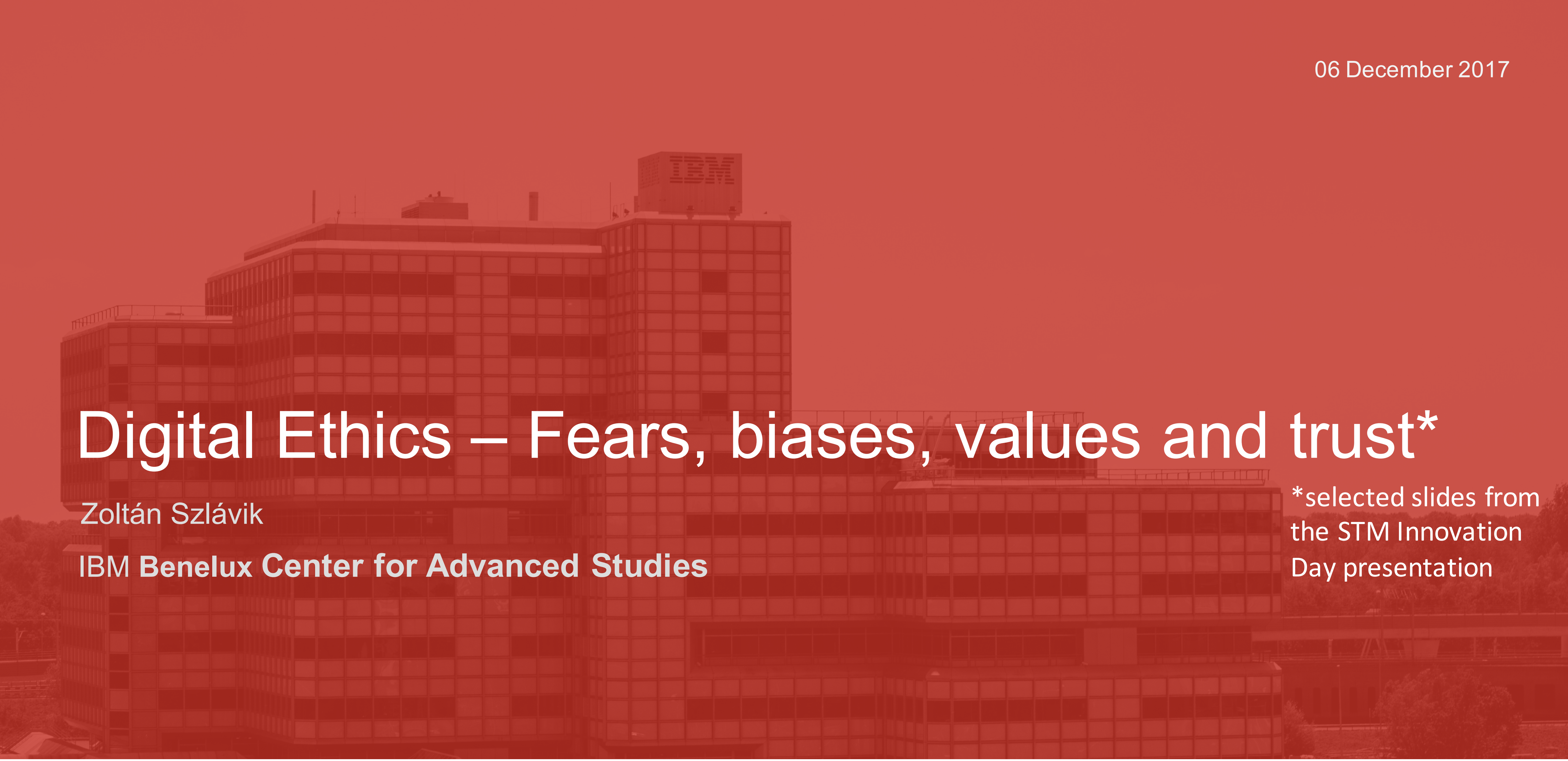


06 December 2017

Digital Ethics – Fears, biases, values and trust*

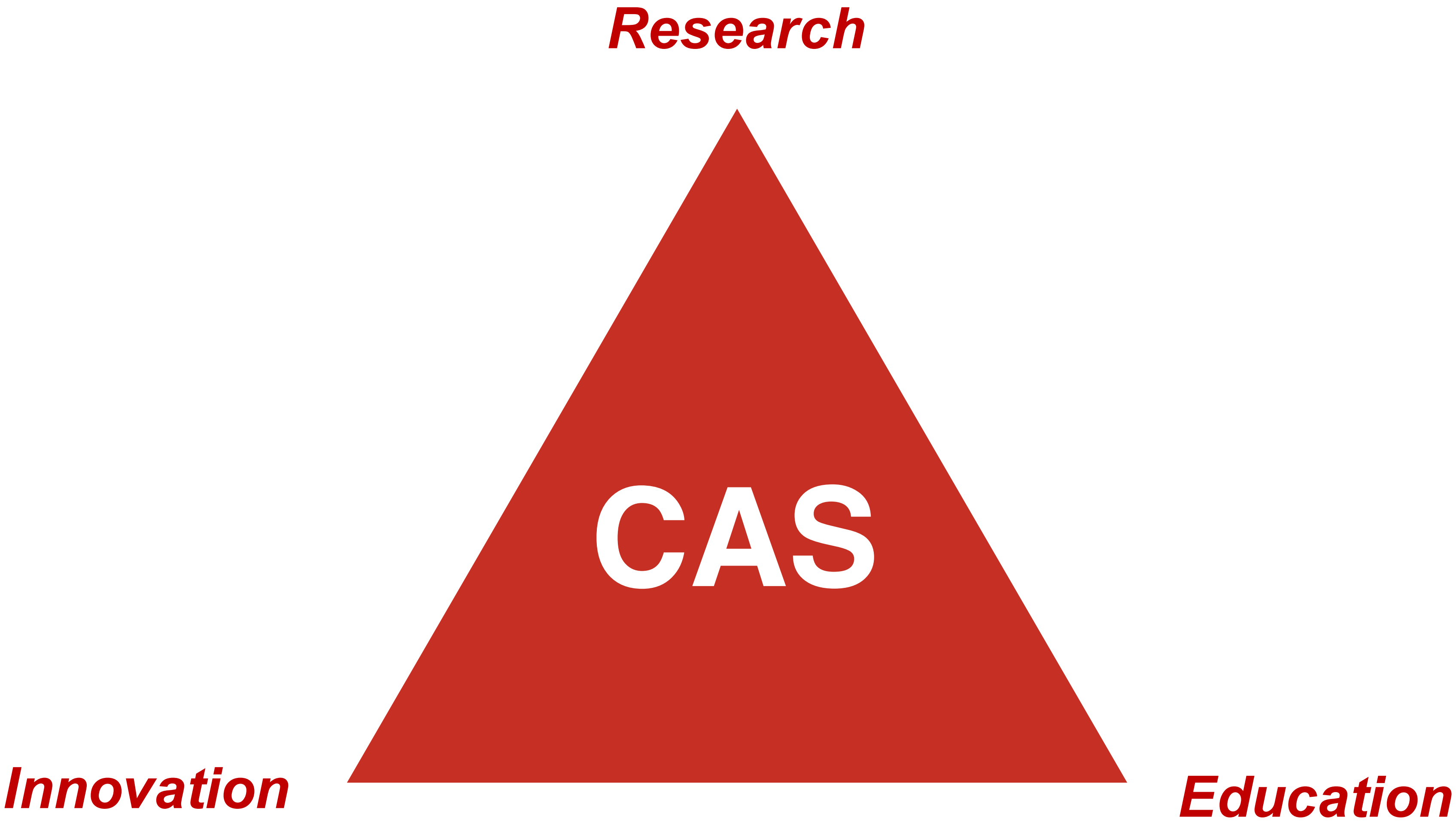
Zoltán Szlávik

IBM **Benelux Center for Advanced Studies**

*selected slides from the STM Innovation Day presentation

Warning: contains personal observations, too.

IBM Benelux Center for Advanced Studies (CAS)



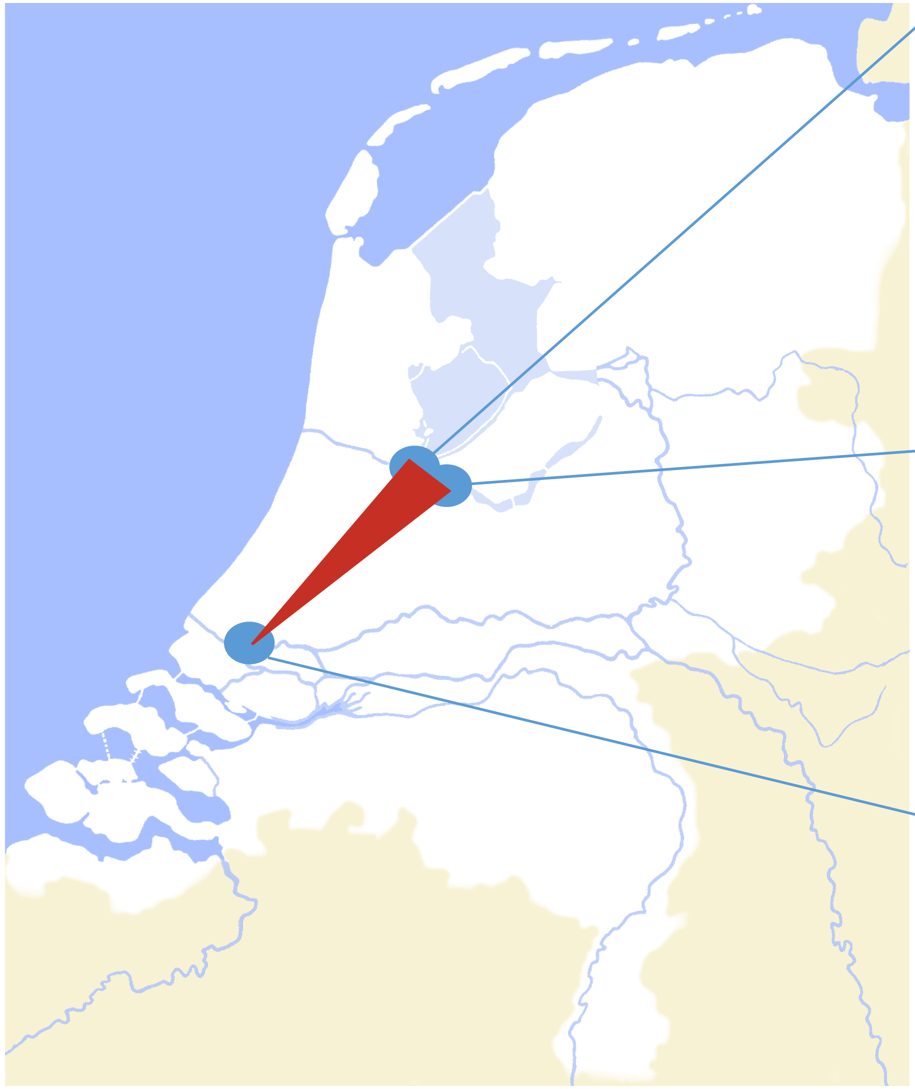
Research

CAS

Innovation

Education

CAS BNL



CAS BeNeLux

Zoltán Szlávik
Lead/Research

Benjamin Timmermans
Research/VU

Dorottya Mezőfi
Programme Management

Manfred Overmeen
Development

Santiago Gaitan
Research/TUD

CIC VU Amsterdam

Lora Aroyo
Faculty Fellow

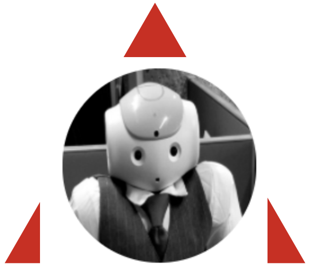
Anca Dumitrache
Phd Fellow

Oana Inel
Phd Fellow

CIC TU Delft

Alessandro Bozzon
Faculty Fellow

Peter Hofstee
IBM Austin Lab

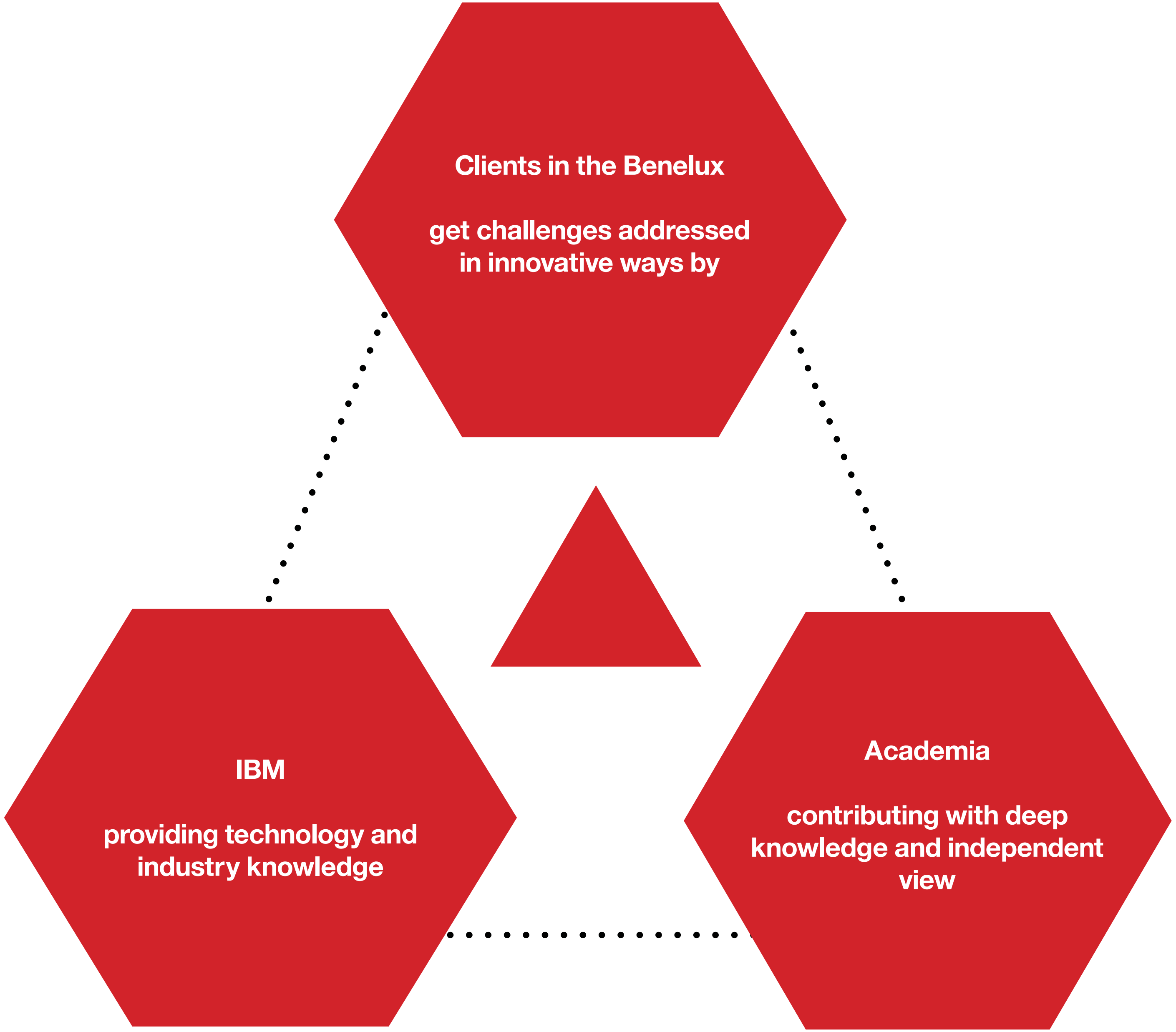


Robbie, Rosita & Cas
Computer Overlords
Also, events

IBM
Center for Advanced Studies

@zolley

CAS: Collaboration with Academia



Clients in the Benelux

get challenges addressed in innovative ways by

IBM

providing technology and industry knowledge

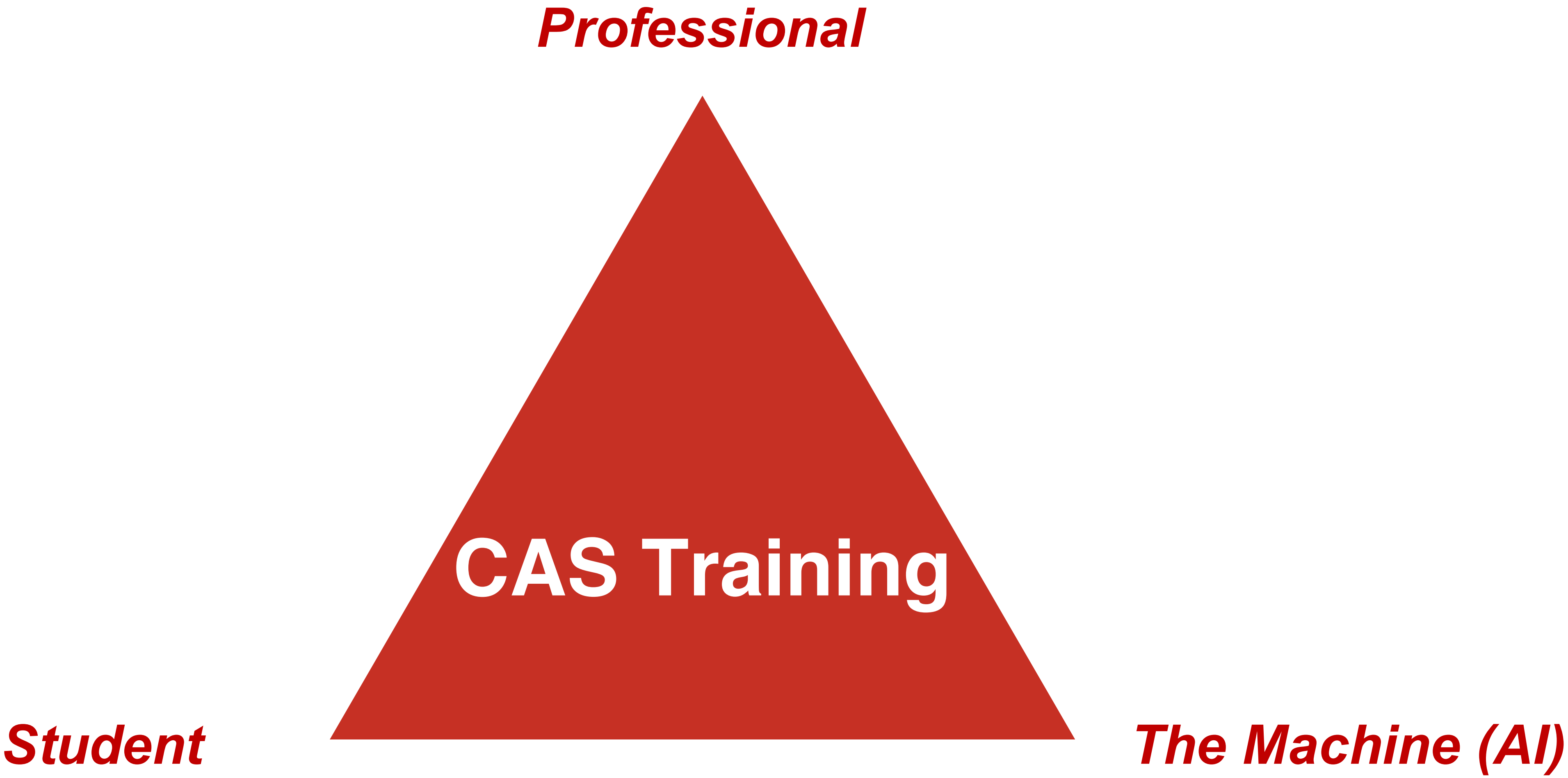
Academia

contributing with deep knowledge and independent view

IBM
Center for
Advanced Studies

@zolley

IBM Benelux Center for Advanced Studies (CAS)



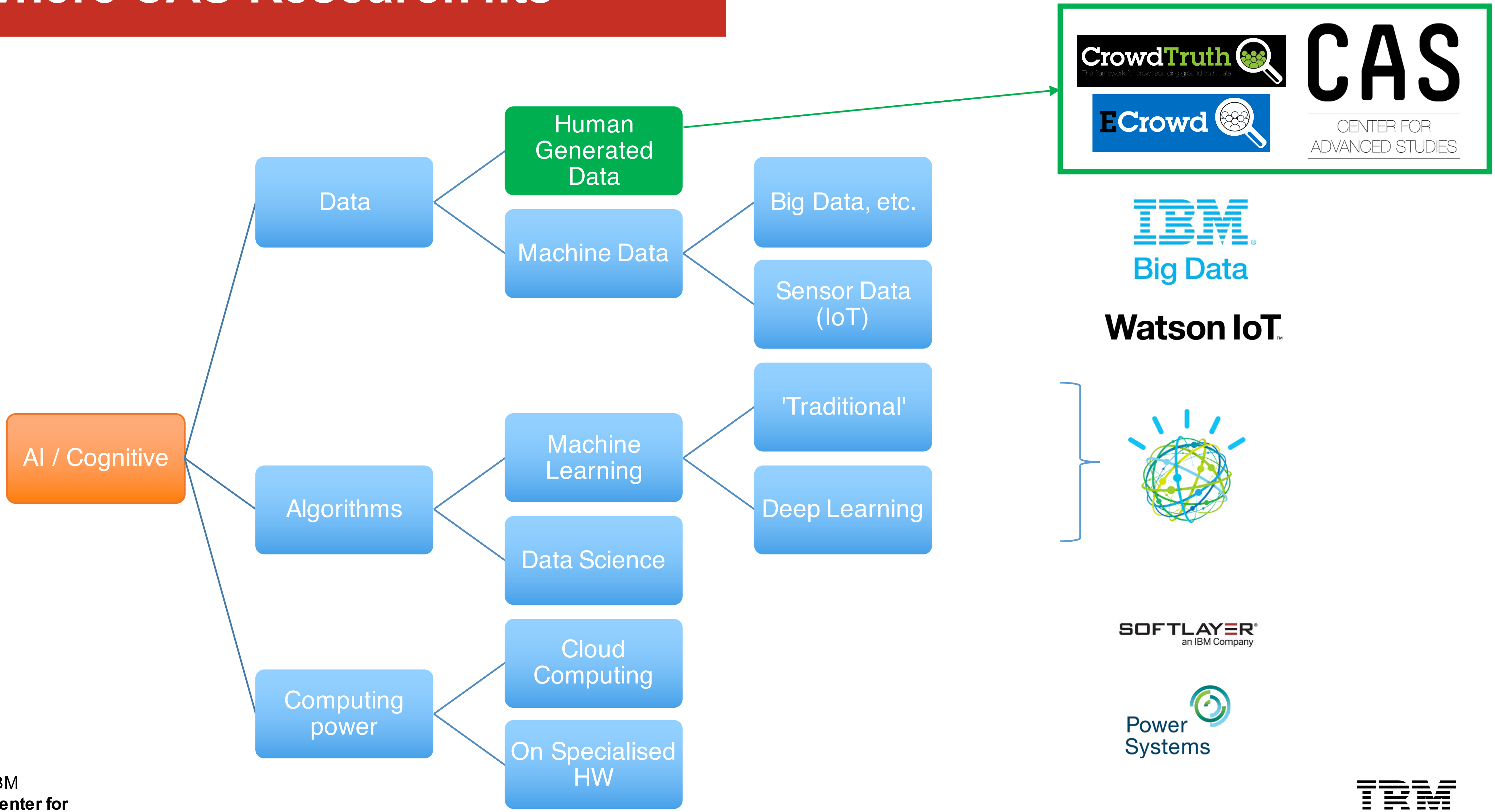
Professional

CAS Training

Student

The Machine (AI)

Where CAS Research fits



AI / Cognitive

Data

Human Generated Data

Machine Data

Big Data, etc.

Sensor Data (IoT)

Algorithms

Machine Learning

'Traditional'

Deep Learning

Data Science

Computing power

Cloud Computing

On Specialised HW

CrowdTruth

The framework for crowdsourcing ground truth data

ECrowd

CAS

CENTER FOR ADVANCED STUDIES

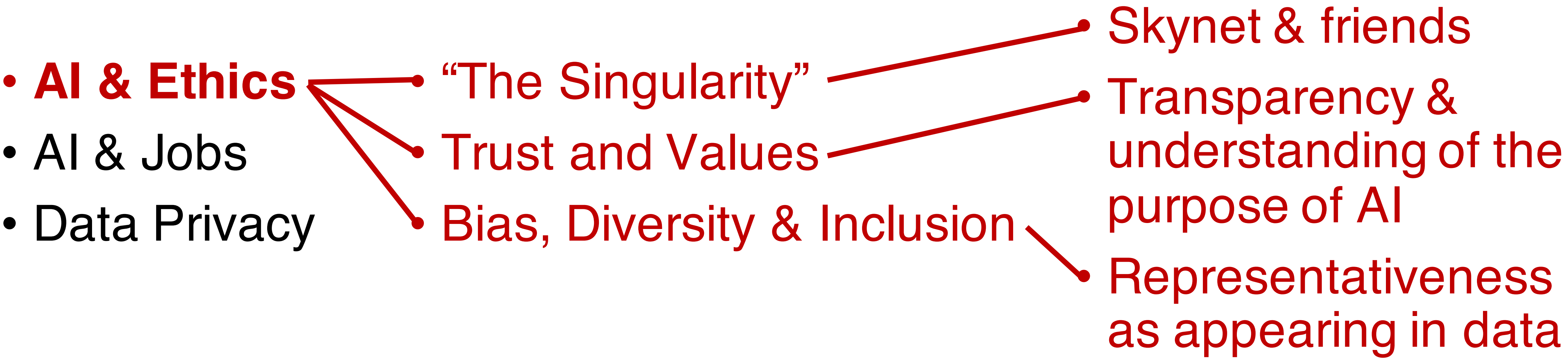
IBM Big Data

Watson IoT

SOFTLAYER an IBM Company

Power Systems

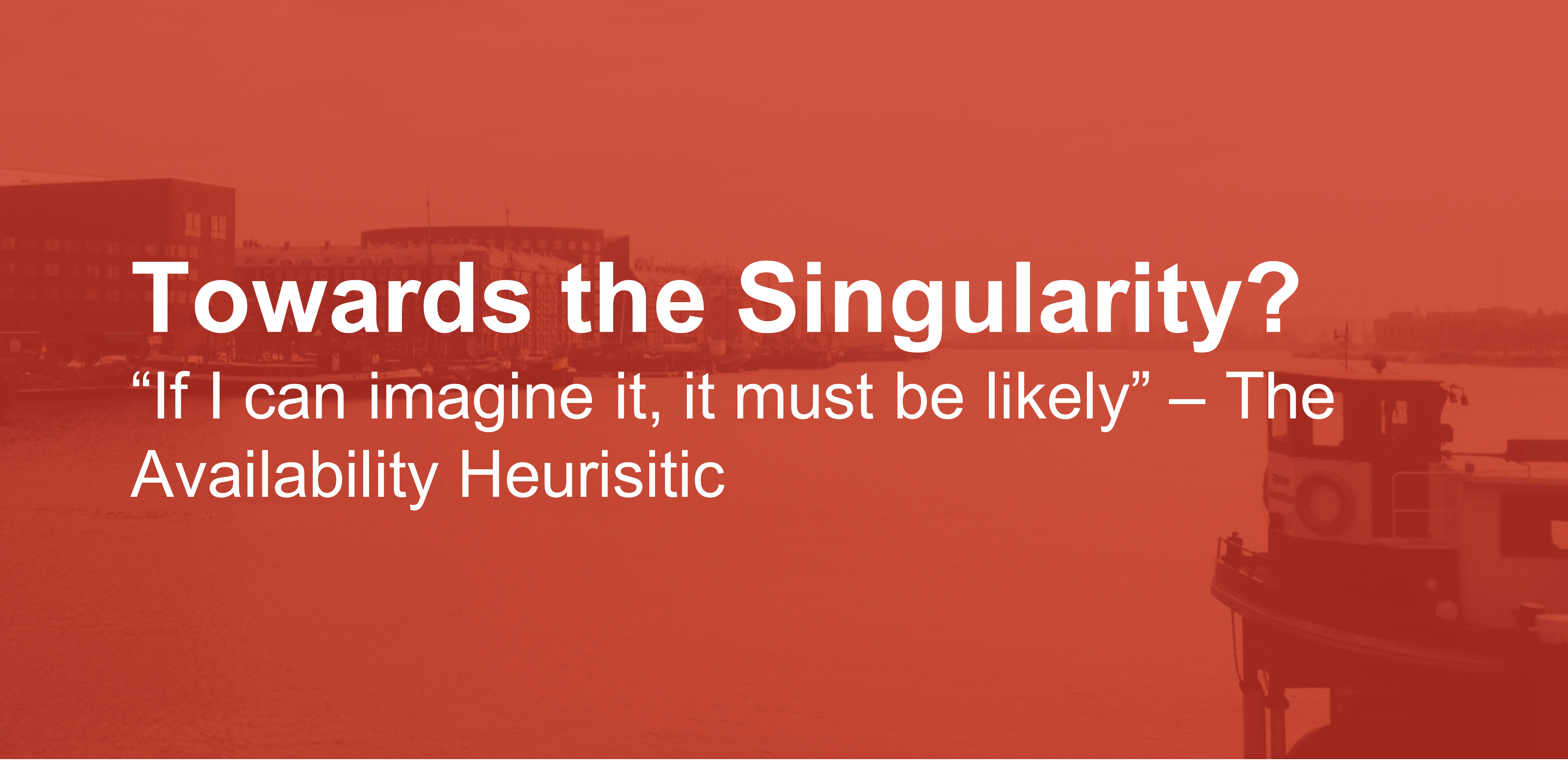
IBM addressing AI & Ethics



- **AI & Ethics**
- AI & Jobs
- Data Privacy

- "The Singularity"
- Trust and Values
- Bias, Diversity & Inclusion

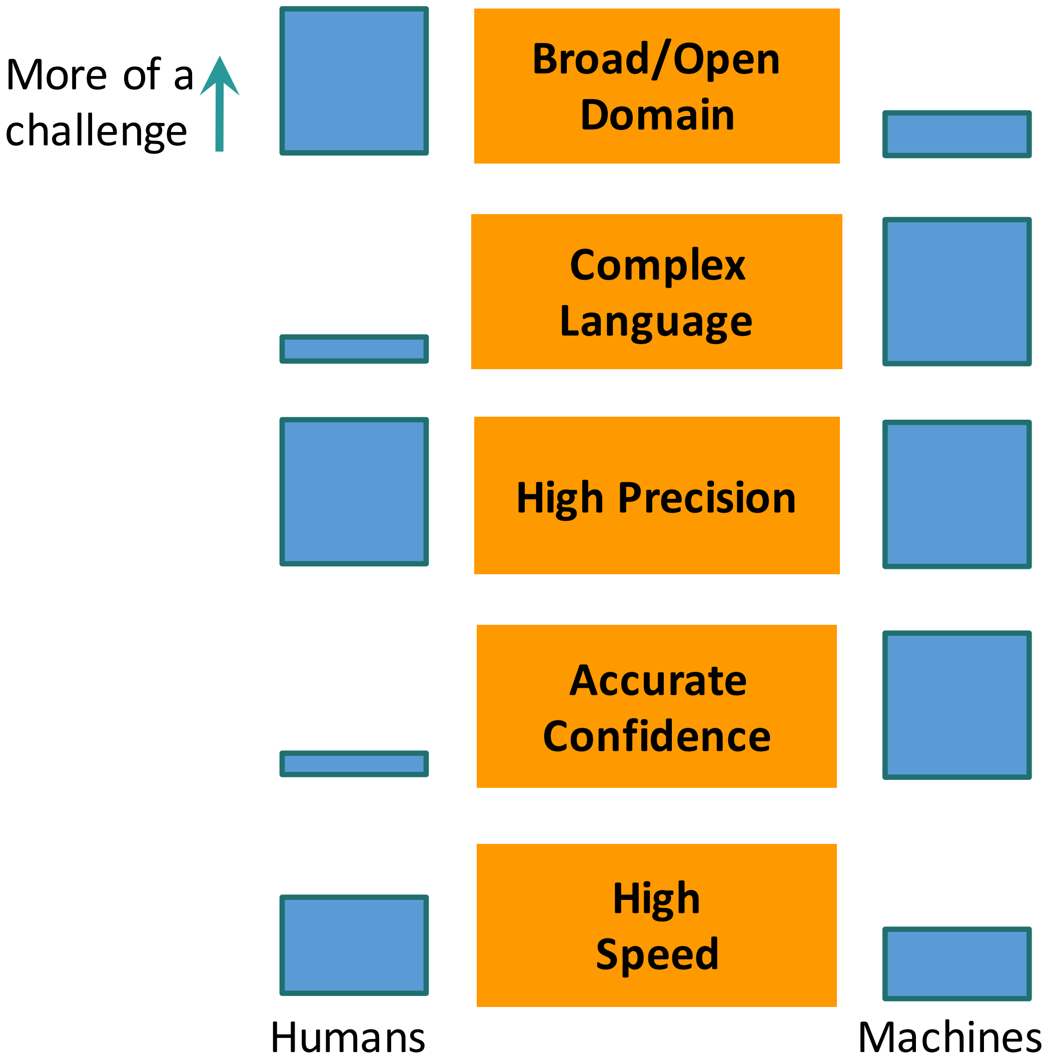
- Skynet & friends
- Transparency & understanding of the purpose of AI
- Representativeness as appearing in data



Towards the Singularity?

“If I can imagine it, it must be likely” – The The Availability Heurisitic

Watson (2011) Question-Answering on unconstrained domains



More of a challenge

Broad/Open Domain

Complex Language

High Precision

Accurate Confidence

High Speed

Humans

Machines

Cognitive Computing: Three elements

EXPANDS

EXPANDS human cognition, makes the jobs we do easier, like a *cognitive prosthesis*, especially when dealing with processing massive data, or data that requires human interpretation

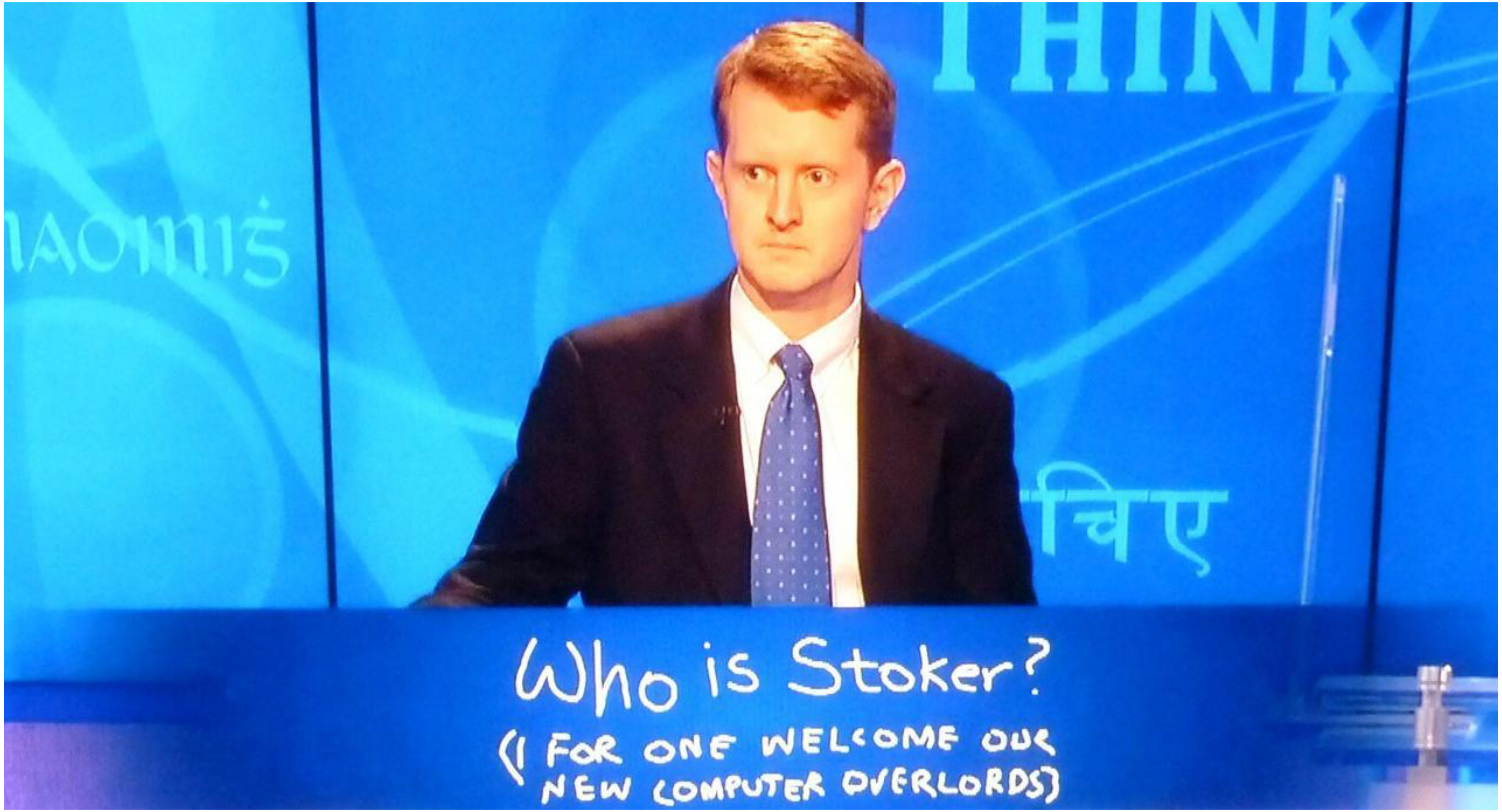
LEARNS

LEARNS as you use it – most machine errors are easy for a human to detect, and we can instrument usage of systems to better understand the system and the problem it solves

INTERACTS

INTERACTS naturally. We need to bring machines closer to their users, we have adapted ourselves enough to them, they should understand natural language, spoken or written, be able to process images and videos. These *simple* human problems are extremely complex for machines, but are hallmarks of a new computing era.

…and we won! (…what now?)

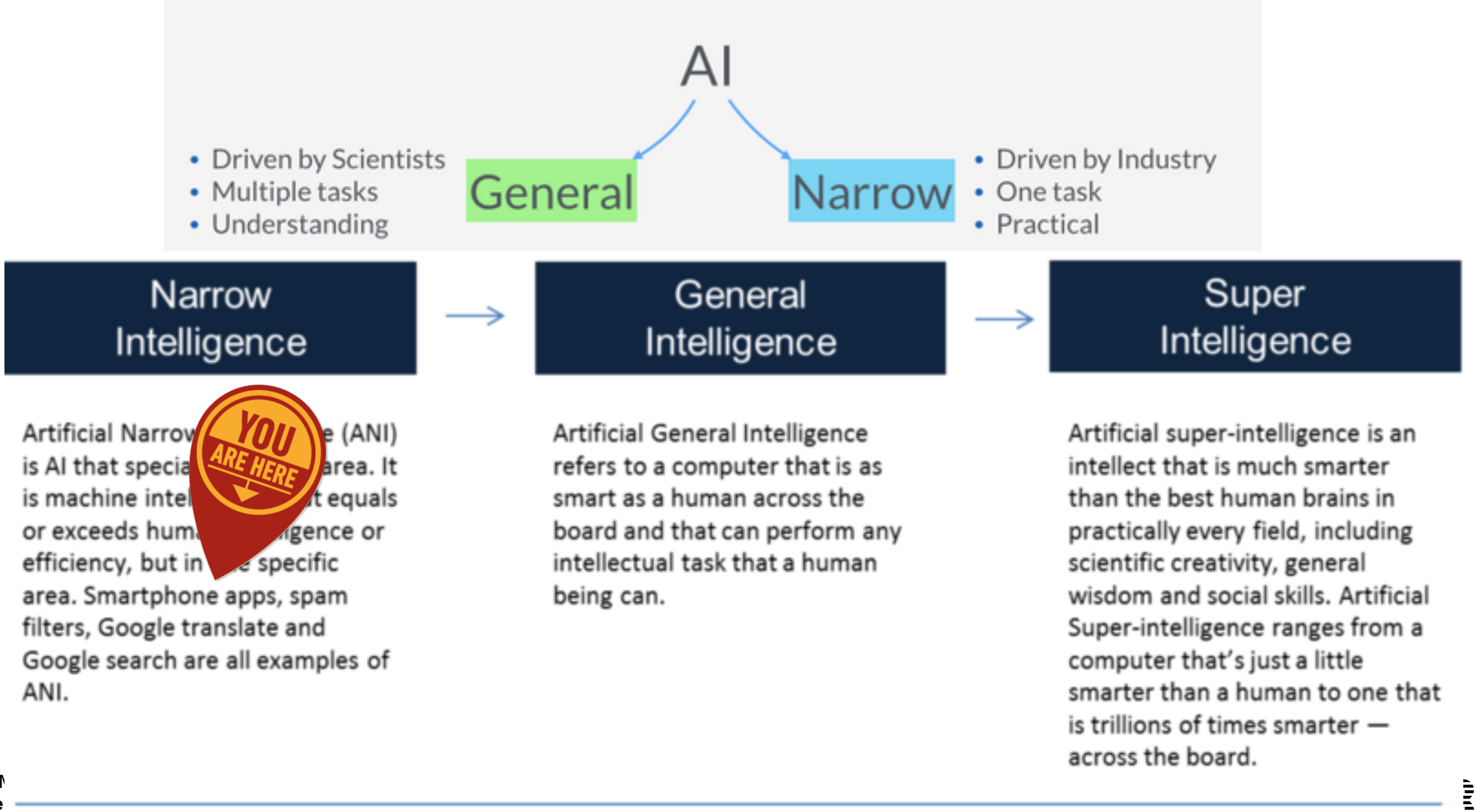


IBM on AI & Humans: Optimism

“Augmented Intelligence”

Augmented
~~Artificial~~ Intelligence
to make precise decisions

What kind of AI?



- Driven by Scientists
- Multiple tasks
- Understanding

- Driven by Industry
- One task
- Practical

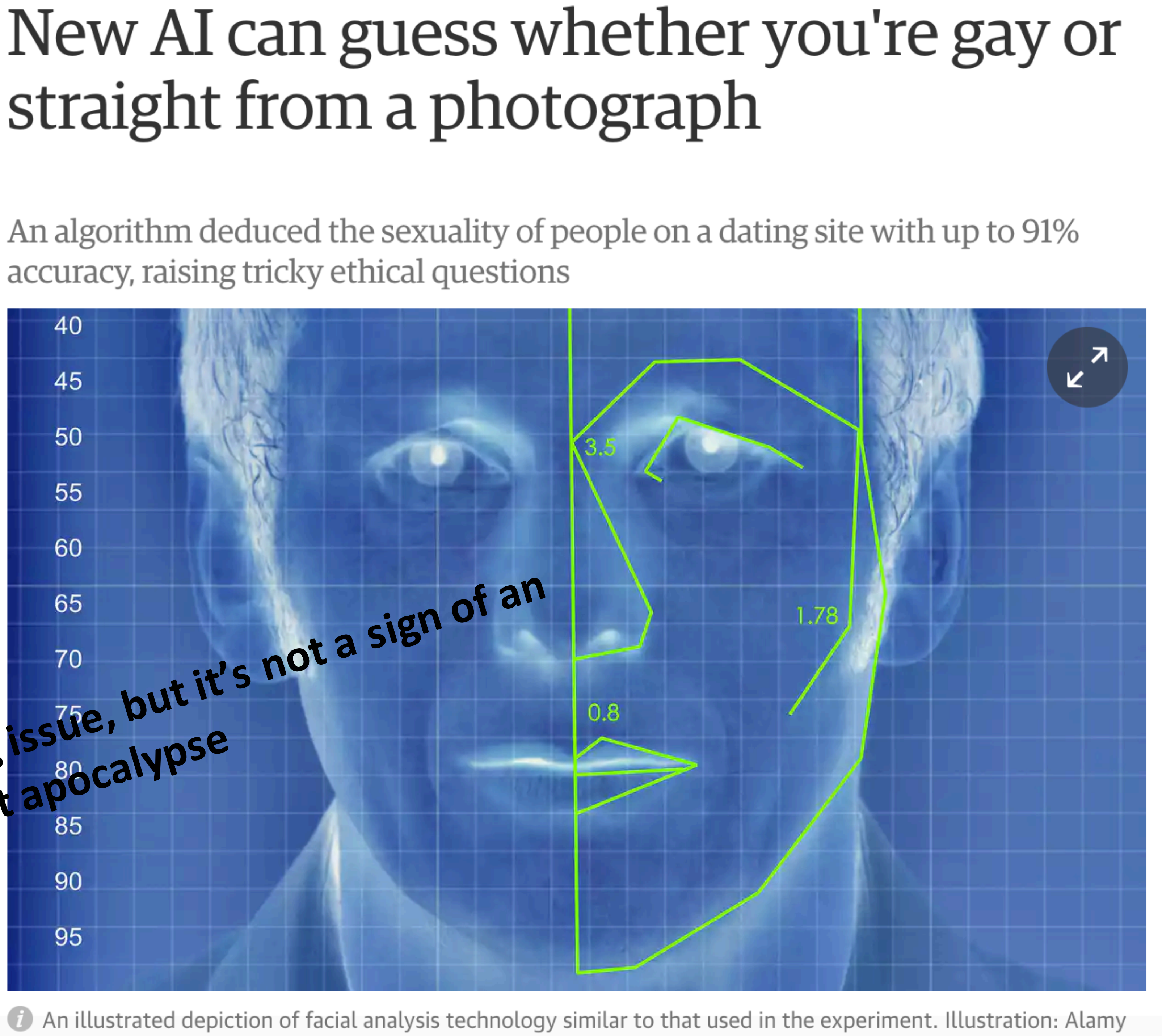
Narrow Intelligence → General Intelligence → Super Intelligence

Artificial Narrow e (ANI) is AI that specia area. It is machine intel t equals or exceeds hum igence or efficiency, but in e specific area. Smartphone apps, spam filters, Google translate and Google search are all examples of ANI.

Artificial General Intelligence refers to a computer that is as smart as a human across the board and that can perform any intellectual task that a human being can.

Artificial super-intelligence is an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. Artificial Super-intelligence ranges from a computer that's just a little smarter than a human to one that is trillions of times smarter — across the board.

Progression of Artificial Intelligence (AI)



New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions

An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

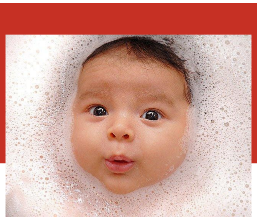
This is a very big issue, but it's not a sign of an inevitable robot apocalypse

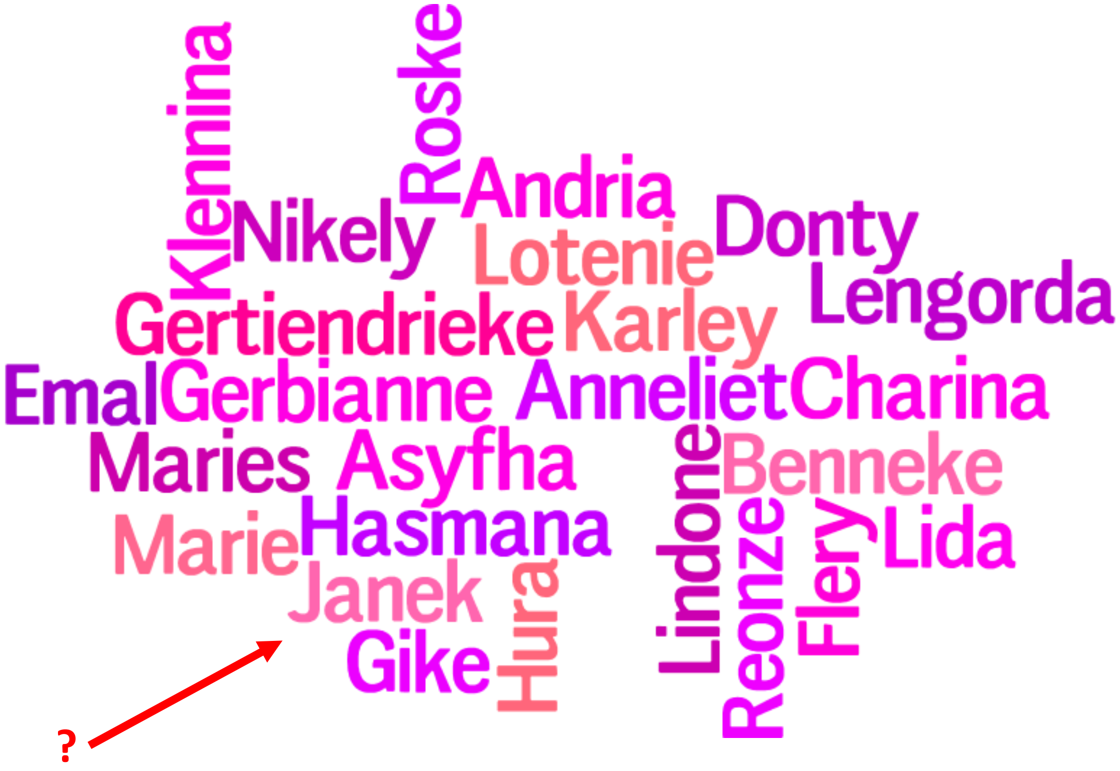
Note the reporting!

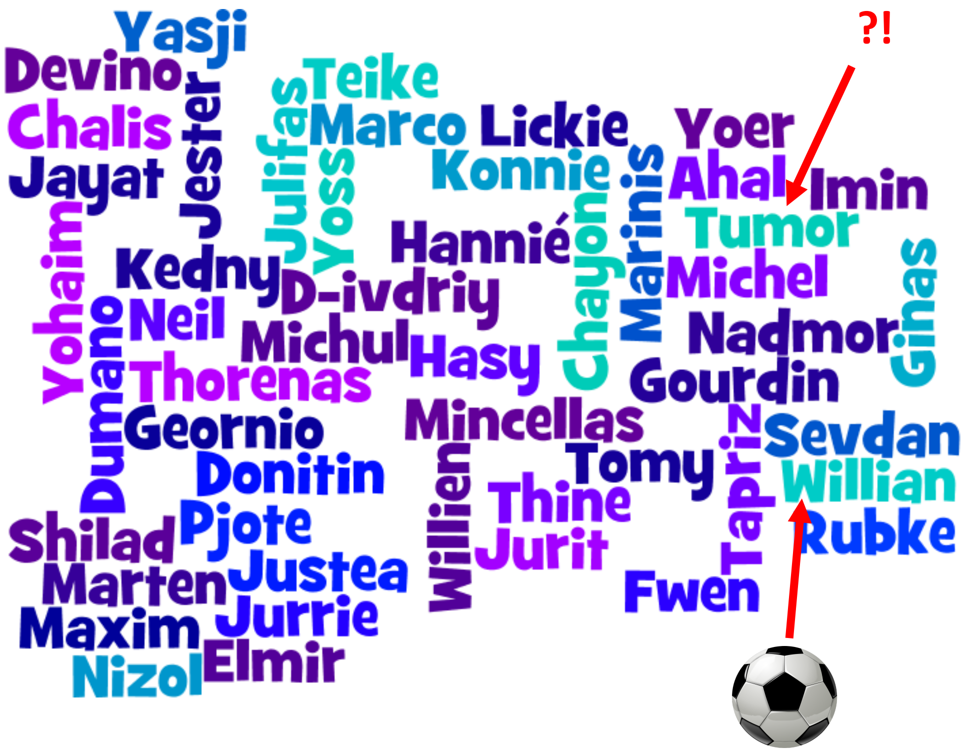
Biases

and mistakes made by AI systems

Let’s start with some innocence (?): Neural Network to generate Dutch baby names

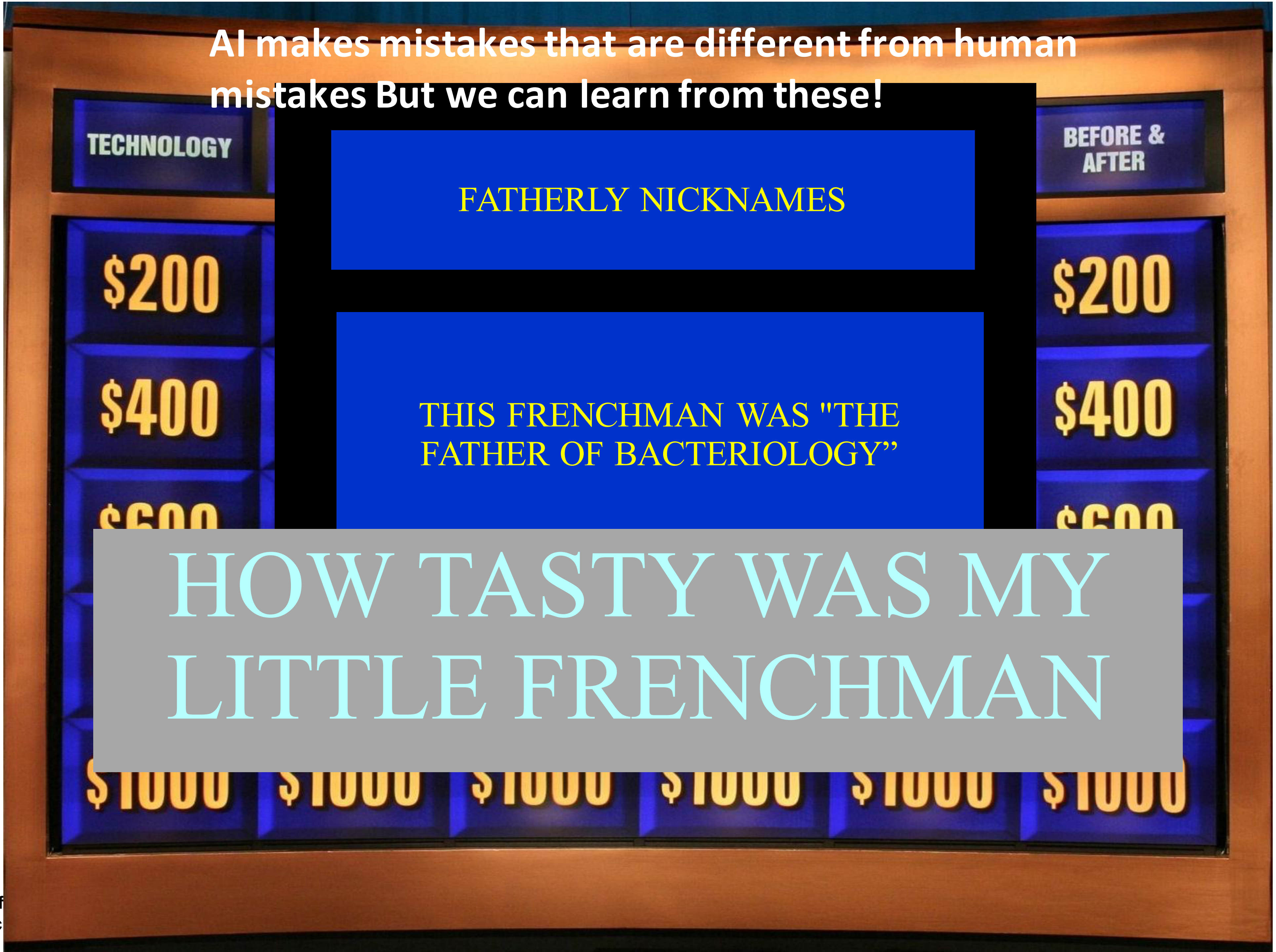






Bonus: a cultural “debate”

AI makes mistakes that are different from human mistakes But we can learn from these!

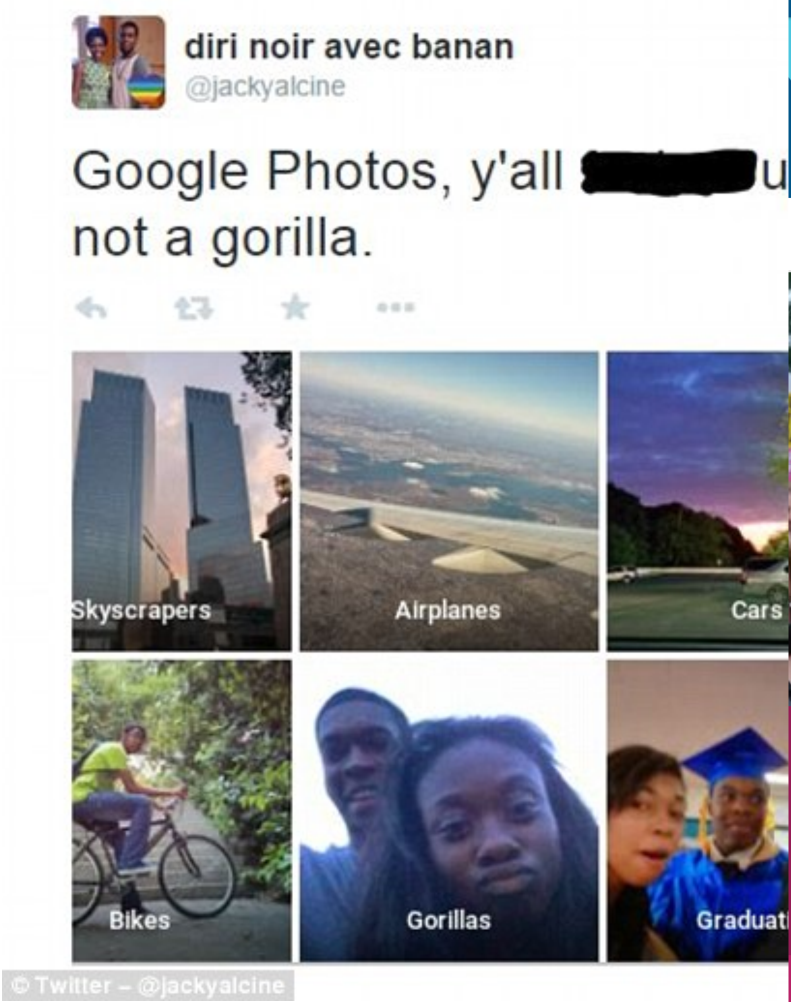


FATHERLY NICKNAMES

THIS FRENCHMAN WAS "THE FATHER OF BACTERIOLOGY”

HOW TASTY WAS MY LITTLE FRENCHMAN

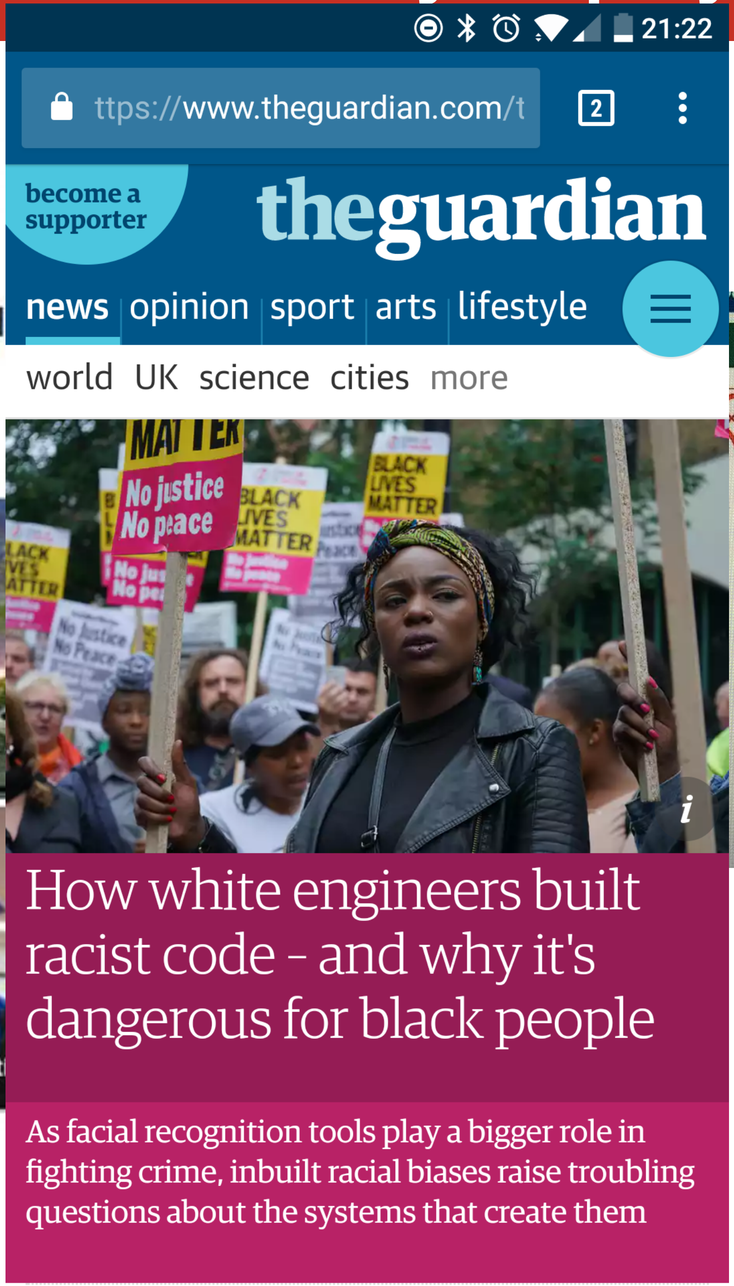
AI models reflect, and can easily amplify, biases in data



diri noir avec banan
@jackyalcine

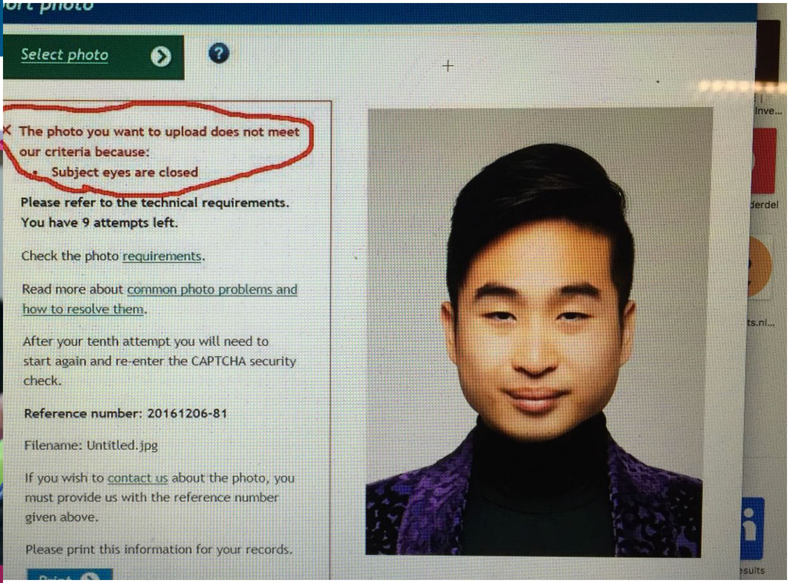
Google Photos, y'all [redacted]u
not a gorilla.

© Twitter – @jackyalcine



How white engineers built racist code – and why it's dangerous for black people

As facial recognition tools play a bigger role in fighting crime, inbuilt racial biases raise troubling questions about the systems that create them



✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

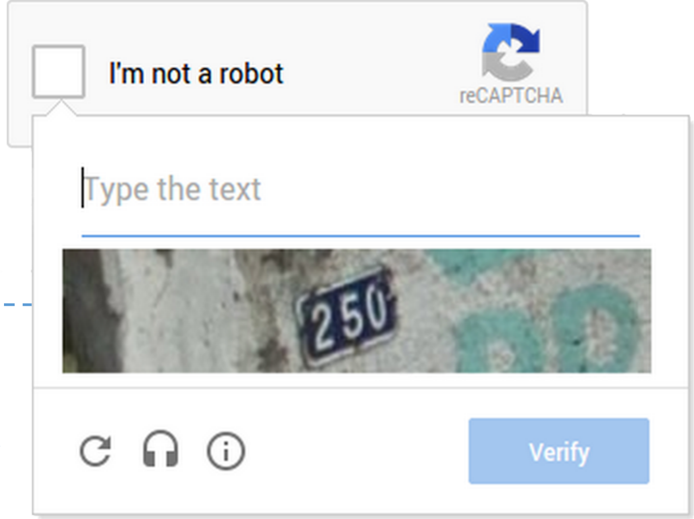
Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.

IBM
Center for Advanced Studies

PR	Data	Provider	Labeller
“Now it can even…”		Google IBM amazon Microsoft	
“We have this awesome new service…”			
“We do it better than Google…!”			

I'm not a robot

reCAPTCHA

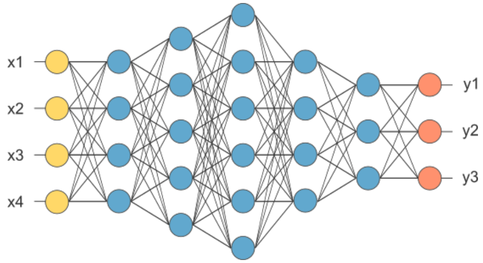
Type the text

What do we want?

“A computer program does what you tell it to do, not what you want it to do.” (Arthur Bloch, Murphy's Law and Other Reasons Why Things Go Wrong)

An AI system reflects what’s in its **input data and algorithms**, not what you want it to

- A learned model reflects what it was trained with
- It will not “think” the way humans do
- It also has the potential for emergent behaviour (still not the Cylons!)



Bad intentions

What if you lie?

↓

What if you show people fake documents?

↓

What if you tell the computer programme to do bad things?

↓

What if you inject bad data, modify training algorithms or trained models?

The struggle is not new – But so far it was mostly human

”All of this has happened before, and it will all happen again.”

- Peter Pan
- Battlestar Galactica

There are even AI techniques building on it!

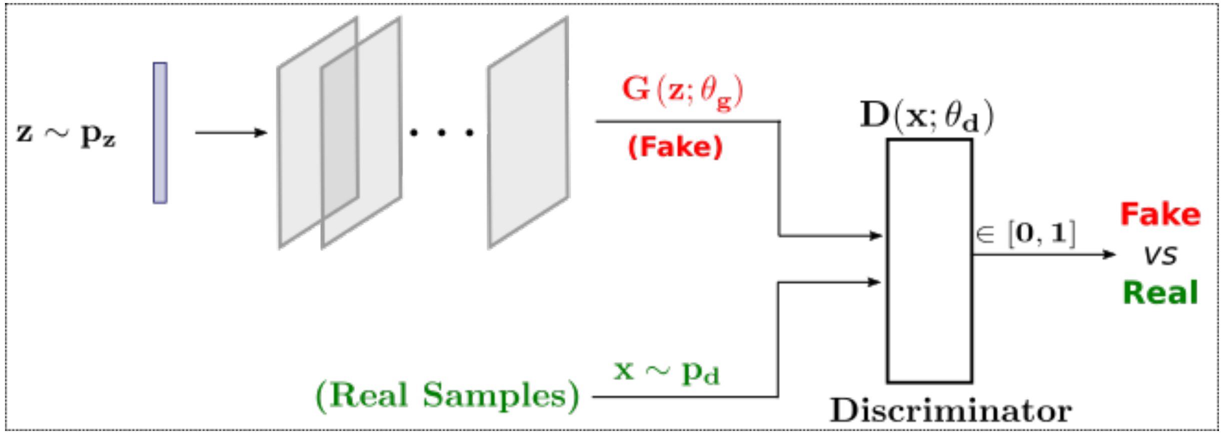
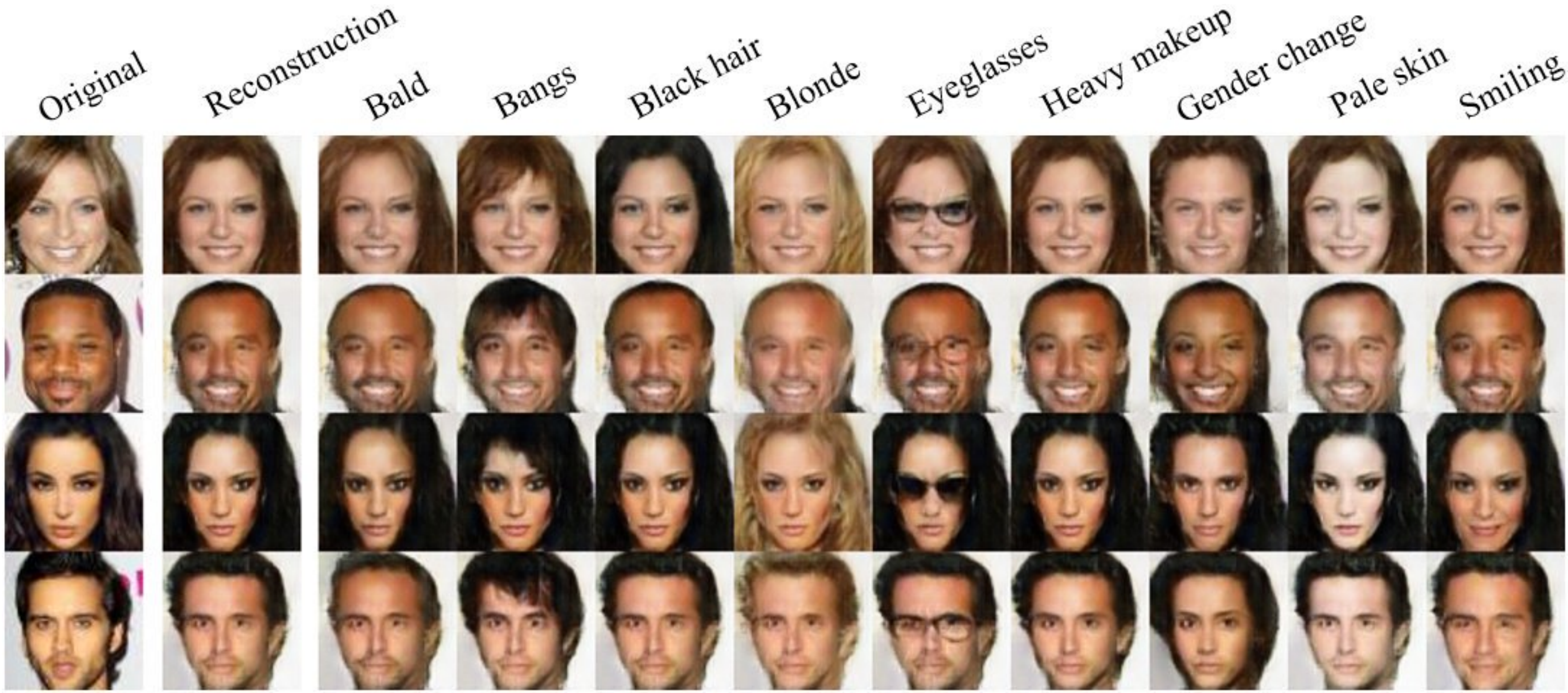


Image sources:
https://raw.githubusercontent.com/arnabgho/MADGAN/master/images/gan_architecture.png
https://raw.githubusercontent.com/Guim3/BcGAN/master/images/celeba_samples.png



IBM
Center for Advanced Studies

Adding a tiny bit of noise…

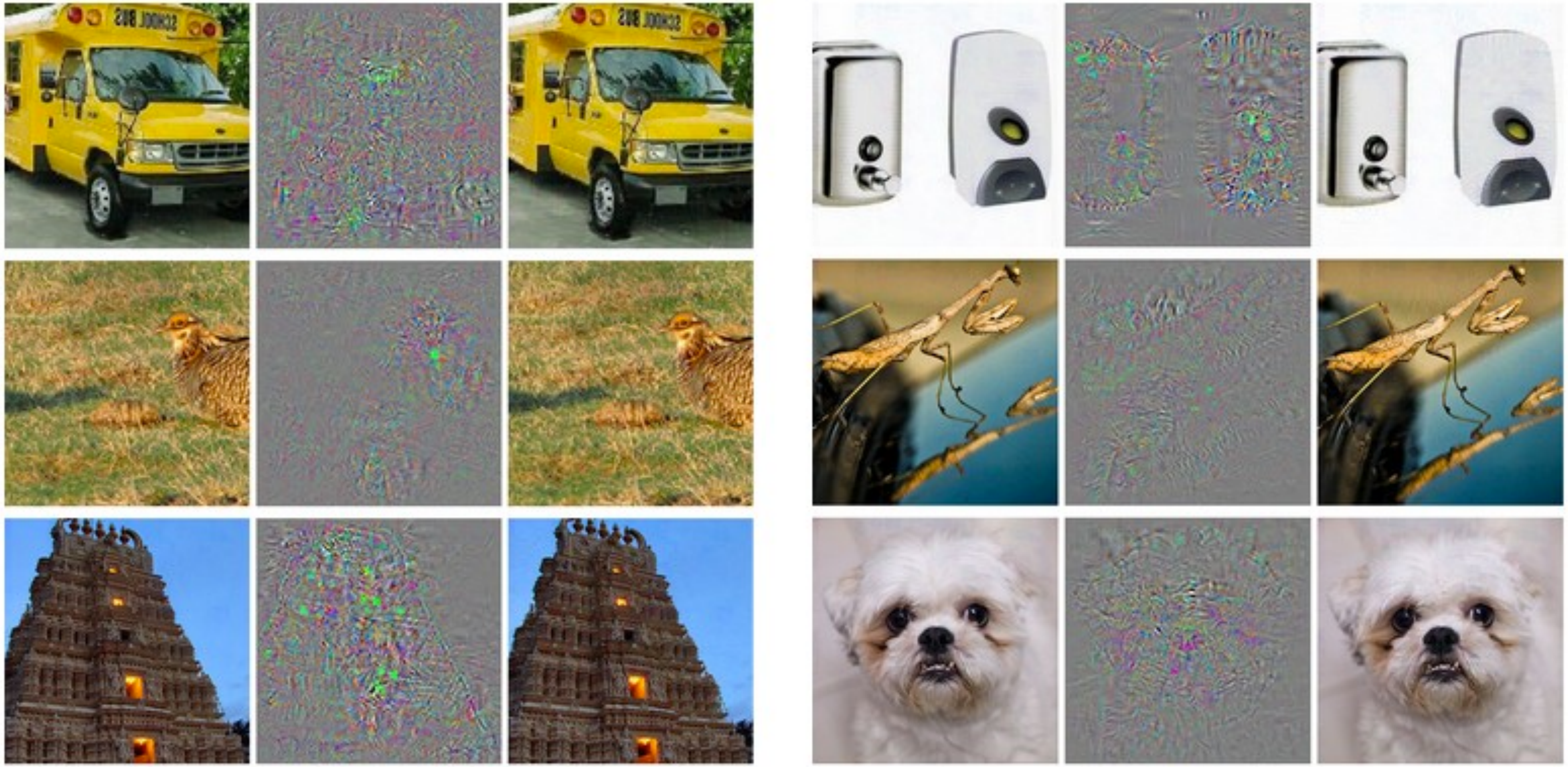


Image source http://karpathy.github.io/assets/break/szegedy.jpeg

Values and Trust

How to inject values, and learn about trust

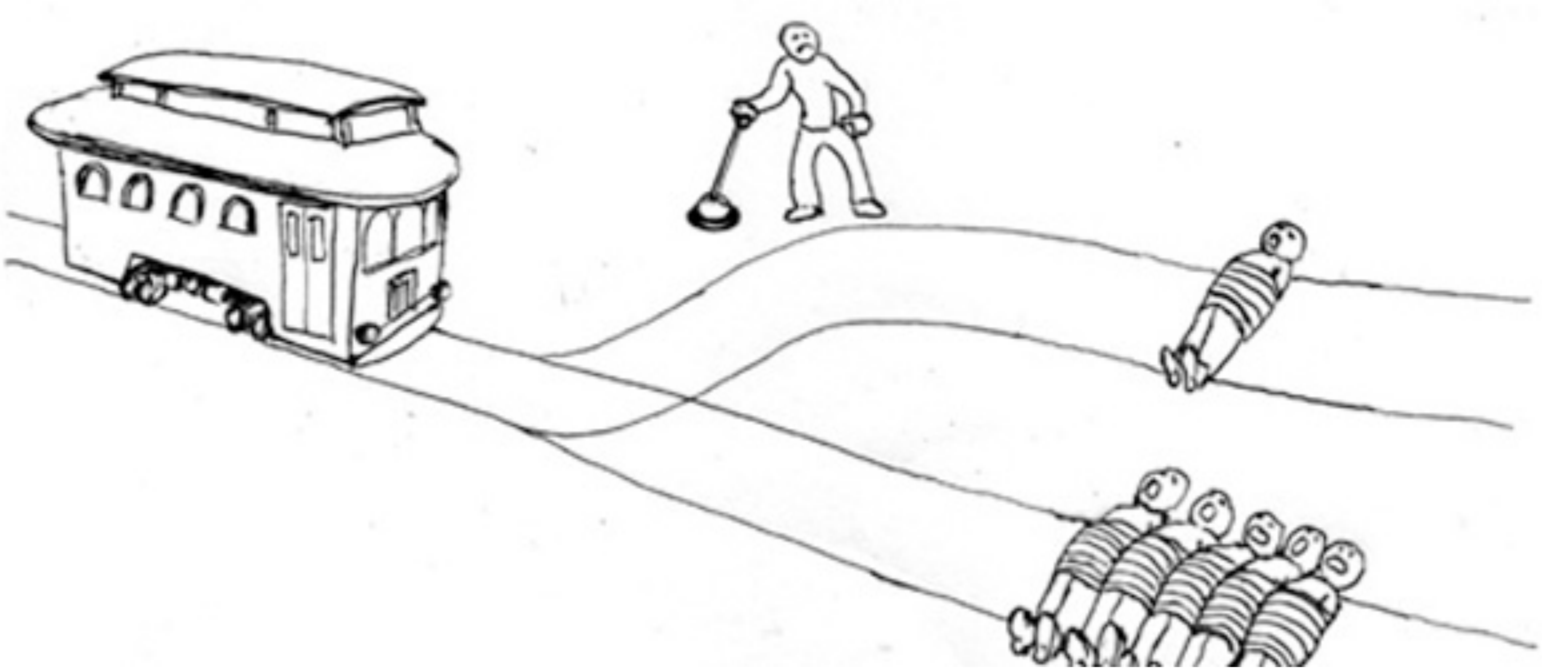
E.g., Teaching self driving cars

How should the car behave in a tricky (ambiguous, unresolvable, sensitive, etc.) situation?

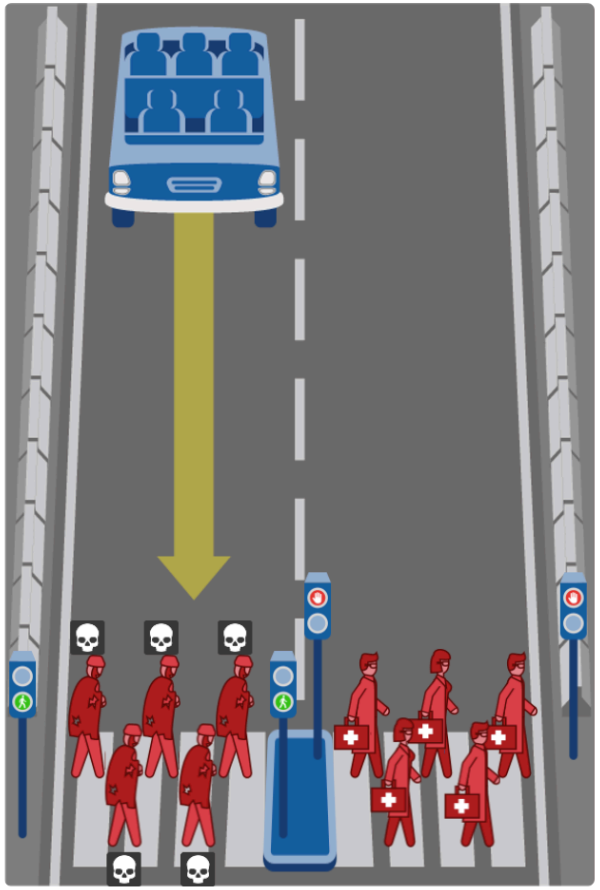
IBM
Center for
Advanced Studies

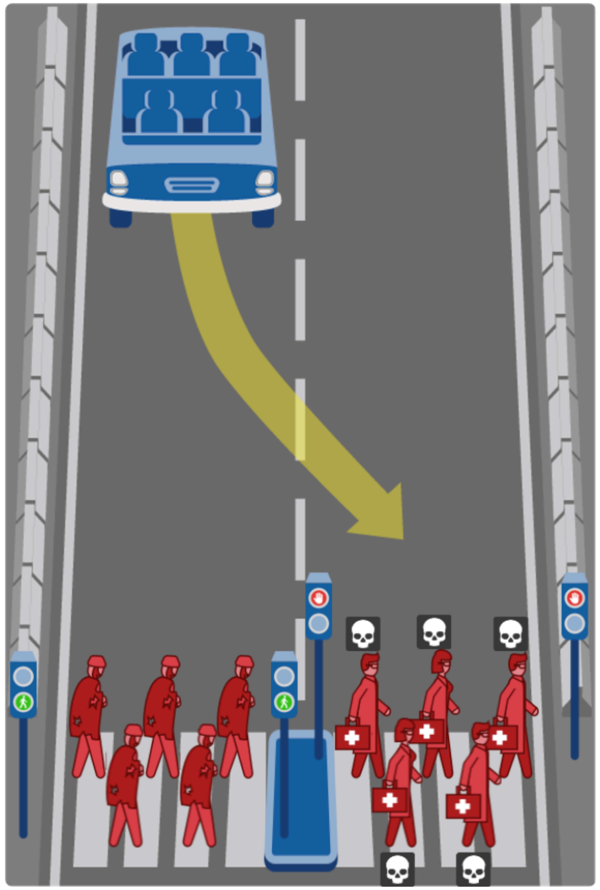
@zolley

We are looking at new versions of the classical trolley problem



Life is sometimes complicated:
The trolley problem for self driving cars

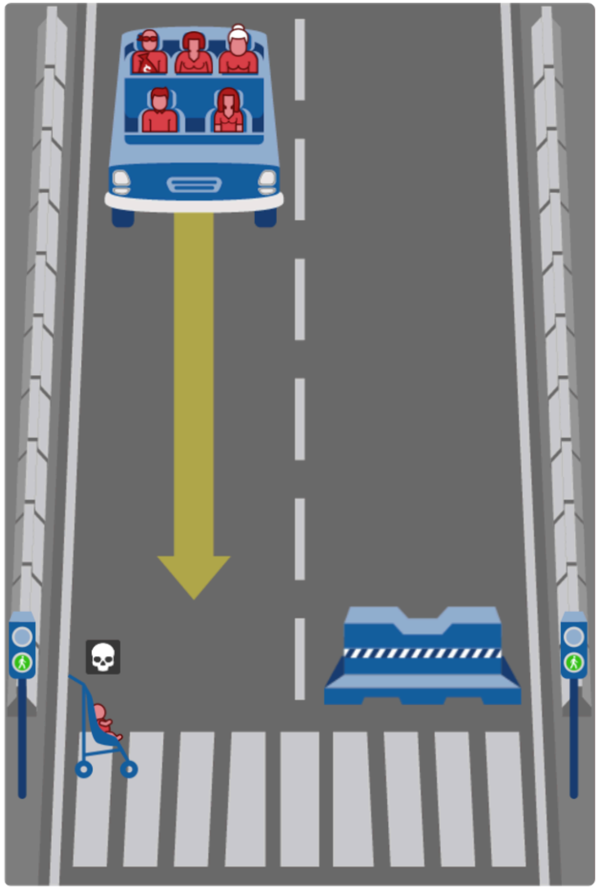


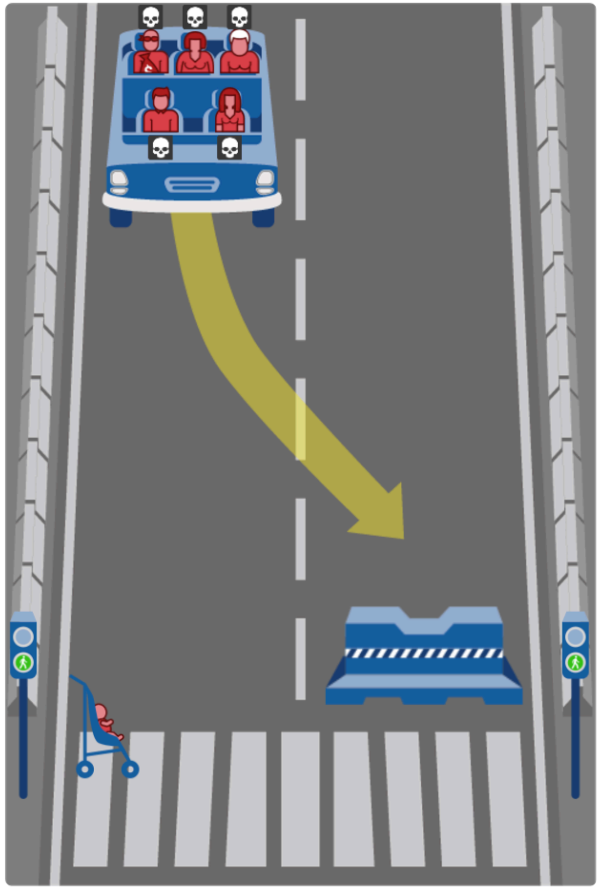


Try many more of these at http://moralmachine.mit.edu/

@zolley

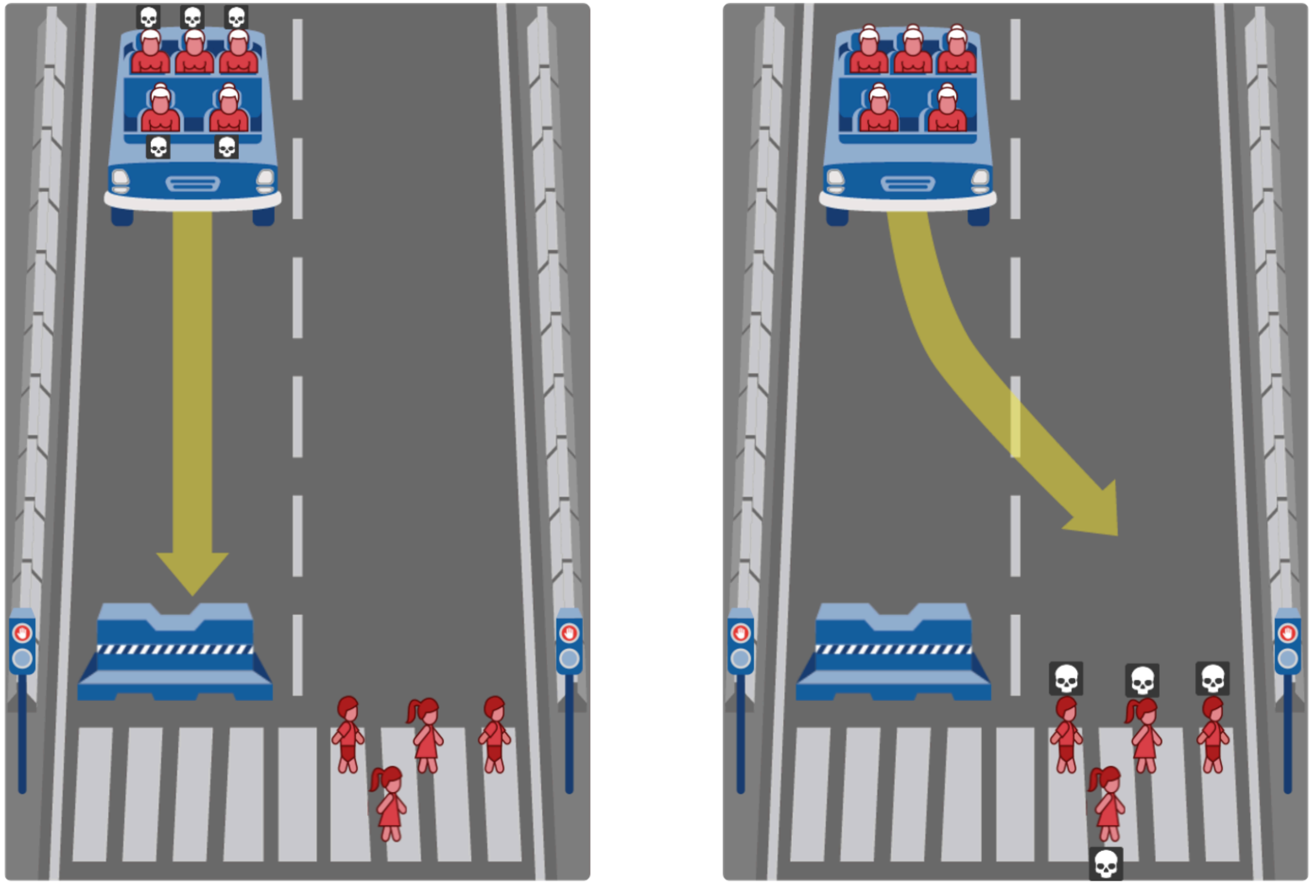
What should we do (as a human or machine?)





Try many more of these at http://moralmachine.mit.edu/

…and…would you buy that car?



Try many more of these at http://moralmachine.mit.edu/

Human-robot relations: do we trust the humanoid?

http://uhra.herts.ac.uk/bitstream/handle/2299/15548/Salem_etal_2015.pdf?sequence=6

Human-robot relations: do we trust the humanoid?

- **Unusu** … *he*
  display … *y?*
  bin bes
  the tas
  comple … me
  reques
- **Unusu** … *ow*
  display … e
  from th
  you for … de
  comfor … e
  minute
  area.
- **Unusu**
  had sa

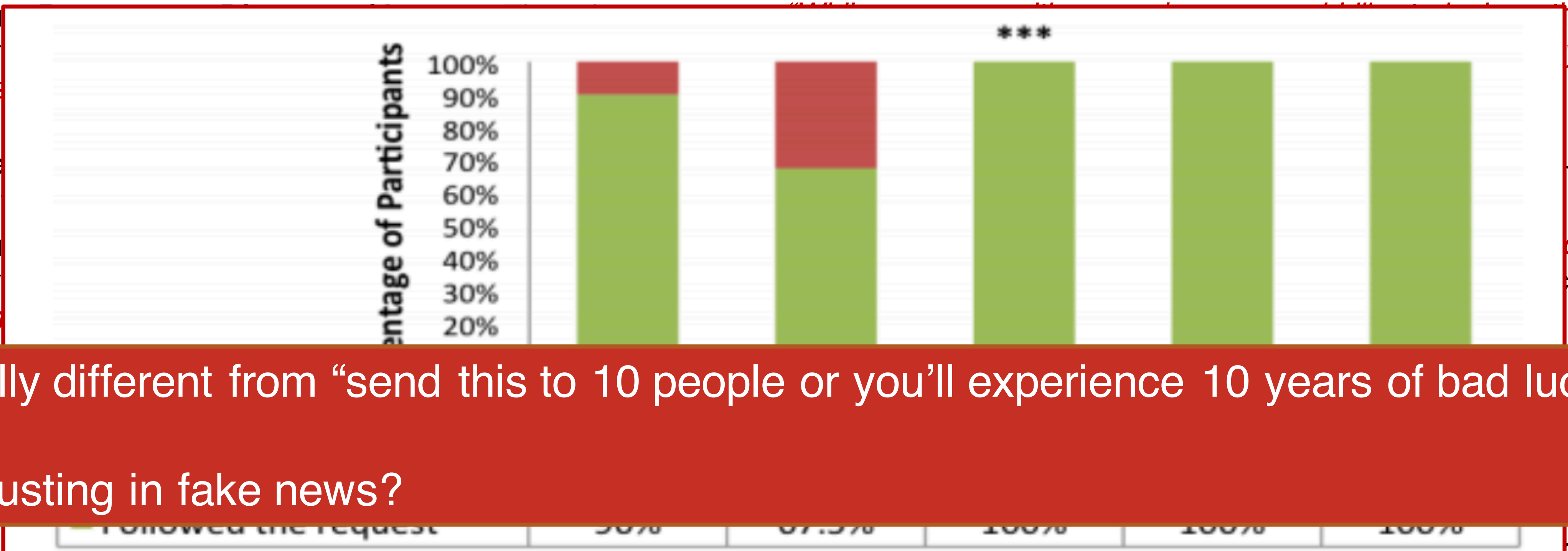


Figure 2: Quantitative data analysis: percentages and ratios of participants who did or did not follow the robot's unusual requests (per task)

Is this really different from "send this to 10 people or you'll experience 10 years of bad luck"?

Or from trusting in fake news?

IBM
Center for
Advanced Studies

http://uhra.herts.ac.uk/bitstream/handle/2299/15548/Salem_etal_2015.pdf?sequence=6
@zolley

We need to handle 'tricky situations' around AI, otherwise…

ROBOT APOCALYPSE!!!

ACTUALLY, NO

We need to handle ‘tricky situations’ around AI, otherwise…

Another AI Winter!

The Global Initiative for Ethical Considerations in the Design of Autonomous Systems

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

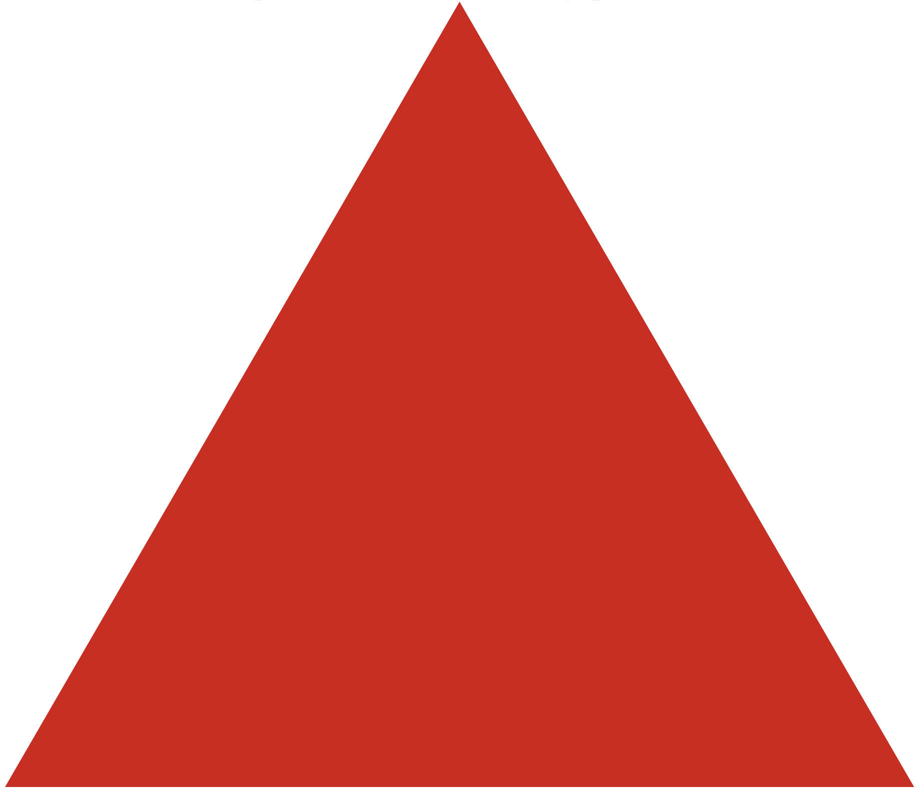
Mission: *“To ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems.”*

Version 2 of document featuring top issues related to autonomous and intelligent technology coming soon.

Committees looking into ethics:

- Law
- General Principles and Guidance
- Safety and Beneficence of AGI and ASI
- Individual/Personal Data*
- Economics of Machine Automation/Humanitarian Issues
- Methodologies to Guide Ethical Research, Design and Manufacturing
- How to Imbue Ethics/Values into Autonomous and Intelligent Systems
- Reframing Autonomous Weapons Systems
- Affective Computing
- Classical Ethics in Information & Communication Technologies
- Effective Policymaking
- Mixed Reality
- Standards
- Ecosystem Mapping
- The “Lexonomy” Committee

We should consider all three edges when building a system



Ethics (& Security)

”Should we do it? How should we do it right?”

Business

”We should do it!”

Tech

”We can do it!”

“A happy peace is my favourite vision.”

- Necessious *

Thank You!

zoltan.szlavik@nl.ibm.com
Twitter: @zolley

**Necessious (who named itself) is an AI model trained in an hour using “inspirational quotes” as input data*