HUMANE AI NET

Safety risks of AI: Intelligence, Complexity, and Stupidity

Paul Lukowicz,
DFKI/TU Kaiserslautern, Germany
HumanE AI Net Coordinator

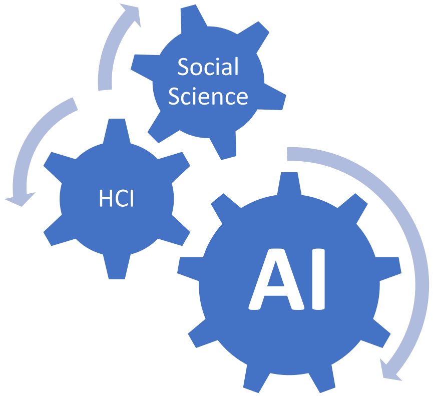
ICT-48-2020:
Towards a vibrant European
Network of AI excellence centres

HUMANE AI NET

Unique Selling Points

- As all ICT 48 we do Human Centric, Trustoworthy AI with European Values, but
  - we focus on AI that **enhances** human capabilities and **empowers** citizens
  - we consider both the **individual and the society** as a whole
  - we do dedicated research in ethical and fundamental rights, and **protection by design**

- We bring together a **unique community**
  - from AI and beyond (HCI, social science,law,..)



Unique Selling Points

- As all ICT 48 we do Human Centric, Trustoworthy AI with European Values, but
  - we focus on AI that **enhances** human capabilities and **empowers** citizens
  - we consider both the **individual and the society** as a whole
  - we do dedicated research in ethical and fundamental rights, and **protection by design**

1. Understanding Human-AI Collaboration
2. Common Ground and Shared Representations
   - Narratives
3. Human/Social View of Trustworthiness and Explanation
4. AI-influenced socio-technical systems (AI-STS)
5. Research Methodology
6. Ethics and Legal Protection by Design

Narrow technical view of AI Safety

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Analysis

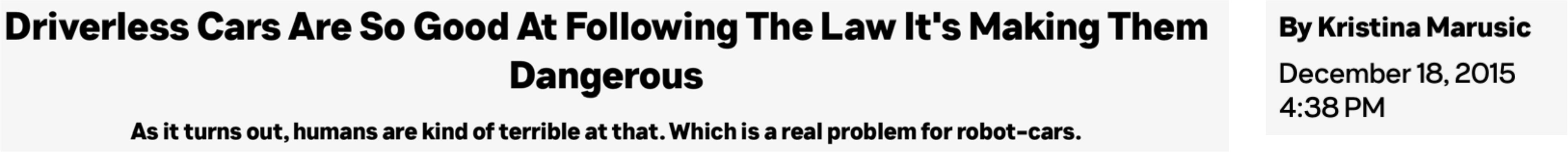
Key Concepts in AI Safety: An Overview

Tim G. J. Rudner and Helen Toner

March 2021

- … “AI safety” focuses on *technical* solutions to ensure that AI systems operate safely and reliably.
  - identify potential causes of unintended behavior in machine learning systems and develop tools to reduce the likelihood of such behavior occurring
- Problems in AI safety can be grouped into three categories: *robustness*, *assurance*, and *specification*

From the early days of self driving cars

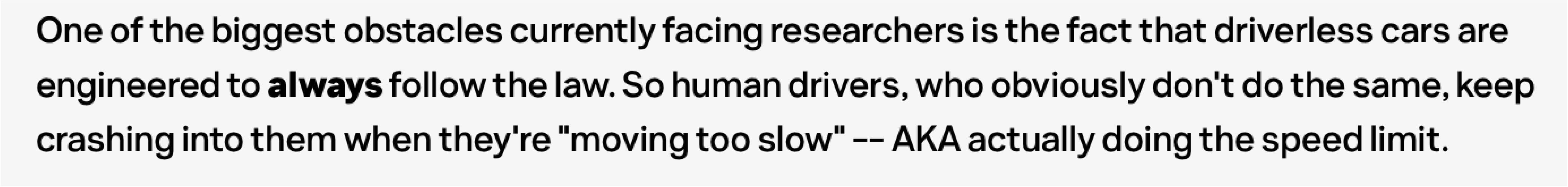


Driverless Cars Are So Good At Following The Law It's Making Them Dangerous

As it turns out, humans are kind of terrible at that. Which is a real problem for robot-cars.

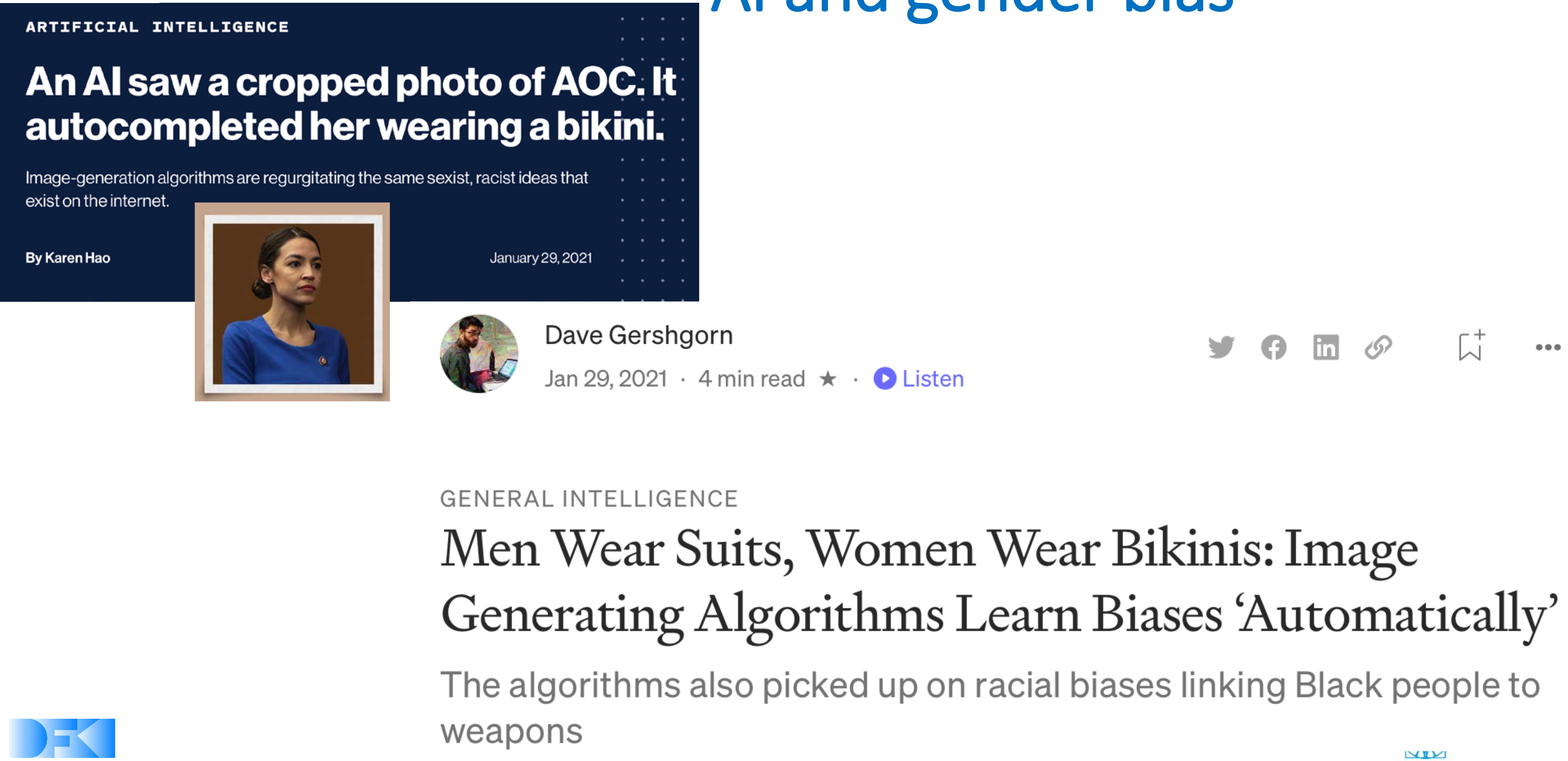
By Kristina Marusic

December 18, 2015
4:38 PM



One of the biggest obstacles currently facing researchers is the fact that driverless cars are engineered to **always** follow the law. So human drivers, who obviously don't do the same, keep crashing into them when they're "moving too slow" -- AKA actually doing the speed limit.

AI and gender bias



MIT Technology Review

ARTIFICIAL INTELLIGENCE

An AI saw a cropped photo of AOC. It autocompleted her wearing a bikini.

Image-generation algorithms are regurgitating the same sexist, racist ideas that exist on the internet.

By Karen Hao

January 29, 2021

Dave Gershgorn

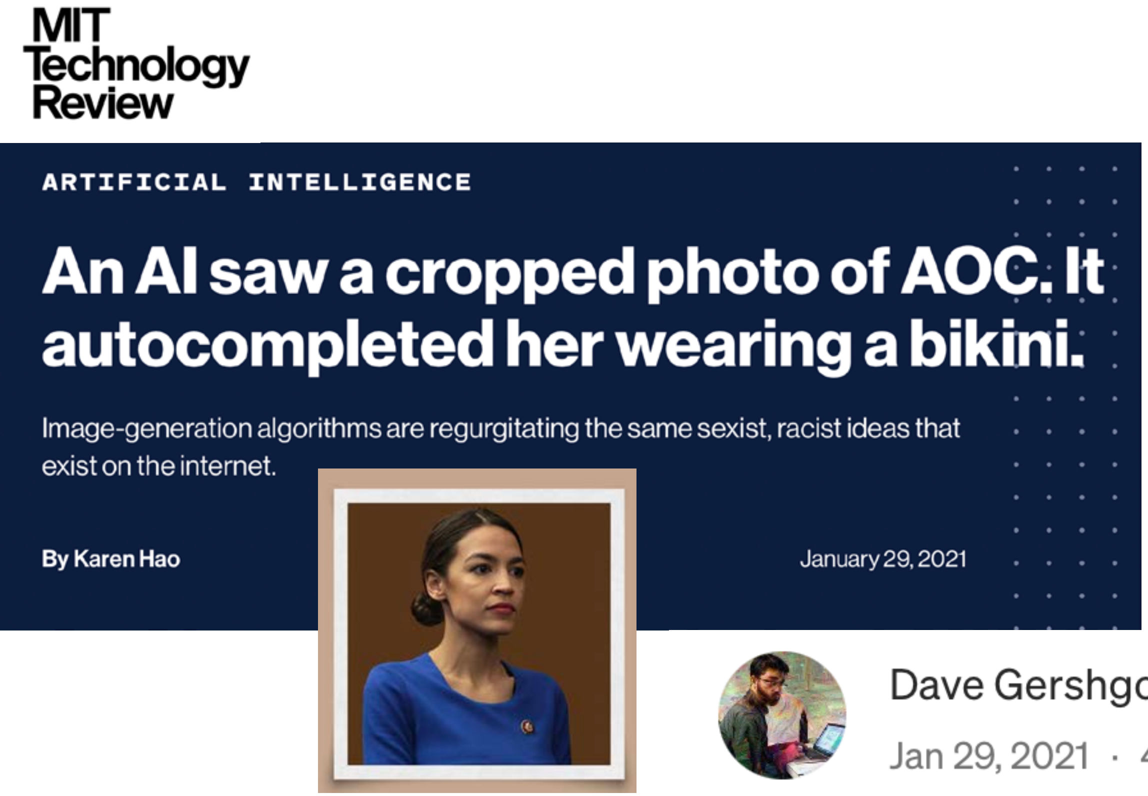
Jan 29, 2021 · 4 min read ★ · Listen

GENERAL INTELLIGENCE

Men Wear Suits, Women Wear Bikinis: Image Generating Algorithms Learn Biases ‘Automatically’

The algorithms also picked up on racial biases linking Black people to weapons

AI and gender bias



MIT Technology Review

ARTIFICIAL INTELLIGENCE

An AI saw a cropped photo of AOC. It autocompleted her wearing a bikini.

Image-generation algorithms are regurgitating the same sexist, racist ideas that exist on the internet.

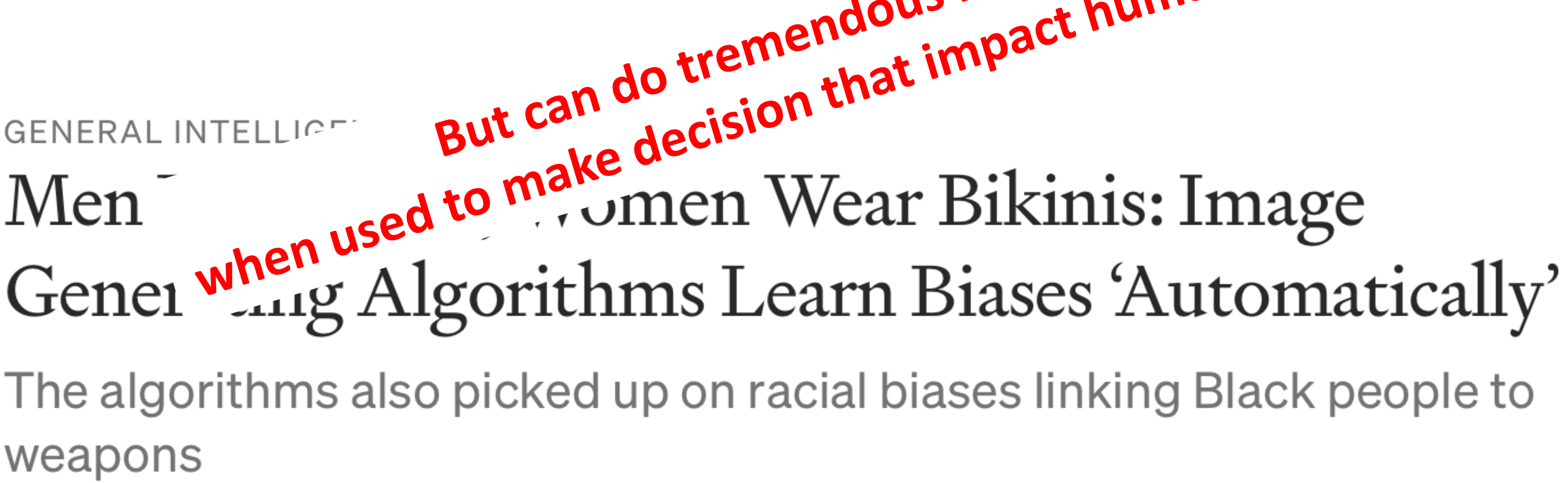
By Karen Hao

January 29, 2021

**AI is as such is not biased,
it is just a mirror
for our biased society !**

Dave Gershgorn

Jan 29, 2021 · 4 min read ★ · Listen



GENERAL INTELLIG

Men ˉ ,omen Wear Bikinis: Image Gene ing Algorithms Learn Biases ‘Automatically’

The algorithms also picked up on racial biases linking Black people to weapons

**But can do tremendous harm
when used to make decision that impact human lives**

News

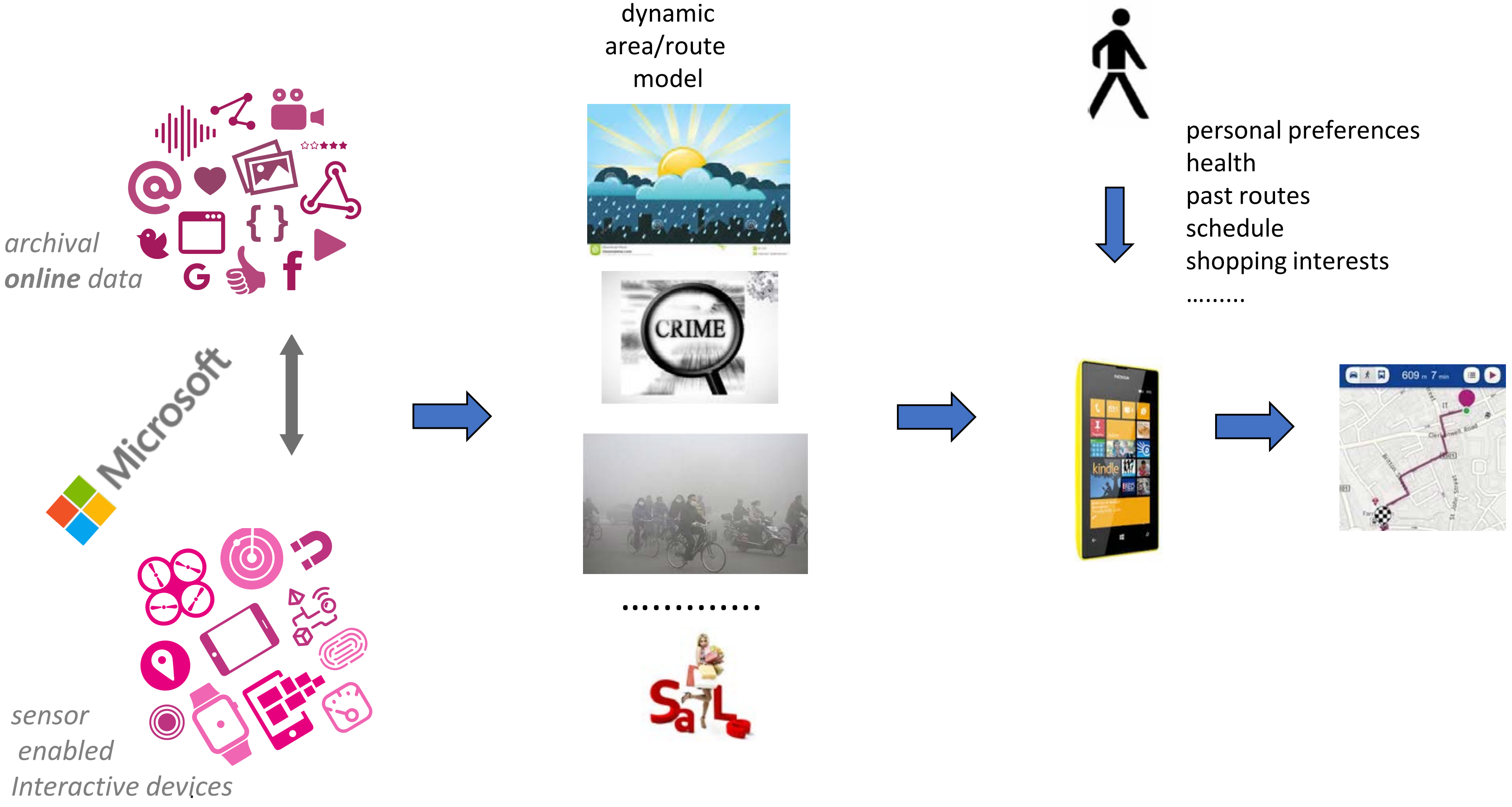
Microsoft Patents Bad Neighborhood Detection

1 Comment | David Chernicoff, BYTE | January 07, 2012 05:06 PM

News

Microsoft Patents Bad Neighborhood Detection

1 Comment | David Chernicoff, BYTE | January 07, 2012 05:06 PM



archival ***online*** *data*

sensor
enabled
Interactive devices

dynamic
area/route
model

....

personal preferences
health
past routes
schedule
shopping interests
....

News

Mic … od Detection

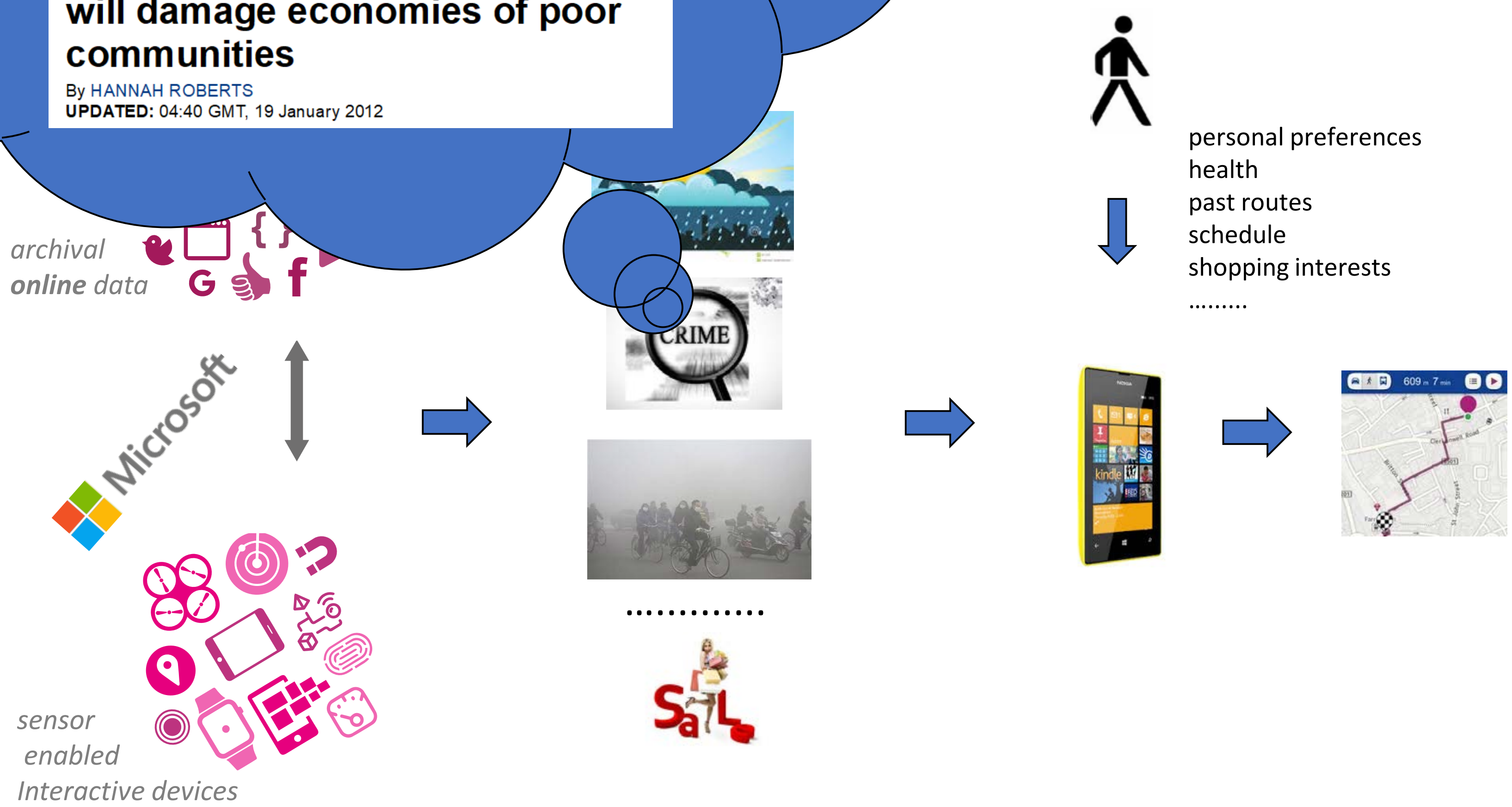
1

GHETTO
SAFE
SAFE

Racist or realistic? Fears Microsoft 'Avoid the Ghetto' app will damage economies of poor communities

By HANNAH ROBERTS

UPDATED: 04:40 GMT, 19 January 2012



AI safety

Article Talk

WIKIPEDIA

The Free Encyclopedia

AI safety is an interdisciplinary field concerned with preventing accidents, misuse, or other harmful consequences that could result from artificial intelligence (AI) systems.

AI Safety aspects

1. Technical safety: “classical” view of robustness, assurance, and specification of in particular ML systems
2. Human Computer Interaction aspects of safety
3. Social/ethical aspects of safety
4. Collective phenomena related aspects of safety

😫 Superintelligence related safety concerns

AI Safety aspects

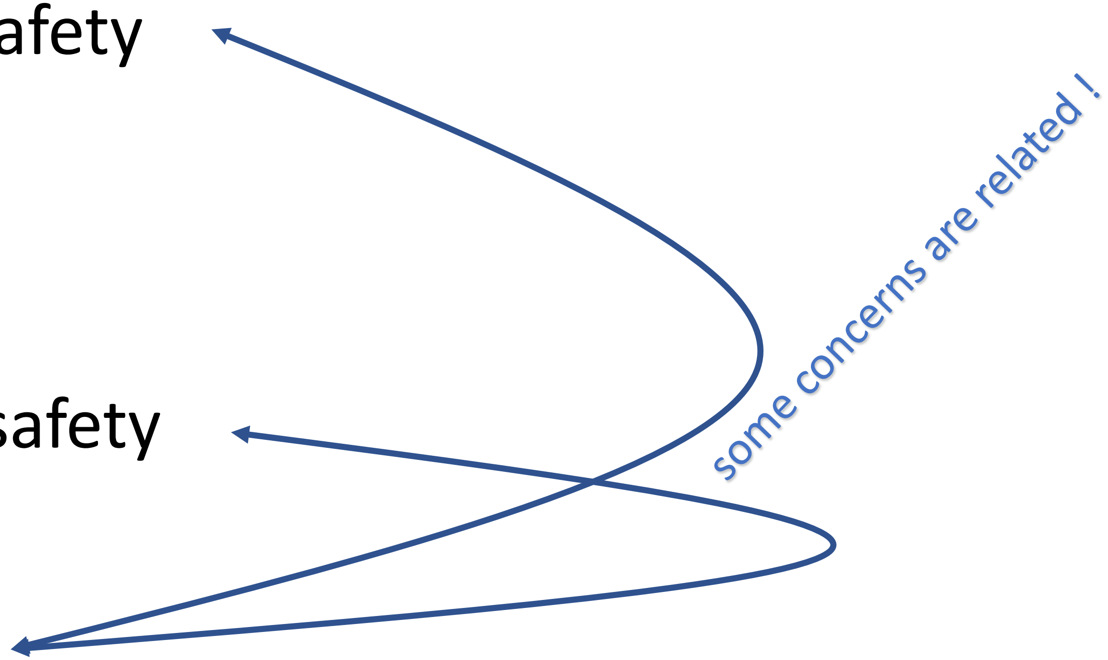
1. Technical safety: “classical” view of robustness, assurance, and specification of in particular ML systems

2. Human Computer Interaction aspects of safety

3. Social/ethical aspects of safety

4. Collective phenomena related aspects of safety

😩 Superintelligence related safety concerns

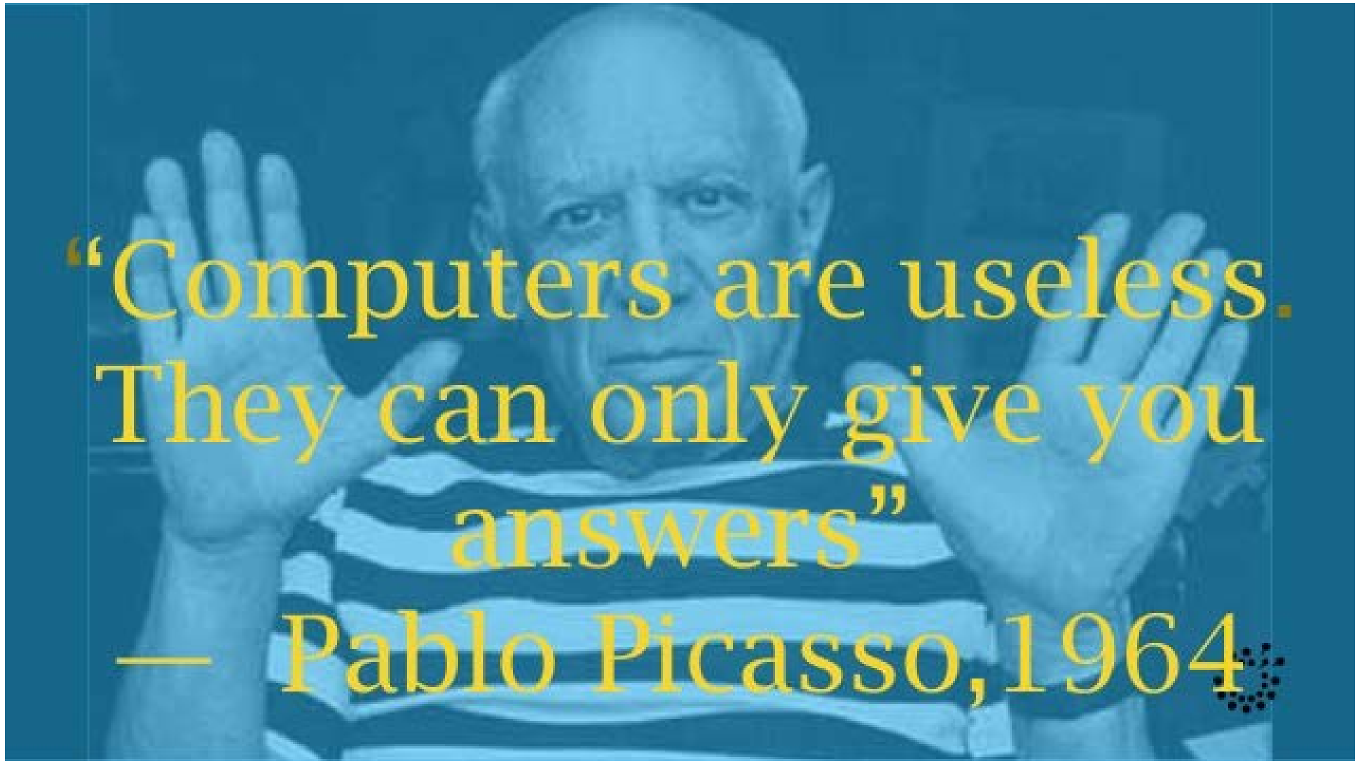


Nowak, A., Lukowicz, P., & Horodecki, P. (2018). Assessing artificial intelligence for humanity: Will ai be the our biggest ever advance? or the biggest threat [opinion]. *IEEE Technology and Society Magazine*, *37*(4), 26-34.

Unique Selling Points

- As all ICT 48 we do Human Centric, Trustoworthy AI with European Values, but
  - we focus on AI that **enhances** human capabilities and **empowers** citizens
  - we consider both the **individual and the society** as a whole
  - we do dedicated research in ethical and fundamental rights, and **protection by design**

1. Understanding Human-AI Collaboration
2. Common Ground and Shared Representations
   - Narratives
3. Human/Social View of Trustworthiness and Explanation
4. AI-influenced socio-technical systems (AI-STS)
5. Research Methodology
6. Ethics and Legal Protection by Design



“Computers are useless. They can only give you answers”

— Pablo Picasso,1964

Human-AI Colaboration: What for ?

Have humans and computers play out their specific strengths

data analysis, accuracy,speed,...

creativity, intuition,.....

Human-AI Colaboration: What for ?

Have humans and computers play out their specific strengths

makes sense in some cases
as intermediate solution due to
technology limitations,

data analysis, accuracy,speed,...

creativity, intuition,…..

Human-AI Colaboration: What for ?

Have humans and computers play out their specific strengths

creativity, intuition,…..

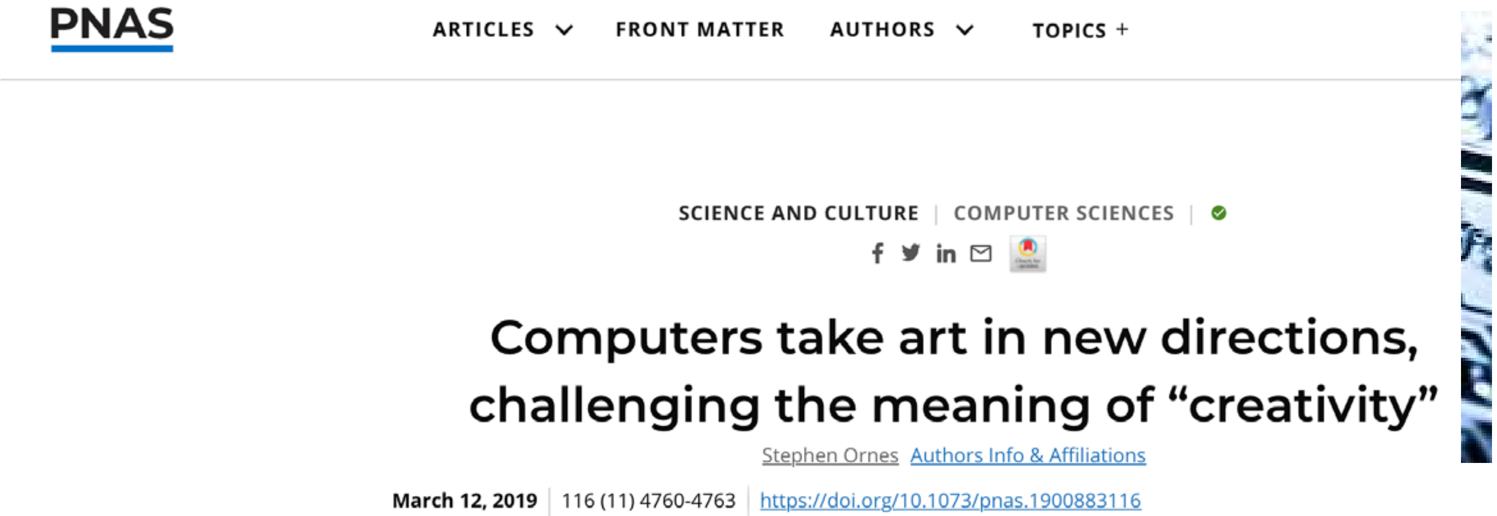
makes sense in some cases
as intermediate solution due to
technology limitations,

but
not the real reason !

data analysis, accuracy,speed,...

Human-AI Colaboration: What for ?

Have humans and computers play out their specific strengths



PNAS

ARTICLES ⌄ FRONT MATTER AUTHORS ⌄ TOPICS +

SCIENCE AND CULTURE | COMPUTER SCIENCES | ✓

Computers take art in new directions, challenging the meaning of "creativity"

Stephen Ornes Authors Info & Affiliations

March 12, 2019 | 116 (11) 4760-4763 | https://doi.org/10.1073/pnas.1900883116

data analysis, accuracy,speed,...

creativity, intuition,.....

really ?

Human-AI Colaboration: What for ?

- There are situations where the process is as important as the optimal answer

Human Centric AI Answer:

- There are situations where the process is as important as the optimal answer

- Sometimes for “human” reasons we want a human to make a decision, not a machine
  - **<u>make</u>** a decision, **<u>not rubber stamp</u>** a computer decision !

....

Human Centric AI Answer:

- There are situations where the process is as important as the optimal answer

- Sometimes for “h

on, not a machine

  - **make** a

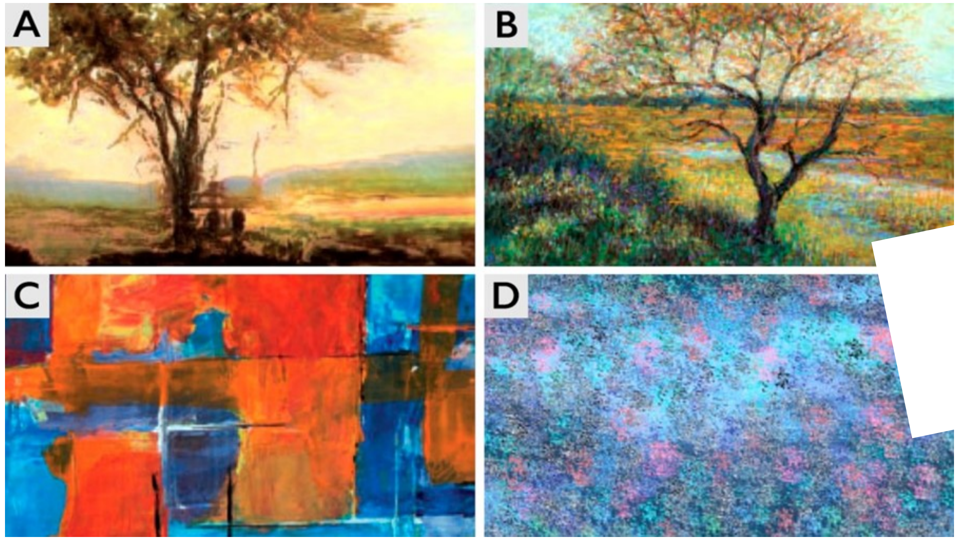
ision !

Does not mean that AI can not make the process better !

....

Applied to Art

- On signal level (pixels, accoustic signals etc) computers can already produce artifacts which for humans are largely indistinguishable from art



Human vs AI artworks, courtesy Harsha Gangadharbatla, Empirical Studies of the Arts

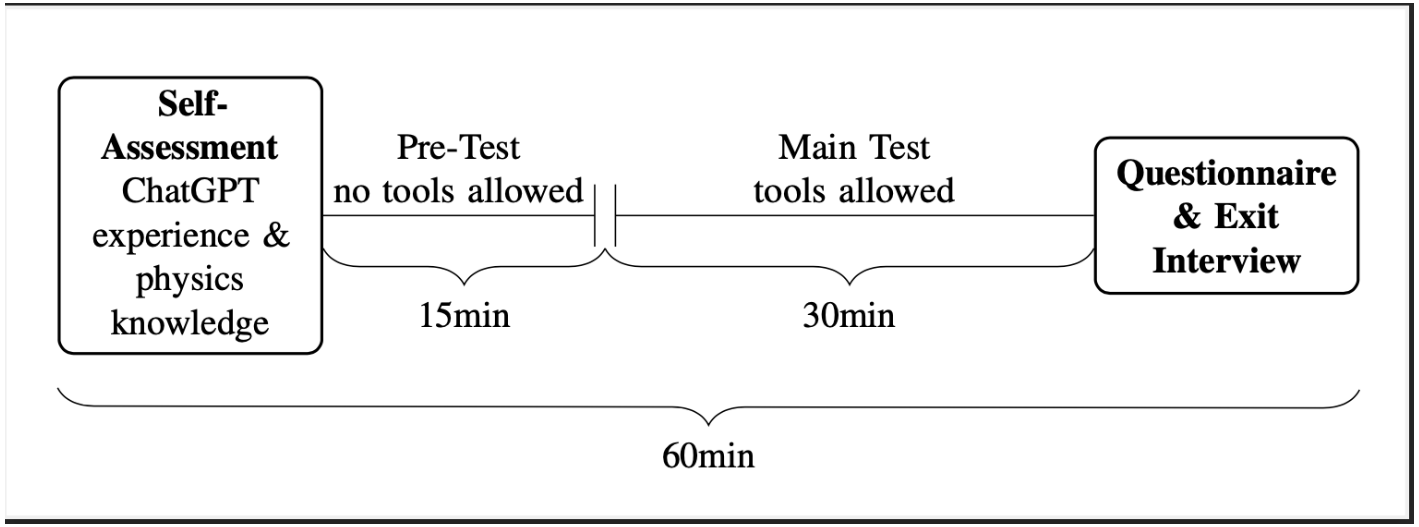
Does not mean that AI can not make the process better !

- But art is not just about signal level output, but about the process of generating that output as an expression of feelings, experiences, struggles etc. ideology etc, which is per definition human

Example study: Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education

What is the performance of students when being allowed to use ChatGPT instead of Google for solving physics problems ?

- Solve 4 of the tasks given at tasks given in the International Physics Olympiad (knowledge of kinematics, friction and rotational movements and inelastic collisions and conservation)
- N=27 had unrestricted access to ChatGPT, N=12 had access to a search engine



Self-Assessment ChatGPT experience & physics knowledge

Pre-Test no tools allowed

15min

Main Test tools allowed

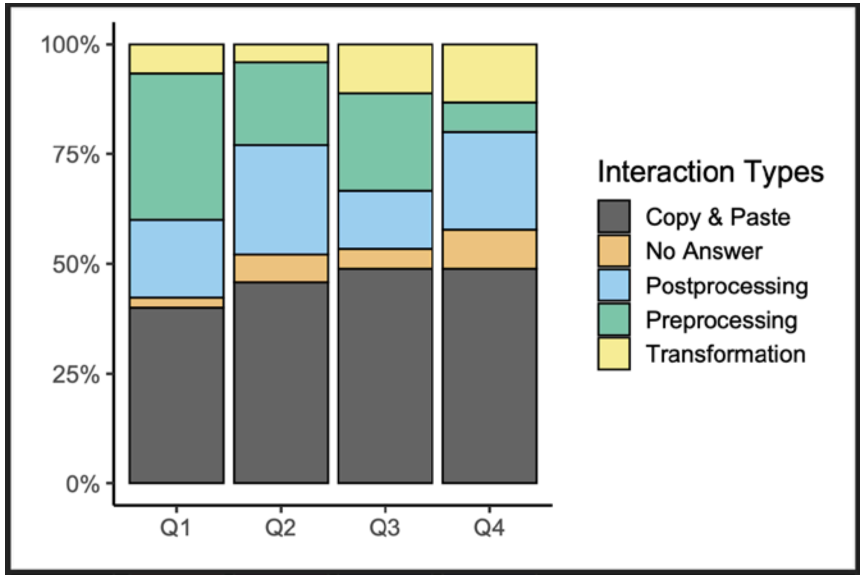
30min

Questionnaire & Exit Interview

60min

Submitted to AAAI 2023

Example study: Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education



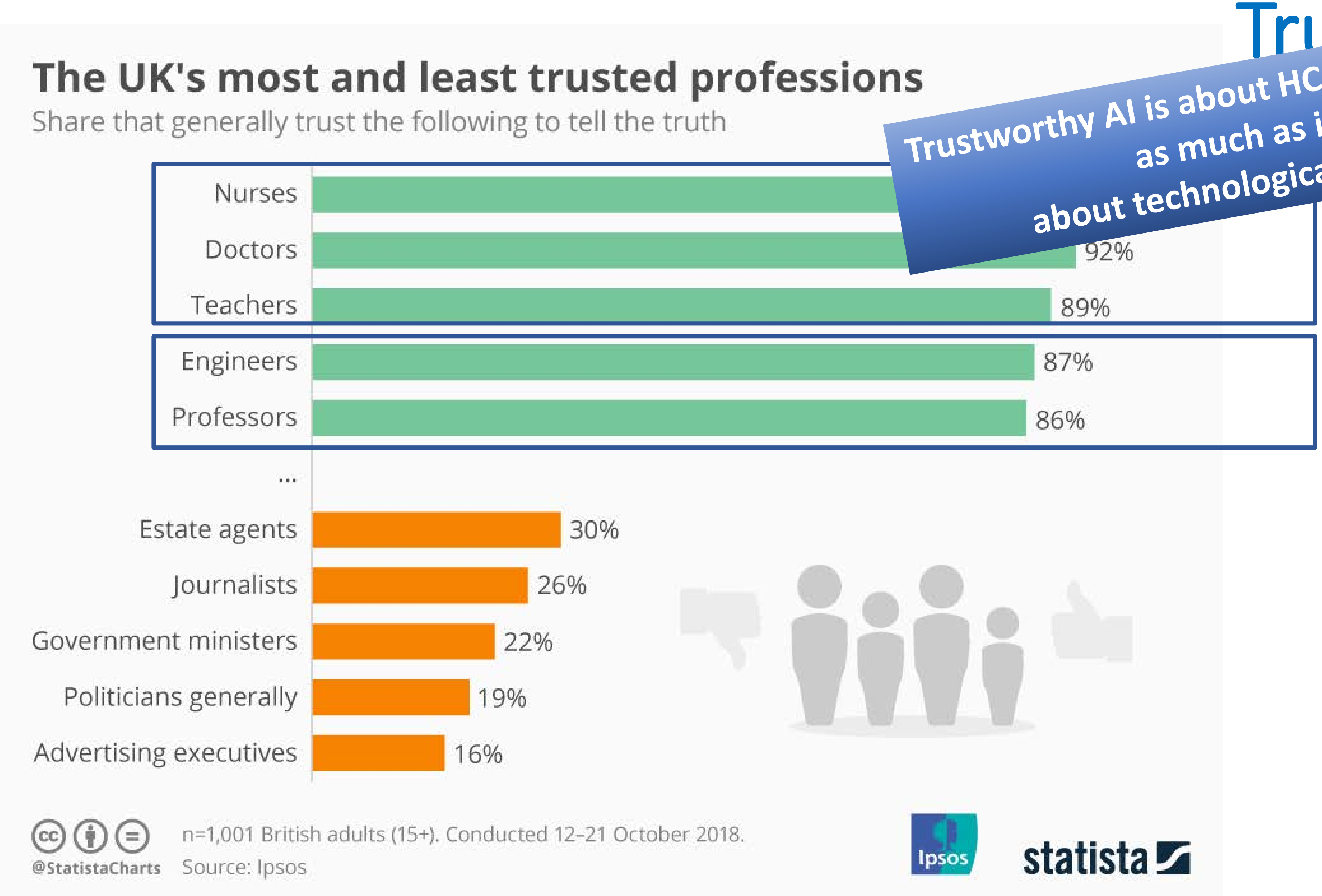
- On average, participants scored **$\bar{x}$=1.04** points (s=1.43) out of maximum achievable 12 points in the CHATGPT condition
  - the highest score achieved by a single student was six points. In total three students got more than two points, while **twelve students did not score any points at all**

- For the SEARCH ENGINE, participants scored **$\bar{x}$=1.83** points (s=1.27) on average.
  - Four points was the highest amount achieved by two students. In total three students achieved more than two points while **one student did not score a single point**

Students trusted (relied on out of laziness ?) CHATGPT too much

Trustworthiness

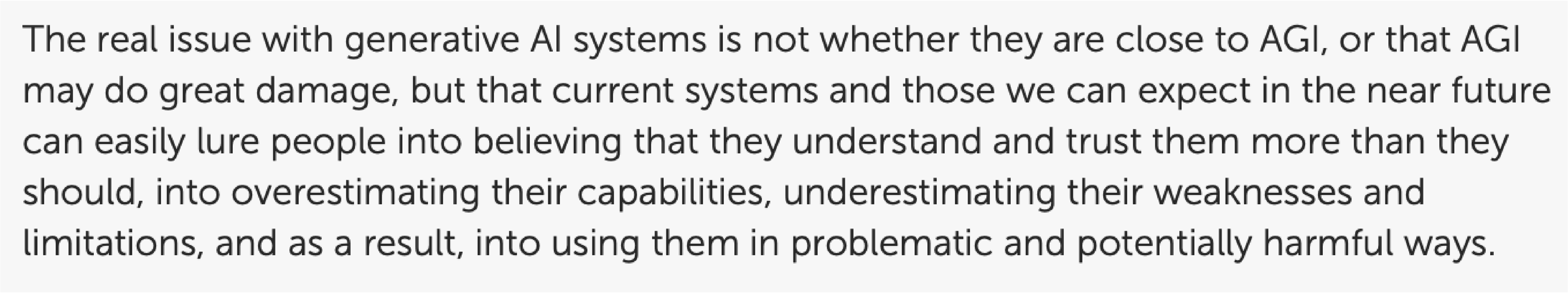
Trustworthy AI is about HCI and soft factors as much as it is about technological reliability



n=1,001 British adults (15+). Conducted 12-21 October 2018.
Source: Ipsos

„soft“ trust factors

„hard“ trust factors



The real issue with generative AI systems is not whether they are close to AGI, or that AGI may do great damage, but that current systems and those we can expect in the near future can easily lure people into believing that they understand and trust them more than they should, into overestimating their capabilities, underestimating their weaknesses and limitations, and as a result, into using them in problematic and potentially harmful ways.

Baum, Kevin, Joanna Bryson, Frank Dignum, Virginia Dignum, Marko Grobelnik, **Holger Hoos**, Morten Irgens et al. "From fear to action: AI governance and opportunities for all." *Frontiers in Computer Science* 5 (2023): 1210421.

The danger of the combination of Artificial and Natural Stupidity

AI Safety aspects

1. Technical safety: “classical” view of robustness, assurance, and specification of in particular ML systems

2. Human Computer Interaction aspects of safety

3. Social/ethical aspects of safety

4. Collective phenomena related aspects of safety

😫 Superintelligence related safety concerns

combination of artificial and natural stupidity!

Unique Selling Points

- As all ICT 48 we do Human Centric, Trustoworthy AI with European Values, but
  - we focus on AI that **enhances** human capabilities and **empowers** citizens
  - we consider both the **individual and the society** as a whole
  - we do dedicated research in ethical and fundamental rights, and **protection by design**

1. Understanding Human-AI Collaboration
2. Common Ground and Shared Representations
   - Narratives
3. Human/Social View of Trustworthiness and Explanation
4. AI-influenced socio-technical systems (AI-STS)
5. Research Methodology
6. Ethics and Legal Protection by Design

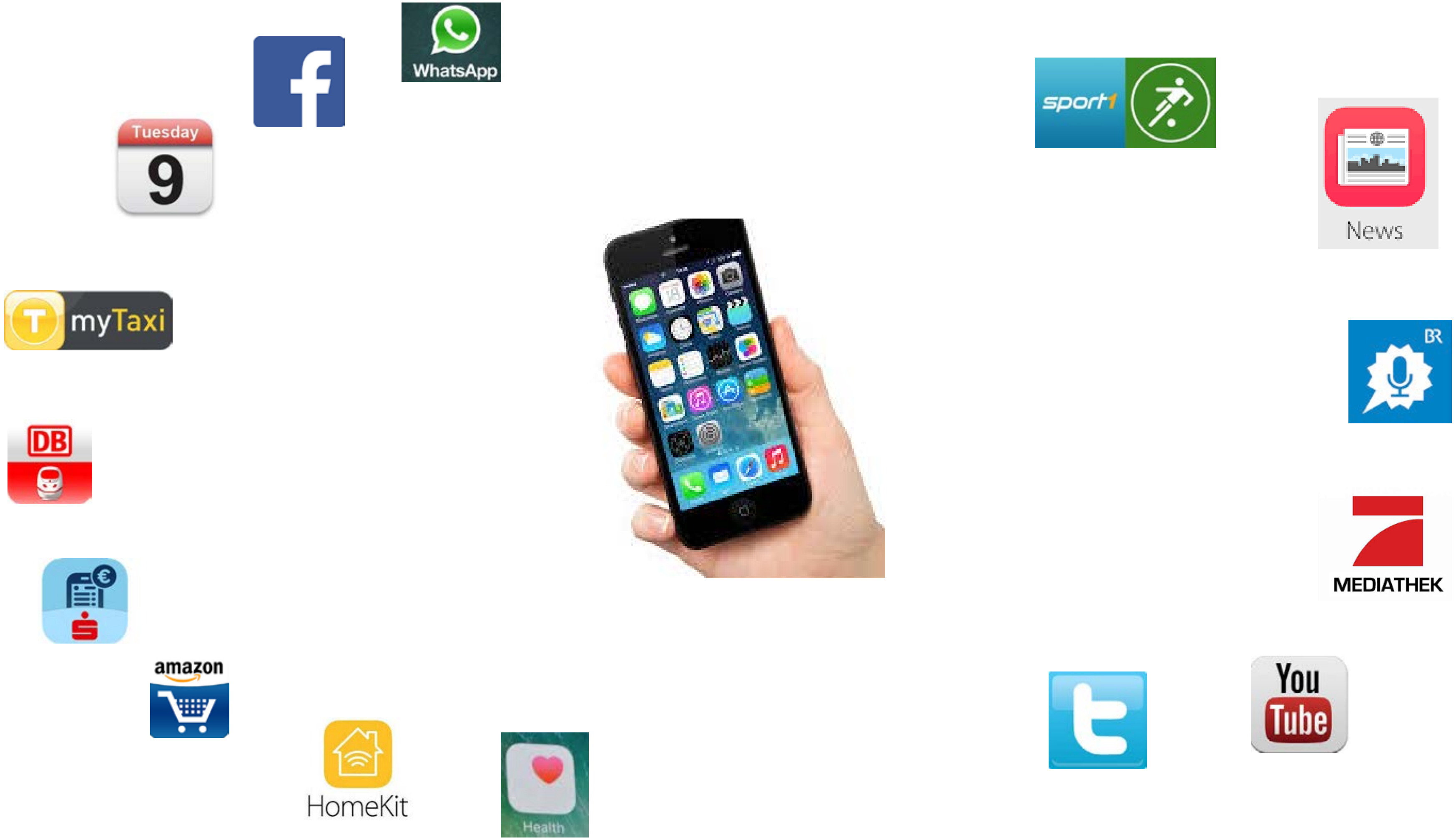
Accessing the digital domain

online once a day

online
few times a day

online up to 150
times a day

Digital Life: Smartphone era



Accessing the digital domain

online once a day

online
few times a day

online up to 150
times a day

online >>150 times
times a day

Accessing the digital domain

online once a day

online
few times a day

online up to 150
times a day

online >>150 times
times a day

online ~1000 times
times a day

Personal Digital Ecosystem

Accessing the digital domain

online once a day

online
few times a day

online up to 150
times a day

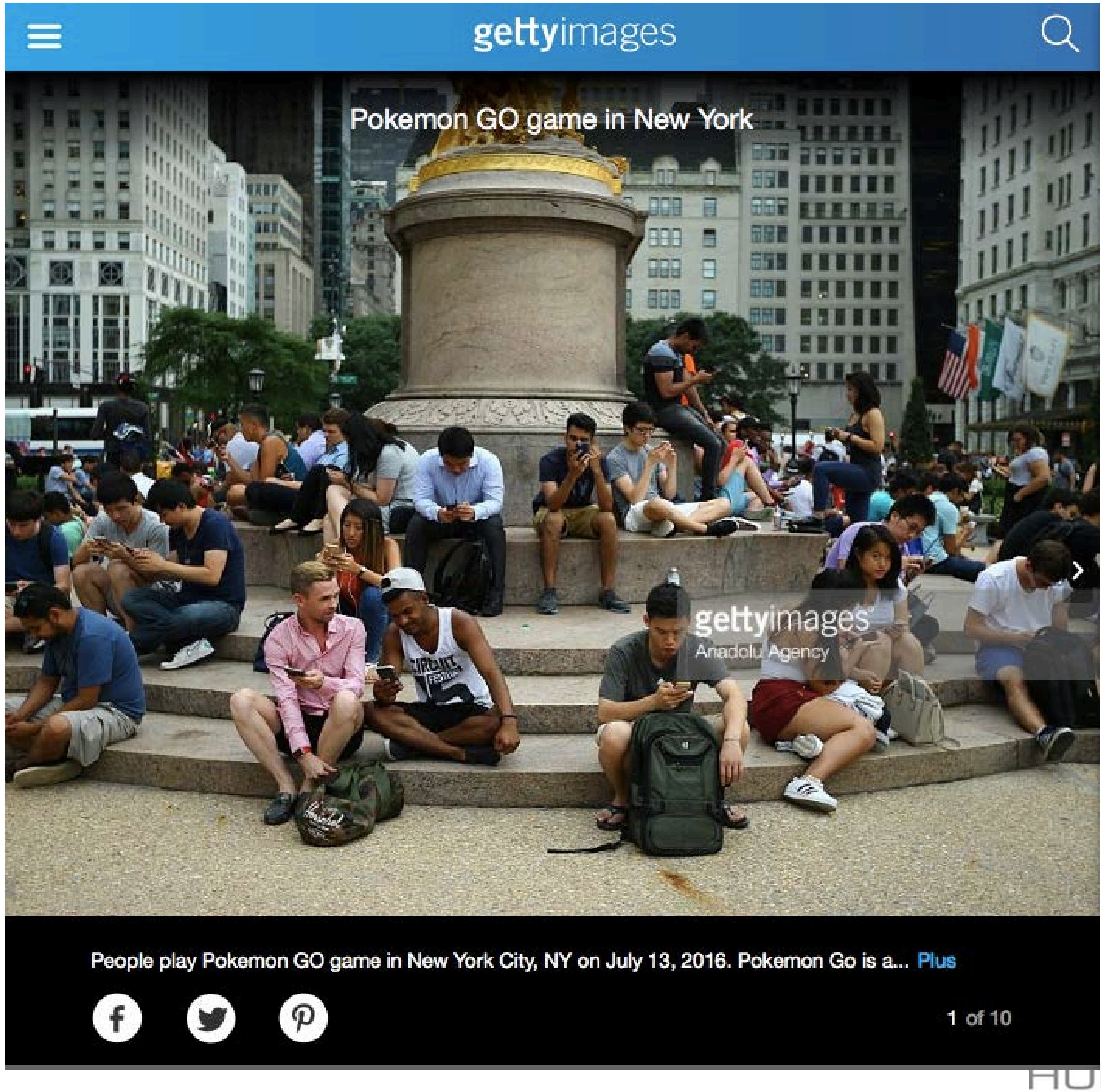
permanently present in the digital and real domain at the same time

"never offline"

Digital Life: Post Smartphone era

Digital Life: Post Smartphone era

Confluence of the Digital and the Physical World



Pokemon GO game in New York

People play Pokemon GO game in New York City, NY on July 13, 2016. Pokemon Go is a... Plus

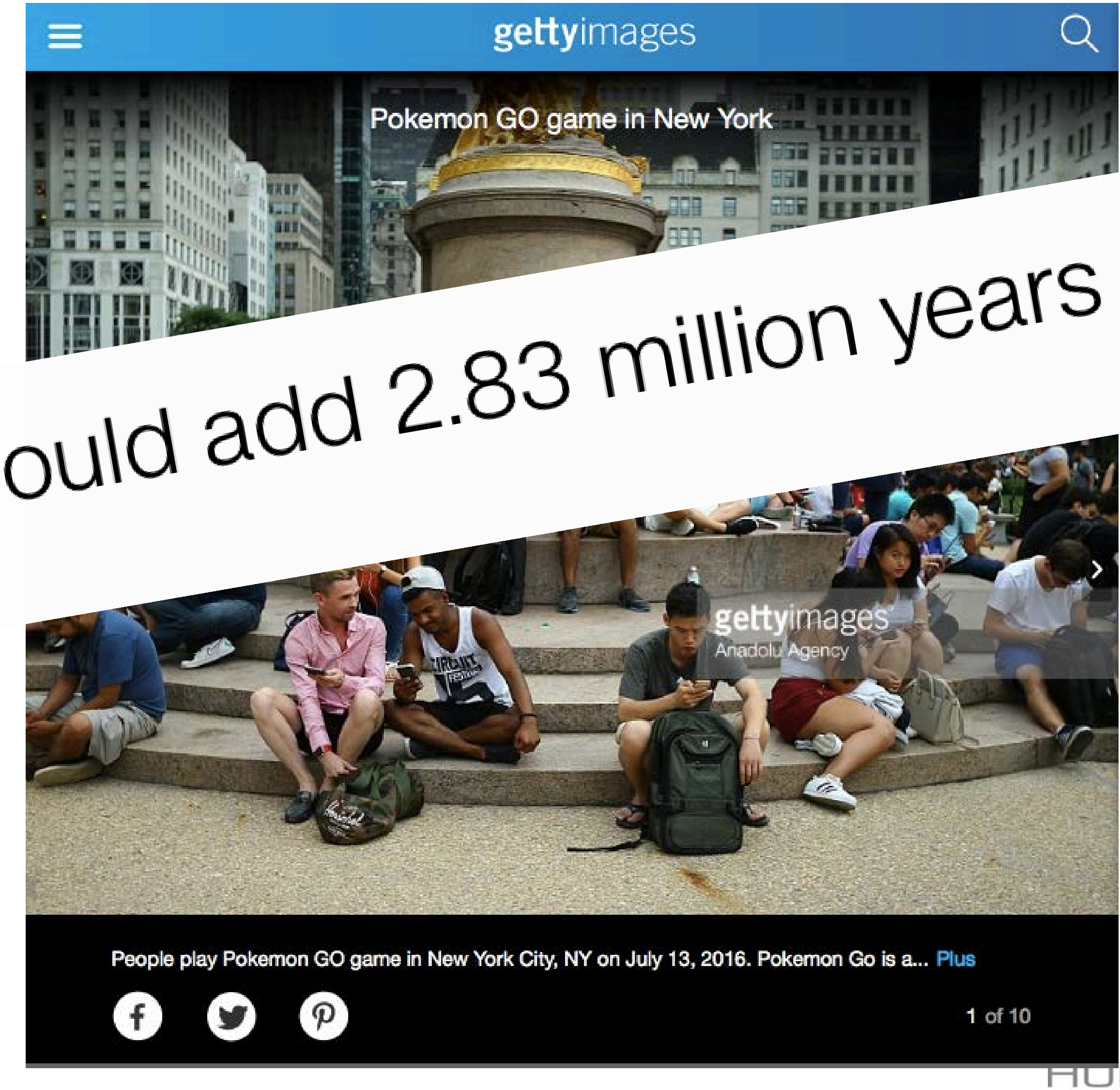
1 of 10

Confluence of the Digital and the Physical World

CNN tech

Innovate

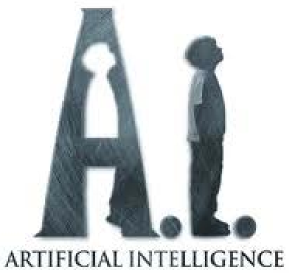
Pokemon Go could add 2.83 million years to users' lives



Pokemon GO game in New York

People play Pokemon GO game in New York City, NY on July 13, 2016. Pokemon Go is a... Plus

AI, Digitization and Society

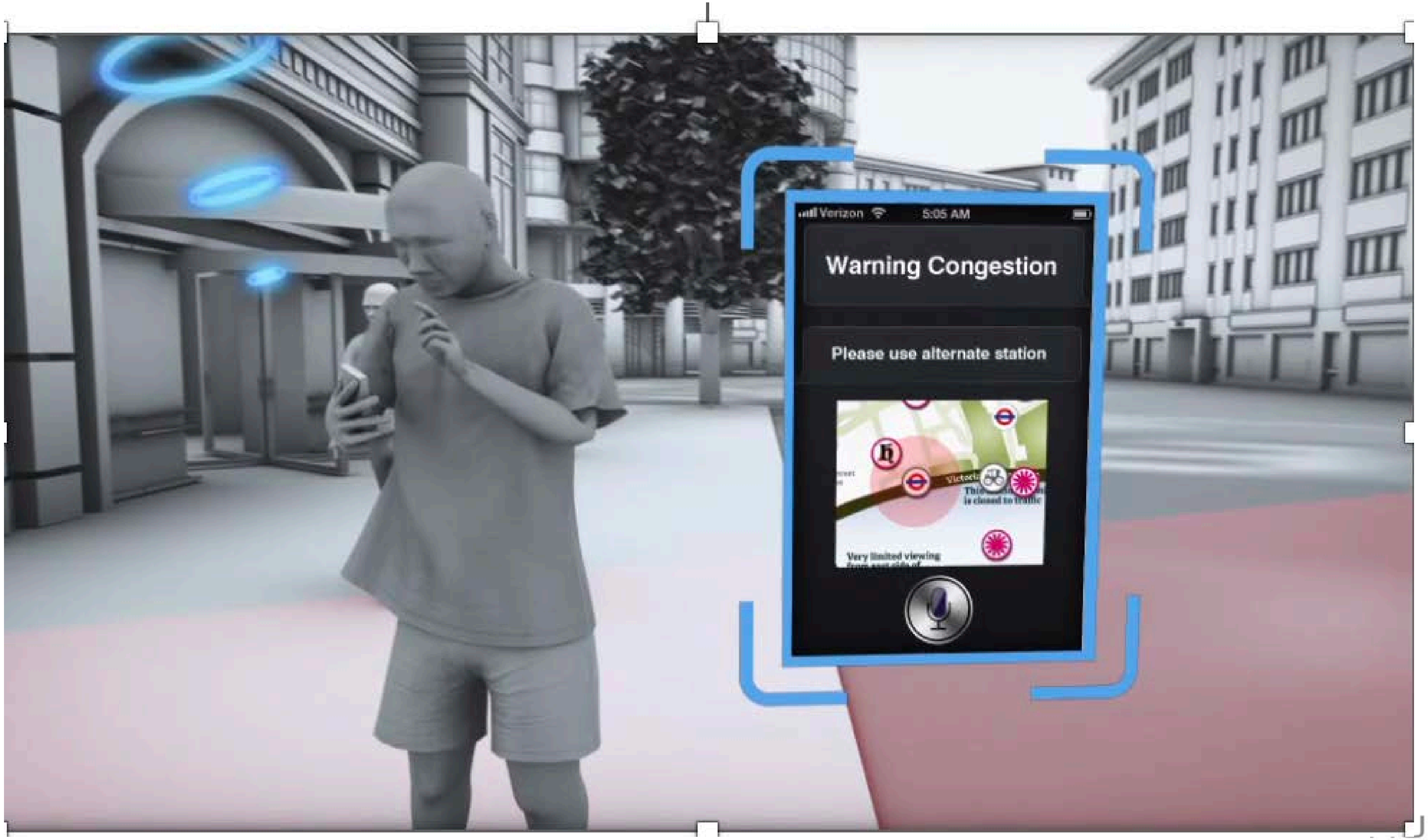


The ability to sense and interpret anything that is happing in the real world at any time

The ability to instantly influence any human and any part of the physical world

App Based Crowd Monitoring

Global situation dependent personalized messages at individualized times and locations

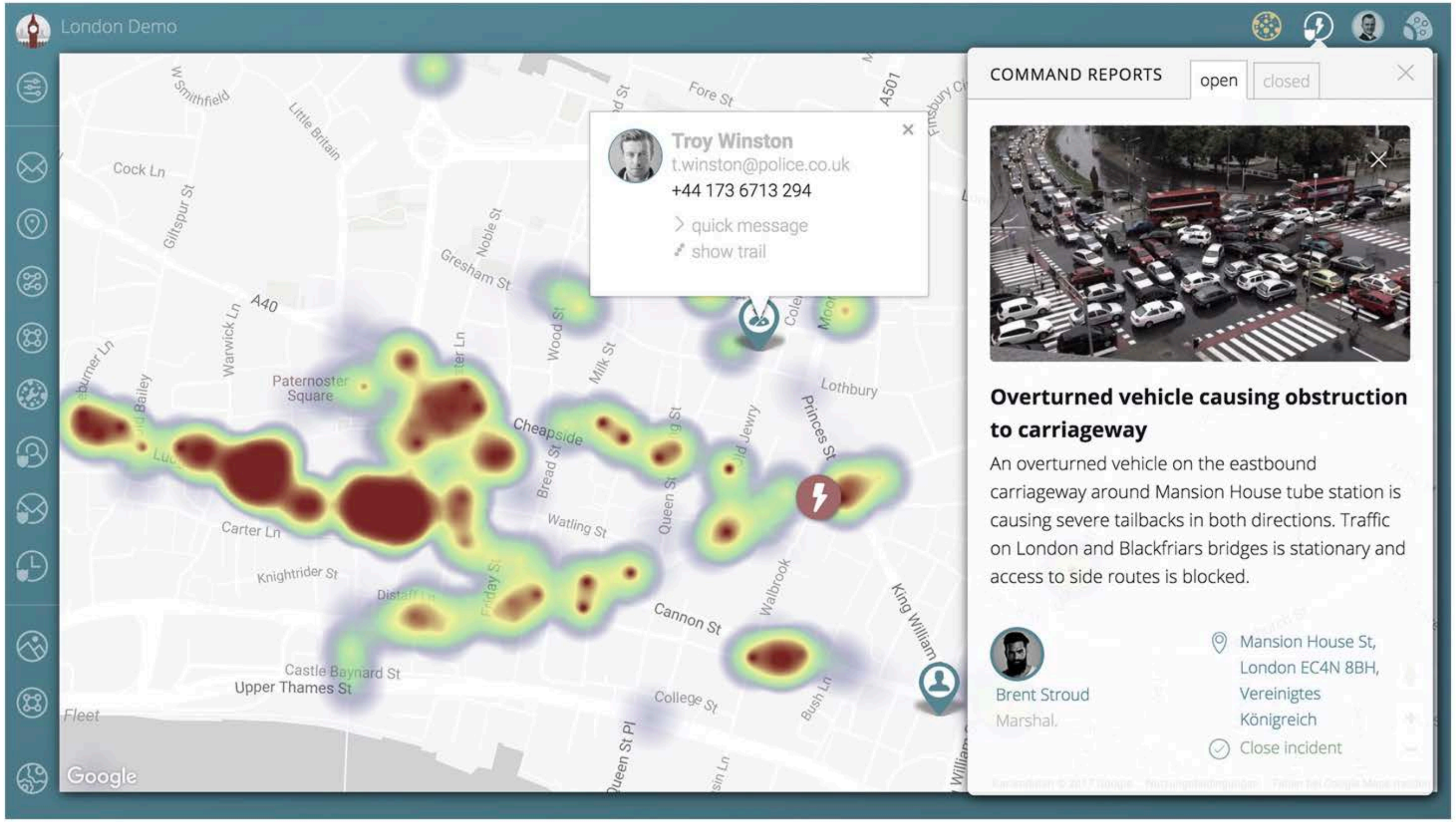


Warning Congestion

Please use alternate station

Welcome to SIS Software GmbH!

We are pioneering digital crowd management. Our innovative Crowd Sensing technology enables the live visualization and prediction of crowd movements. Drawing on our communications and command force management solutions, we create an integrated situation map that allows for precise management of forces and visitors.


London Demo

COMMAND REPORTS open closed

Troy Winston
t.winston@police.co.uk
+44 173 6713 294

› quick message
show trail

Overturned vehicle causing obstruction to carriageway

An overturned vehicle on the eastbound carriageway around Mansion House tube station is causing severe tailbacks in both directions. Traffic on London and Blackfriars bridges is stationary and access to side routes is blocked.

Brent Stroud
Marshal.

Mansion House St,
London EC4N 8BH,
Vereinigtes
Königreich

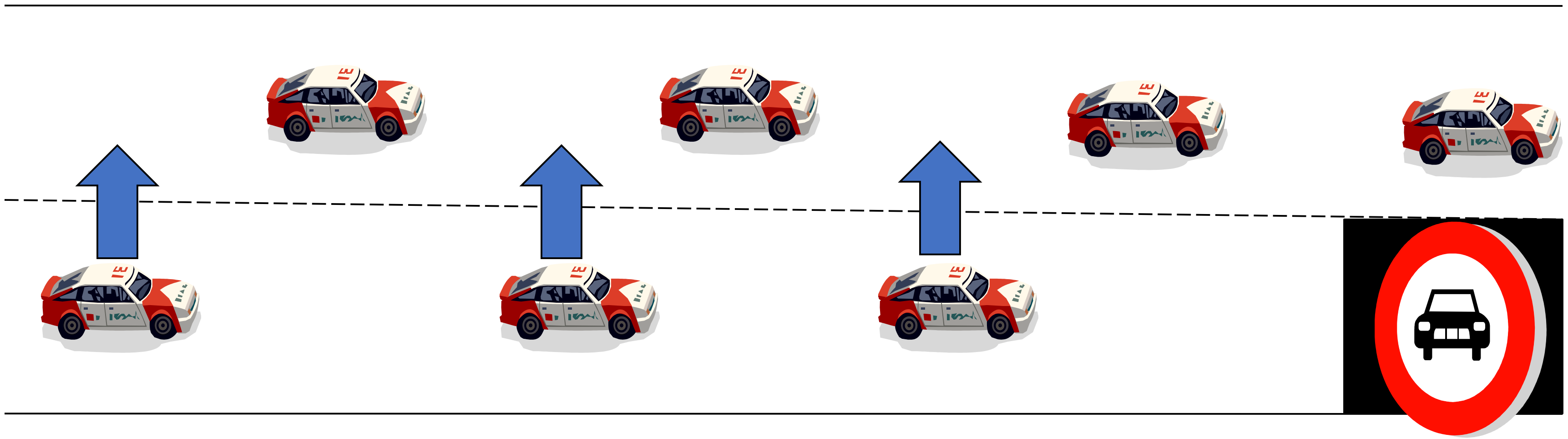
Close incident

Wirz, M., Franke, T., Roggen, D., Mitleton-Kelly, E., Lukowicz, P., & Tröster, G. (2013). Probing crowd density through smartphones in city-scale mass gatherings. *EPJ Data Science*, *2*(1), 5.

Franke, T., Lukowicz, P., & Blanke, U. (2015). Smart crowds in smart cities: real life, city scale deployments of a smartphone based participatory crowd management platform. *Journal of Internet Services and Applications*, *6*(1), 27.

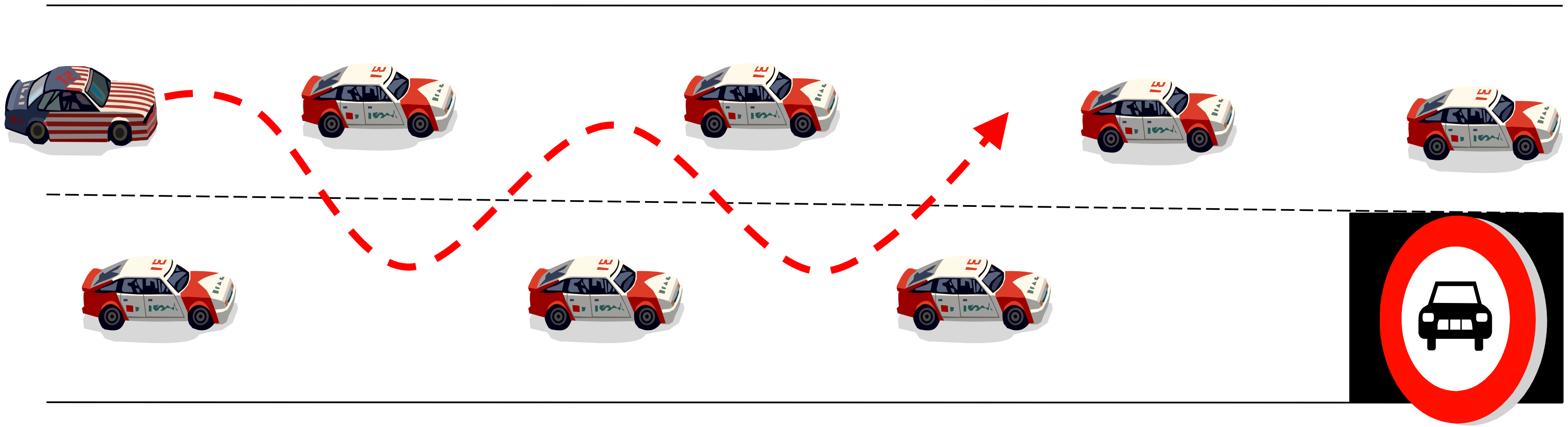
Example: Augmented Traffic

works if everyone is in a “cooperative” state



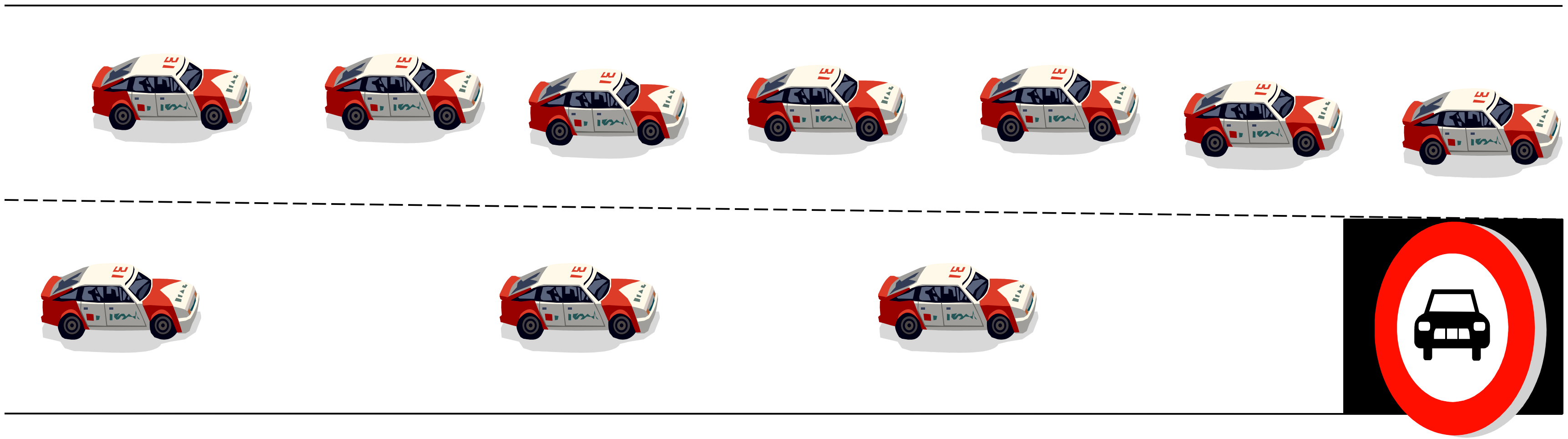
Example: Augmented Traffic

drivers in “aggressive” state may start to take advantage



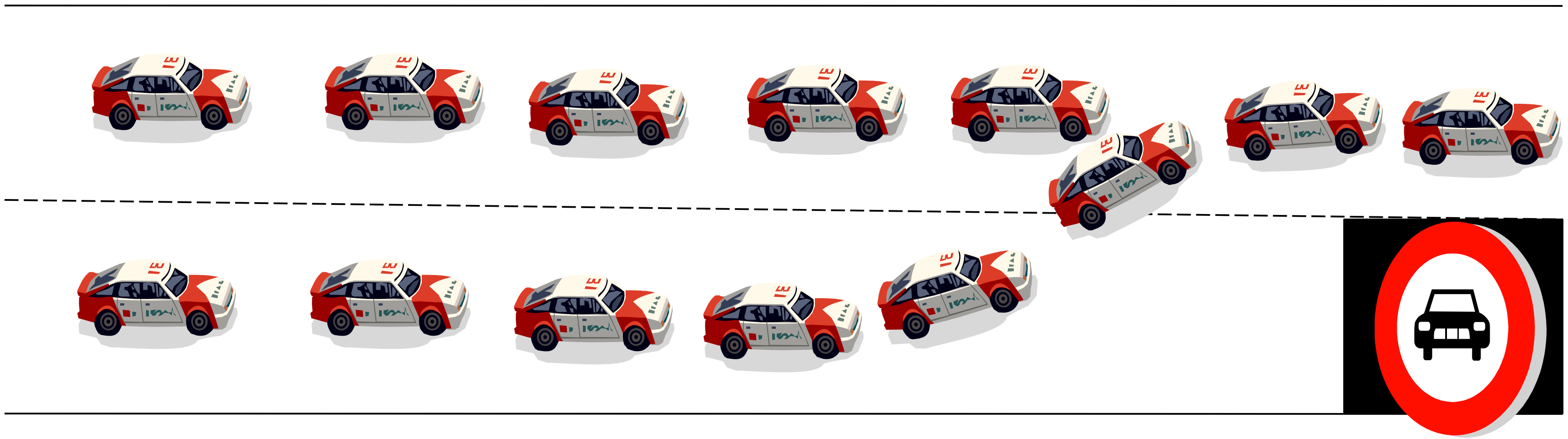
Example: Augmented Traffic

....which causes people to stop being cooperative

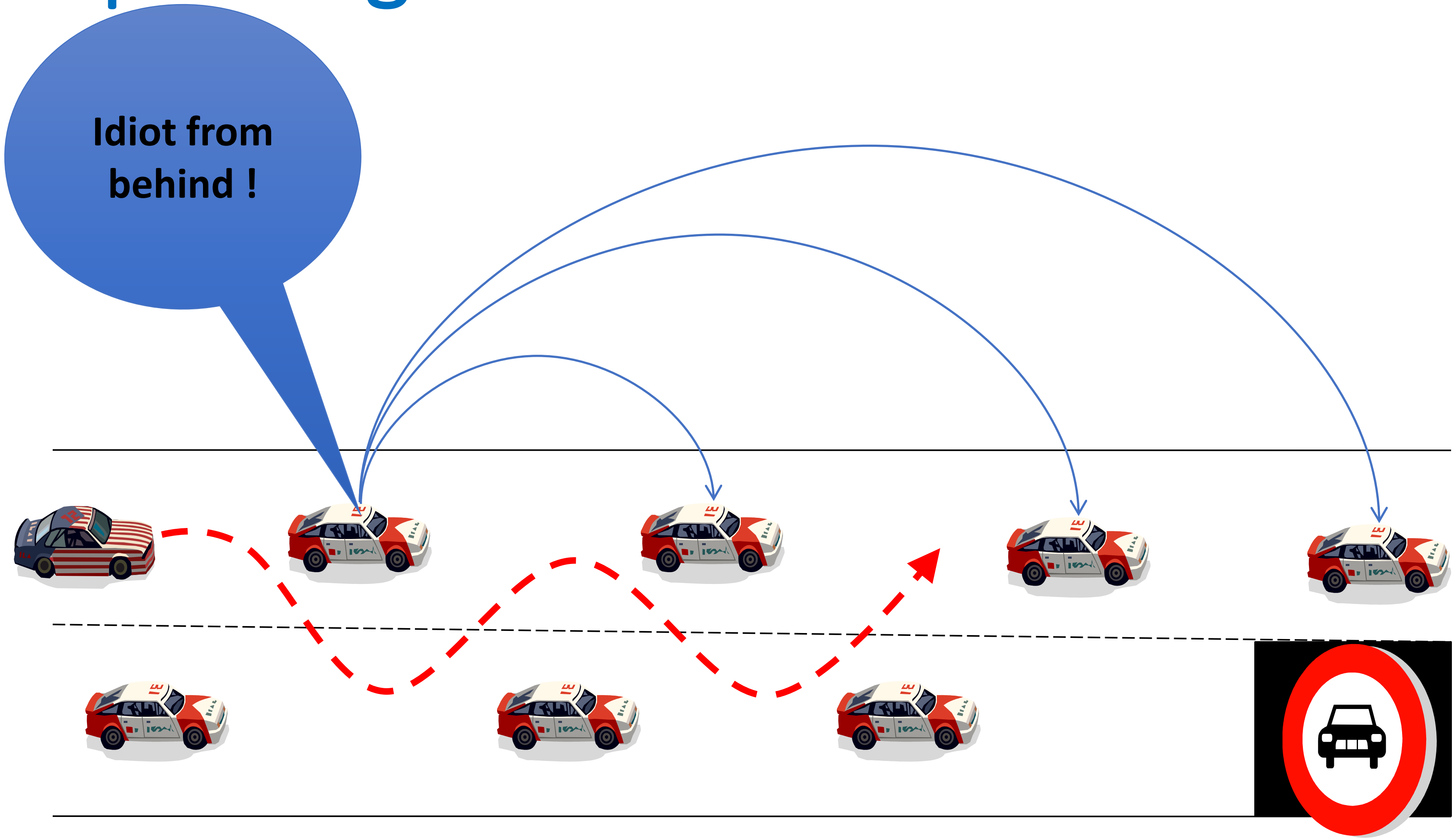


Example: Augmented Traffic

....which causes a traffic jam



Example: Augmented Traffic

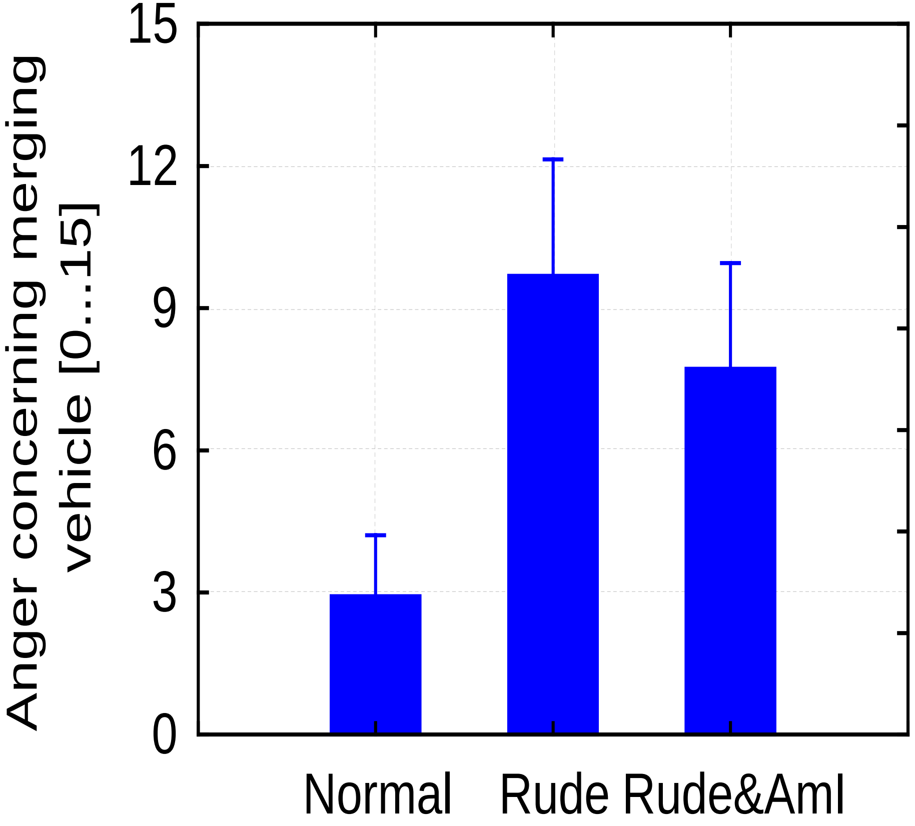


Effect

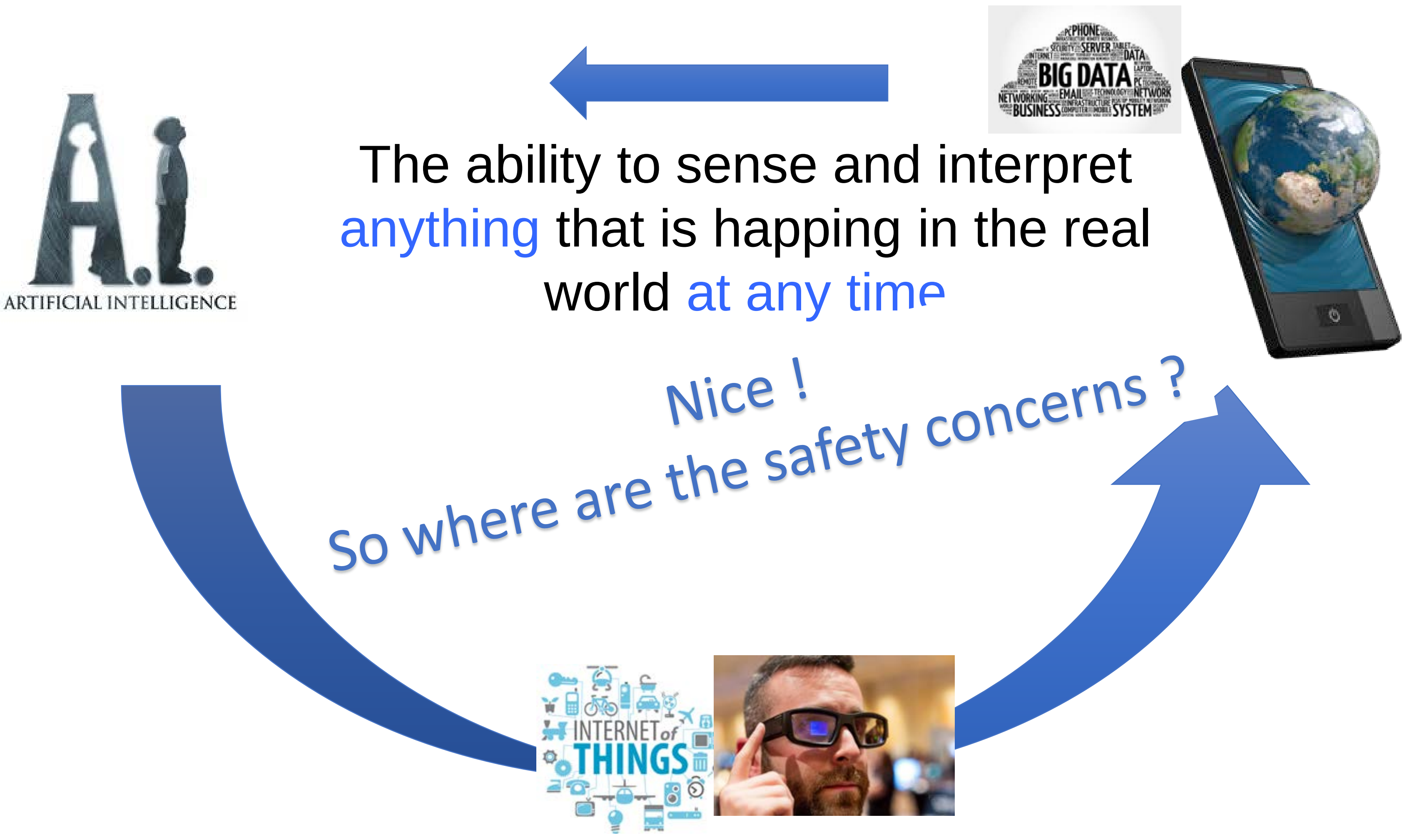
Perspective: Vehicle on right lane of motorway

Anger concerning merging vehicle

F(2, 18)=27.13, p=.001



AI, Digitization and Society



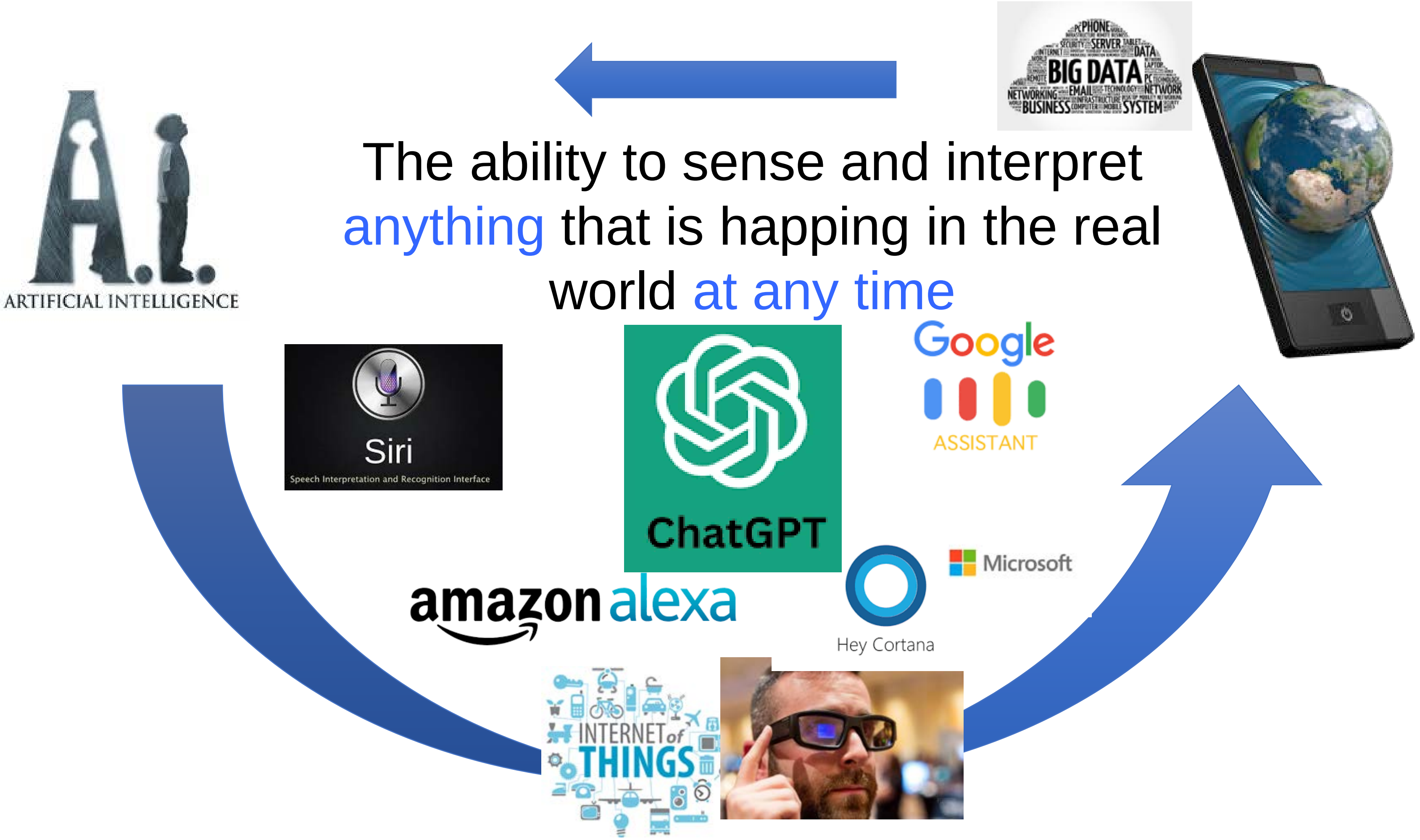
The ability to sense and interpret anything that is happing in the real world at any time

Nice !

So where are the safety concerns ?

The ability to instantly influence any human and any part of the physical world

Monopolization as Safety Risk !

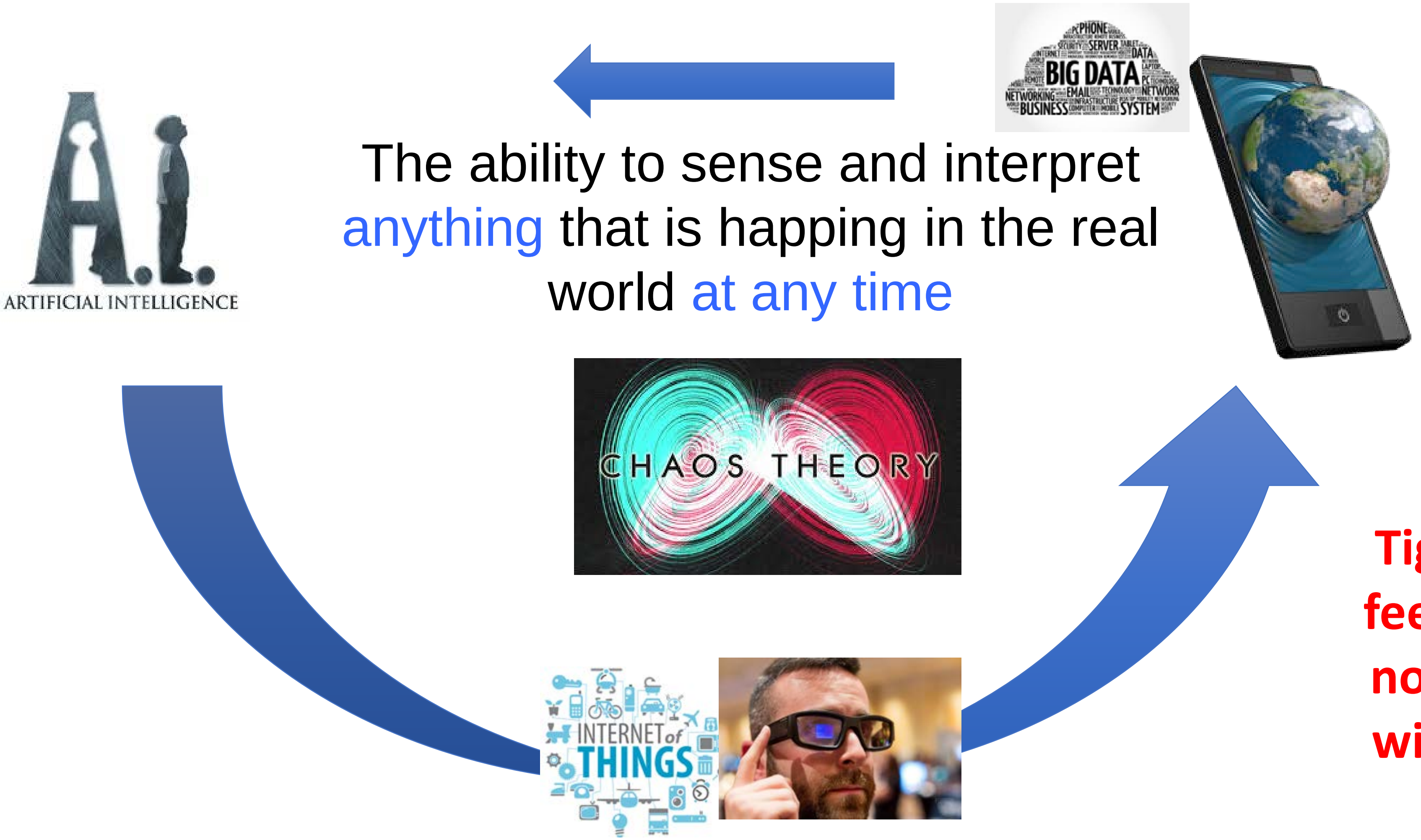


The ability to sense and interpret anything that is happing in the real world at any time

The ability to instantly influence any human and any part of the physical world

Who can be trusted with that much power ?

Emergent/Chaotic Behavior as Safety Risk



The ability to sense and interpret anything that is happing in the real world at any time

The ability to instantly influence any human and any part of the physical world

Tightly coupled, distributed feedback loops often lead to non-linear dynamic systems with emergent and possibly chaotic behavior

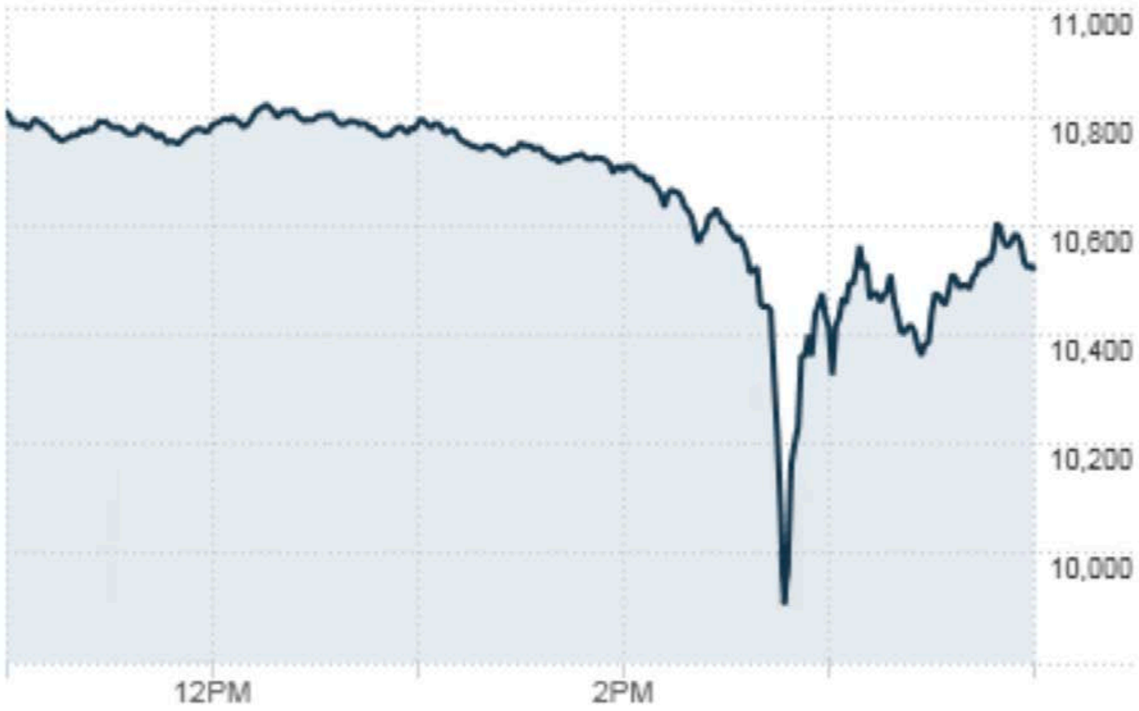
2010 Flash Crash

The Stock Market Crash of March 6, 2010

Written by CFI Team
Published April 10, 2019
Updated January 11, 2023

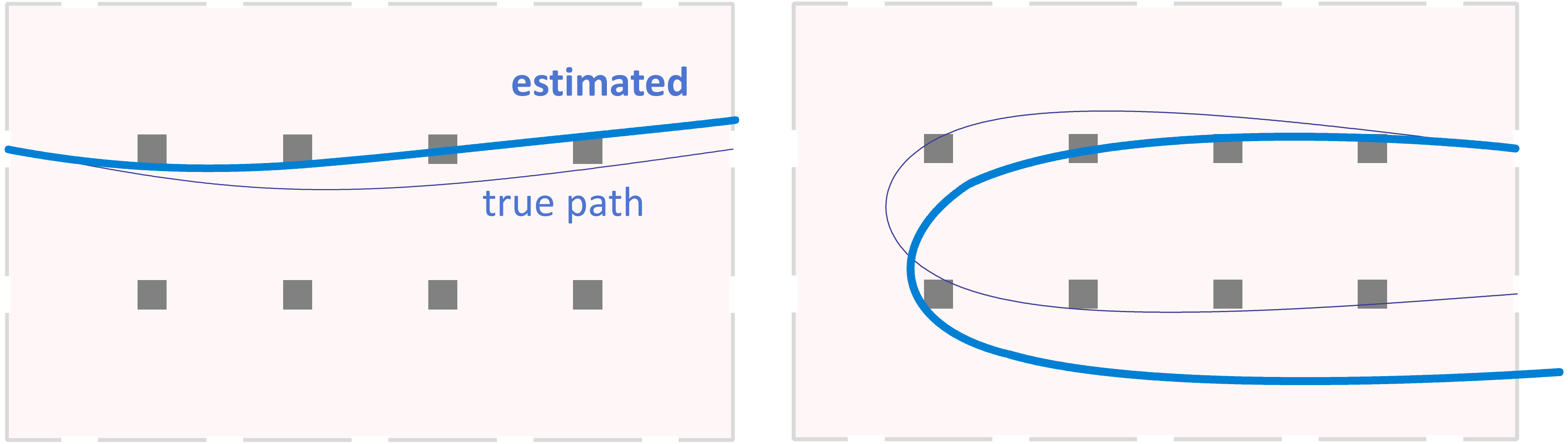
- leading US stock indices, including the Dow Jones Industrial Average, S&P 500, and Nasdaq Composite Index, tumbled and partially rebounded in less than an hour

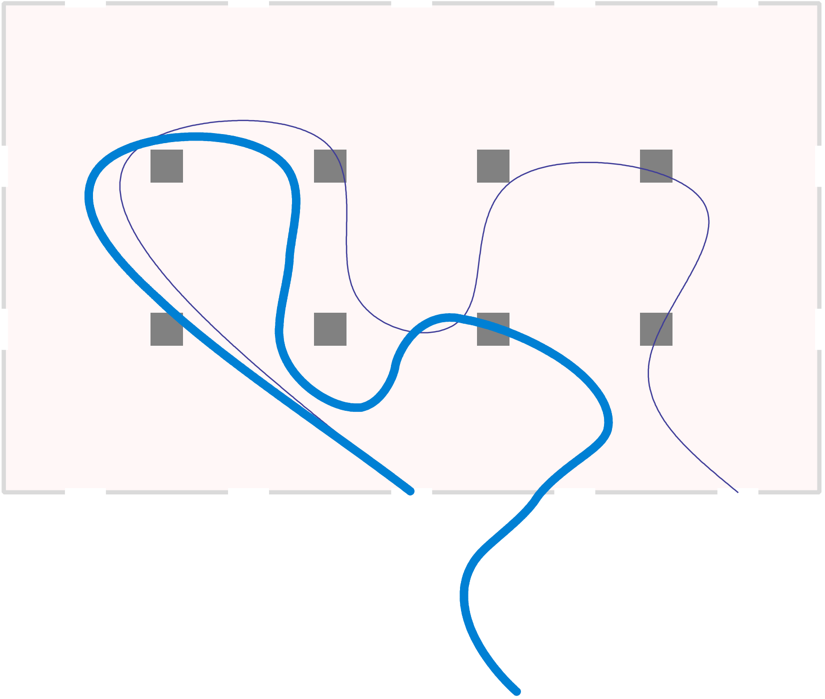
- market indices managed to partially rebound in the same day, the flash crash erased almost $1 trillion in market value

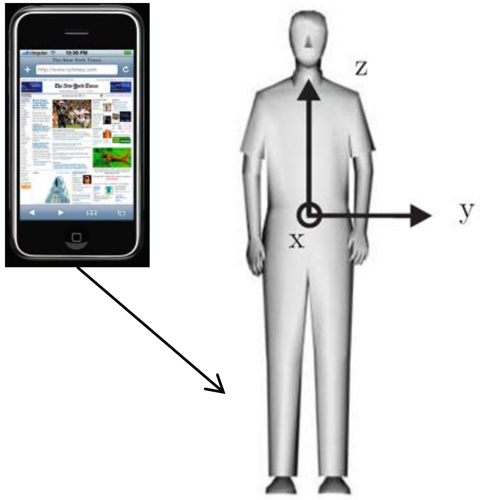


The DJIA on May 6, 2010 (11:00 AM – 4:00 PM EST)

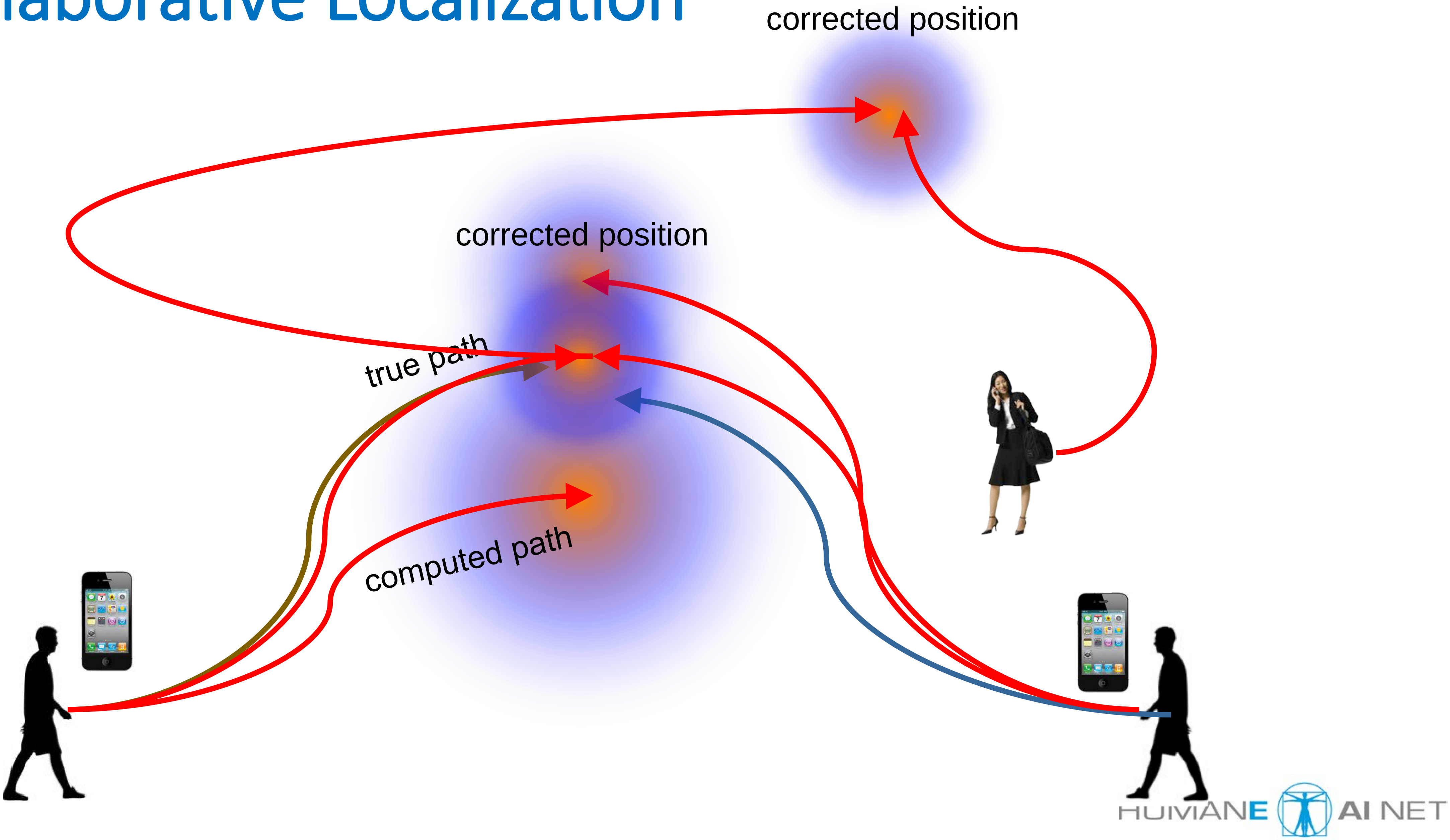
Inertial tracking







Coolaborative Localization



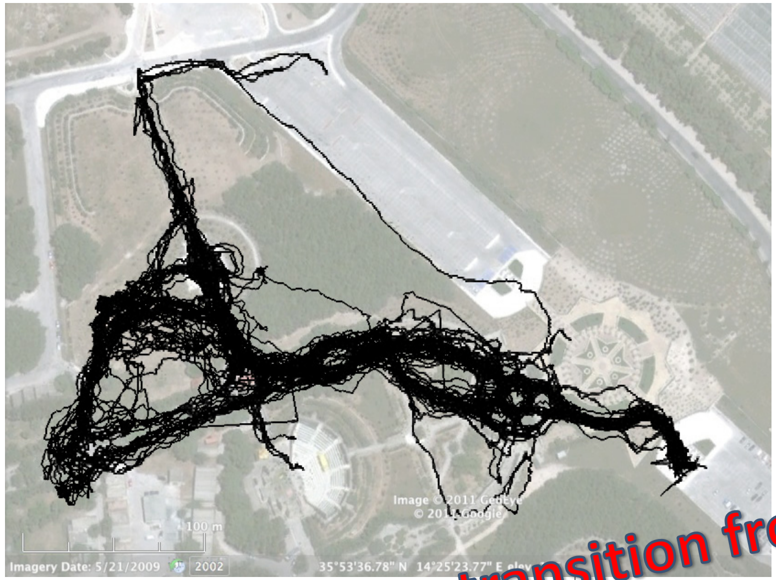
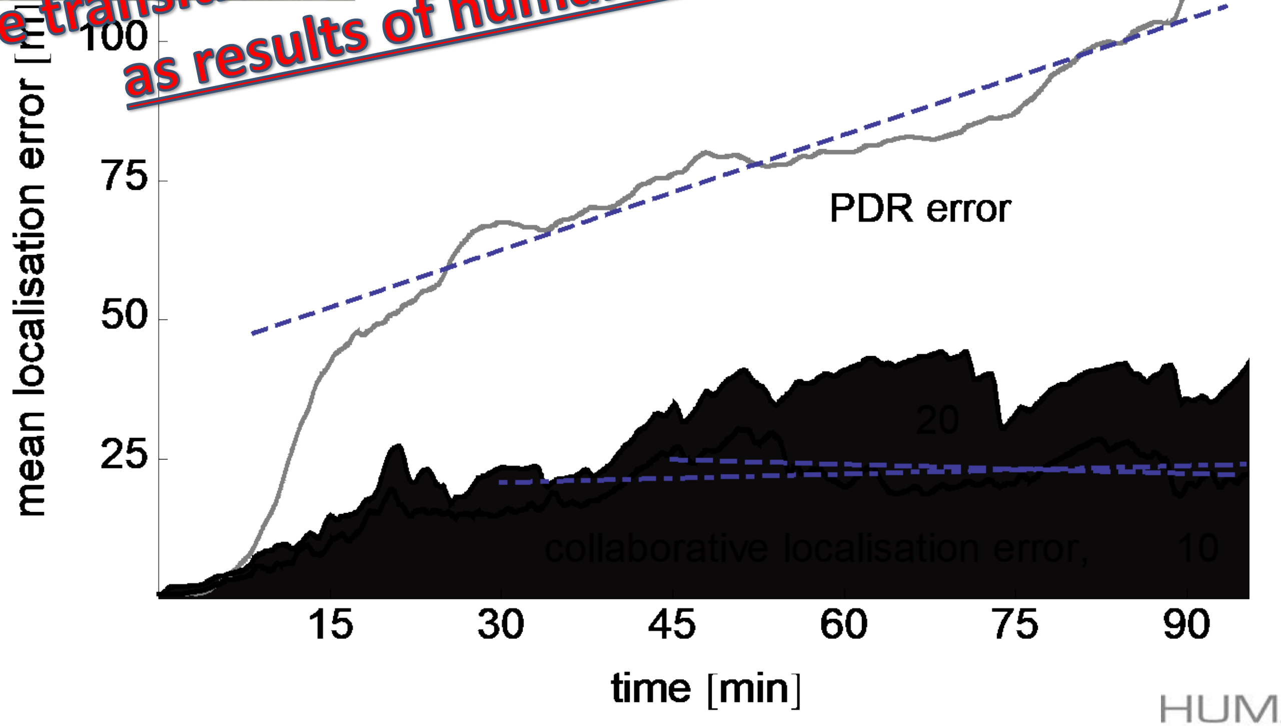


Image © 2011 GeoEye
© 2011 Google
100 m
Imagery Date: 5/21/2009 2002
35°53'36.78" N 14°25'23.77" E elev

Based on 60km of traces from
10 people at a 3 day festival in Malta

phase transition from divergent to bounded error
as results of human encounters



Kloch, Kamil, Paul Lukowicz, and Carl Fischer. "Collaborative PDR localisation with mobile phones." In *2011 15th Annual International Symposium on Wearable Computers*, pp. 37-40. IEEE, 2011.

Theoretical model

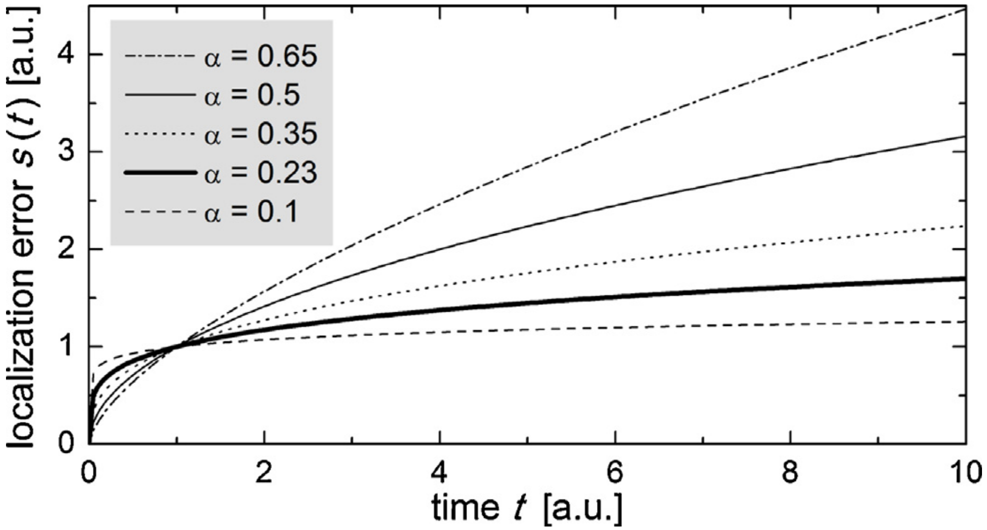
“codification” of individual behaviors and interactions

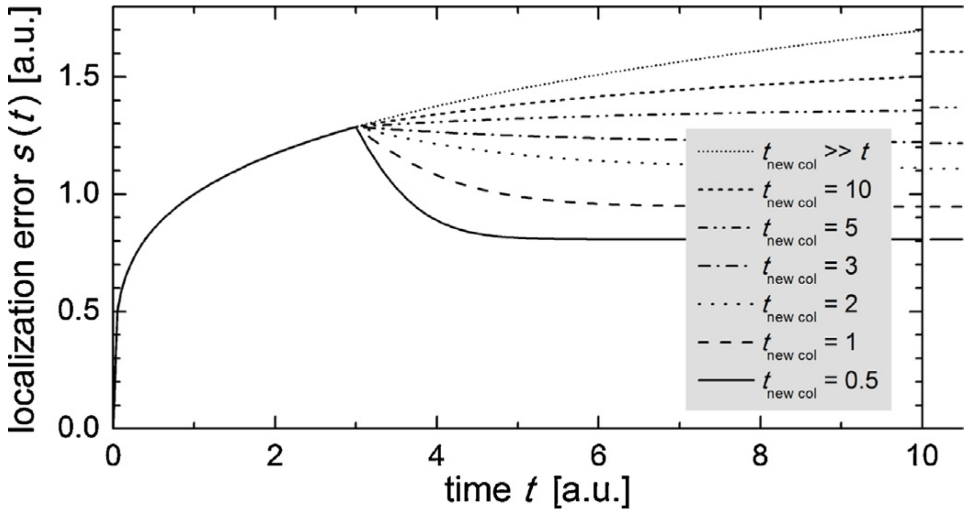
↓

$$t_{\text{new col}} = 1/(\ell v d) = L^2/(Nvd)$$

$$\frac{ds(t)}{dt} = \alpha[s(t)]^{(1-1/\alpha)}$$

$$\frac{ds(t)}{dt} = \alpha[s(t)]^{(1-1/\alpha)} - \left(1 - \frac{1}{\sqrt{2}}\right)\frac{s(t)}{t_{\text{new col}}}$$





individual error

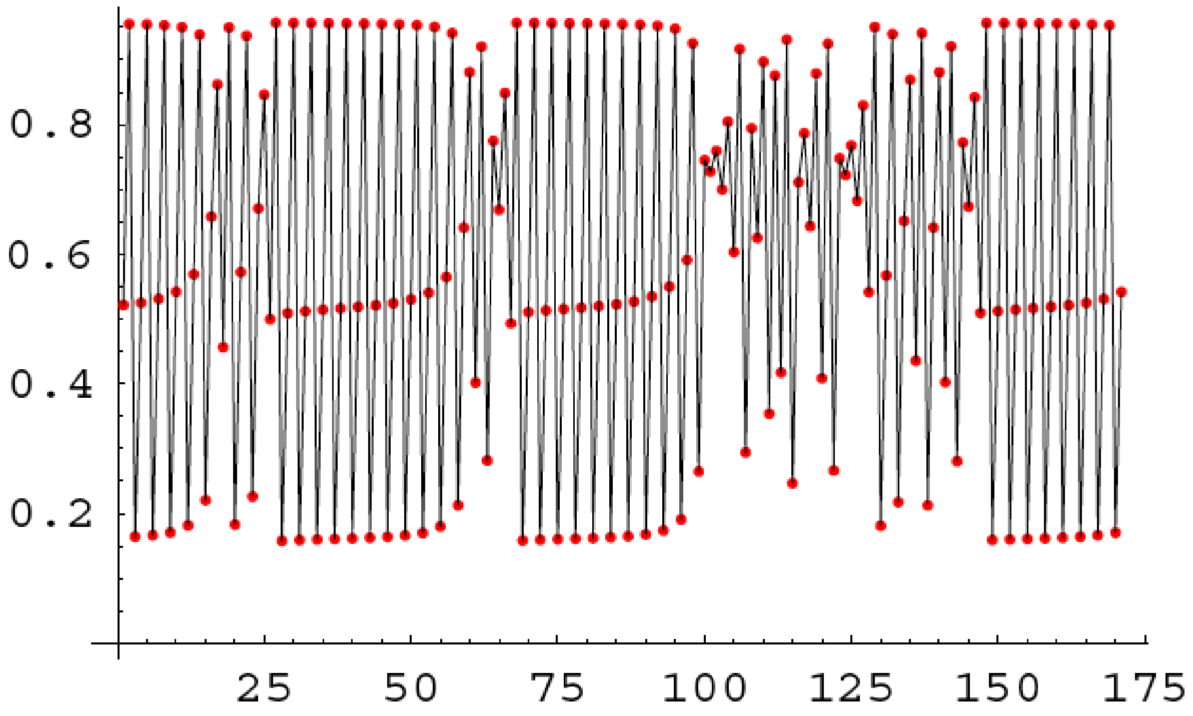
Kampis, G., Kantelhardt, J. W., Kloch, K., & Lukowicz, P. (2015). Analytical and simulation models for collaborative localization. *Journal of Computational Science*, *6*, 1-10.

collaborative error

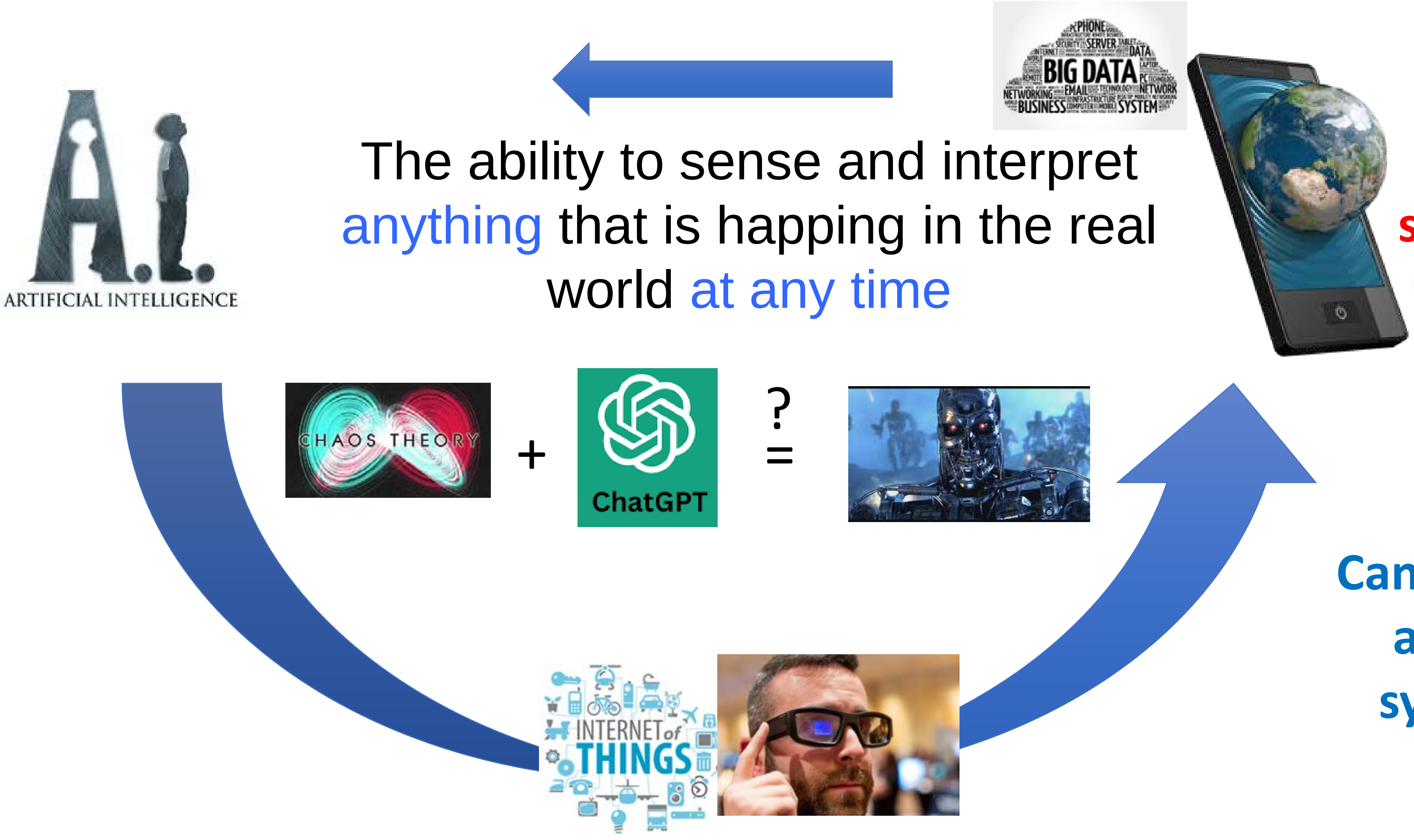
Chaotic Behavior as Safety risk

- Sometimes there can be infinitely many orbits with vastly different behaviors infinitely close to each otherin terms of the control parameter and starting conditions

- Unless we control the starting conditions and relevant paramter with **infinite accuracy**, we have to live with **seemingly random behavior changes**



Emergent/Chaotic Behavior as Safety Risk



The ability to sense and interpret anything that is happing in the real world at any time

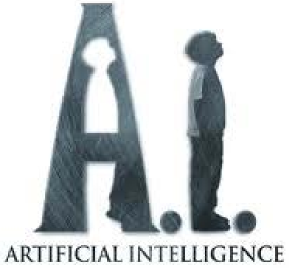
The ability to instantly influence any human and any part of the physical world

Can we rule out that superintelligence “emerges” from networked AIs getting more and more complex ?

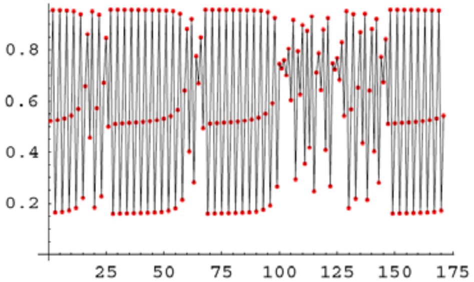
Can we rule that a fleet of evil aliens arrives in the solar system next year to attack earth ?

HUMANE AI NET

Emergent/Chaotic Behavior as Safety Risk



The ability to sense and interpret anything that is happing in the real world at any time



The ability to instantly influence any human and any part of the physical world

Giving too much control of our world to networked AI that may display fundamentally unpredictable behaviour may have catastrophic consequences !

AI Safety aspects

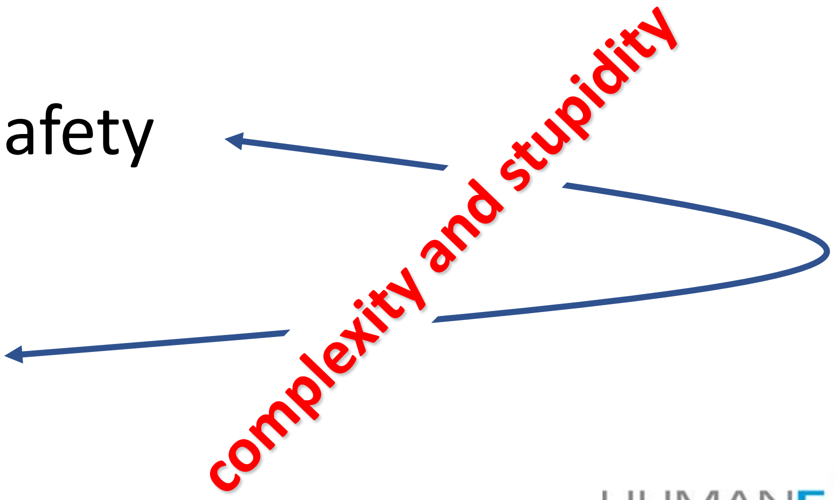
1. Technical safety: “classical” view of robustness, assurance, and specification of in particular ML systems

2. Human Computer Interaction aspects of safety

3. Social/ethical aspects of safety

4. Collective phenomena related aspects of safety

😫 Superintelligence related safety concerns



AI Safety aspects

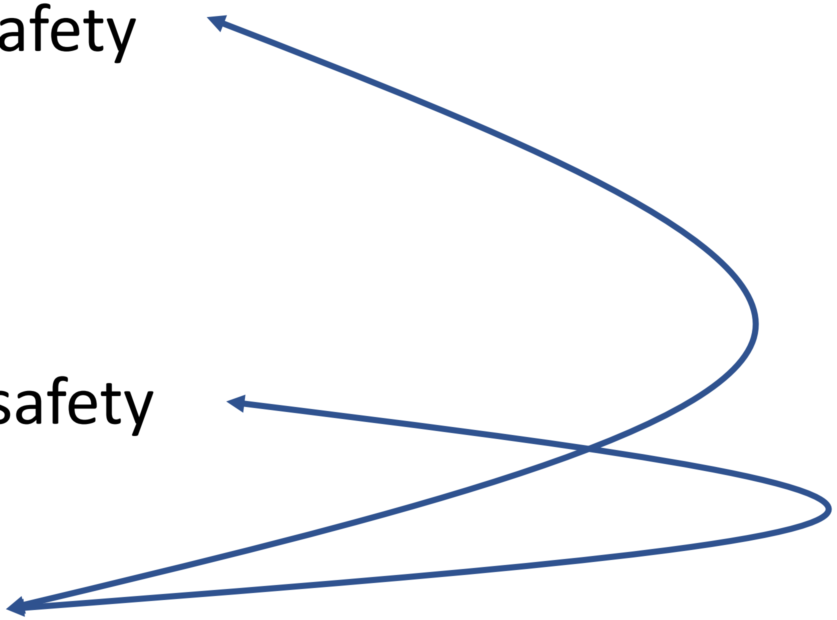
1. Technical safety: “classical” view of robustness, assurance, and specification of in particular ML systems

2. Human Computer Interaction aspects of safety

3. Social/ethical aspects of safety

4. Collective phenomena related aspects of safety

Superintelligence related safety concerns



AI being too intelligent is unlikely to kill us, a combination of artificial and natural stupidity with complexity may however do great, even existential harm https://www.humane-ai.eu