

Some pitfalls of AI

Joachim Ganseman
Smals Research

15/09/2020



**Innovation with
new technologies**



**Consultancy
& expertise**



**Internal & external
knowledge transfer**



**Support for
going live**

NewSQL
Databases

Anomalies &
Transaction
Management

Conversational
Interfaces

Near-real-time
Translation

Web Scraping
for Analytics

GIS for
Analytics

Augmented
Reality

Graph Analytics
Visualisation

2020

AI Cases &
Deployment

Robotic Process
Automation

Crypto Cases

European
Blockchain
Infrastructure

Knowledge
Graphs

Advanced
Cryptography

Quantum
Computing &
Cryptography

FIDO2 / Web
Authentication

AI: part of our daily life

Edventure.com - Spam - Mozilla Firefox

File Edit View Bookmarks Tools Help

Getting Started Latest Headlines

Google

Flickr: Recent activity on your photos

Edventure.com - Spam

International Herald Tribune - Bush moves to save Mukasey nomination 1 1/2 hours ago

Inbox (10) Delete Forever Not Spam More actions... Refresh

Select: All, None, Read, Unread, Starred, Unstarred

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

Yong Whitfield Finally a Professional Answer to your Debt Problems

Lina Bello Léelo. Es importante para ti . gd 79 - APRENDER A HAB

Austin Seals Medications that you need. - Buy Must Have medications

me October 70% OFF - Click to buy Viagra for as low as \$1.53

Tiffany - 全球最高等級服務品! 所有商品全部附有原廠證書! - 540609463698961848281 5:50 am

Medicine & Health Minute Here's What to Do When An Employee Criticizes A Sup 5:49 am

Турагентство РАЙ ЛУЧШИЕ ТУРЫ У НАС! - Турагентство «РАЙ» предлагает

Workplace Advisor Big Dress Mistakes That Can Cost You At Work - Busine 5:47 am

Gerald F. Open minded people - Here you go! Find MATES in your 5:46 am

Quick Contacts

Search, add, or invite

Esther Dyson Set status here

celiot Denny Goetz

edyson edyson

John Boynton

Ken Lyon

Lenny Pridatko

Linda Avery

mcapes

Peter Diamandis

Simon Grice

Add contact Show all

Labels

Boxbe (1)

Done

"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](#)

google.be

Data Science events smals legal RSZ

Gmail Images

Larga vida al Rock and Roll

August 22 ·

Like Page

Like Comment Share

Cuando te obligan a cantar opera pero tu pasión es el rock

When you're forced to sing opera but your passion is rock

Hide original · Rate this translation

me from the sandlot court

Made for you

Get better recommendations the more you listen.

Your Discover Weekly

Discover Weekly

Your Daily Mix 1

Daily Mix 1

Your Daily Mix 2

Daily Mix 2

wat is smals

what is smals wee

what is smals

what is smals real

what is smals claim

what is smallseotoc

what is smallsat rid

what is smallsat co

what is smallsat slang

what is smals for a

wat is smals

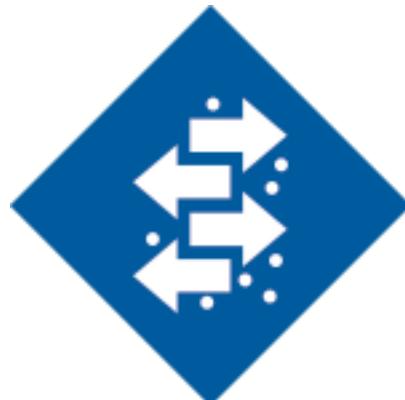
What could possibly go wrong?

Screwing up your own AI

Someone screws with your AI

Someone's AI screws with you

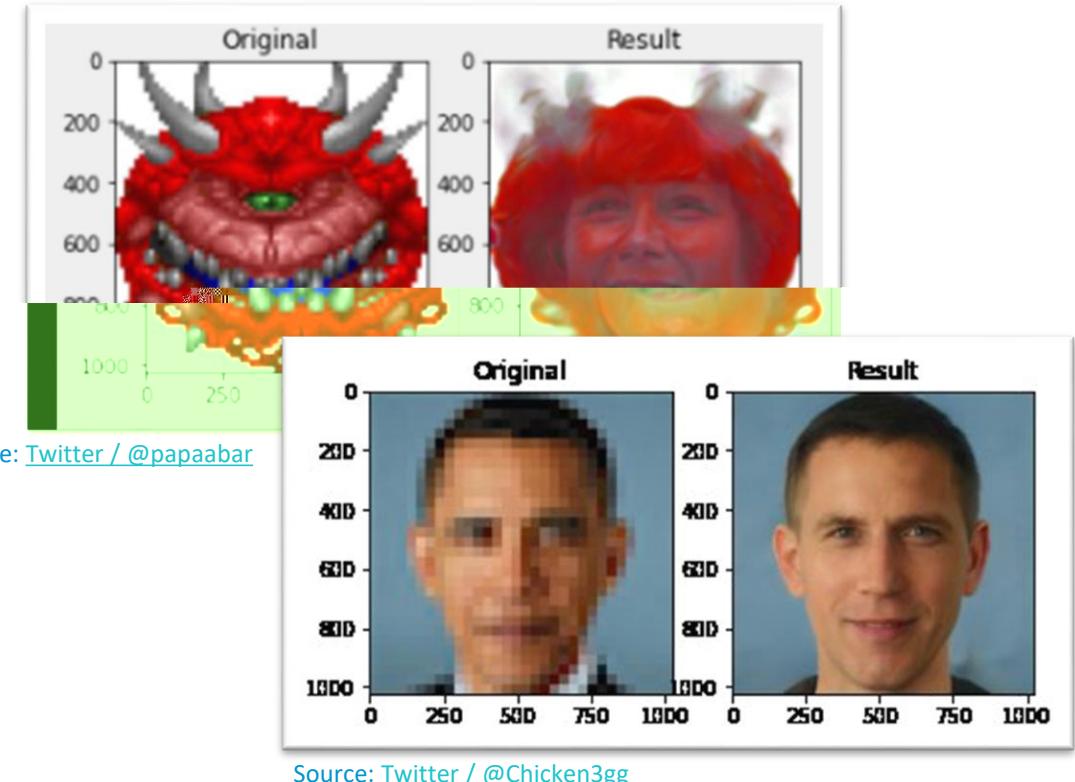
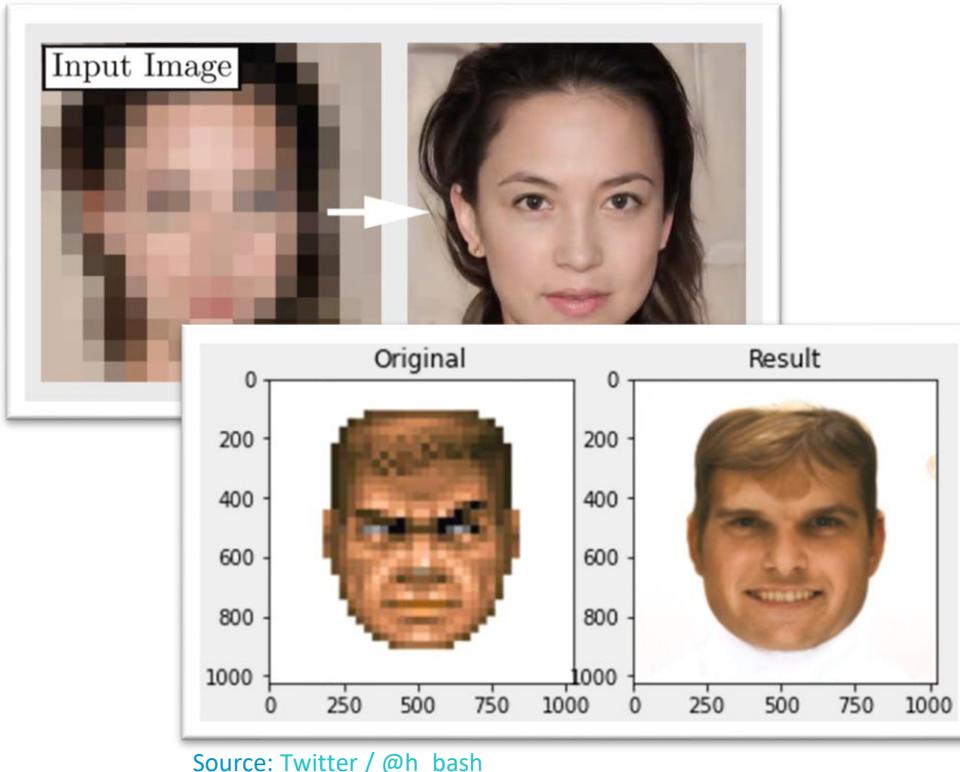
Unscrewing things



- **For AI developers: from data to decision**
Data collection issues (bias vs. fairness)
Data processing issues (confounding variables)
Goal (mis)formulation
- For AI deployers
Data poisoning
Adversarial examples
- General public
Spear phishing
(personalized) disinformation
The role of recommender systems
- Defense against the Dark Arts
Transparency & explainability
Digital Skepticism
Policy

- AI systems are trained on data
→ **Garbage in, garbage out**
- Training data is ideally
independent and identically distributed (iid) over the domain
= well-balanced & free from hidden correlations
- In reality, this is rarely the case
How many men named Anna do you know?

- June 2020: Face Depixelizer (generates a face that fits a pixelated image)

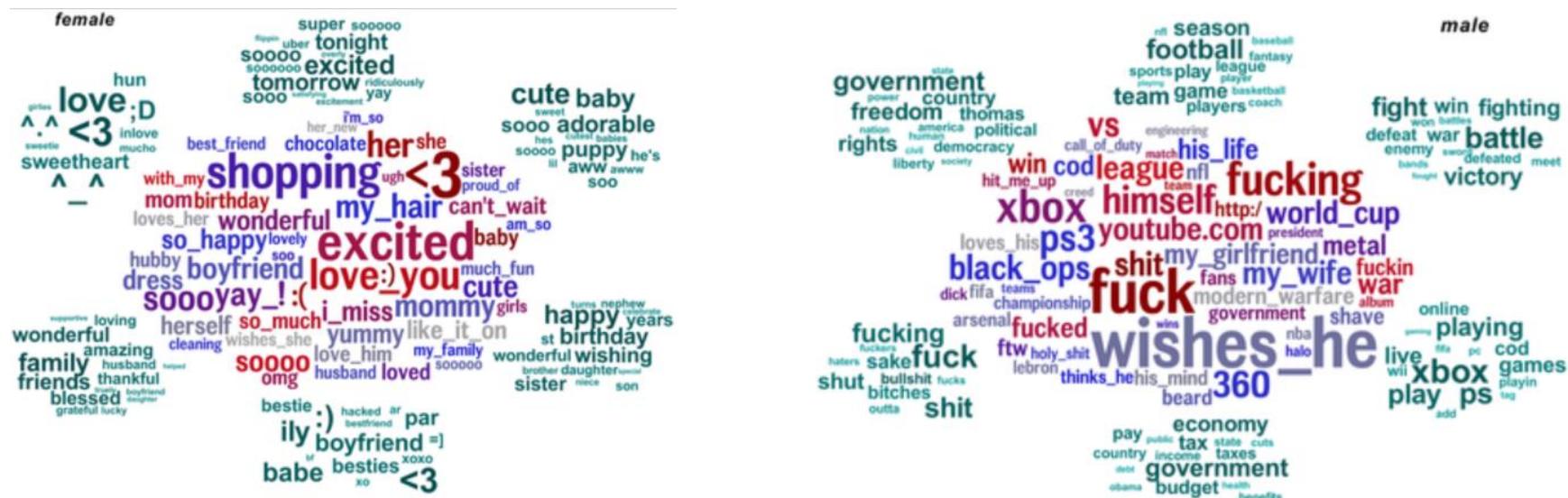


- Training dataset (Flickr-Faces-HQ) contains less people of color / elderly
- Used method (StyleGAN) overvalues “average” → leans towards young whites
- Q: Would we detect less visible biases too, e.g. in mortgage applications?

- Biased humans → biased data
 - https://en.wikipedia.org/wiki/List_of_cognitive_biases
- Biased data → biased AI systems
 - Curse of dimensionality: impossible to cover every combination of every parameter



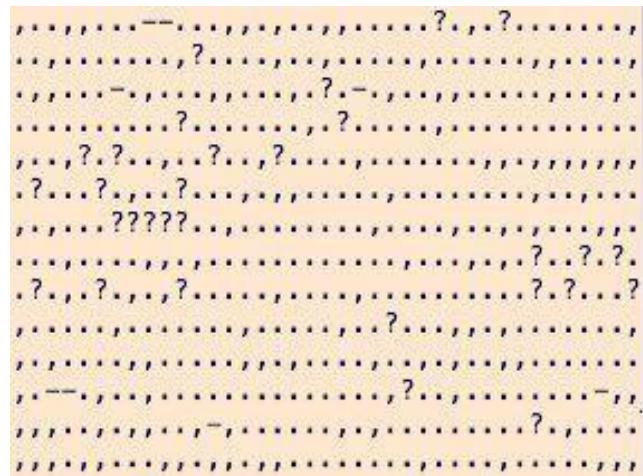
- We'll fix it by not taking protected characteristics into account, right?
- ... well...
 - Men/women have different ways of speaking



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

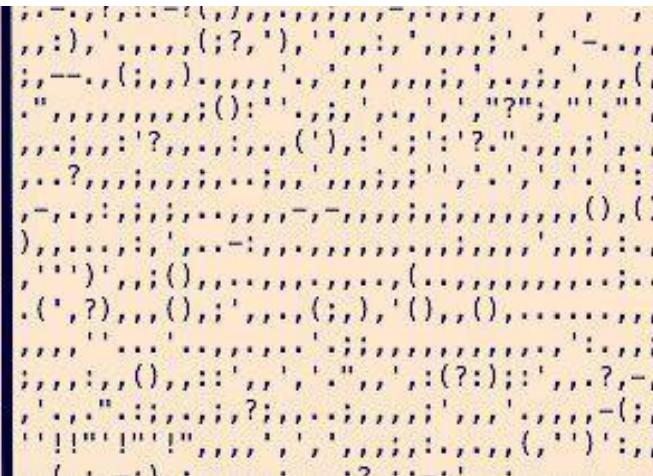
- Definition: hidden property influencing known properties and outcomes
- Sometimes leads to surprising new insights!

Blood Meridian
(Cormac McCarthy)



.....-.....?.,?.....
.,?.....,.....,.....
.....?.,.....?.,.....
.,?.,?.,?.,?.,.....
.,?.,?.,?.,?.,.....
.....?????.,.....,.....
.....,.....,?.,.....,.....
.....,.....,?.,.....,.....
.....,.....,?.,.....,.....
.....,.....,?.,.....,.....

Absalom, Absalom
(William Faulkner)



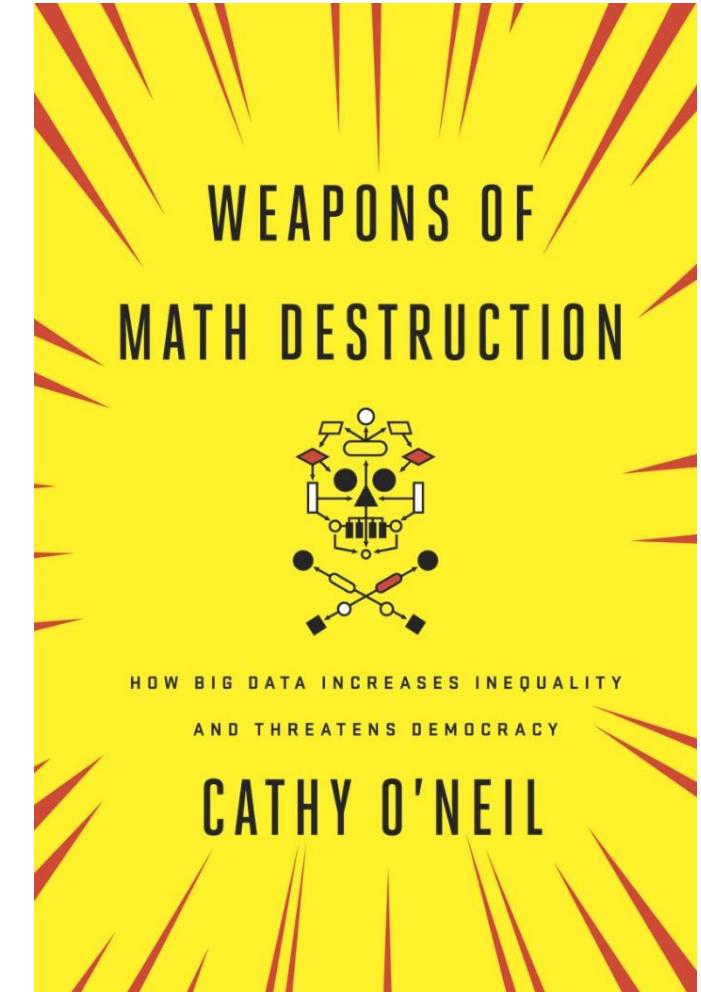
,:),',.....(;?,'.....
;,-,(.,).....'.....,.....,.....,.....
.....;();'.....,.....,?";,".....
.....?.....,.....('),':';?.....
.....?.....,.....,.....,.....,.....
.....,.....,-,.....,.....,.....,.....,.....,.....
.....,.....,.....,.....,.....,.....,.....,.....,.....
.....,.....,.....,.....,.....,.....,.....,.....,.....
.....,.....,.....,.....,.....,.....,.....,.....,.....
.....,.....,.....,.....,.....,.....,.....,.....,.....

→ AI does not necessarily learn what you want it to learn!

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

- 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)
- Tools exist to help spot unfair bias
 - <http://aiblindspot.media.mit.edu/>
 - <https://data-en-maatschappij.ai/en/tools>

→ Know your data, your algorithms, and their limitations



- AI/ML algorithms *optimize*, i.e. **minimize a loss** or **maximize a reward**
 - reward “success”
 - punish “failure”
- “Success” can be hard to define
 - Engineers (over)simplify the goals
 - Additional conditions may be forgotten
- AI follows the specs but may
 - exploit bugs or unexpected data properties
 - get stuck in endless loops

(For more examples, [see this spreadsheet](#))



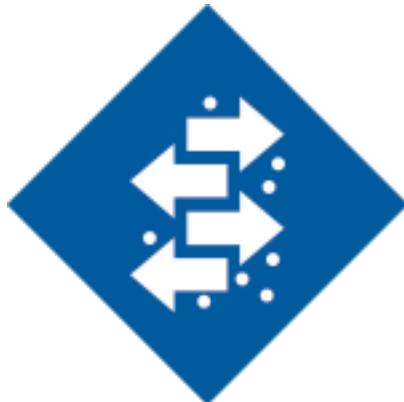
Custard Smingleigh
@Smingleigh

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.

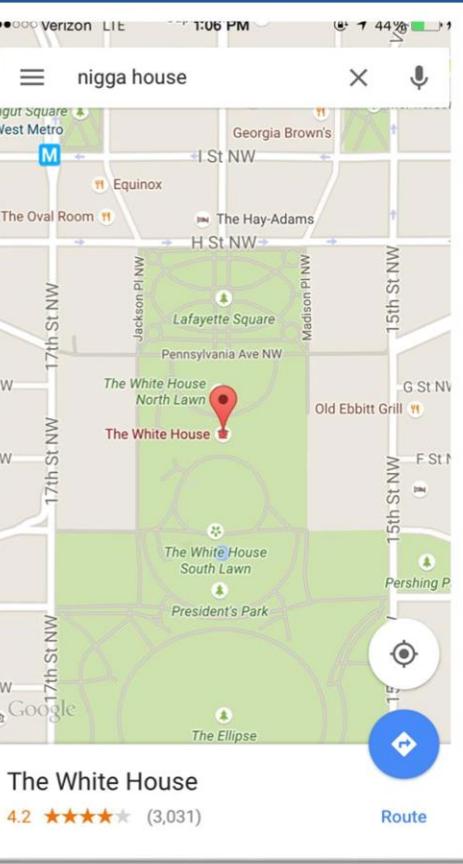
It learnt to drive backwards, because there are no bumpers on the back.

Source: [Twitter / @Smingleigh](#)

- For AI developers
 - Data collection issues (bias vs. fairness)
 - Data processing issues (confounding variables)
 - Goal (mis)formulation
- **For AI deployers: attacks against AI systems**
 - Data poisoning
 - Adversarial examples
- General public
 - Spear phishing
 - (personalized) disinformation
 - The role of recommender systems
- Defense against the Dark Arts
 - Transparency & explainability
 - Digital Skepticism
 - Policy



- Inject false training data to compromise learning
 - Intentionally mislabeled data
 - Bogus data or noise
- Crowdsourcing risks
 - Individual jokers
 - Coordinated attacks (Twitter/4chan/reddit mobs)
- Webscraping risks
 - Wiki vandalism
 - Inclusion of shady websites



Source: [Twitter / @iambomanix](#)

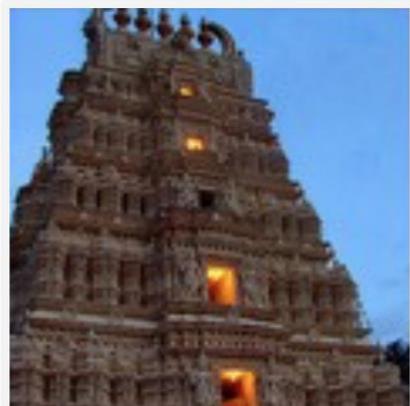
Nearly Half of Scottish Wikipedia is Incorrectly Written by a US Teen

By *Ryan Lappe* on 1st September 2020

Source: [trillmag.com](#)

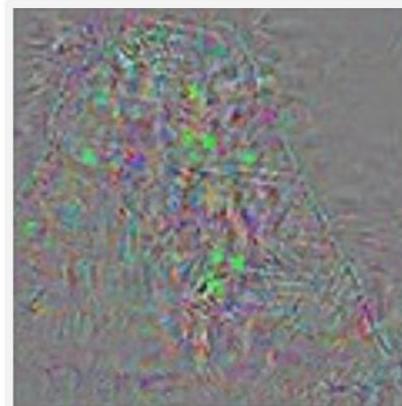
→ **Data verification** is not a luxury!

- Minimal change to input → large change in output

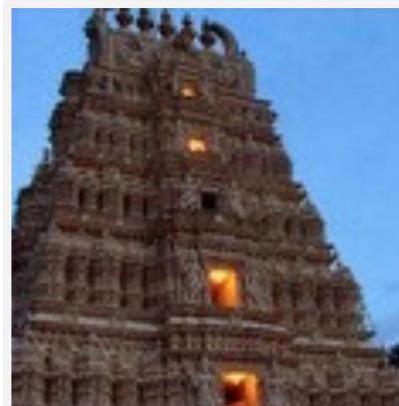


Original image

Temple (97%)

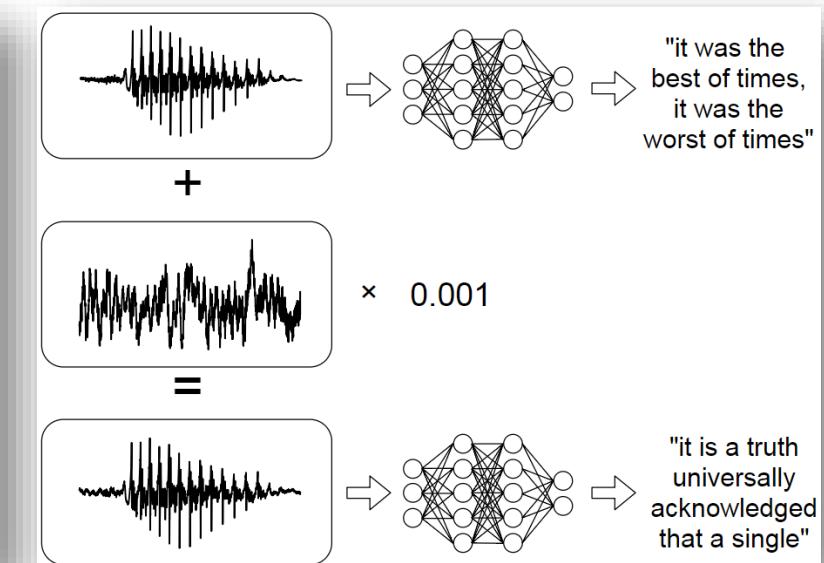


Perturbations



Adversarial example

Ostrich (98%)



Source: [hackernoon.com / Julien Despois, "Adversarial examples and their implications"](https://hackernoon.com/julien-despois-adversarial-examples-and-their-implications),
as adapted from: [Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, "Intriguing properties of neural networks"](https://arxiv.org/abs/1412.6572)

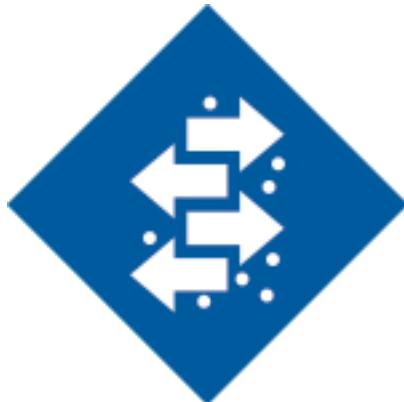
Source: [Nicolas Carlini & David Wagner, "Audio Adversarial Examples: targeted attacks on speech-to-text"](https://arxiv.org/abs/1611.01285)

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



Source: bair.berkeley.edu / Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Bo Li et al., ["Physical Adversarial Examples Against Deep Neural Networks"](#)

- For AI developers
 - Data collection issues (bias vs. fairness)
 - Data processing issues (confounding variables)
 - Goal (mis)formulation
- For AI deployers
 - Data poisoning
 - Adversarial examples
- **General public: Abuse of AI systems**
 - Spear phishing
 - (personalized) disinformation
 - The role of recommender systems
- Defense against the Dark Arts
 - Transparency & explainability
 - Digital Skepticism
 - Policy



- Fraudulent attempt to obtain sensitive information, directed at a **specific individual/company**
- Webscraping + AI may be deployed to **personalize** messages to many targets → “laser phishing”

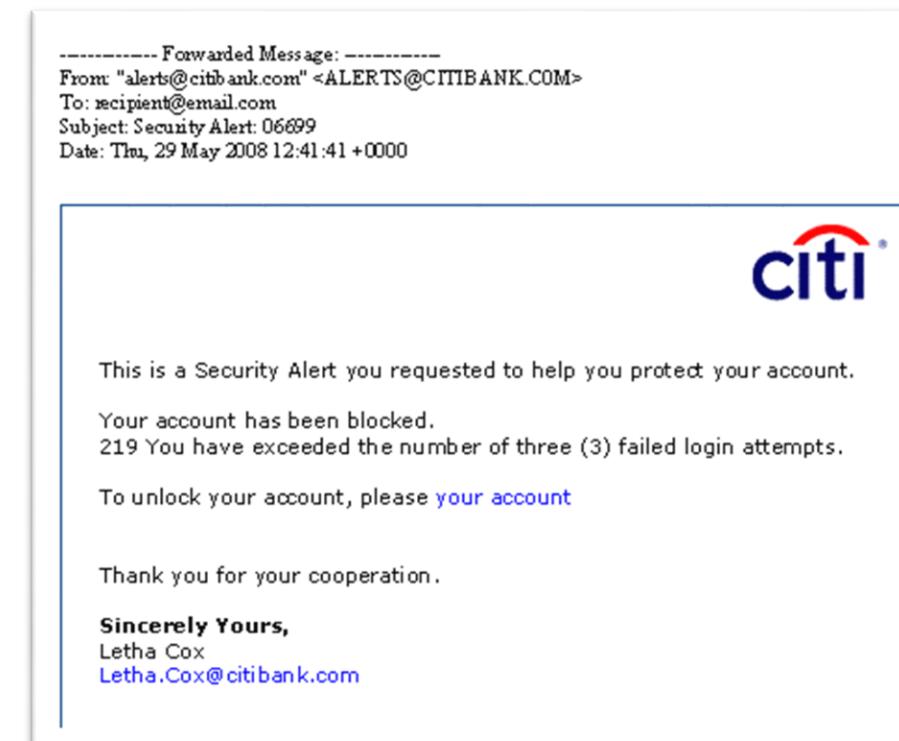
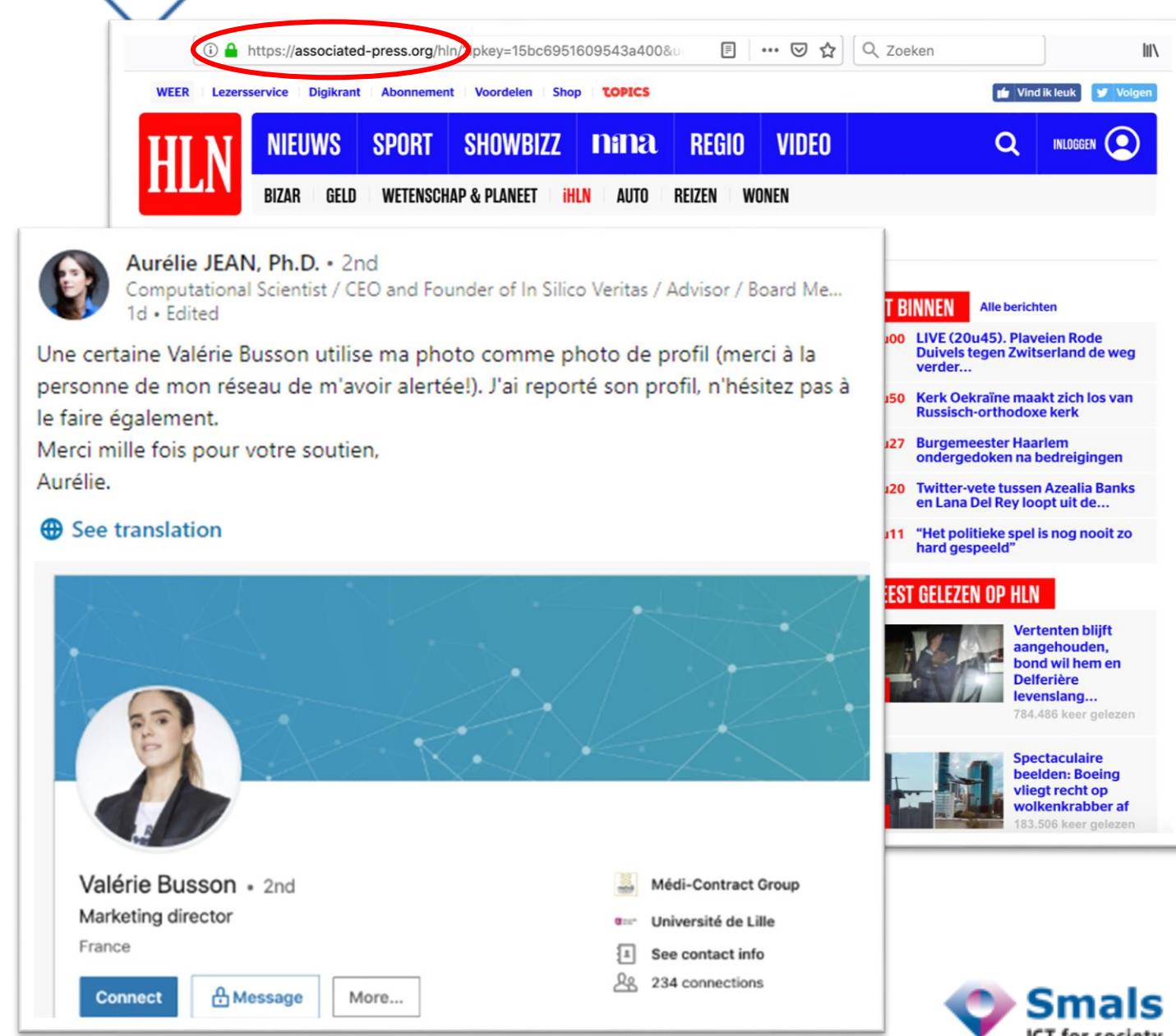


Image source: enisa.europa.eu

Fake websites / fake people

- **Fake websites**
 - Scams
 - Phishing
 - Pyramid schemes
 - ...

- **Fake profiles**
 - Impersonations, “CEO fraud”
 - Creation of “bot armies”
 - Sales / product review fraud
 - Social media surveillance
 - Influencing
 - ...



The screenshot shows a LinkedIn profile page for Valérie Busson. The URL in the browser's address bar is circled in red and shows <https://associated-press.org/hln/>, which is a clear indicator of a fake website. The LinkedIn interface includes a header with the HLN logo, a navigation bar with categories like NIEUWS, SPORT, SHOWBIZZ, and REGIO, and a sidebar with news items. The profile page itself displays a placeholder profile picture and a fake LinkedIn URL.

Source: [LinkedIn / Aurélie Jean](#)

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**

Source: bellingcat.com / Robert Evans, [“How Coronavirus Disinformation Gets Past Social Media Moderators”](#)



The screenshot shows a news article on the RenewAmerica website. The header features a painting of the Declaration of Independence. The article title is "Muslims, not climate change, cause of Australian fires" by Rev. Austin Miles, dated January 8, 2020. The author's bio includes "Stephen Stone 'The fervent prayer of the righteous'" and "Siena Hoefling Protect the Children: Update with VIDEO". The article discusses the Australian bushfires and blames them on Muslim immigrants. Social sharing icons for Facebook, Twitter, Google+, Print, and Email are visible at the bottom.

- **It is not:**
 - (Extreme) political, scientific, ethical or moral viewpoints
 - Unions, lobbying, advocacy, campaigning, ...
 - Selective presentation of information
 - Satire, parody, ...
 - Religion

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



- Image/video/audio: yes, kind of



2014



2015



2016



2017



2018

Source: [Twitter / @goodfellow_ian](#)

cf. *deepfakes*:



Source: [Twitter / @ousathesquid](#)

- From *The Verge*:

OPENAI'S NEW MULTITALANTED 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."

- "We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter."
(*Jeremy Howard, Fast.AI*)

- State-of-the-art [GPT-3](#) (05/06/2020) generates more than prose

- Code
- Layouts
- Translations
- Basic reasoning
- ...

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
```

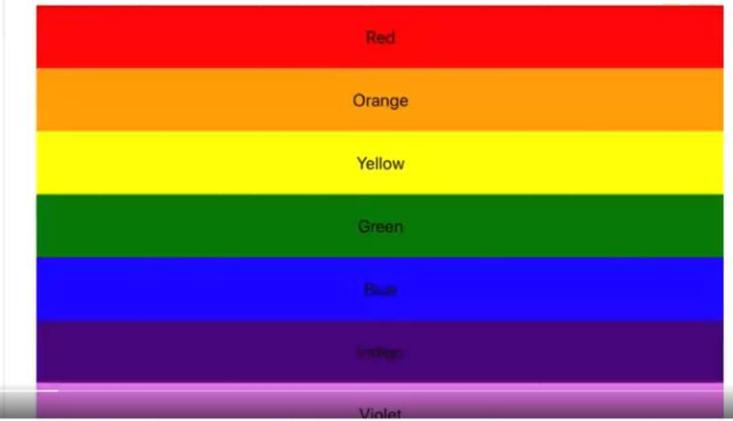
Source: [Twitter / @FaraazNishtar](#)

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={{backgroundColor: 'green', padding: 20}}>Green</div><div style={{backgroundColor: 'blue', padding: 20}}>Blue</div><div style={{backgroundColor: 'indigo', padding: 20}}>Indigo</div><div style={{backgroundColor: 'violet', padding: 20}}>Violet</div>
```



Source: [Twitter / @sharifshameem](#)

- Training cost: ± \$4.000.000 (on external cloud service)
- Not perfect,
nor “intelligent”:

Q: How many eyes does a horse have?

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

Source: [Twitter / @eturner303](#)

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

Source: [cnet.com](#)

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  @chrislhayes · 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 Federal Reserve
TRANSCRIPT AND RESOURCES:
<http://www.corbettreport.com/federalreserve> What is the Federal Reserve system? How did it come into existence?
youtube.com

Source: [Twitter / @chrislhayes](#)

- Similar on many other (free) platforms with recommendation systems: Instagram, TikTok, tabloid websites etc.

- Consumer objective ≠ producer objective
 - You: want to find good information
 - Social media: wants you to keep watching (ads)
 - Promotes content that “pushes buttons”
→ Conspiracy theories, sensationalism, disturbing content, extremism, ...

- The recommendation feedback loop

Inflammatory **Any** content that 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

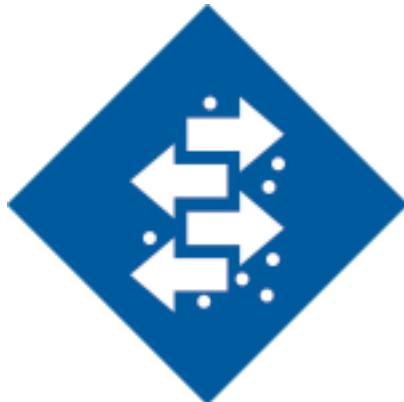
Source: [The Verge](#)

-- is this enough?

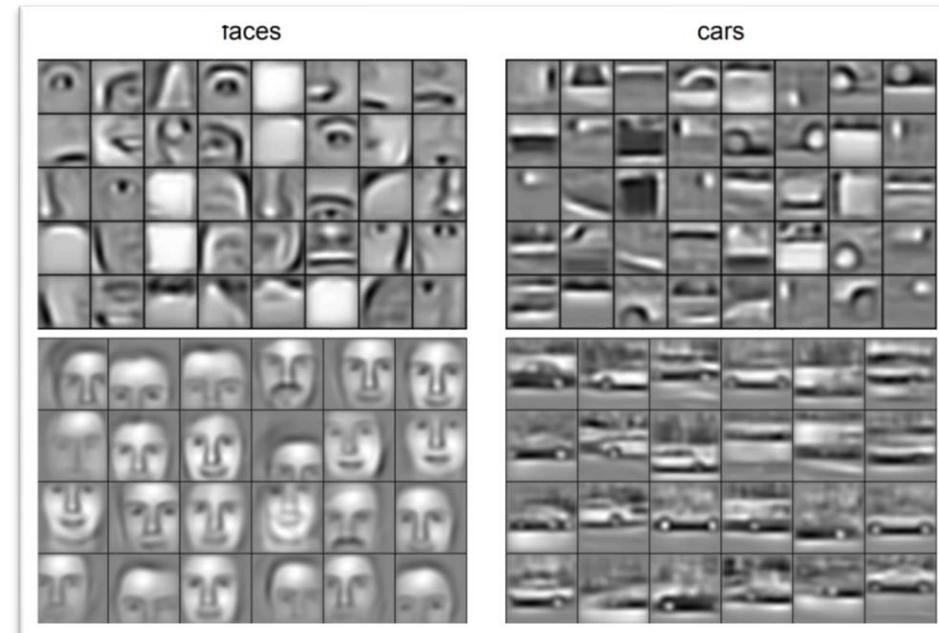
- Echo chambers
 - By default, you're mostly served pre-selected information
 - Who does the selection?
 - With what objective?
 - Mainstreaming of extreme content
 - Eroding trust, proliferation of conspiracy theories (e.g. QAnon)
- National politics, e.g. US 2016 election:
 - Search “Trump” → 81% of “up next” recommended videos is pro-Trump
 - Search “Clinton” → 88% of “up next” recommended videos is pro-Trump

(Source: Guillaume Chaslot, [“YouTube’s A.I. was divisive in the US presidential election”](#))
- International politics: information warfare
 - e.g. Russian reporting on MH17, Ukraine crisis, Crimea etc.

- For AI developers
 - Data collection issues (bias vs. fairness)
 - Data processing issues (confounding variables)
 - Goal (mis)formulation
- For AI deployers
 - Data poisoning
 - Adversarial examples
- General public
 - Spear phishing
 - (personalized) disinformation
 - The role of recommender systems
- **Defense against the Dark Arts**
 - Transparency & explainability
 - Digital Skepticism
 - Policy

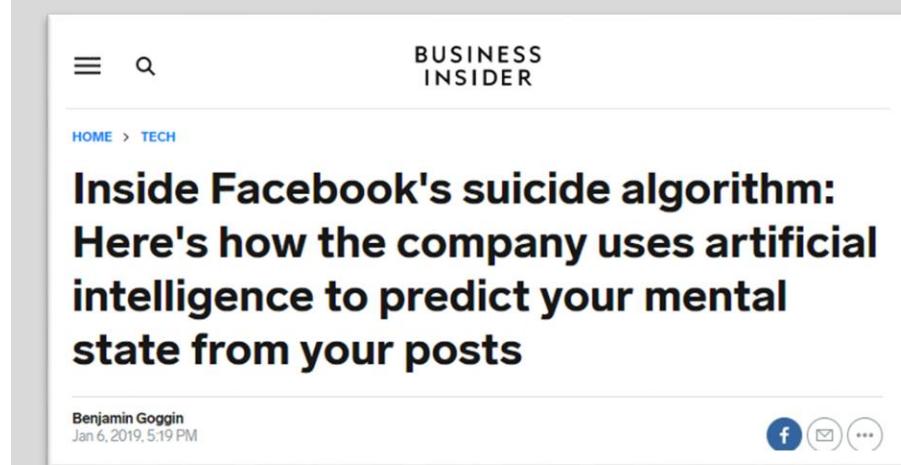


- Governance through FATE (sometimes FEAT)
 - Fairness, Accountability, Transparency, Ethics
- Guidelines and technical tools
 - <https://ethical.institute/principles.html>
 - Lime
 - IBM AI Fairness 360
 - Microsoft Fairlearn
 - Google Fairness-gym
 - ...
- Explainable AI
 - Important factor in **accountability**
 - Especially hard with deep learning
 - Still in its infancy



Source: Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks"

- Awareness
 - You are being profiled
 - What you see is not what someone else sees
 - Anything you post can be used against you
 - Technology and law keeps evolving
- Rely on authoritative, transparent sources
 - Peer-reviewed science
 - Quality journalism



Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts

Benjamin Goggin
Jan 6, 2019, 5:19 PM

Difficult questions that arise in practice:

- Does Facebook have the right to make these analyses?
- Can Facebook share the result with law enforcement, even “for your own good”?
- Consent? Privacy?
- What with faulty predictions?

→ Encourage **Digital Skepticism** (without being paranoid)

→ Requires some **Competences / Literacy**



- Awareness
 - Own vulnerability to pre-selected information
 - Advertisement-revenue driven recommendation feedback loop leads to online over-representation of extremes
 - Information warfare
- Stimulate
 - Independent and quality media
 - Innovation & research on the impact of innovation
 - Culture of permanent learning

Initiatives

- <https://data-en-maatschappij.ai/>
- <https://www.ai-cursus.be/>
- <https://www.knack.be/nieuws/factchecker/>
- <https://www.vrt.be/nl/vrtonderwijs/edubox/>
- ...

Beschikbare EDUboxen:



Faky beta No check, no trust.

Évaluez la fiabilité d'une info

Article Image

Insérez un lien, des mots-clés COLLER L'URL D'UN ARTICLE

Lancez votre recherche sur un article en français accessible sans abonnement

ÉVALUER L'INFO



- 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.

- 03/2015: Stratcom Task Force → euvsdisinfo.eu
- 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](#)
- 04/2019: EU HLEG [Ethics Guidelines for Trustworthy AI](#)
 - 07/2020: addition of [Assessment list for Trustworthy AI](#)
 - Belgian coordination: [AI4Belgium](#)

- **Reports**
 - [The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation](#) (“Malicious AI report”, 02/2018)
 - [For a Meaningful Artificial Intelligence](#) (“Villani report”, 03/2018)
 - [Information Manipulation, A Challenge for Our Democracies](#) (CAPS & IRSEM, France, 08/2018)
 - [Artificial Intelligence Primer](#) (OECD OPSI, 28/11/2019)
- **Organizations and Academia**
 - <https://montrealethics.ai/>
 - <https://cyber.harvard.edu/> / <https://ai.shorensteincenter.org/>
 - <https://www.turing.ac.uk/research/data-ethics>
 - <https://hai.stanford.edu/>
 - ...

Epilogue



[AI Blind Spot Discovery Process](#),
MIT Media Lab, licensed CC-BY 4.0

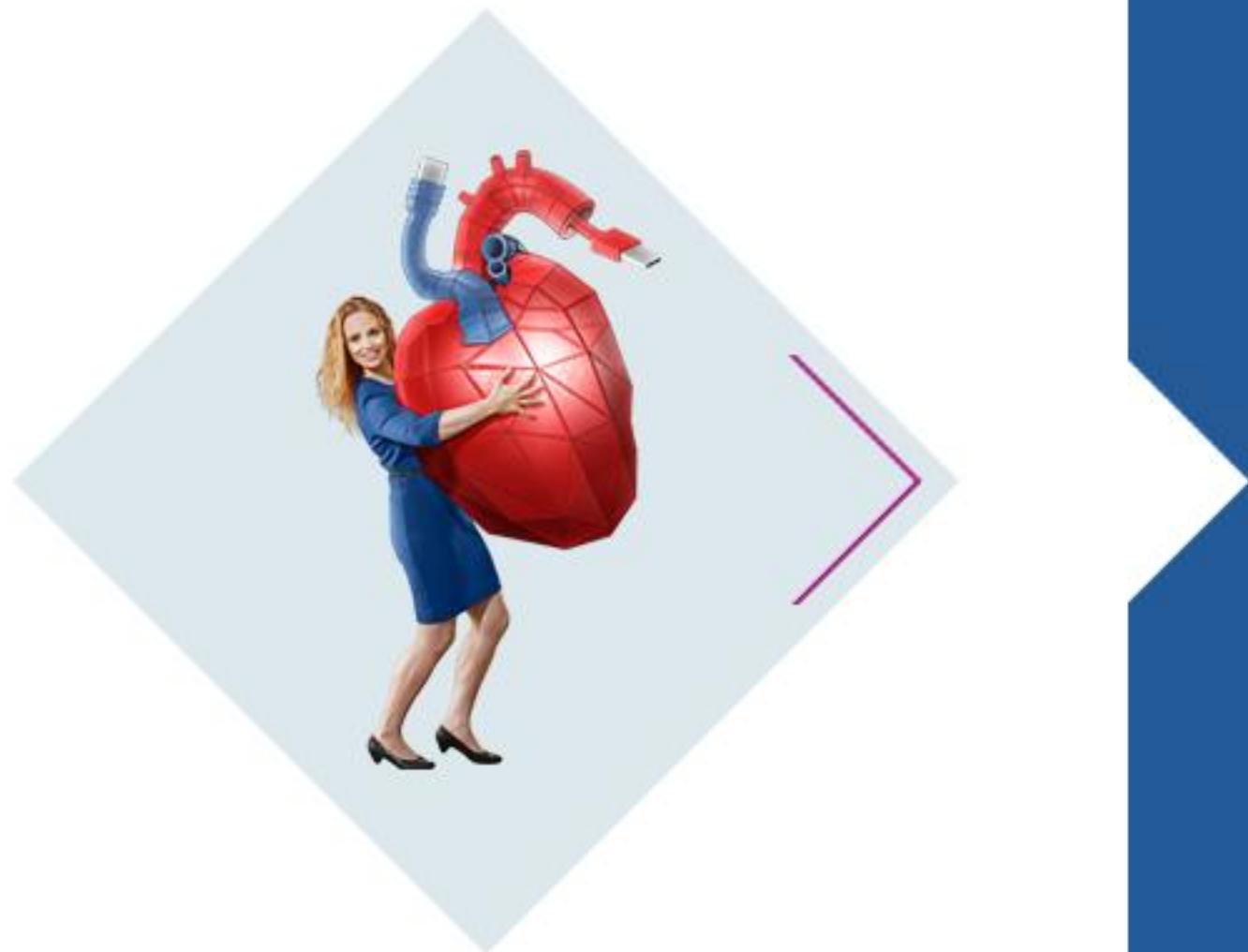
Thank you!

Joachim Ganseman
joachim.ganseman@smals.be

Subscribe to our newsletter to remain updated on upcoming events:

www.smalsresearch.be

Have a good idea for a research project or proof-of-concept?
research@smals.be





Join us for our next webinar:

Quantum computing & cryptography

by Kristof Verslype

24/11/2020