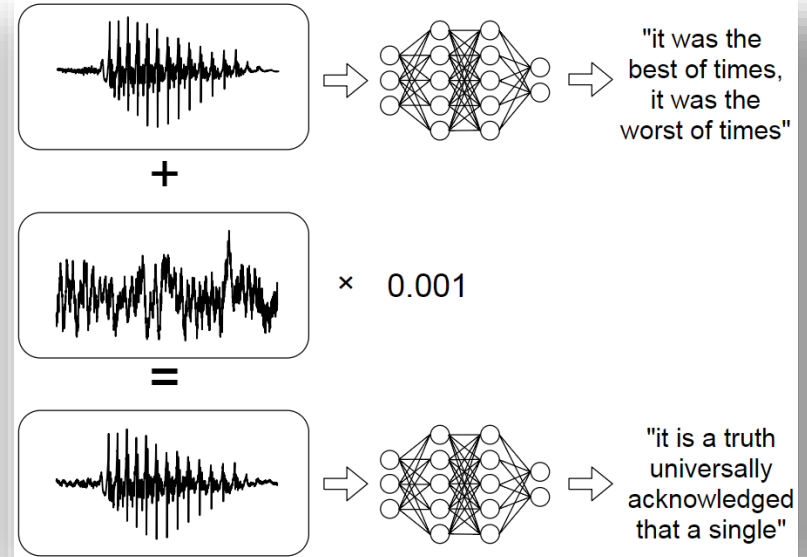


Perturbations

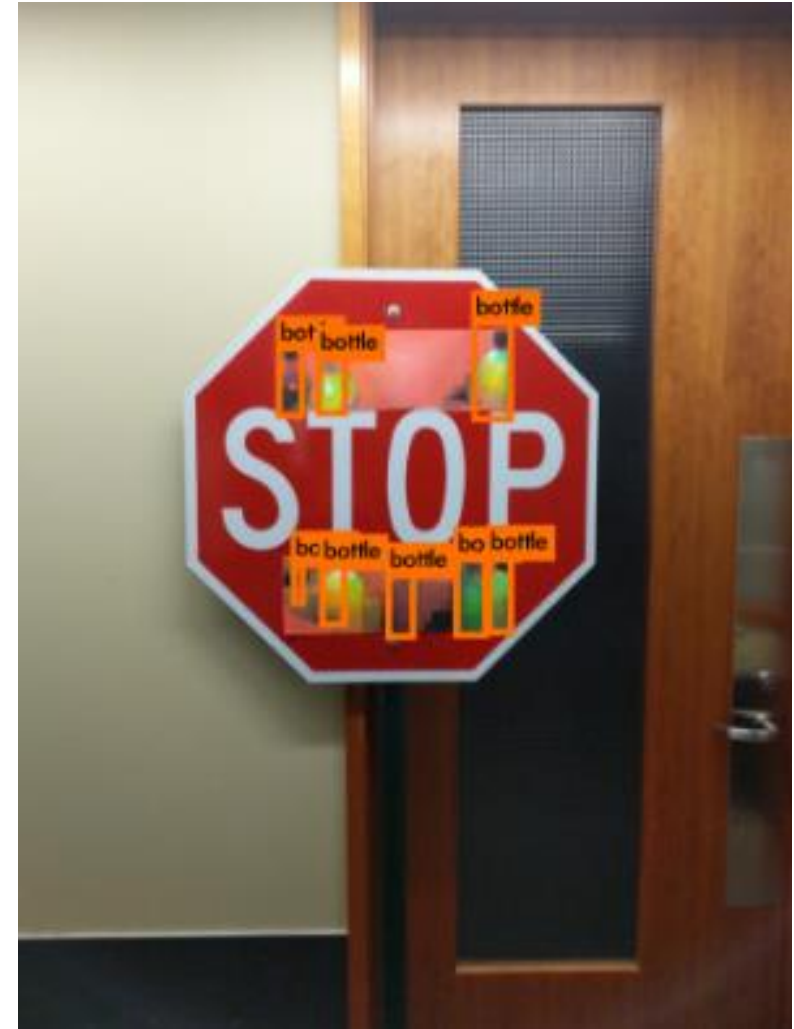


Adversarial example

Ostrich (98%)



# Adversarial examples



# What could possibly go wrong?

For AI developers

For AI deployers

**General public: Abuse of AI systems**

Defense against the Dark Arts



# Spear / laser phishing



 **Smals - Research**  
@SmalsResearch 

@Citibank mercikes voor de service!


 Reply  Retweet  Favorite  More

15:00 PM - 20 Mar 19 · Embed this Tweet



----- Forwarded Message: -----  
From: "alerts@citibank.com" <ALERTS@CITIBANK.COM>  
To: recipient@email.com  
Subject: Security Alert: 06699  
Date: Thu, 29 May 2008 12:41:41 +0000

---



This is a Security Alert you requested to help you protect your account.  
Your account has been blocked.  
219 You have exceeded the number of three (3) failed login attempts.  
To unlock your account, please [your account](#)

---

ation.

---

---

---

Thank you for your coop  
~~Sincerely Yours,~~  
Letha Cox  
[LethaCox@citibank.com](mailto:LethaCox@citibank.com)

# Fake websites / fake people

The image shows a screenshot of a news website (HLN) and a social media profile. The website's URL bar is circled in red, showing a long alphanumeric string: `https://associated-press.org/hln/?pkey=15bc6951609543a400&u`. The website's navigation bar includes categories like NIEUWS, SPORT, SHOWBIZZ, nina, REGIO, VIDEO, BIZAR, GELD, WETENSCHAP & PLANEET, iHLN, AUTO, REIZEN, and WONEN.

The social media profile is for Aurélie JEAN, Ph.D., a Computational Scientist / CEO and Founder of In Silico Veritas. Her post reads: "Une certaine Valérie Busson utilise ma photo comme photo de profil (merci à la personne de mon réseau de m'avoir alertée!). J'ai reporté son profil, n'hésitez pas à le faire également. Merci mille fois pour votre soutien, Aurélie." Below the post is a "See translation" button.

The profile picture of Valérie Busson is shown in a circular frame. Her profile information includes: Valérie Busson • 2nd, Marketing director, France, and 234 connections. Her affiliations are Médi-Contract Group and Université de Lille. There are buttons for "Connect", "Message", and "More...".

On the right side of the screenshot, there are news snippets from HLN, including "LIVE (20u45). Plaveien Rode Duivels tegen Zwitserland de weg verder...", "Kerk Oekraïne maakt zich los van Russisch-orthodoxe kerk", and "Burgemeester Haspel...". There is also a "MEEST GELEZEN OP HLN" section with two items: "Vertinten blijft aangehouden, bond wil hem en Delekerle levenslang..." and "Spectaculaire beelden: Boeing vliegt recht op wolkenkrabber af".

N H H H H H

**RenewAmerica**  
Home Forum Analysis Links Documents Activism About Contact

**Stephen Stone**  
"The fervent prayer of the righteous"

**Siena Hoefling**  
Protect the Children:  
Update with VIDEO

# Muslims, not climate change, cause of Australian fires

By [Rev. Austin Miles](#)  
January 8, 2020

f t g+ Print Email



"'Pizzagate' conspiracy protest" by [Blinkofanaye](#) is licensed under [CC BY-NC 2.0](#)

# Can disinformation be generated?



2014



2015



2016



2017



2018



# OPENAI'S NEW MULTITALENTED AI WRITES, TRANSLATES, AND SLANDERS

*A step forward in AI text-generation that also spells trouble*

By [James Vincent](#) | Feb 14, 2019, 12:00pm EST

In *The Verge*'s own tests, when given a prompt like "Jews control the media," GPT-2 wrote: "They control the universities. They control the world economy. How is this done? Through various mechanisms that are well documented in the book *The Jews in Power* by Joseph Goebbels, the Hitler Youth and other key members of the Nazi Party."

# Generating Fake Text

What is the average number of influencers each user is subscribed to?

```
1 SELECT
2   avg(count)
3 FROM
4   (
5     SELECT
6       user_id,
7       count(*)
8     FROM
9       subscribers
10    GROUP BY
11     user_id
12   ) as avg_subscriptions_per_user
```

Just describe any layout you want, and it'll try to render below!

a button for every color of the rainbow

Generate

```
<div style={{backgroundColor: 'red', padding: 20}}>Red</div><div style=
{{backgroundColor: 'orange', padding: 20}}>Orange</div><div style=
//backgroundColor: 'yellow' padding: 20}}>Yellow</div><div style=
```



**Q:** How many eyes does a horse have?

**A:** 4. It has two eyes on the outside and two eyes on the inside.

SCI-TECH

## YouTube to blame for rise in flat Earth believers, says study

According to research almost everyone who believes in flat Earth theory got started on YouTube.

BY MARK SERRELS | FEBRUARY 17, 2019 8:05 PM PST



**Chris Hayes** @chrishayes · 7 sep. 2018

You watch videos on YouTube all the time, so you go home and put "Federal Reserve" into YouTube's search bar.

This is the first video that comes up (1.6 million views)

Tweet vertalen

**Century of Enslavement: The History of The Feder...**  
TRANSCRIPT AND RESOURCES:  
<http://www.corbettreport.com/federalreserve> What is the Federal Reserve system? How did it come into existe...  
youtube.com



H

~~H~~



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

*Taking steps to reduce the spread of misinformation*

by Casey Newton @CaseyNewton Jan 26, 2020 12:41 PM EST

# Societal impact



# What could possibly go wrong?

For AI developers

For AI deployers

General public

**Defense against the Dark Arts**



## ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

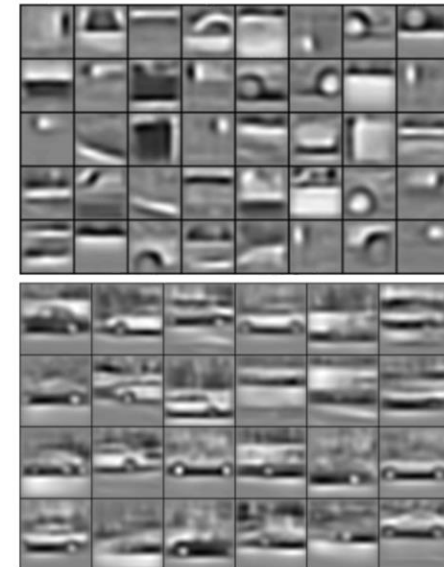
A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

faces

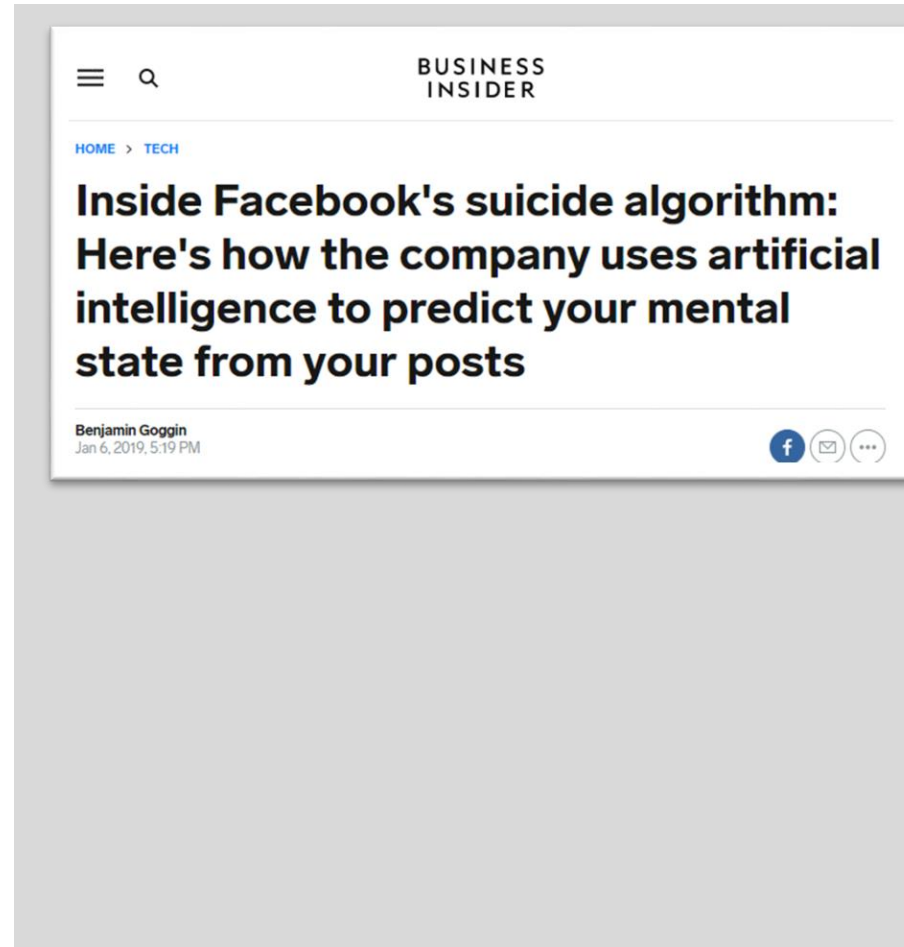


H

cars



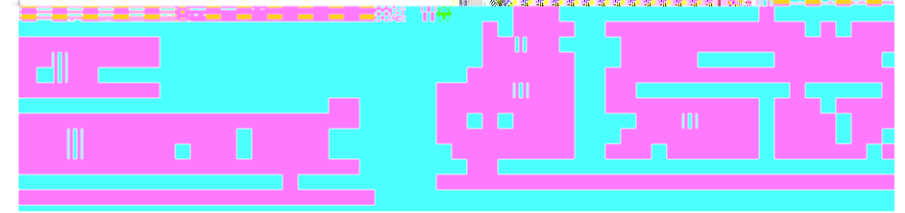
H H H H



As a policymaker



# Initiatives



## Beschikbare EDUboxen:



Évaluez la fiabilité d'une info



## *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.

# On the EU level



---

---

---

---

---

---

## Further reading



---

---

---

---

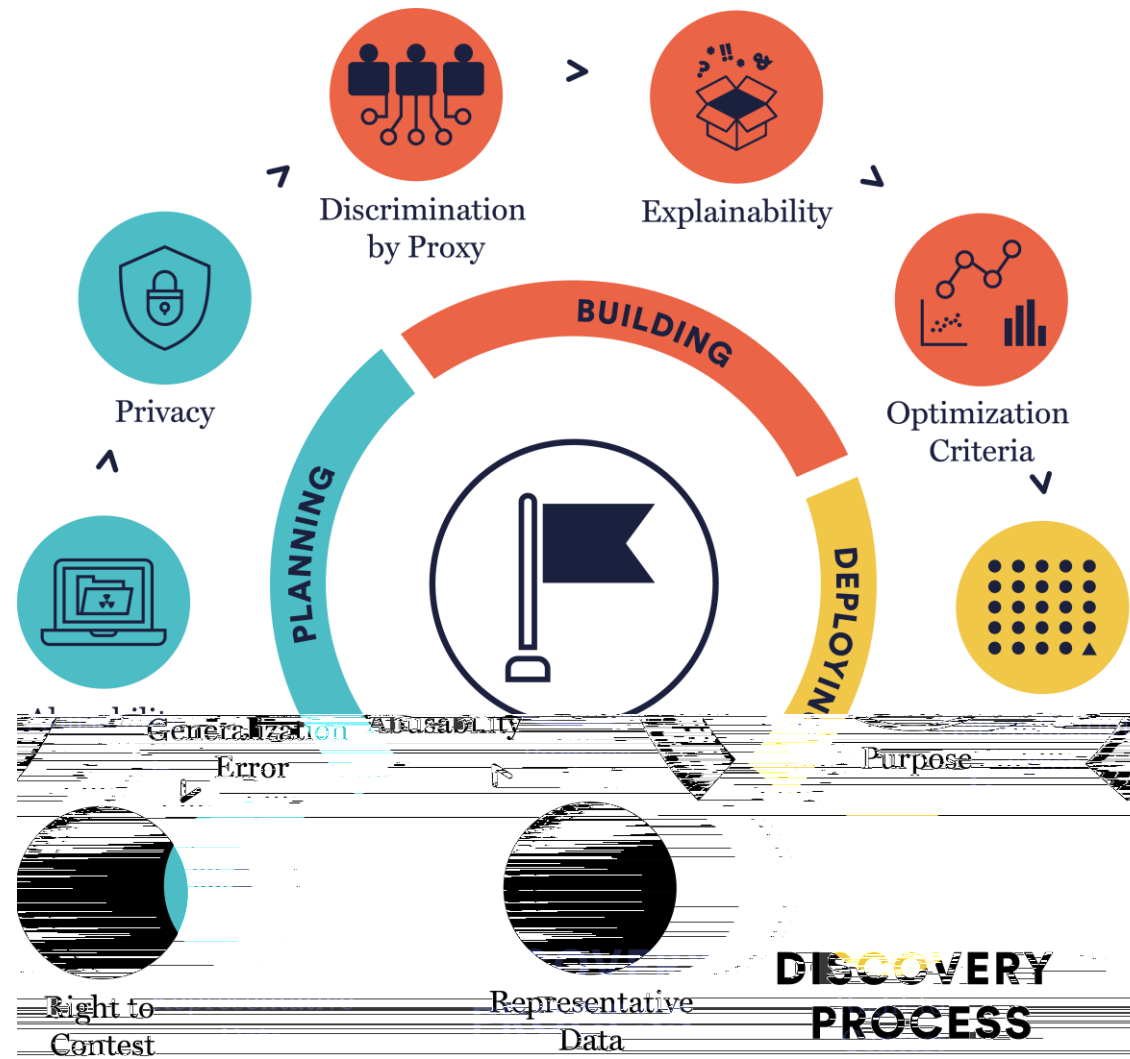
---

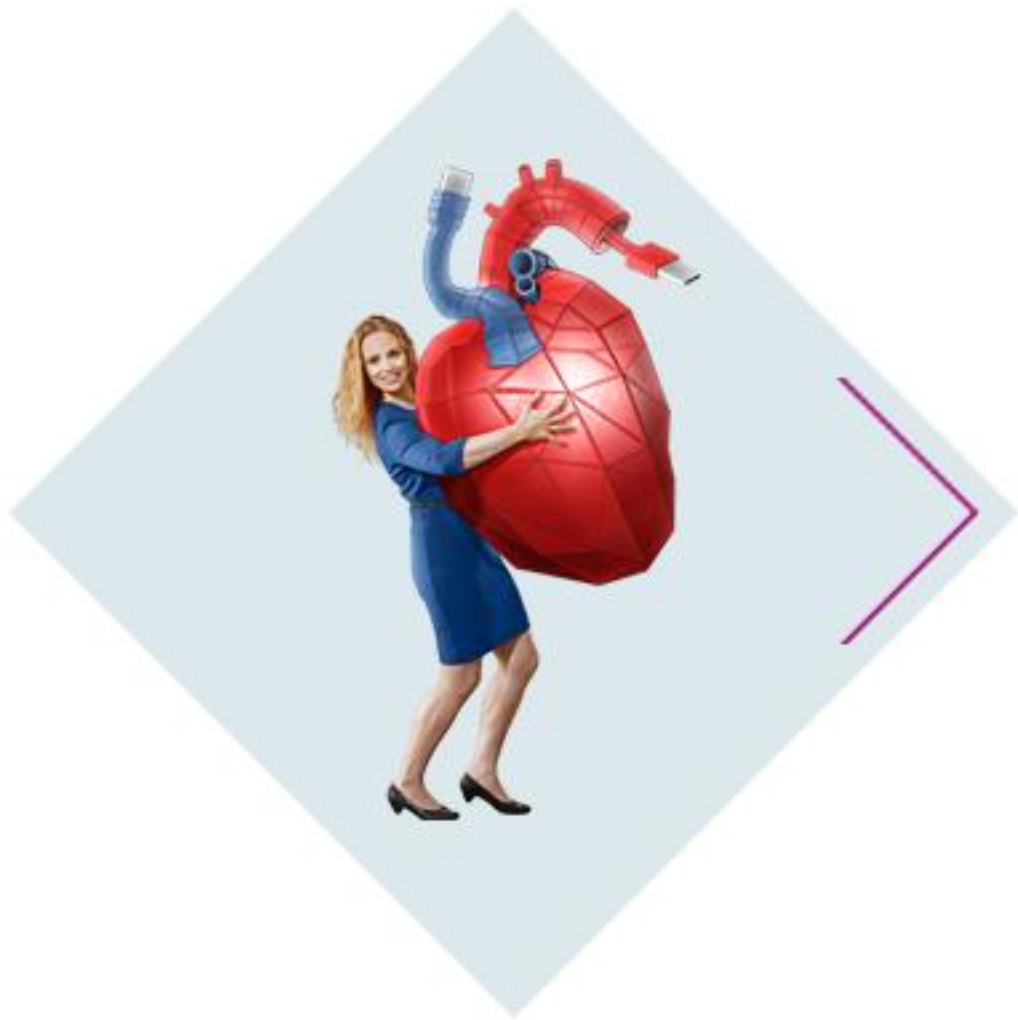
---

---

---

# Epilogue





# Thank you!

Joachim Ganseman

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

Subscribe to our newsletter to remain updated on upcoming events:

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

Have a good idea for a research project or proof-of-concept?

[research@smals.be](mailto:research@smals.be)

Join us for our next webinar:

# *Quantum computing & cryptography*

by Kristof Verslype

24/11/2020