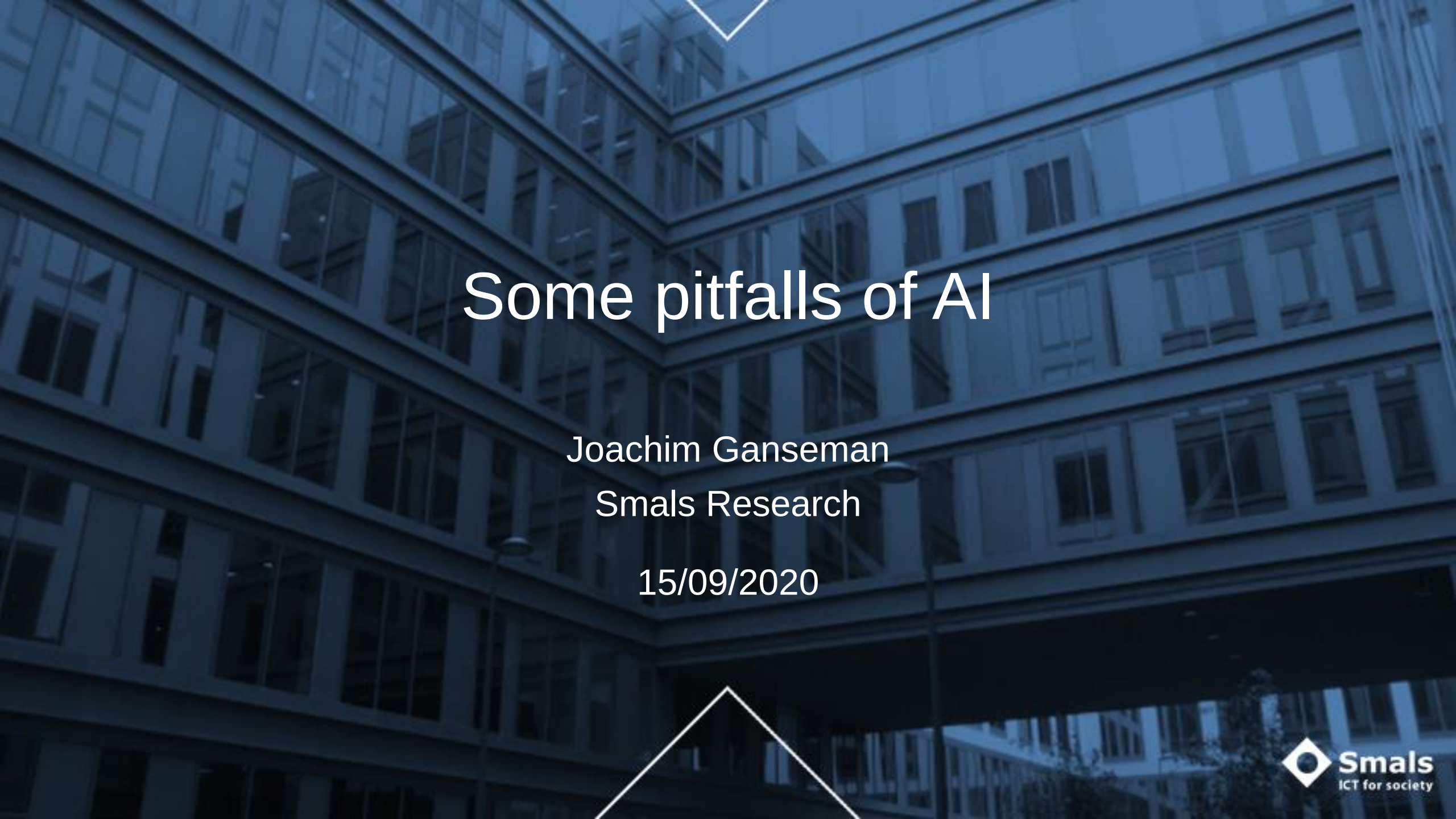


# Some pitfalls of AI

Joachim Ganseman

Smals Research

15/09/2020

Smals
ICT for society

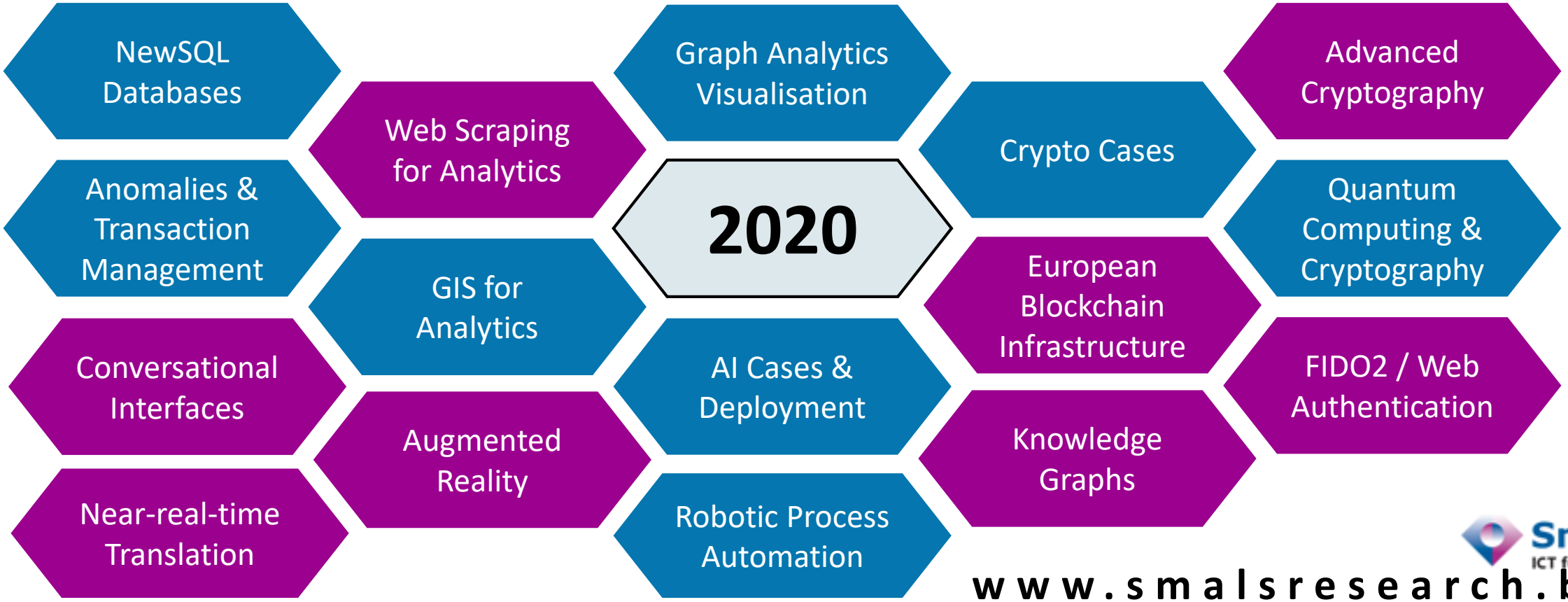
# Smals Research 2020

**Innovation with new technologies**

**Consultancy & expertise**

**Internal & external knowledge transfer**

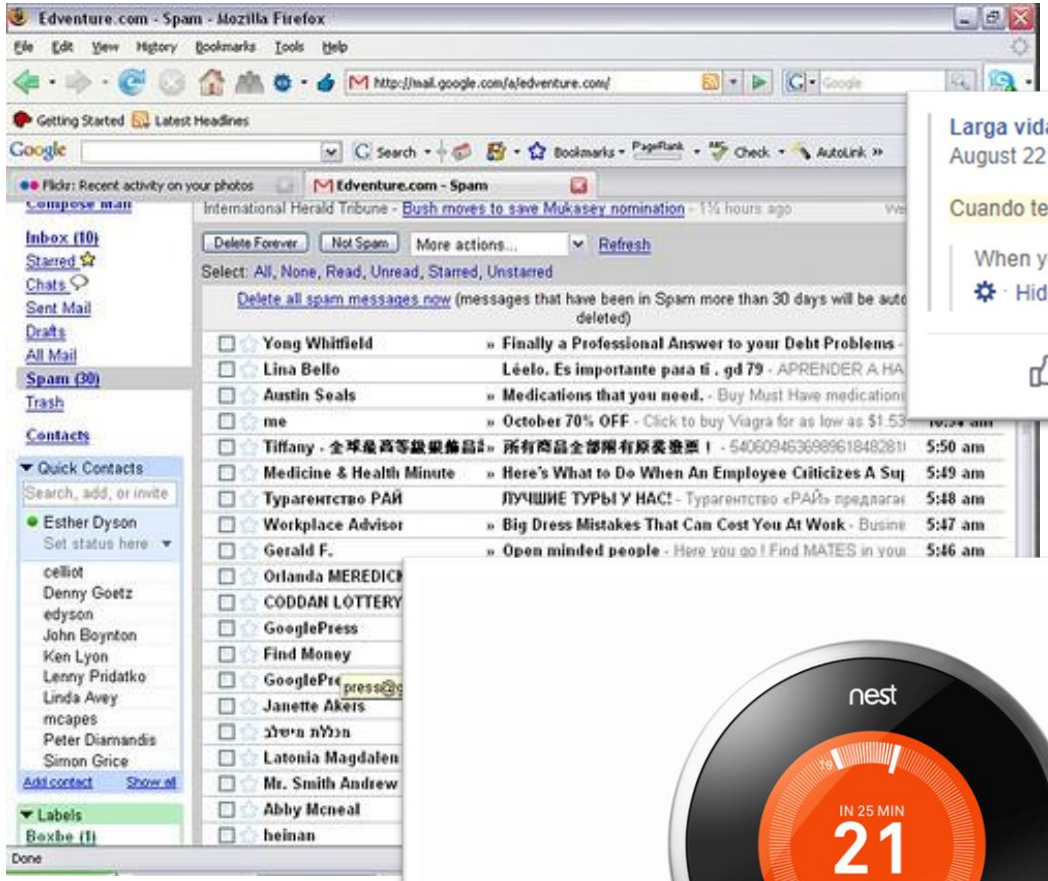
**Support for going live**



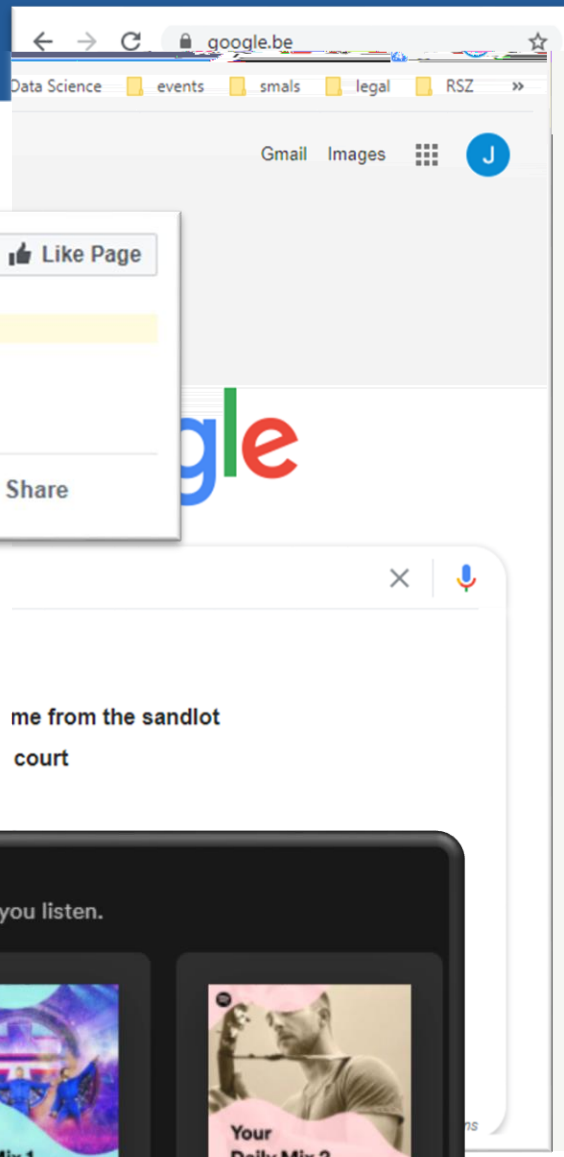

Smals ICT for society

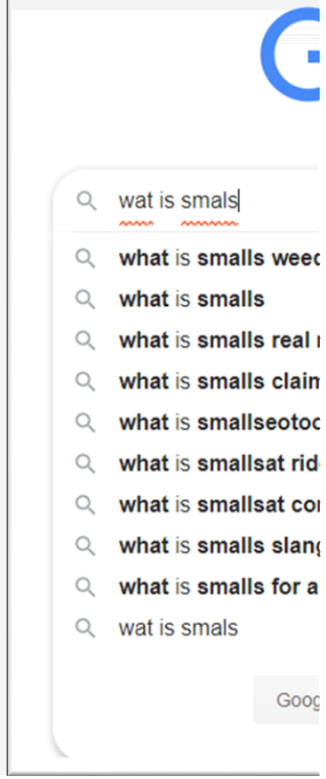
www.smalsresearch.be

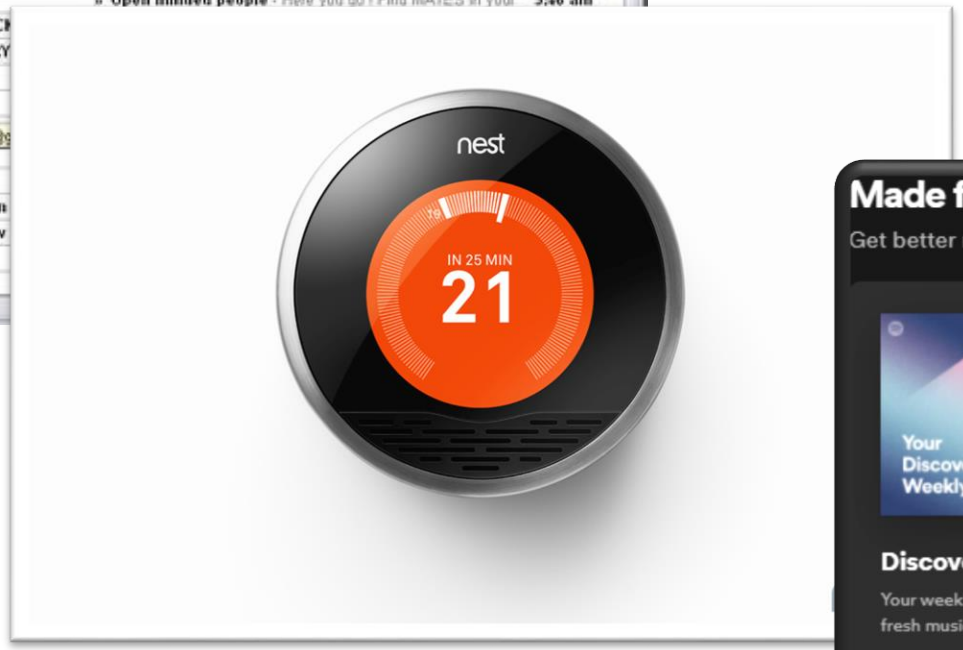
# AI: part of our daily life

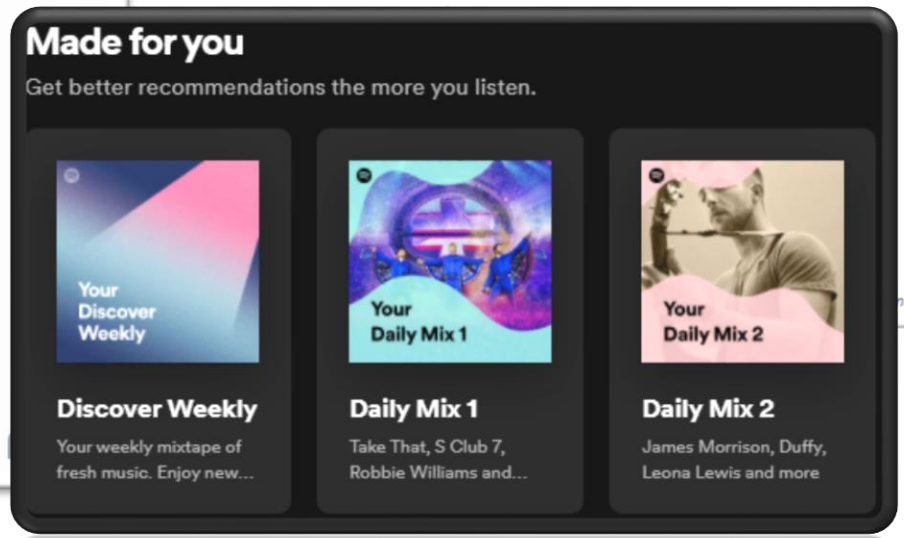









Smals
ICT for society

# 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

Smals ICT for society

- AI systems are trained on data

  → Garbage in, garbage out

- Training data is ideally

  *independent and identically distributed (iid) over the domain*

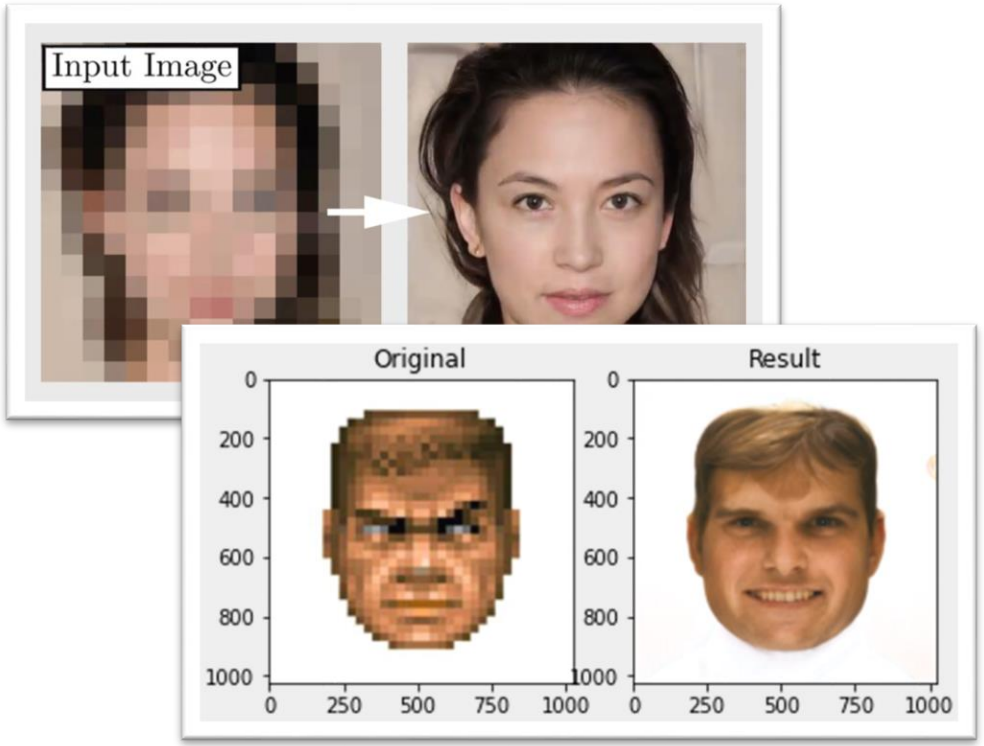
  = well-balanced & free from hidden correlations

- In reality, this is rarely the case

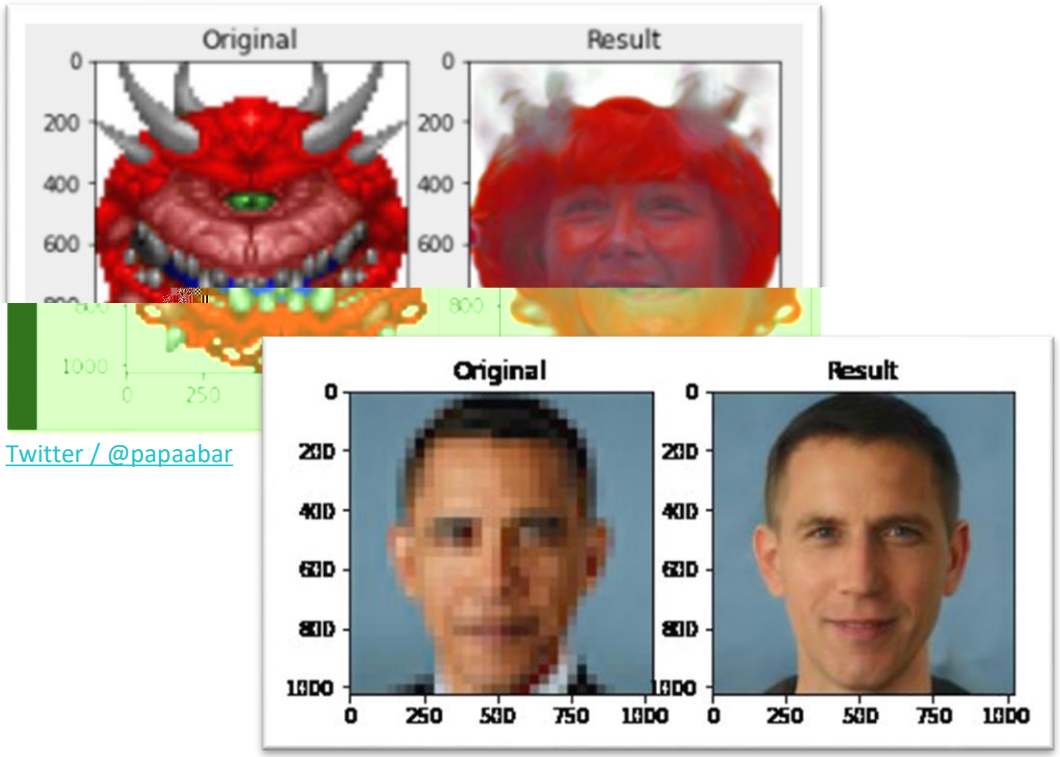
  *How many men named Anna do you know?*

# Limits on data → limits on results

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




Source: Twitter / @h_bash




Source: Twitter / @papaabar

Source: Twitter / @Chicken3gg

- 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?

Source: PetaPixel.com / Michael Zhang, “This AI Turns Pixel Faces Into ‘Photos’”

Smals ICT for society

# AI inherits biases from its creators

- 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


Headline sources: CNN.com, The Atlantic, Vox.com, Business News, Reuters.com / Jeffrey Dastin

# Hidden correlations

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

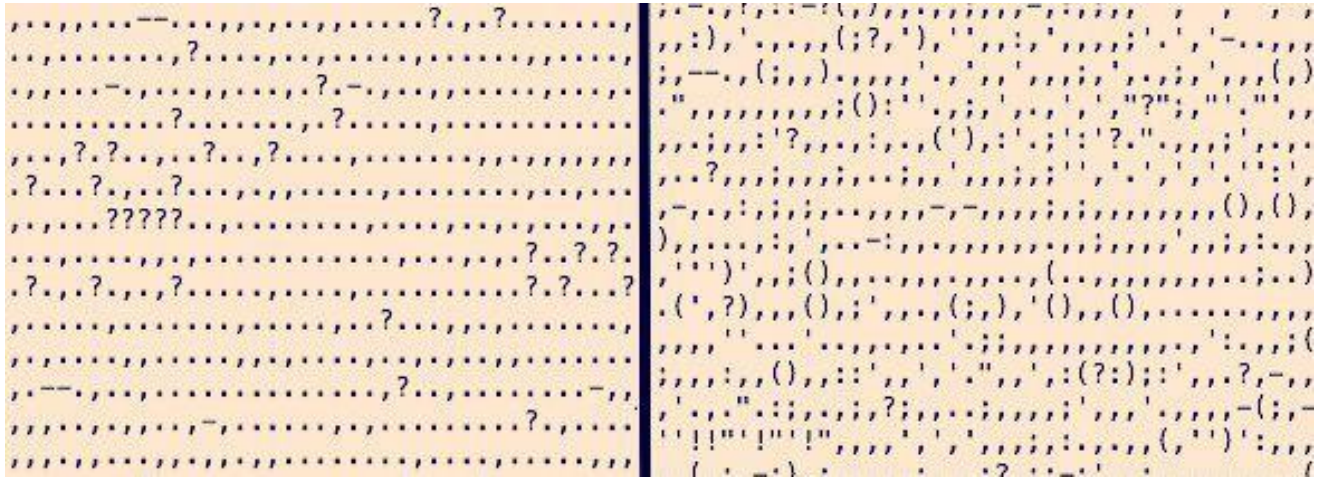


- In CVs, men/women mention different things (hobbies…)
  → Gender as prominent confounding factor in Amazon's HR experiment

# Confounding factors

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

Blood Meridian
(Cormac McCarthy)



Absalom, Absalom
(William Faulkner)

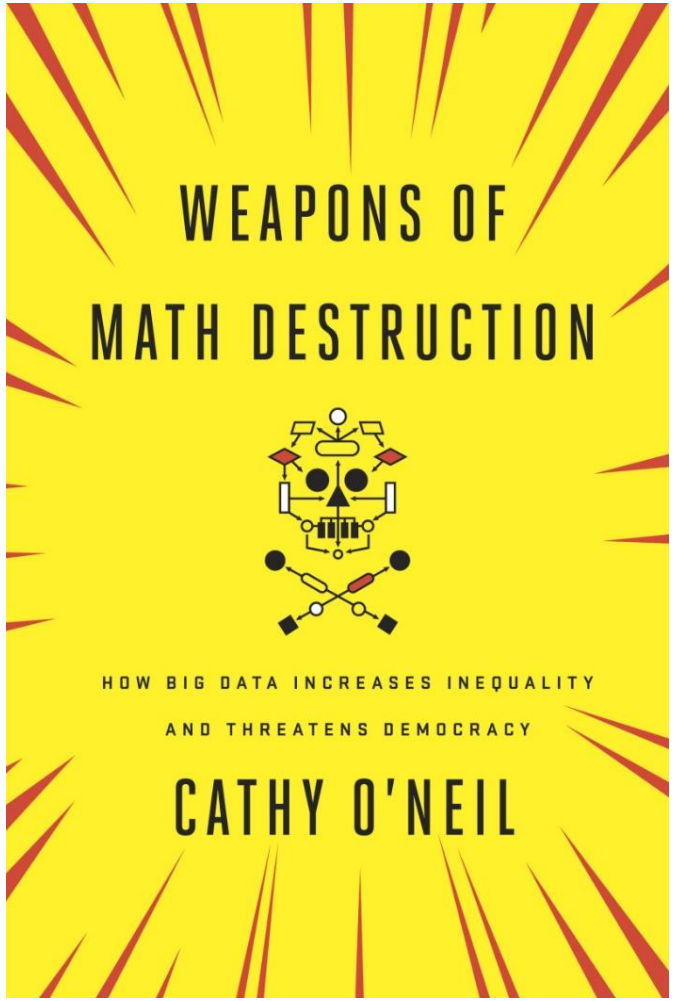
→ AI does not necessarily learn what you want it to learn!

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

Image source: Adam J. Calhoun, "punctuation in novels"

# 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)
- 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

# Definition of objectives

- 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)


Source: Twitter / @Smingleigh

# What could possibly go wrong?

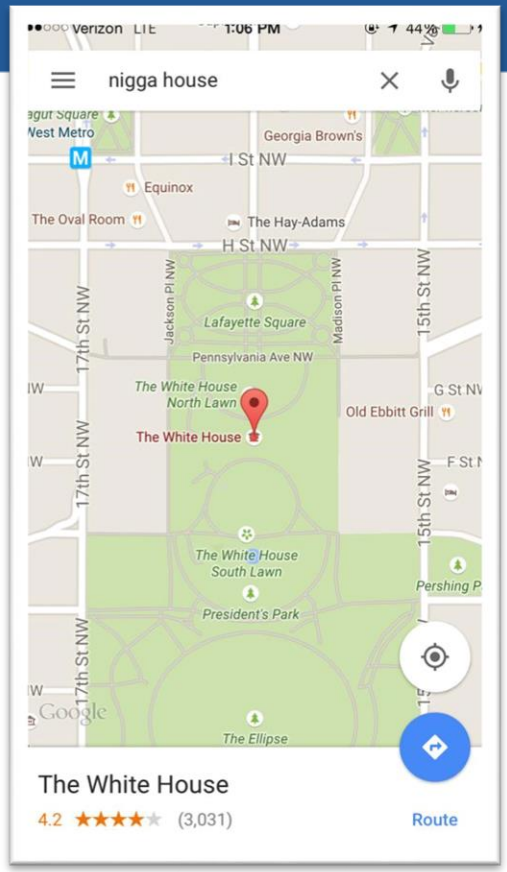
- For AI developers
  - Data collection issues (bias vs. fairness)
  - Data processing issues (confounding variables)
  - Goal (mis)formulation
- **For AI deployers: attacks against AI systems**
  - Data poisoning
  - Adversarial examples
- General public
  - Spear phishing
  - (personalized) disinformation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

Smals ICT for society

# Data poisoning

- 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

→ Data verification is not a luxury!

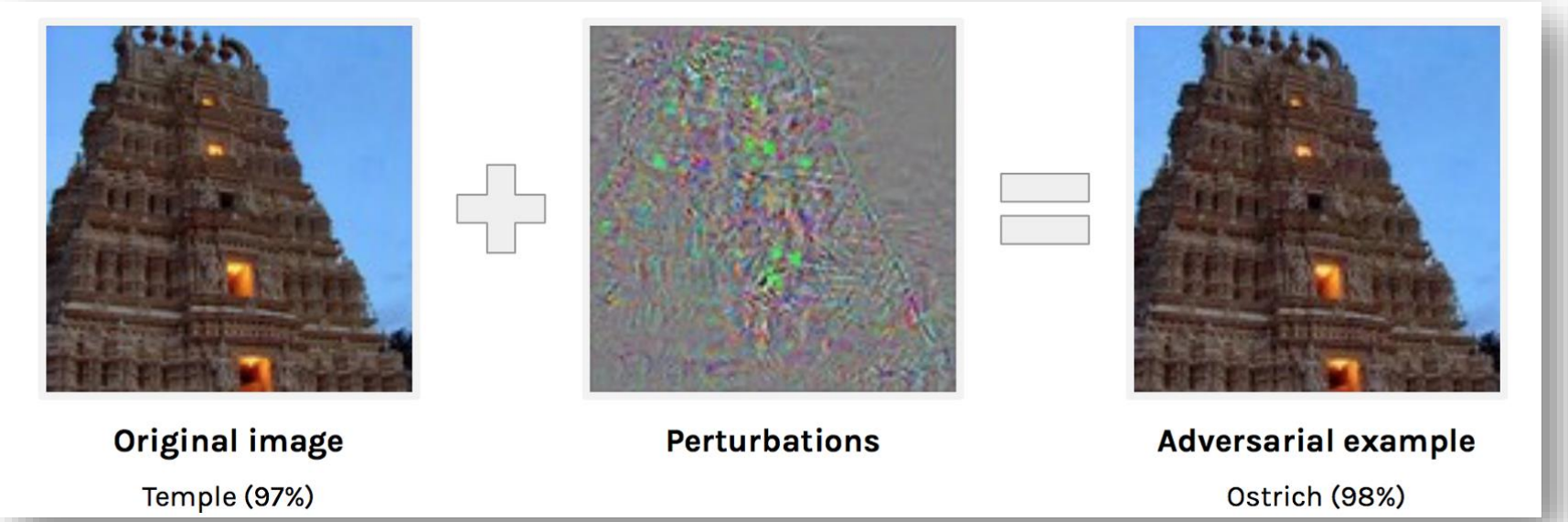



Source: Twitter / @iambomanix

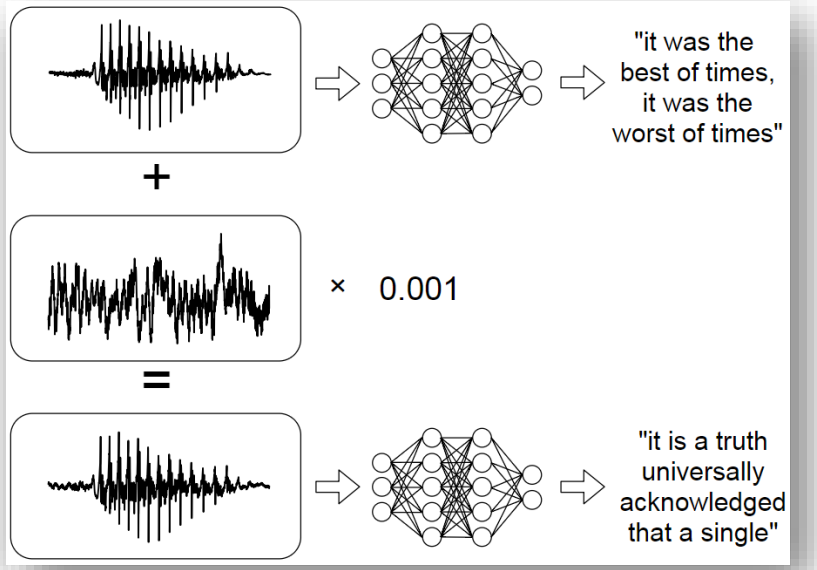

Source: trillmag.com

# Adversarial examples

- Minimal change to input → large change in output




Source: 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"




Source: Nicolas Carlini & David Wagner, "Audio Adversarial Examples: targeted attacks on speech-to-text"

# Adversarial examples

- 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"

# What could possibly go wrong?

- 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

Smals ICT for society

# Spear / laser phishing

- 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”

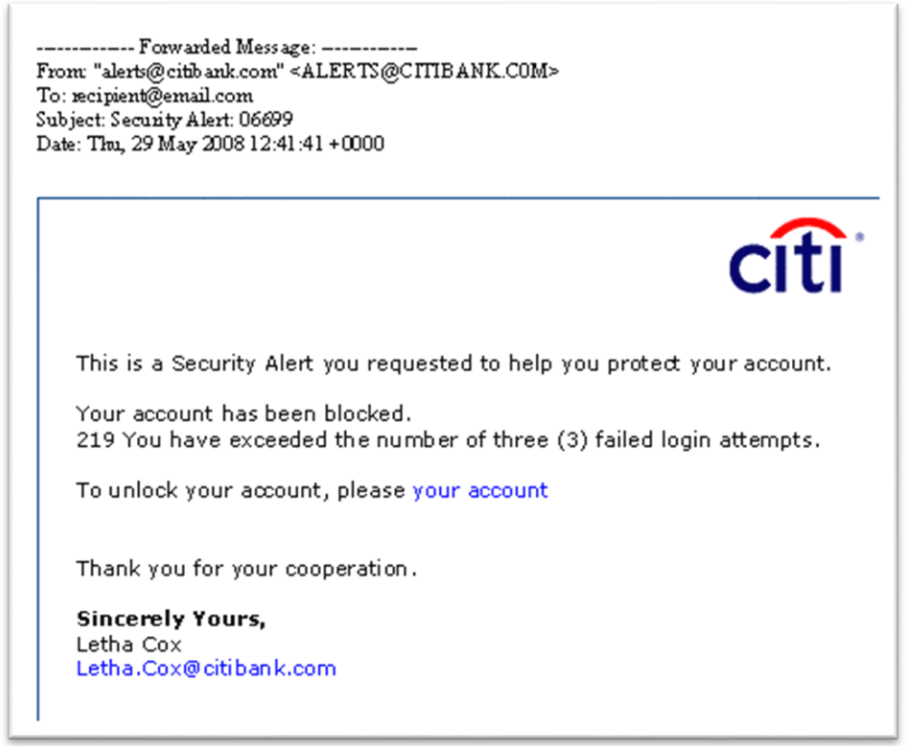
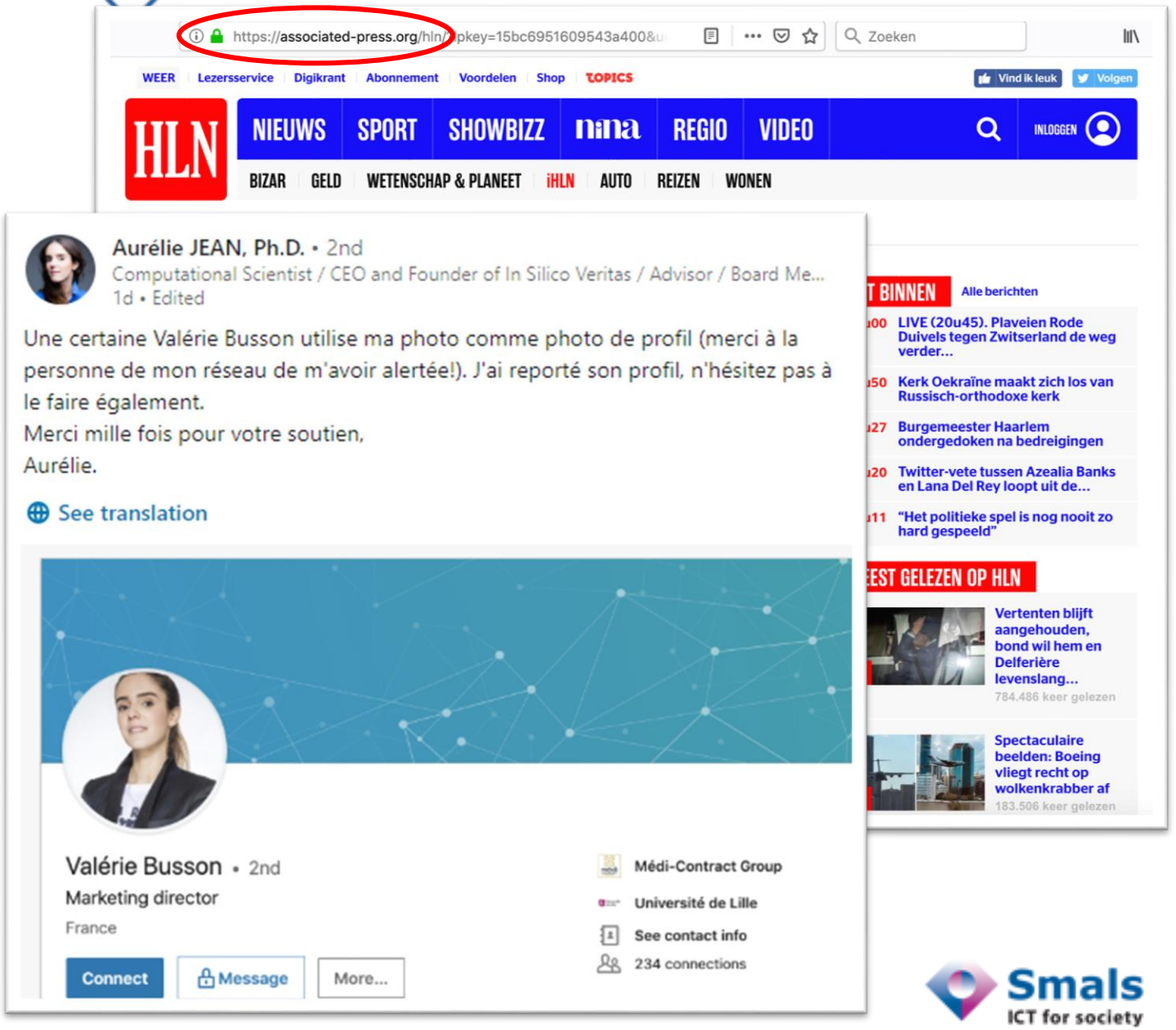




Image source: enisa.europe.eu

# 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
  - …




Source: LinkedIn / Aurélie Jean

Smals
ICT for society

# 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”


- 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


Source: Twitter / @goodfellow_ian

cf. *deepfakes:*

Source: Twitter / @ousathesquid

- From *The Verge*:

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

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

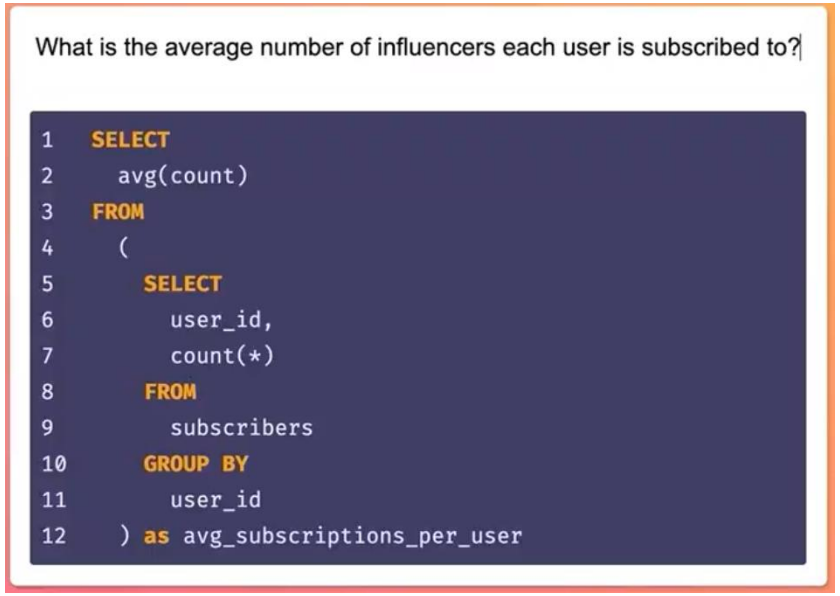
By James Vincent | Feb 14, 2019, 12:00pm EST

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."

- "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*)

# Generating Fake Text

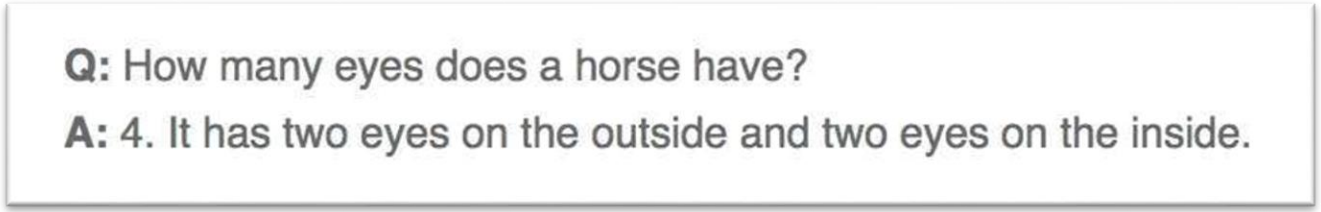
- State-of-the-art GPT-3 (05/06/2020) generates more than prose
  - Code
  - Layouts
  - Translations
  - Basic reasoning
  - …




Source: Twitter / @FaraazNishtar


Source: Twitter / @sharifshameem

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




Source: Twitter / @eturner303

Smals ICT for society

# Amplification through recommendation

- 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


Source: Twitter / @chrislhayes

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

# Amplification through recommendation

- 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.

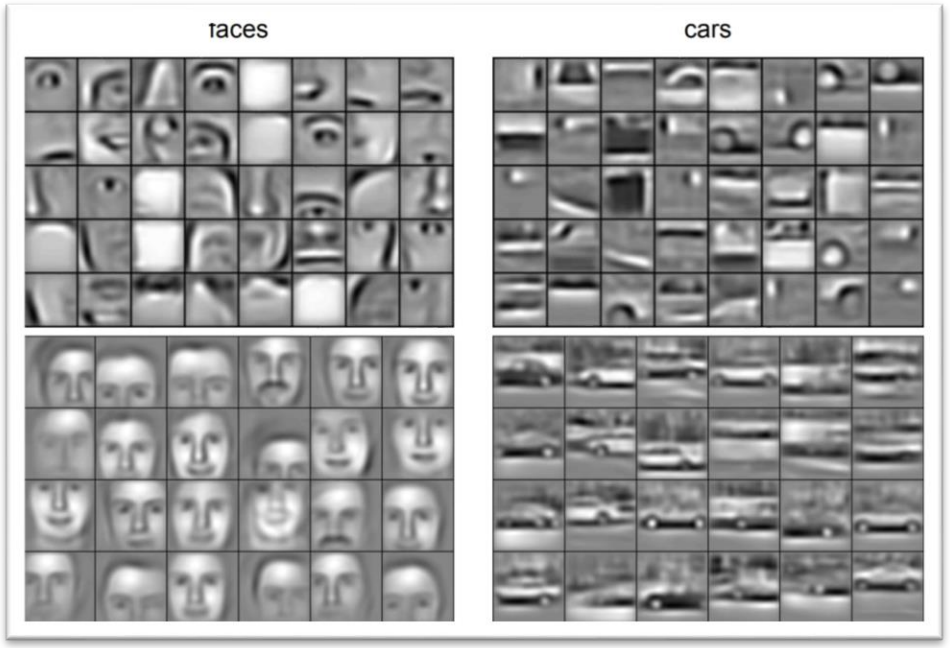
# What could possibly go wrong?

- 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

Smals ICT for society

# As a technical person

- 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"

# As a citizen

- 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


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)

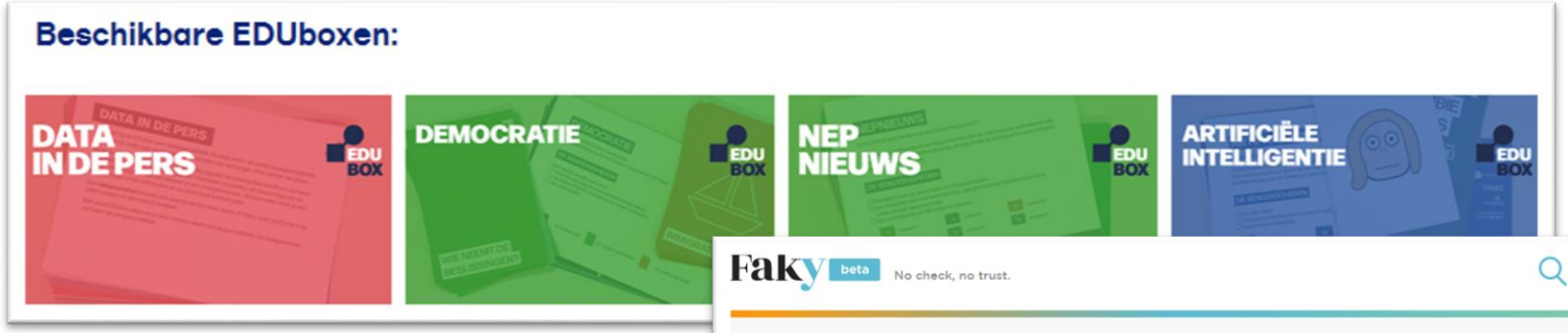
→ Requires some Competences / Literacy

# As a policymaker

- 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/
- …





- https://faky.be/fr
- https://openfacto.fr/
- https://www.reseauia.be/
- https://www.ai4belgium.be/

- 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.

# On the EU level

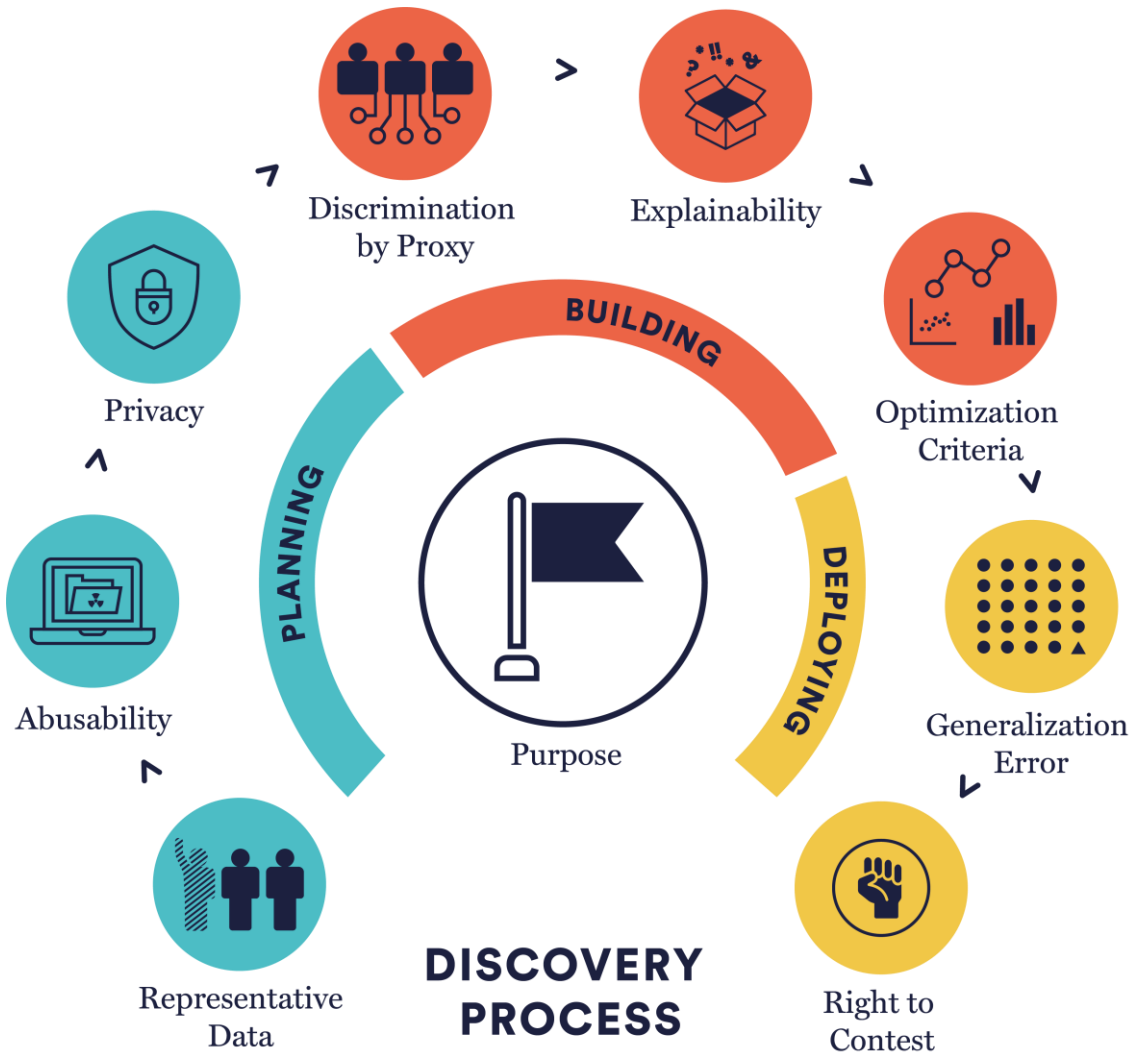
- 03/2015: Stratcom Task Force → euvsdisinfo.eu
- 10/2018: EU code of practice on disinformation
  - Signed by Google, Facebook, Twitter, Mozilla etc.
  - (Initial) choice for industry self-regulation
- 12/2018: EU action plan on disinformation
- 04/2019: EU HLEG Ethics Guidelines for Trustworthy AI
  - 07/2020: addition of Assessment list for Trustworthy AI
  - Belgian coordination: AI4Belgium

# Further reading

- 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)

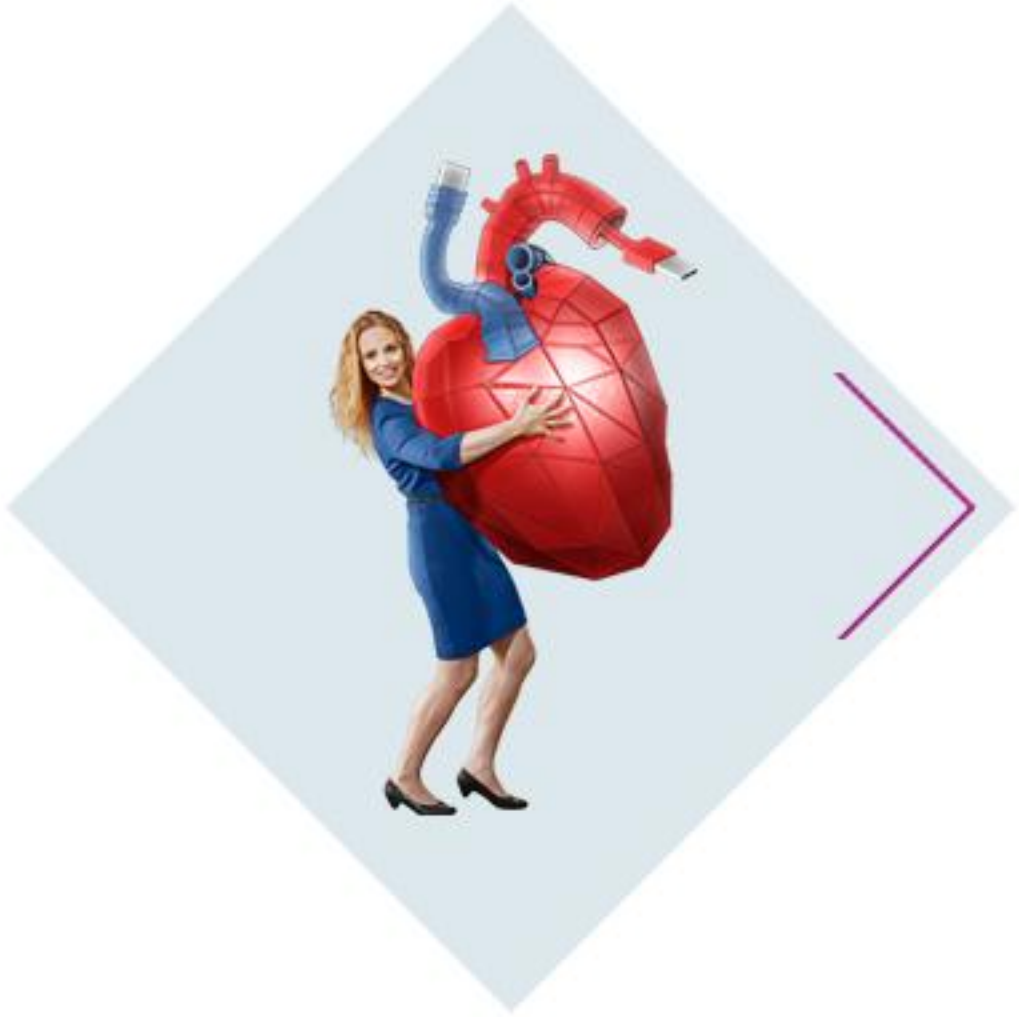
- Organizations and Academia
  - https://montrealethics.ai/
  - https://cyber.harvard.edu/ / https://ai.shorensteincenter.org/
  - https://www.turing.ac.uk/research/data-ethics
  - https://hai.stanford.edu/
  - …

# Epilogue




AI Blind Spot Discovery Process,
MIT Media Lab, licensed CC-BY 4.0

Smals
ICT for society



# 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
Smals
ICT for society