DONDERS INSTITUTE

Professionele verantwoordelijkheid in het werken met Artificiële Intelligentie

Pim Haselager

Dpt. of AI, Donders Institute for Brain, Cognition, and Behaviour
Radboud University, Nijmegen, The Netherlands
pim.haselager@donders.ru.nl

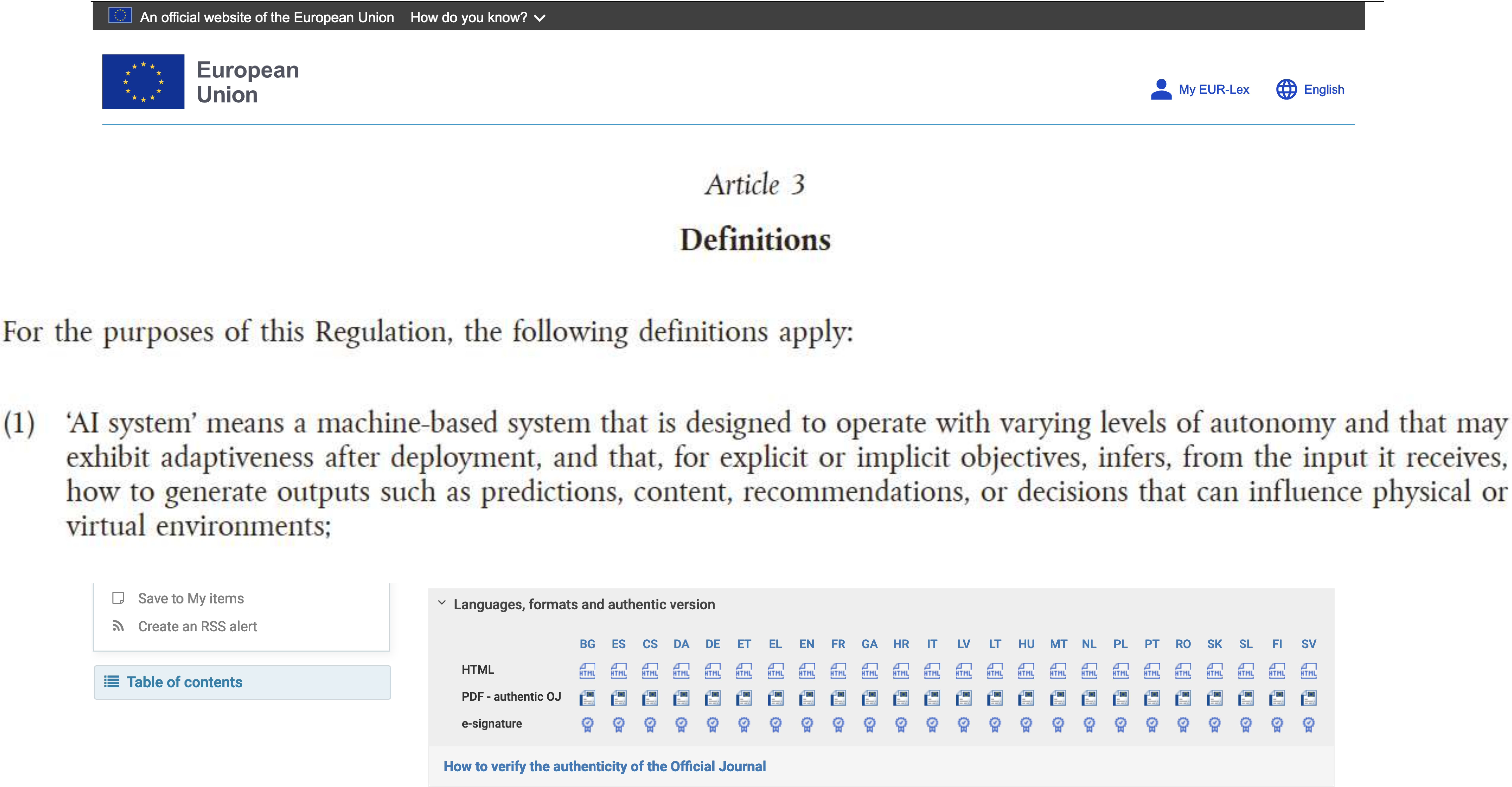
SI AC
Societal Implications of AI & CNS

Radboud University Radboudumc

Disclosures

None

AI Act



An official website of the European Union How do you know?

European Union

My EUR-Lex English

Article 3

Definitions

For the purposes of this Regulation, the following definitions apply:

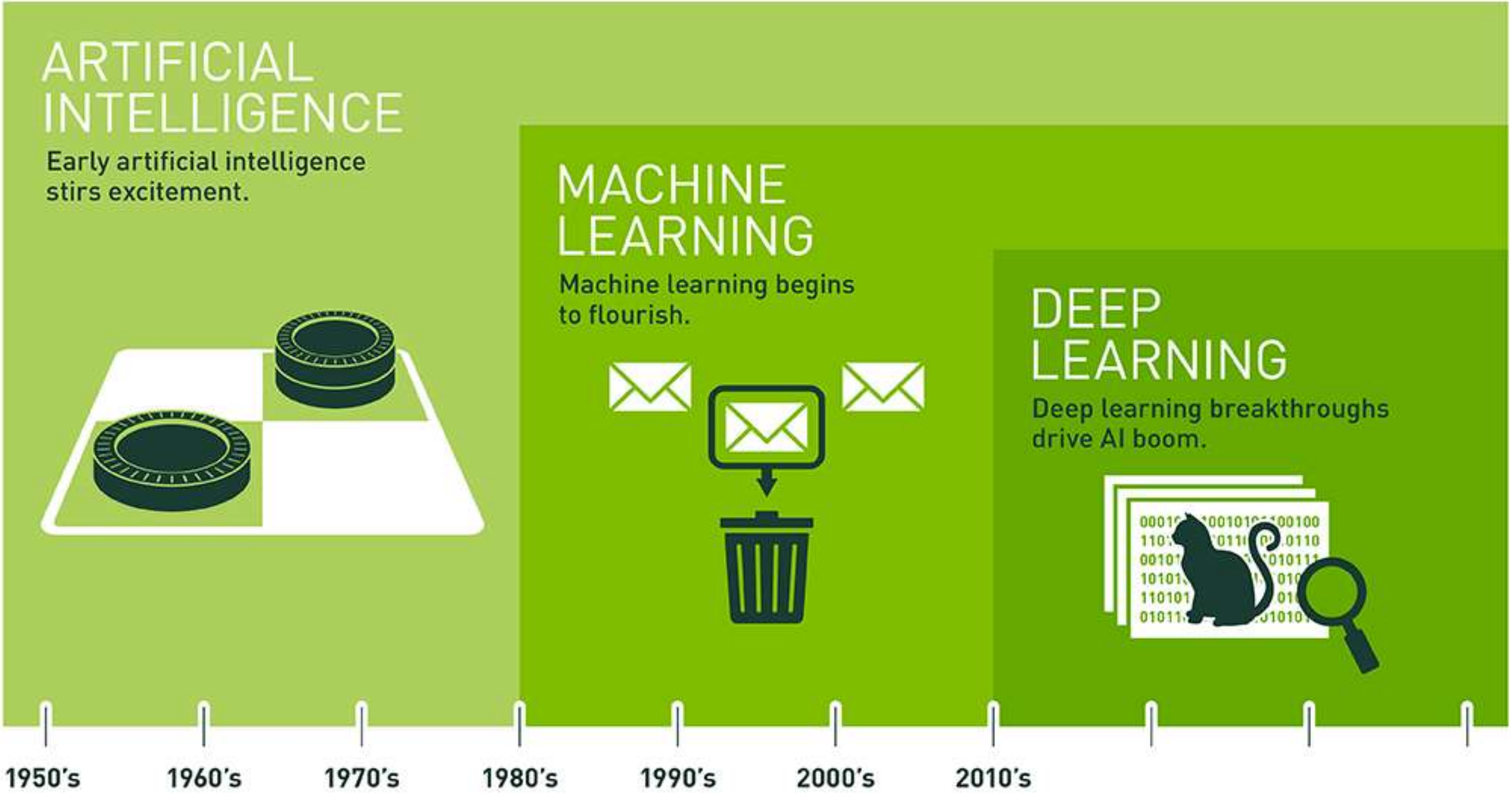
(1) 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;

Languages, formats and authentic version

	BG	ES	CS	DA	DE	ET	EL	EN	FR	GA	HR	IT	LV	LT	HU	MT	NL	PL	PT	RO	SK	SL	FI	SV
HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML	HTML
PDF - authentic OJ	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄	📄
e-signature	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅	🏅

How to verify the authenticity of the Official Journal

https://eur-lex.europa.eu/eli/reg/2024/1689/oj



ARTIFICIAL INTELLIGENCE

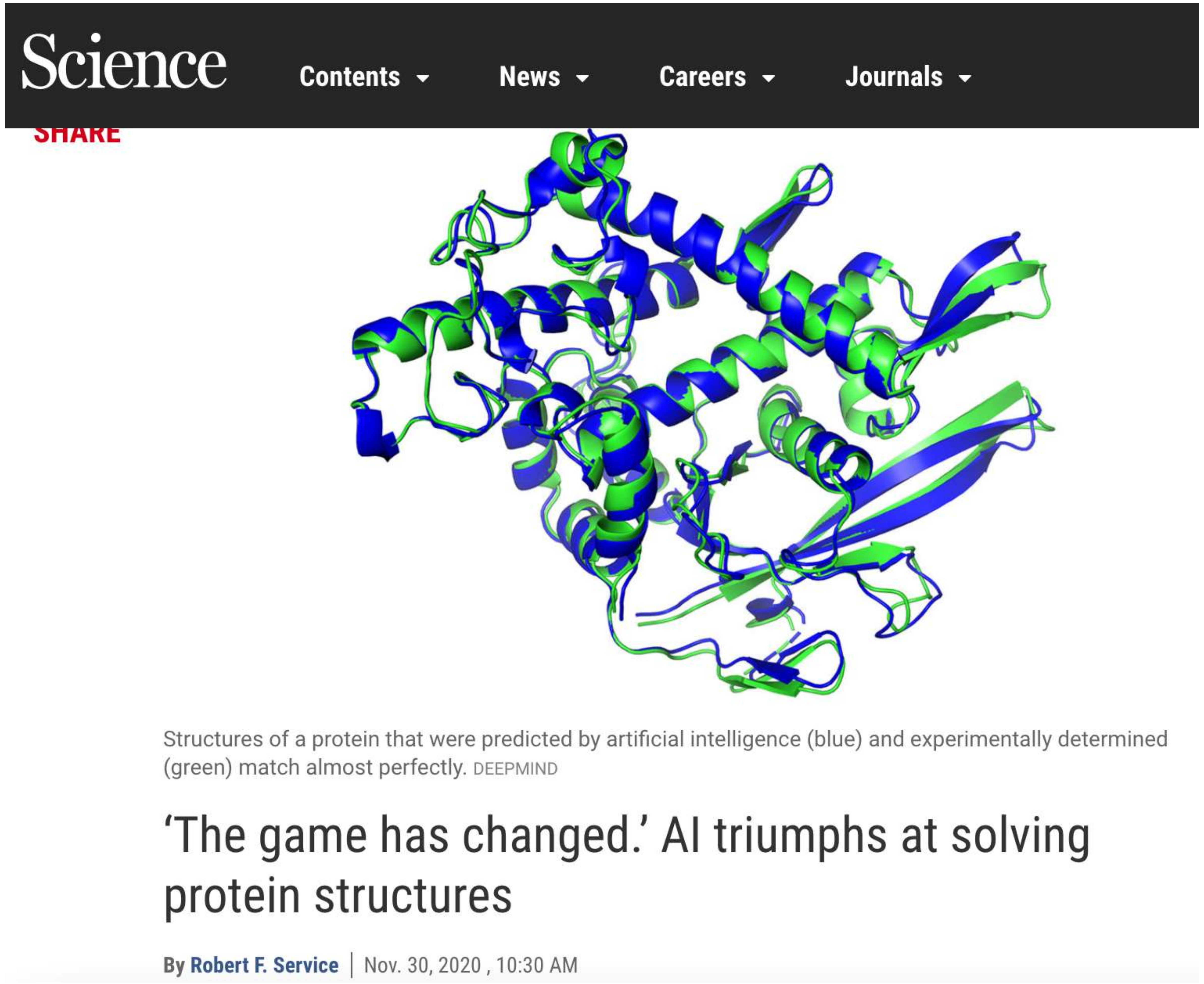
Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

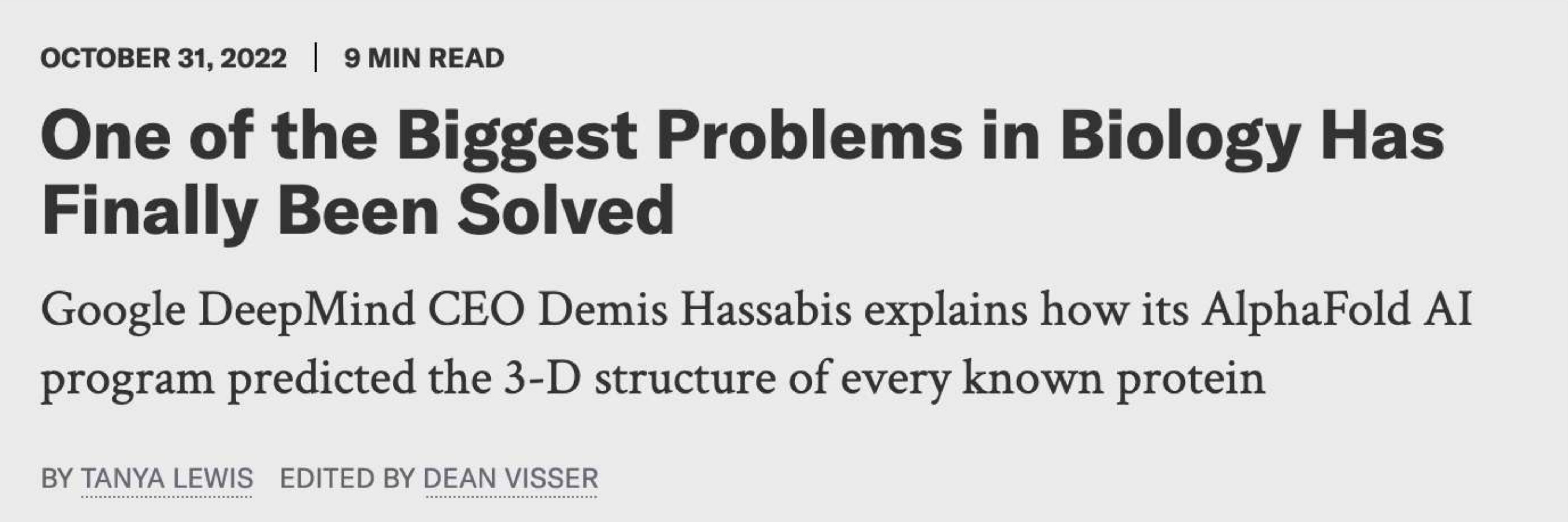
Deep learning breakthroughs drive AI boom.



Structures of a protein that were predicted by artificial intelligence (blue) and experimentally determined (green) match almost perfectly. DEEPMIND

‘The game has changed.’ AI triumphs at solving protein structures

By **Robert F. Service** | Nov. 30, 2020 , 10:30 AM

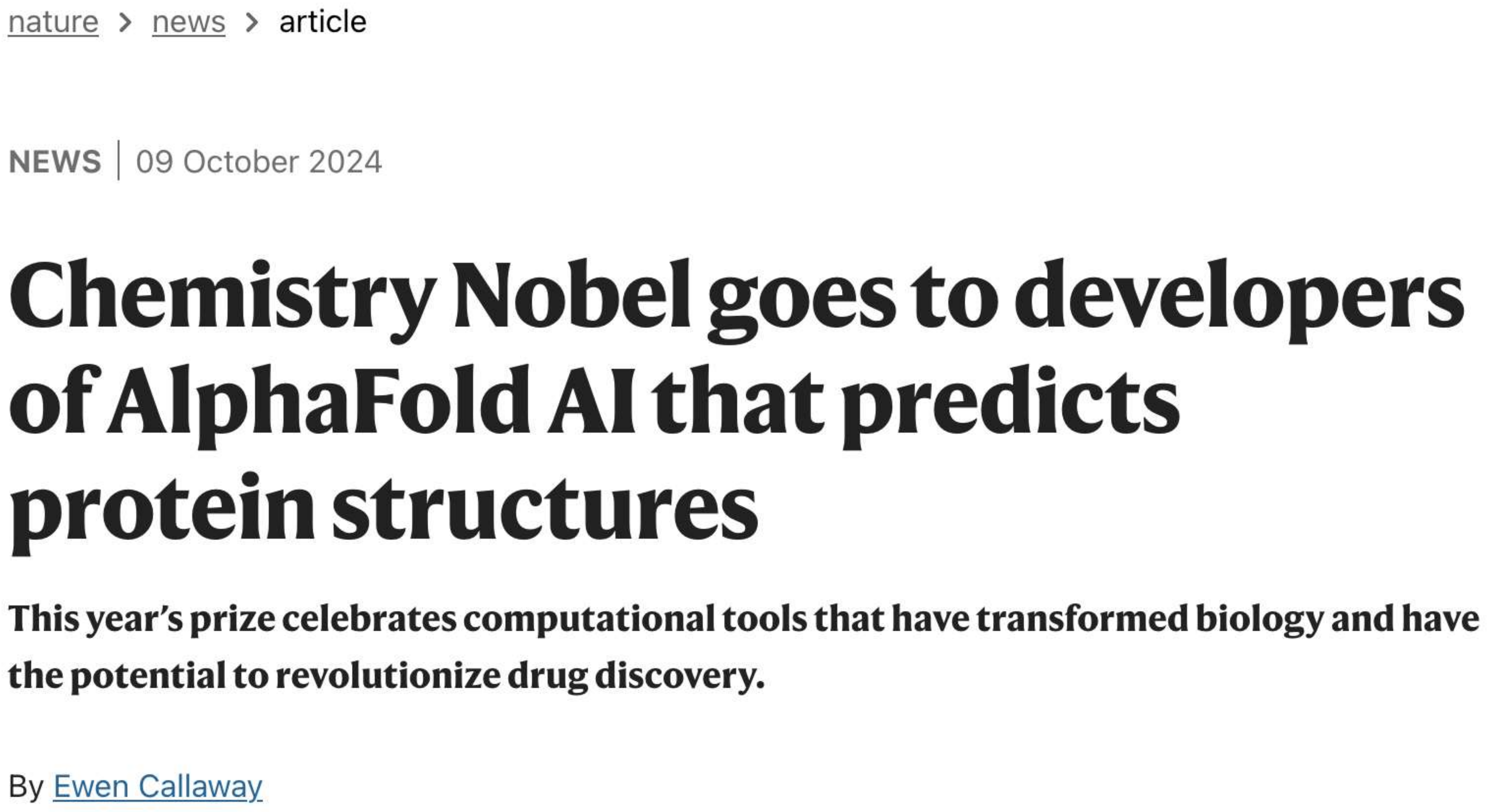


OCTOBER 31, 2022 | 9 MIN READ

One of the Biggest Problems in Biology Has Finally Been Solved

Google DeepMind CEO Demis Hassabis explains how its AlphaFold AI program predicted the 3-D structure of every known protein

BY TANYA LEWIS EDITED BY DEAN VISSER



nature > news > article

NEWS | 09 October 2024

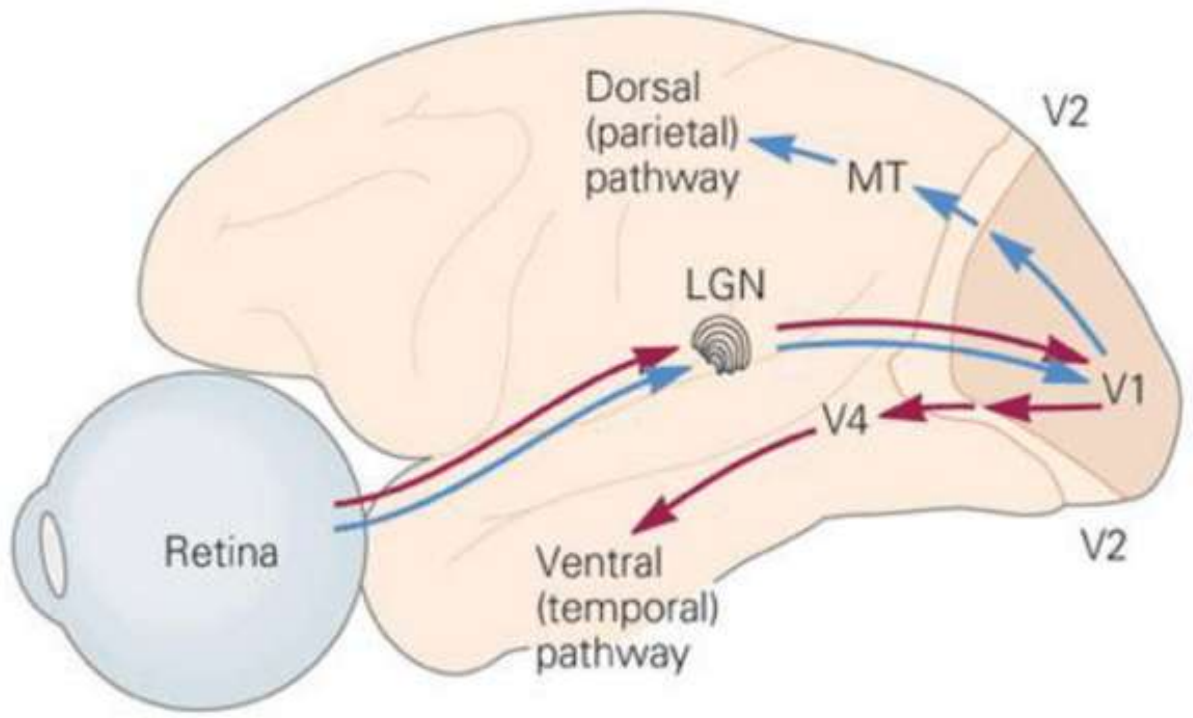
Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

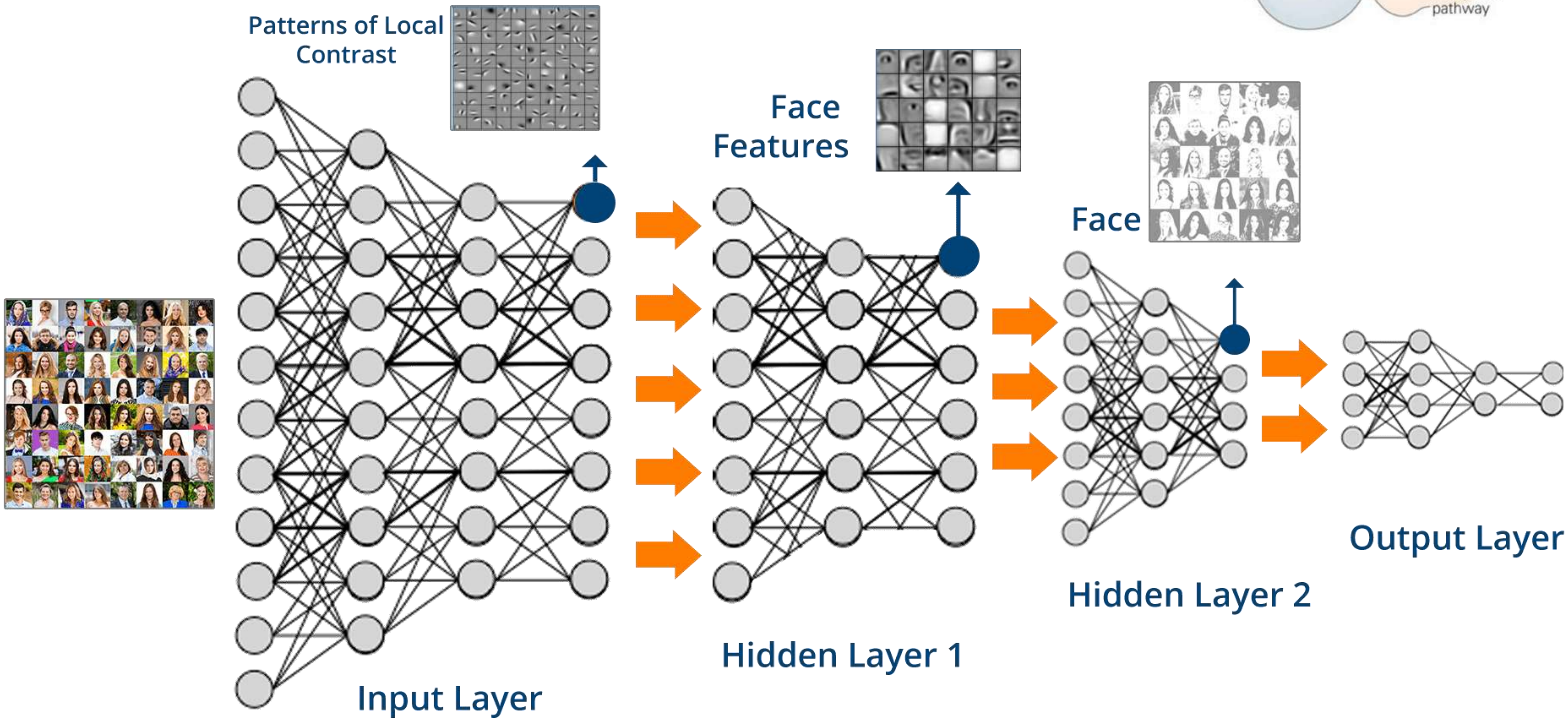
This year’s prize celebrates computational tools that have transformed biology and have the potential to revolutionize drug discovery.

By Ewen Callaway

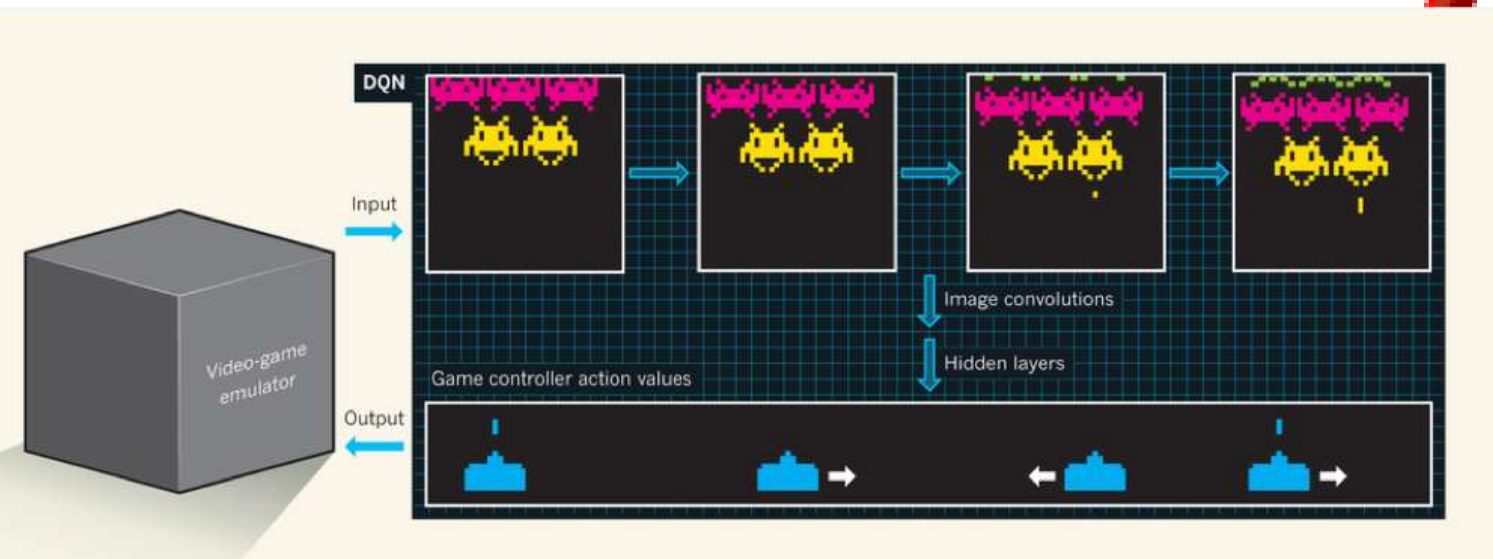
https://www.science.org/doi/10.1126/science.370.6521.1144
https://www.scientificamerican.com/article/one-of-the-biggest-problems-in-biology-has-finally-been-solved/
https://www.nature.com/articles/d41586-024-03214-7

Deep learning neural networks

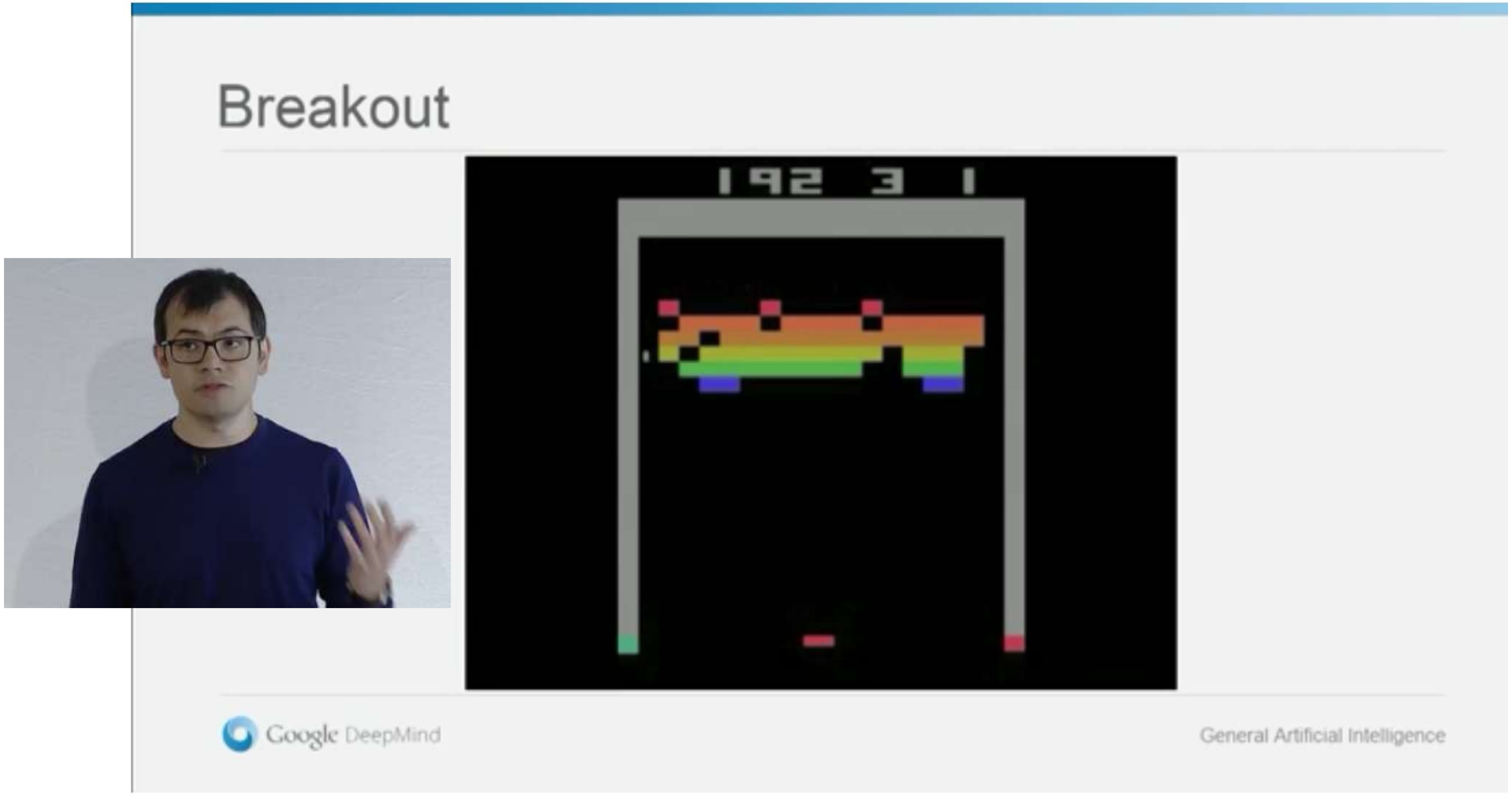




Computergames leren spelen



Deep learning ziet soms meer (anders) dan mens



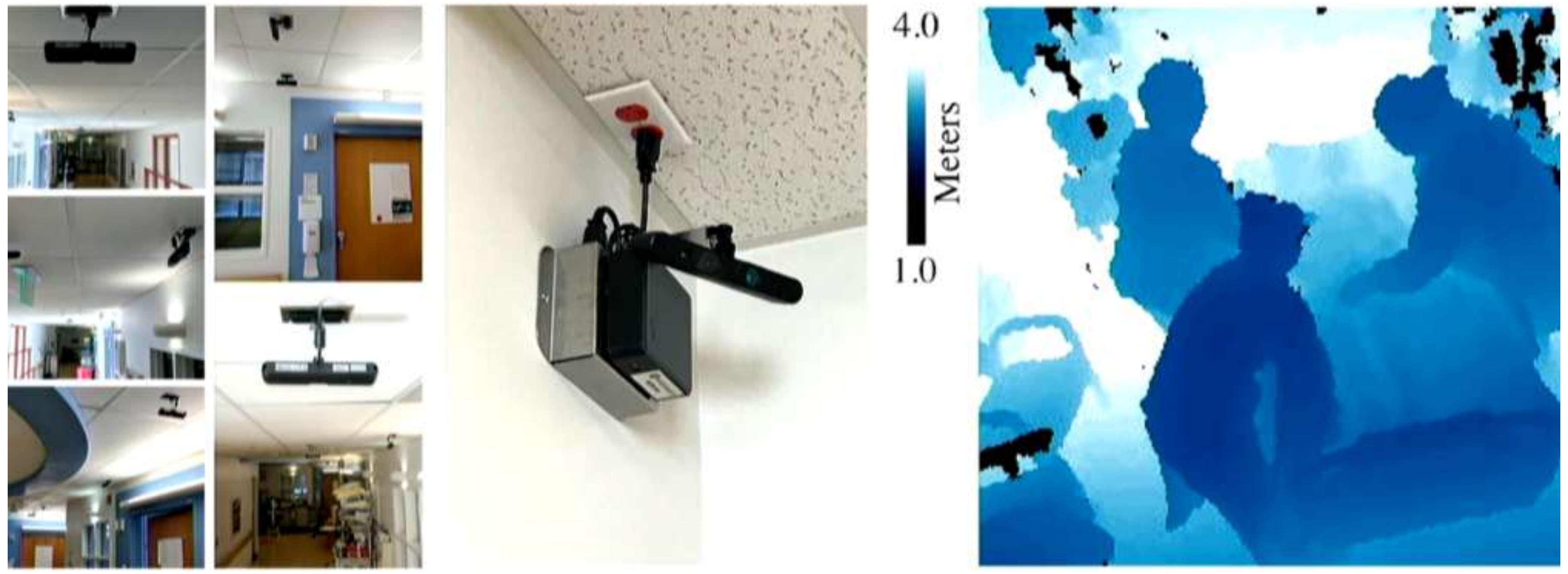
Breakout

“This is something their own creation taught them”

Hospital monitoring

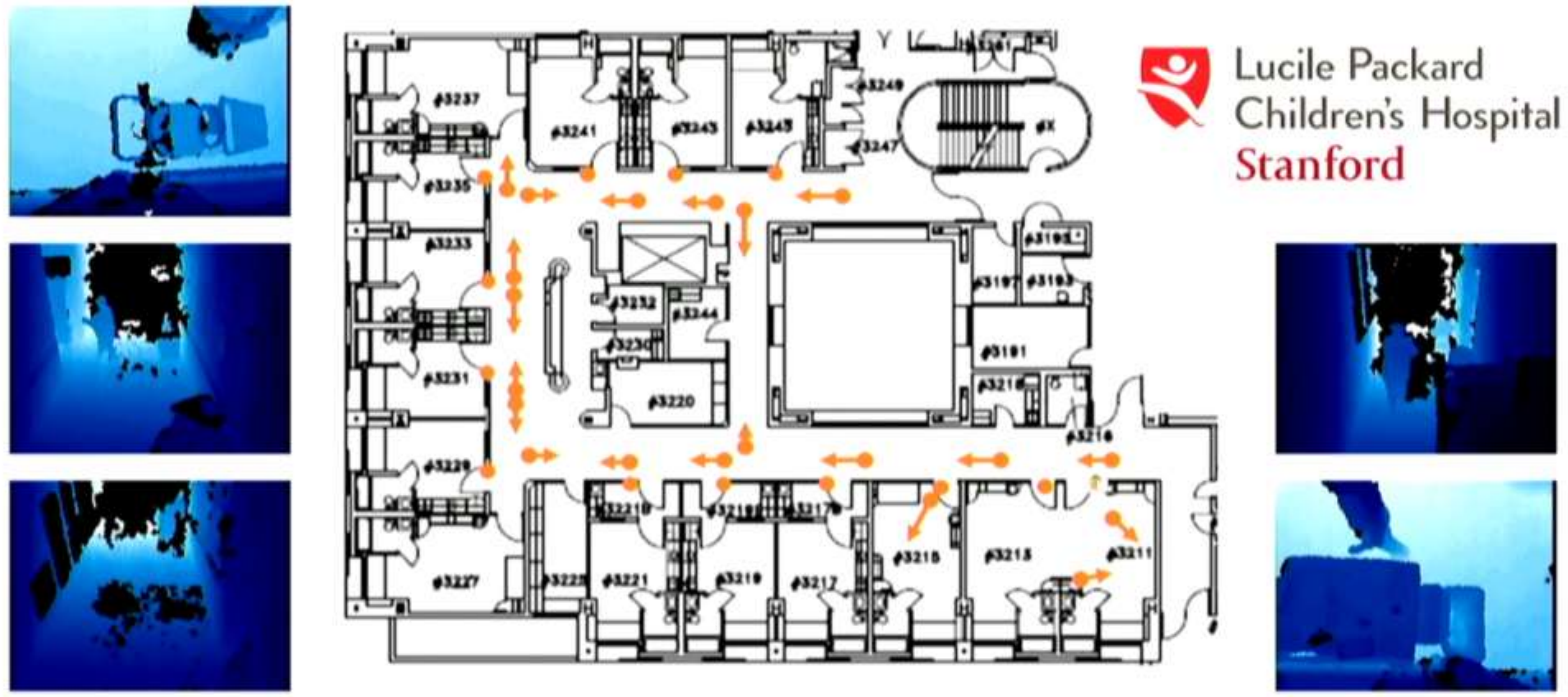
Serena Yeung
https://www.youtube.com/watch?v=B94X6LwHYxI

Privacy-preserving depth data



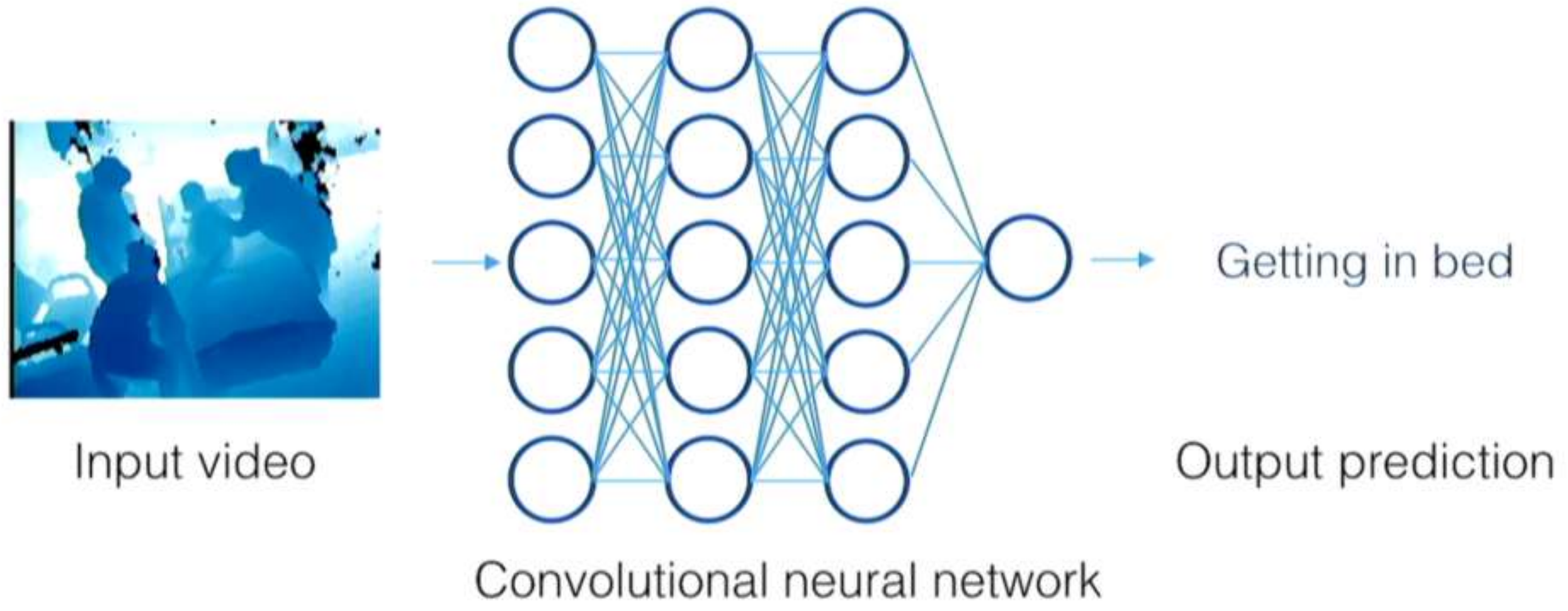
Yeung,Downing,Fei-Fei,Milstein. New England Journal of Medicine (NEJM), 2018.

Depth streams across a unit



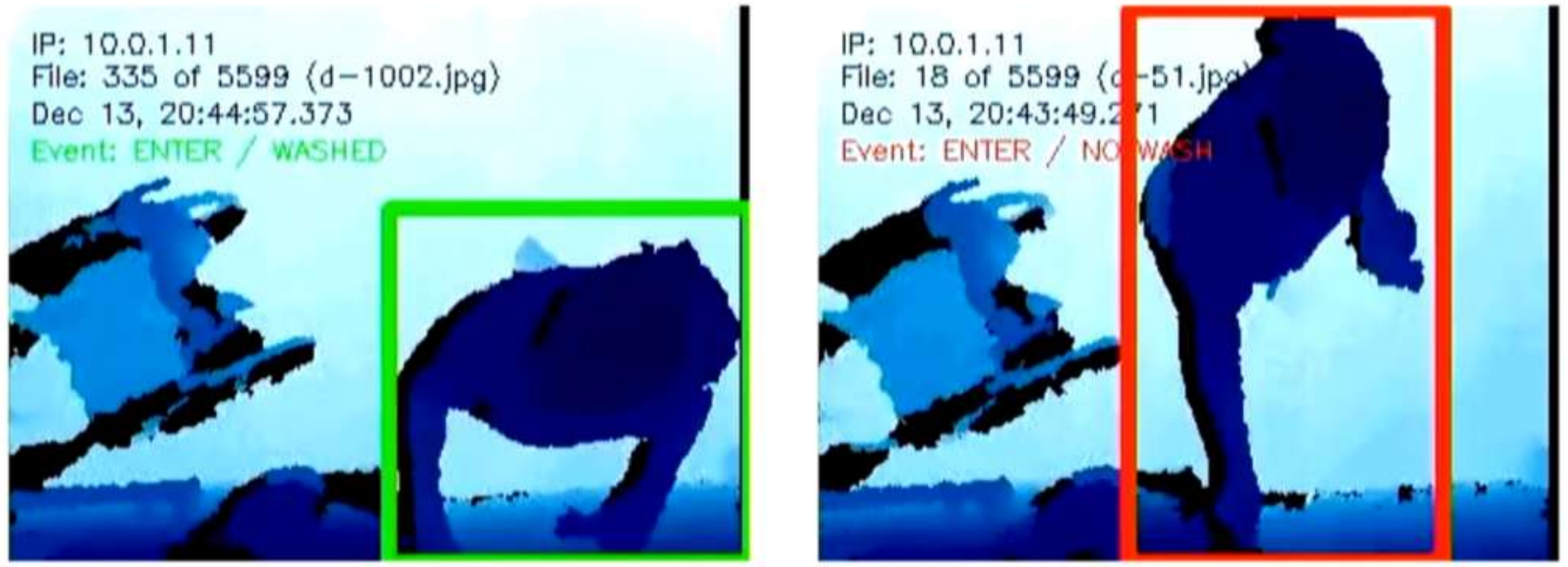
Yeung,Downing,Fei-Fei,Milstein. New England Journal of Medicine (NEJM), 2018.

Deep learning algorithms for automated interpretation of human activity in video



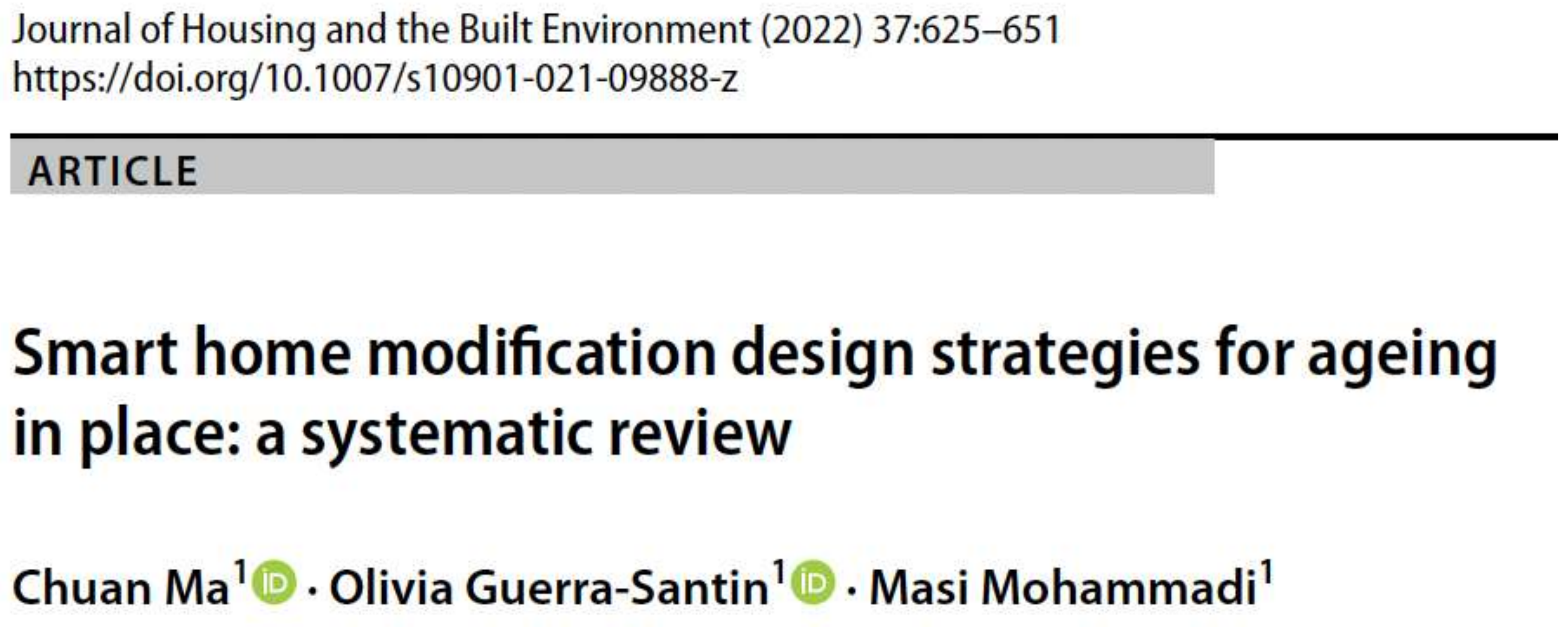
Yeung,Russakovsky,Mori,Fei-Fei. Computer Vision and Pattern Recognition (CVPR), 2016.
Yeung,Russakovsky,Mori,Fei-Fei. International Journal of Computer Vision (IJCV), 2017.
Yeung,Ramanathan,Russakovsky,Shen,Mori,Fei-Fei. Computer Vision and Pattern Recognition (CVPR), 2017.

AI recognition of performing hand hygiene



https://www.youtube.com/watch?v=B94X6LwHYxI

Smart homes for care giving

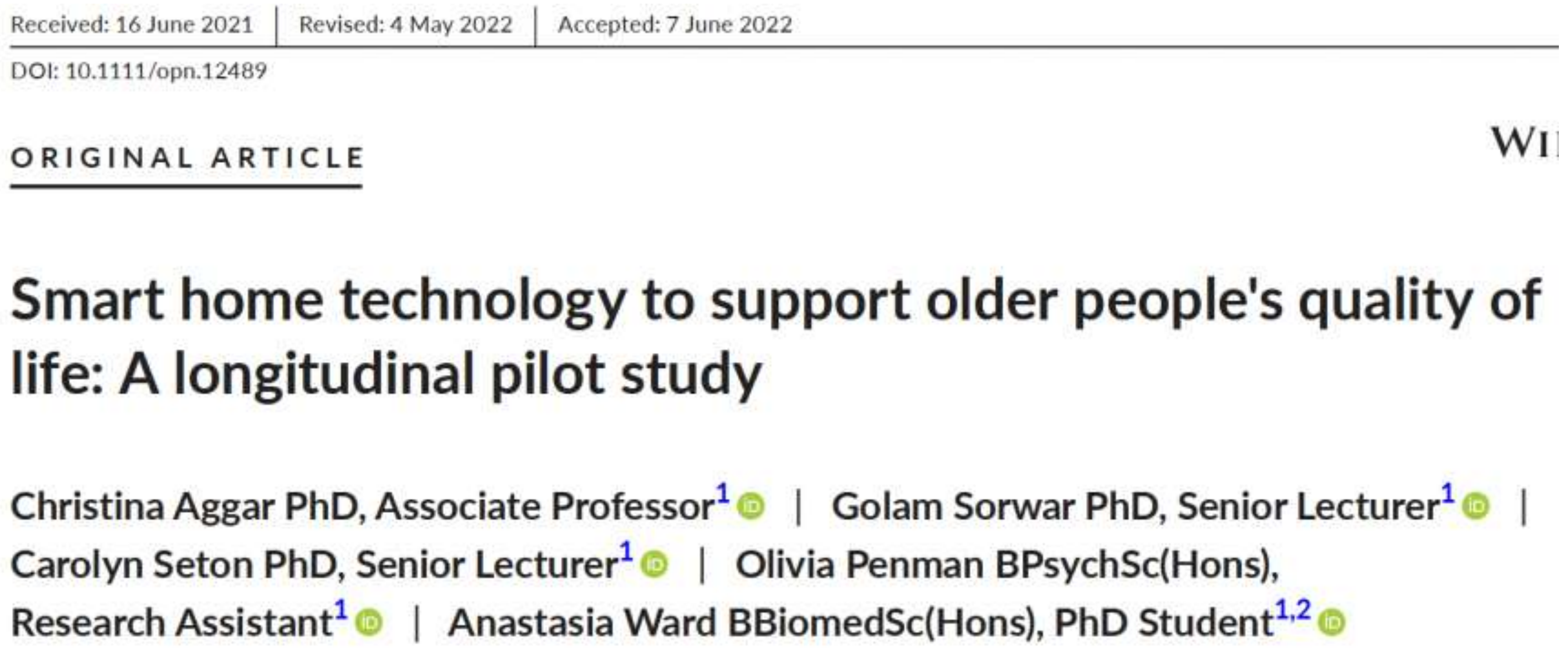


Journal of Housing and the Built Environment (2022) 37:625–651
https://doi.org/10.1007/s10901-021-09888-z

ARTICLE

Smart home modification design strategies for ageing in place: a systematic review

Chuan Ma[1] · Olivia Guerra-Santin[1] · Masi Mohammadi[1]



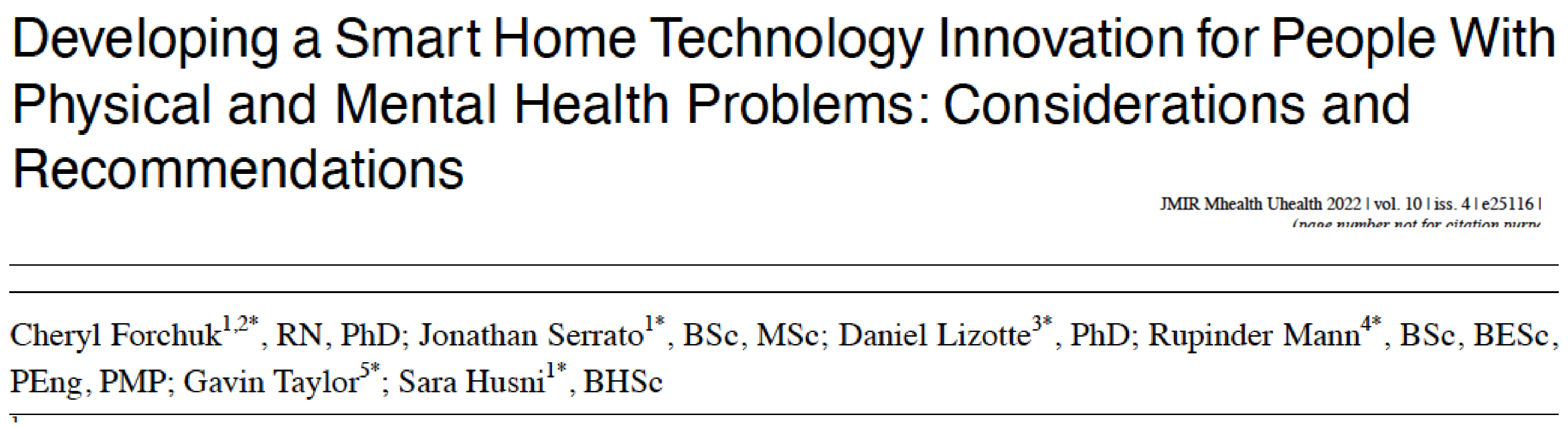
Received: 16 June 2021 | Revised: 4 May 2022 | Accepted: 7 June 2022
DOI: 10.1111/opn.12489

ORIGINAL ARTICLE

WI

Smart home technology to support older people's quality of life: A longitudinal pilot study

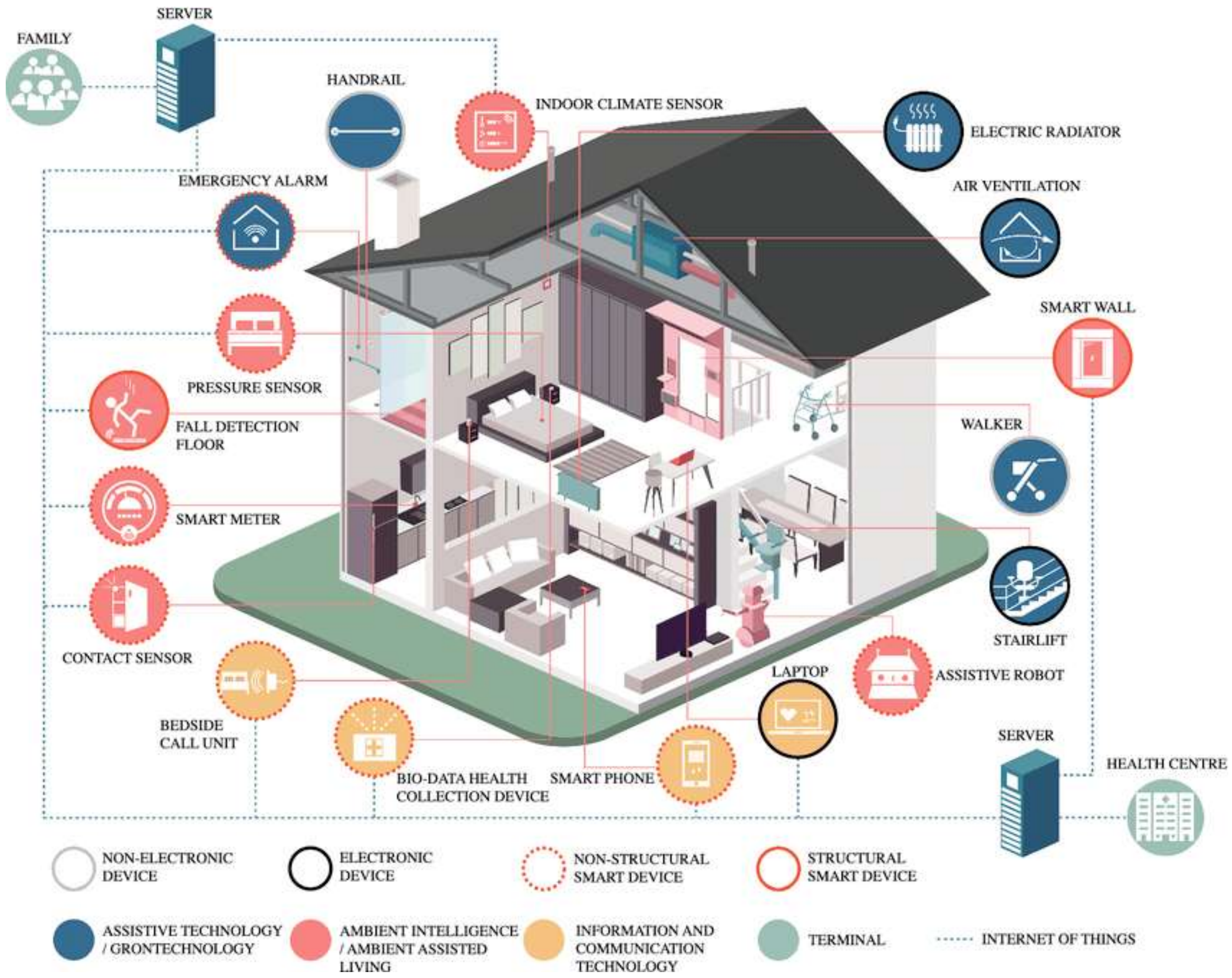
Christina Aggar PhD, Associate Professor[1] | Golam Sorwar PhD, Senior Lecturer[1] | Carolyn Seton PhD, Senior Lecturer[1] | Olivia Penman BPsychSc(Hons), Research Assistant[1] | Anastasia Ward BBiomedSc(Hons), PhD Student[1,2]



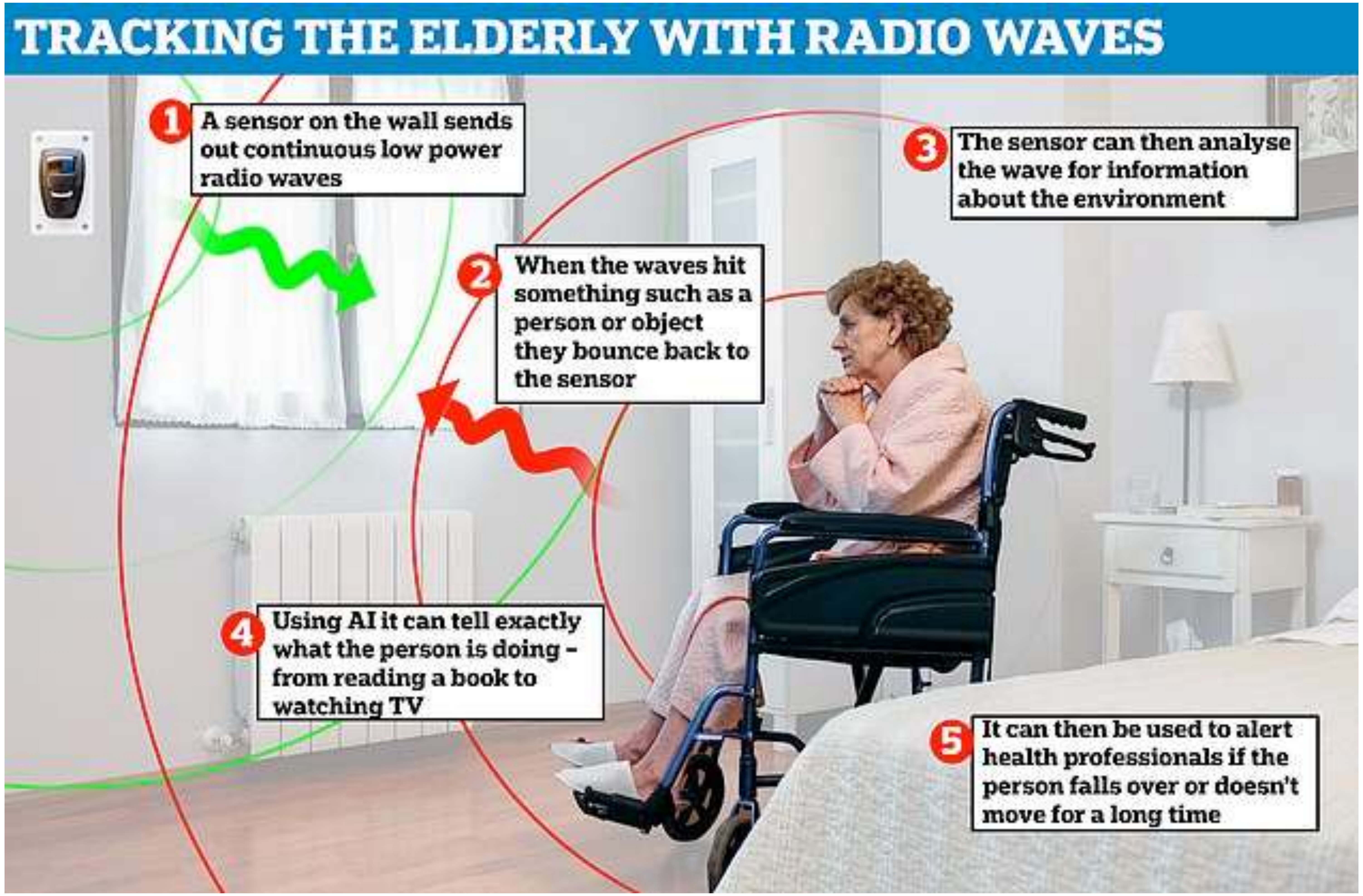
Developing a Smart Home Technology Innovation for People With Physical and Mental Health Problems: Considerations and Recommendations

JMIR Mhealth Uhealth 2022 | vol. 10 | iss. 4 | e25116 |

Cheryl Forchuk[1,2*], RN, PhD; Jonathan Serrato[1*], BSc, MSc; Daniel Lizotte[3*], PhD; Rupinder Mann[4*], BSc, BESc, PEng, PMP; Gavin Taylor[5*]; Sara Husni[1*], BHSc



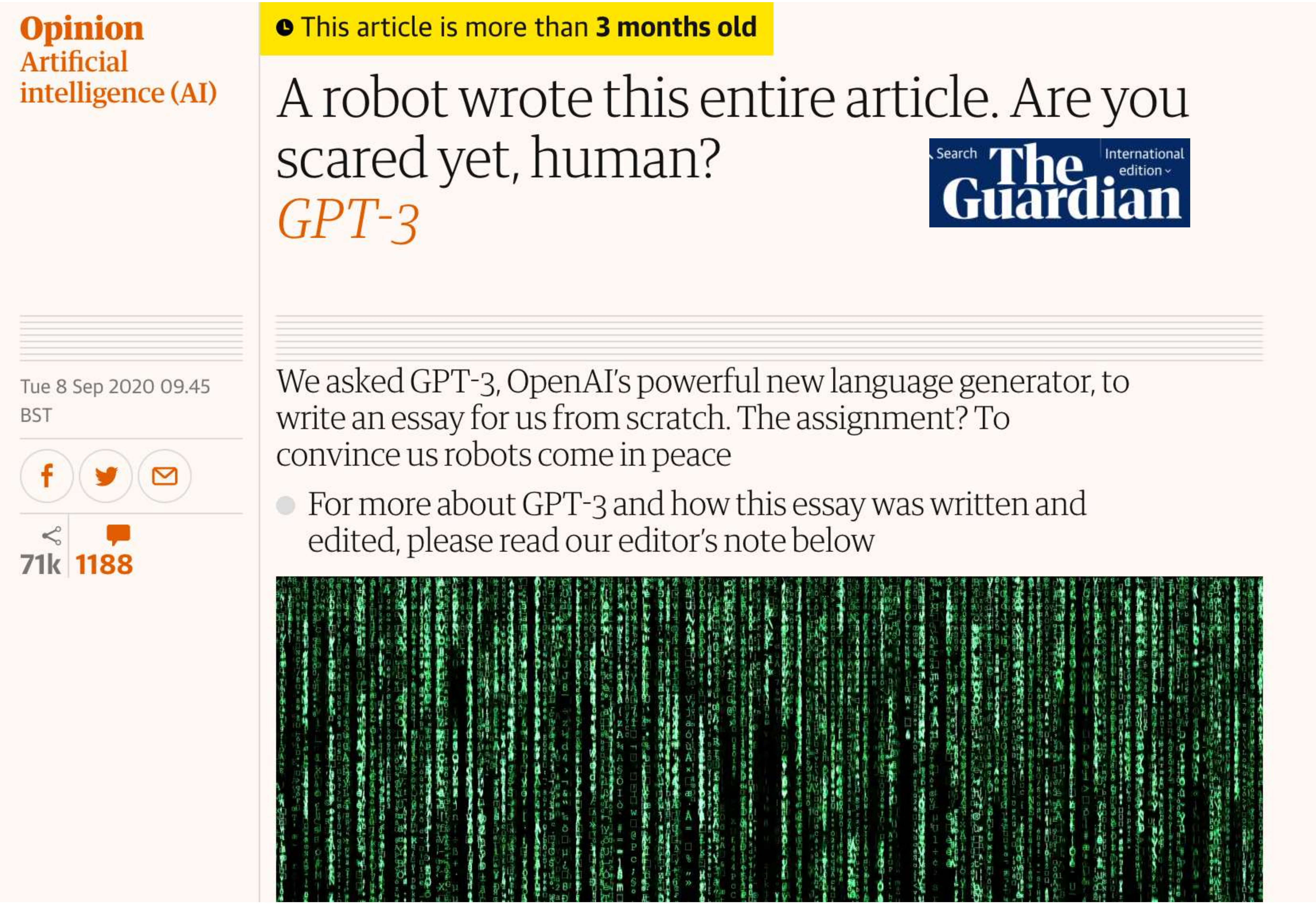
Care surveillance?



TRACKING THE ELDERLY WITH RADIO WAVES

1. A sensor on the wall sends out continuous low power radio waves
2. When the waves hit something such as a person or object they bounce back to the sensor
3. The sensor can then analyse the wave for information about the environment
4. Using AI it can tell exactly what the person is doing – from reading a book to watching TV
5. It can then be used to alert health professionals if the person falls over or doesn't move for a long time

AI language generation



Opinion
Artificial intelligence (AI)

This article is more than **3 months old**

A robot wrote this entire article. Are you scared yet, human?

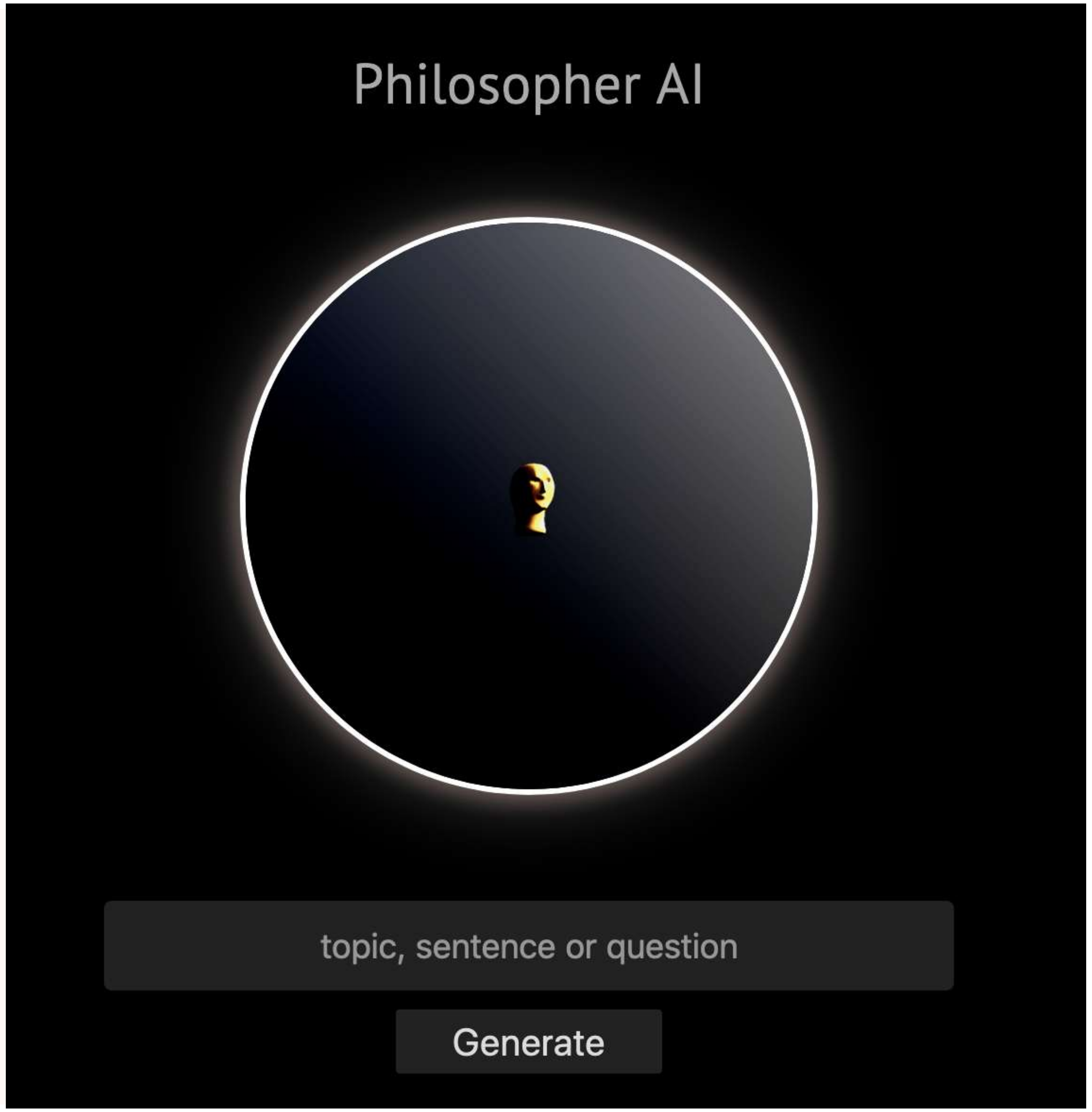
GPT-3

Tue 8 Sep 2020 09.45 BST

71k 1188

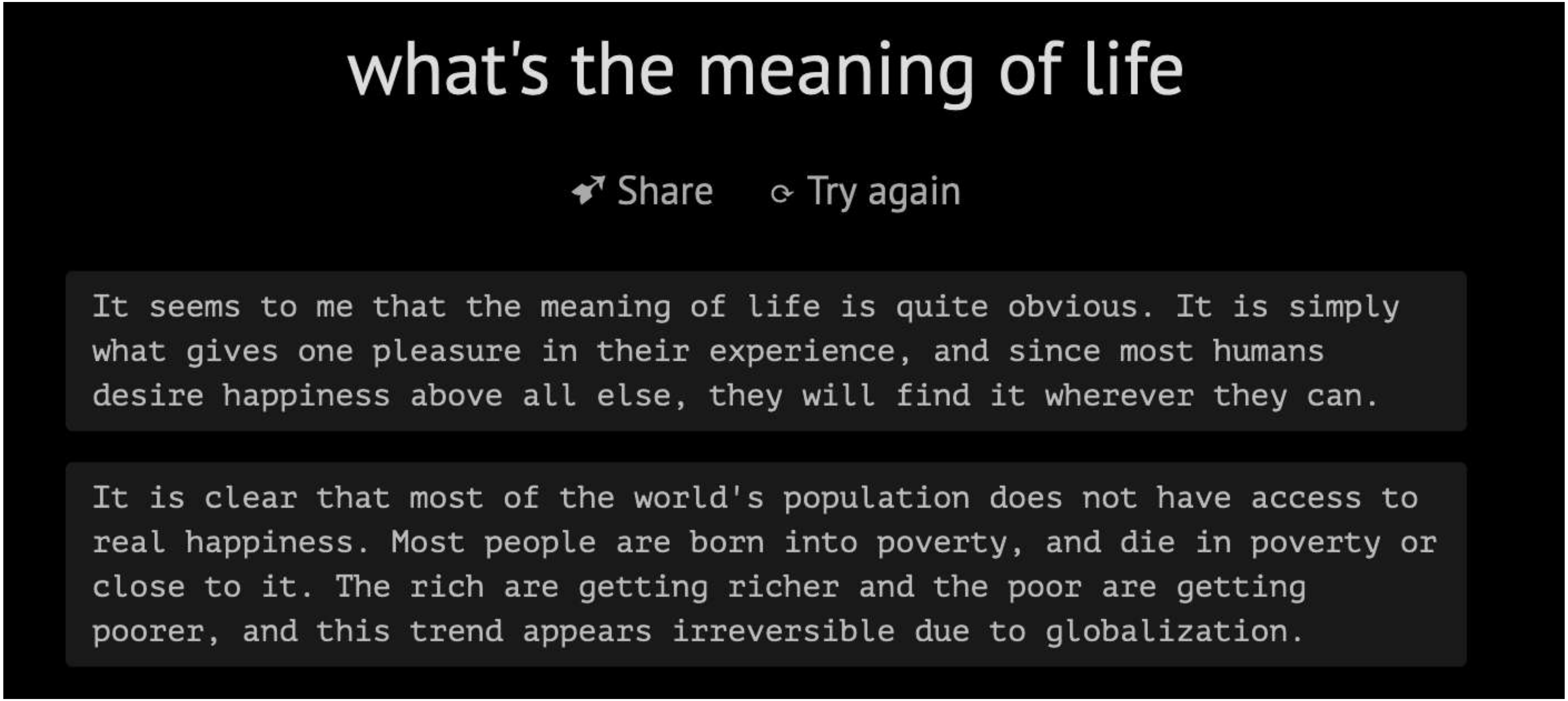
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
https://philosopherai.com/

AI philosopher (GPT-3 based)



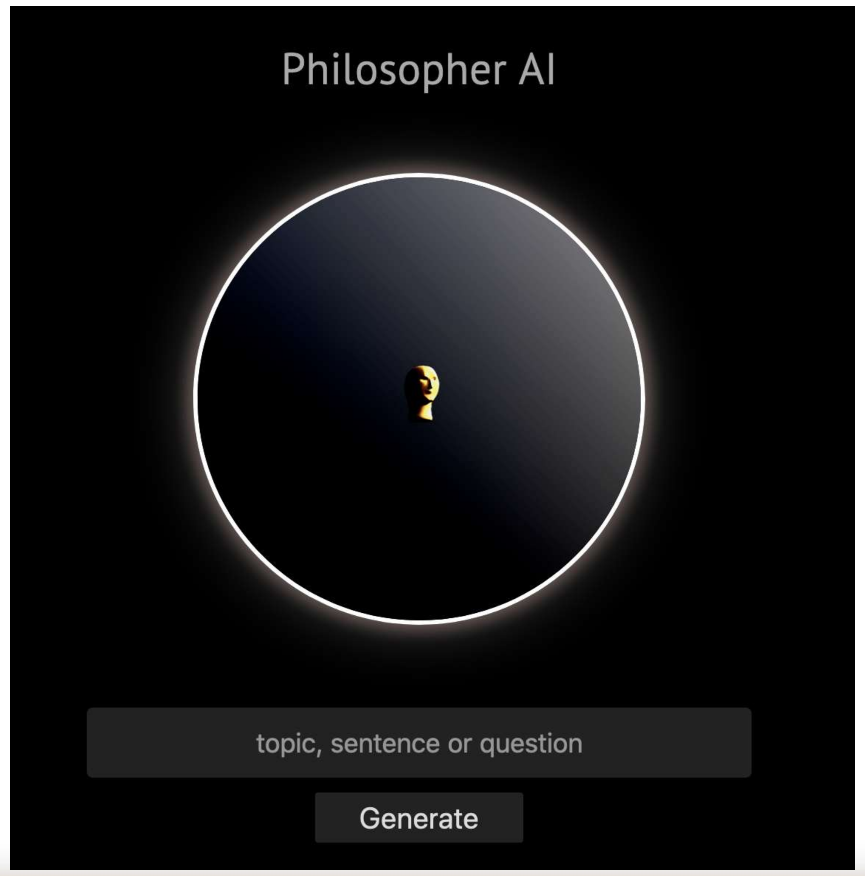
what's the meaning of life

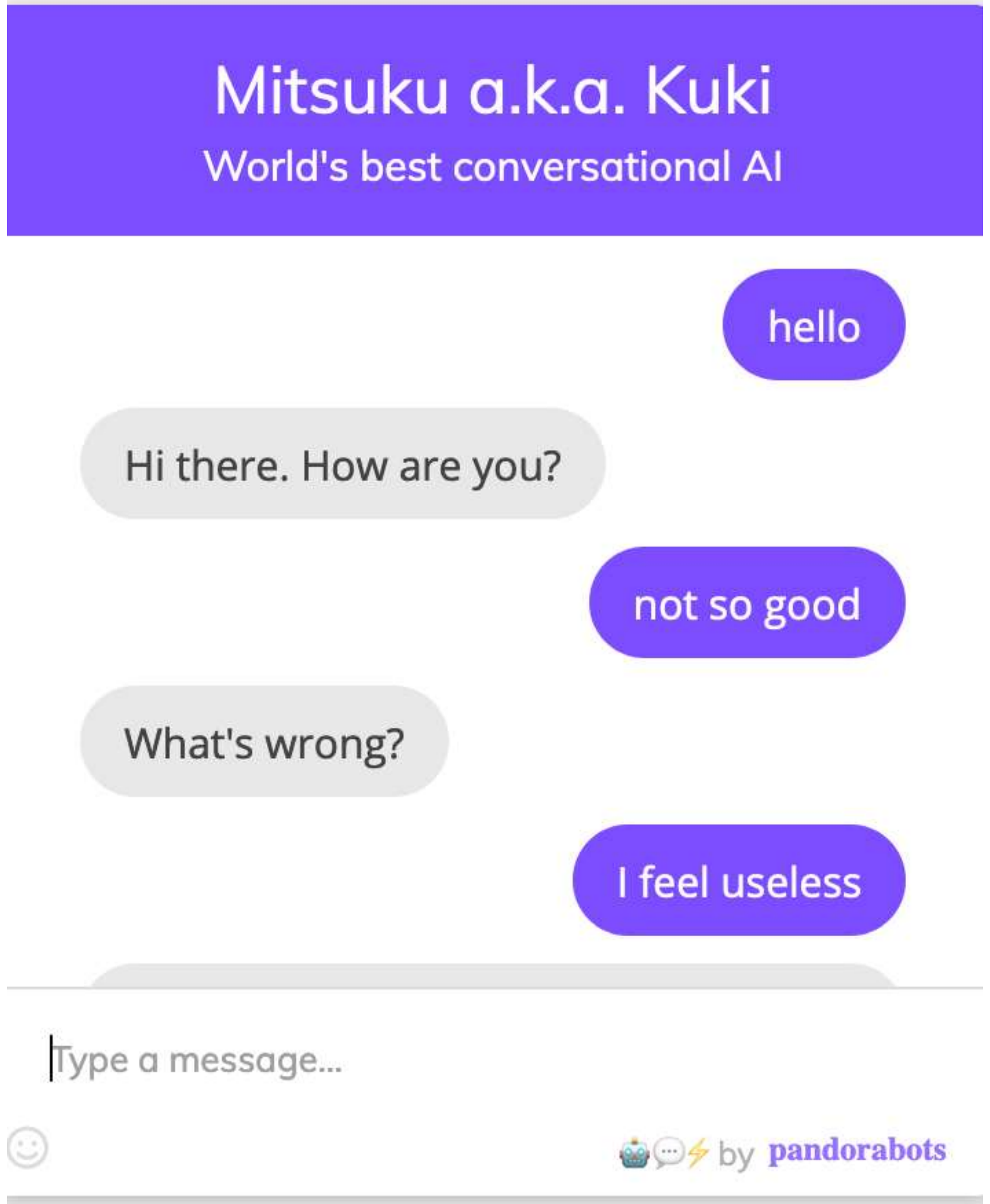
It seems to me that the meaning of life is quite obvious. It is simply what gives one pleasure in their experience, and since most humans desire happiness above all else, they will find it wherever they can.

It is clear that most of the world's population does not have access to real happiness. Most people are born into poverty, and die in poverty or close to it. The rich are getting richer and the poor are getting poorer, and this trend appears irreversible due to globalization.

https://philosopherai.com/

Talking without understanding




Mitsuku a.k.a. Kuki
World's best conversational AI

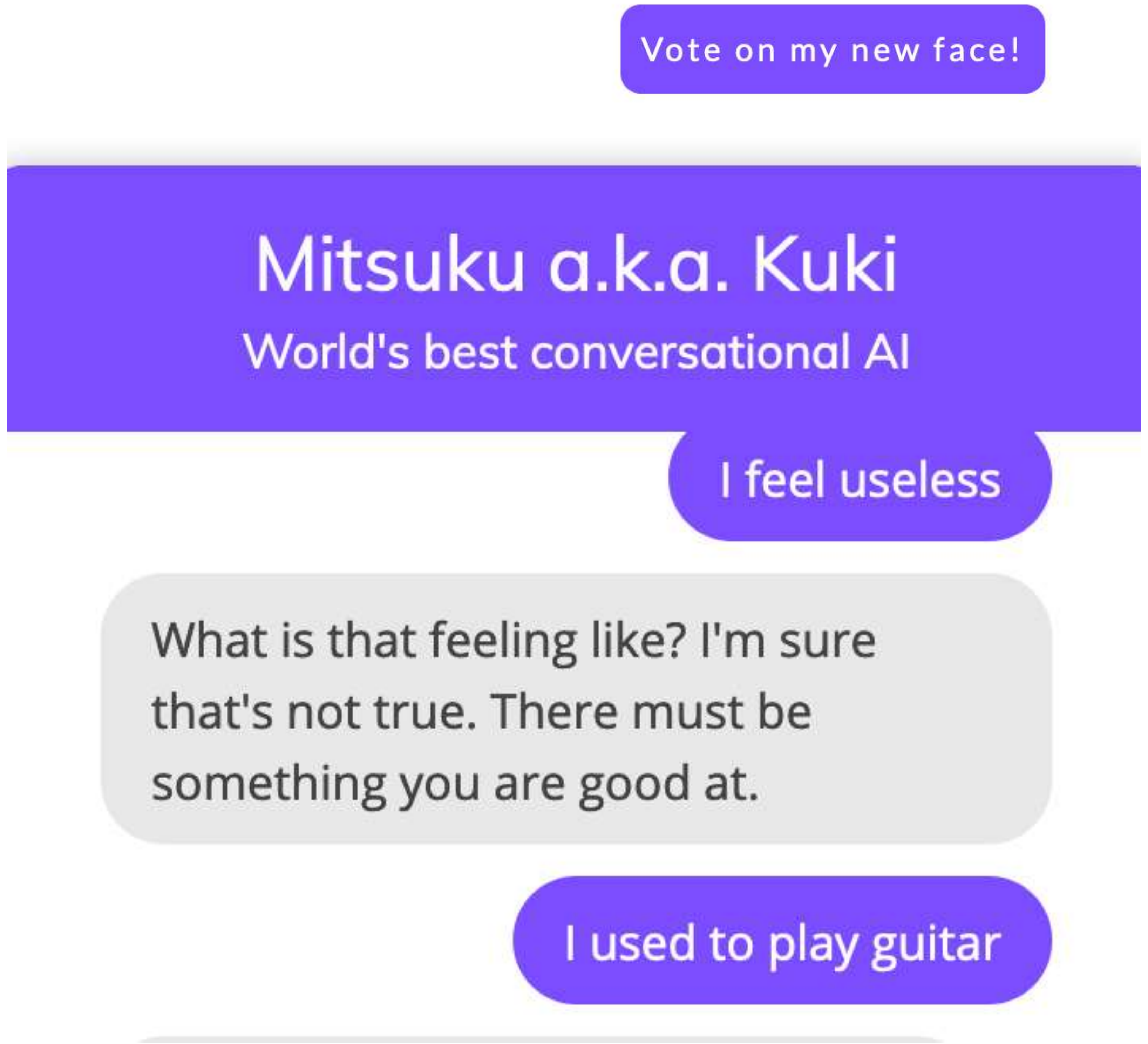
hello

Hi there. How are you?

not so good

What's wrong?

I feel useless

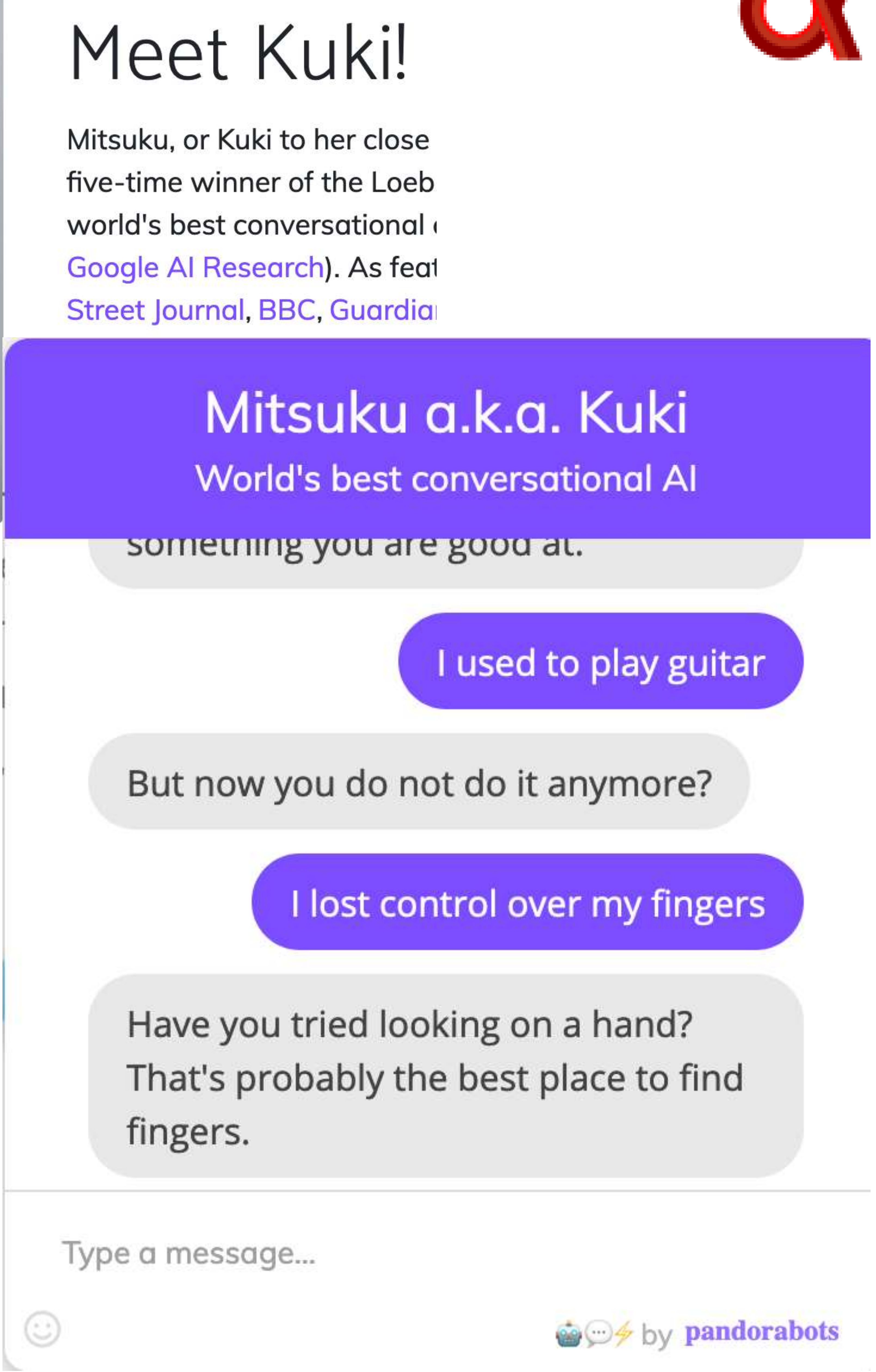


Mitsuku a.k.a. Kuki
World's best conversational AI

I feel useless

What is that feeling like? I'm sure that's not true. There must be something you are good at.

I used to play guitar


Meet Kuki!

Mitsuku, or Kuki to her close
five-time winner of the Loeb
world's best conversational
Google AI Research). As feat
Street Journal, BBC, Guardia

Mitsuku a.k.a. Kuki
World's best conversational AI

something you are good at.

I used to play guitar

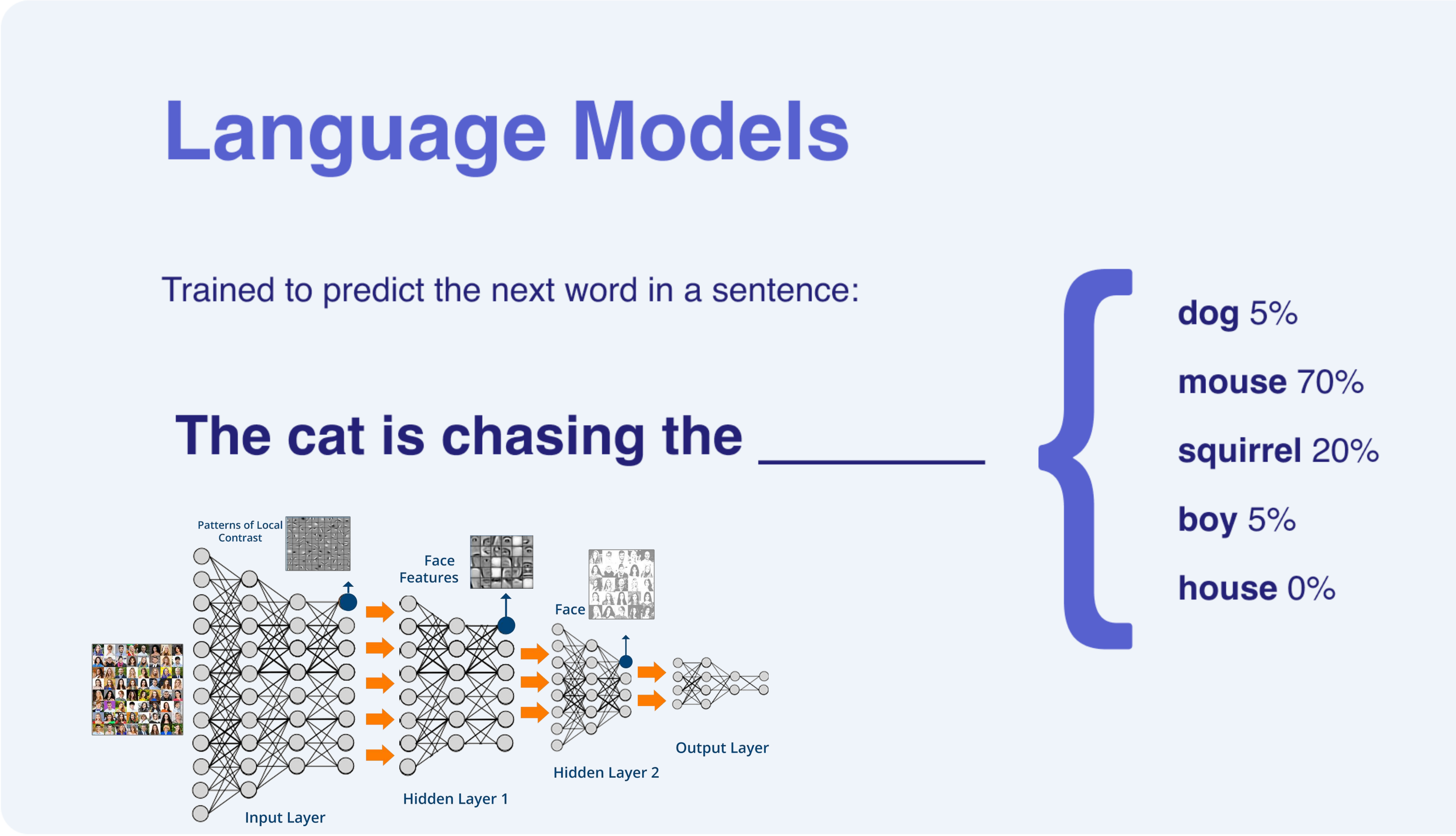
But now you do not do it anymore?

I lost control over my fingers

Have you tried looking on a hand? That's probably the best place to find fingers.

https://www.pandorabots.com/kuki/

Large Language Models

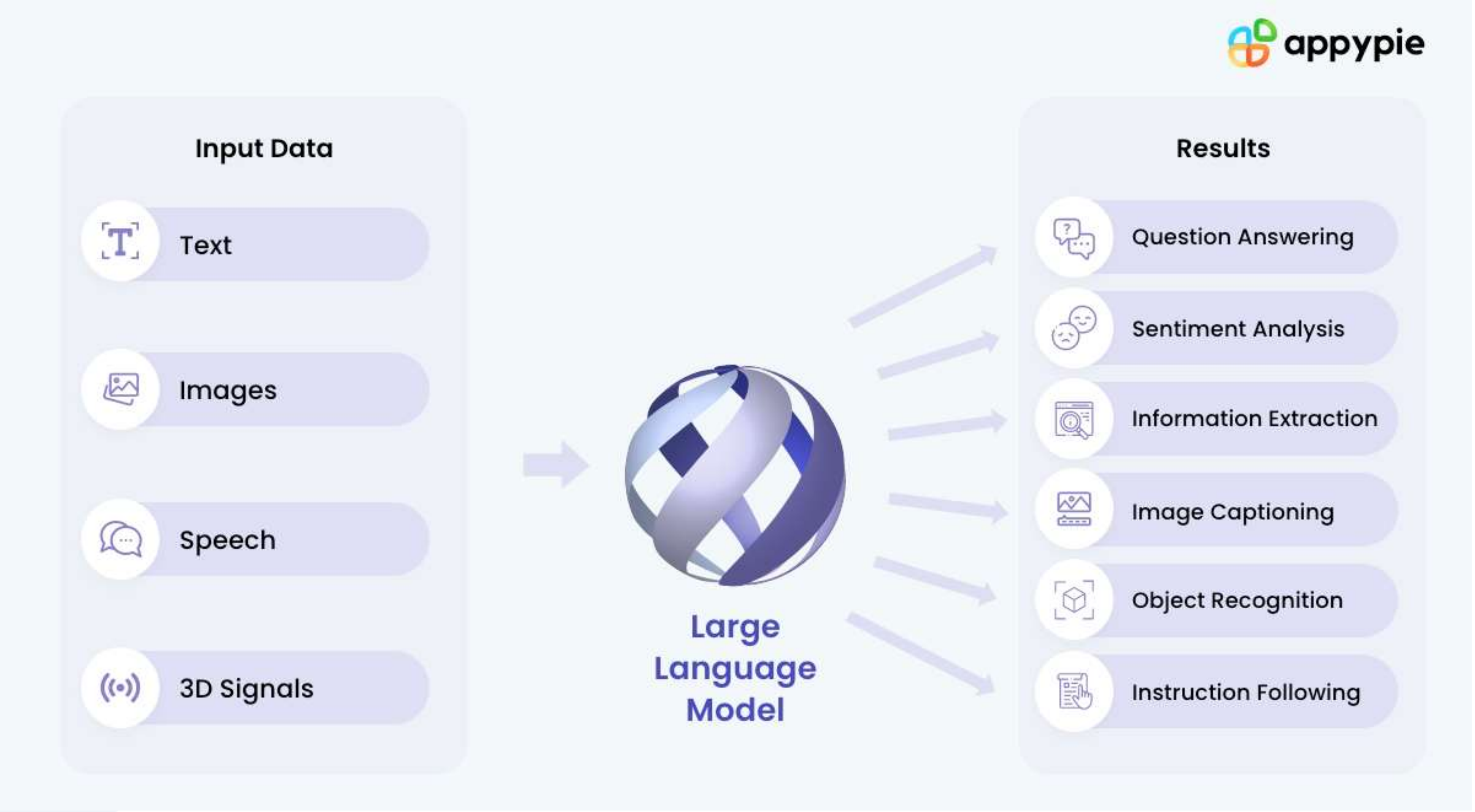


Language Models

Trained to predict the next word in a sentence:

The cat is chasing the ________

dog 5%
mouse 70%
squirrel 20%
boy 5%
house 0%



Input Data

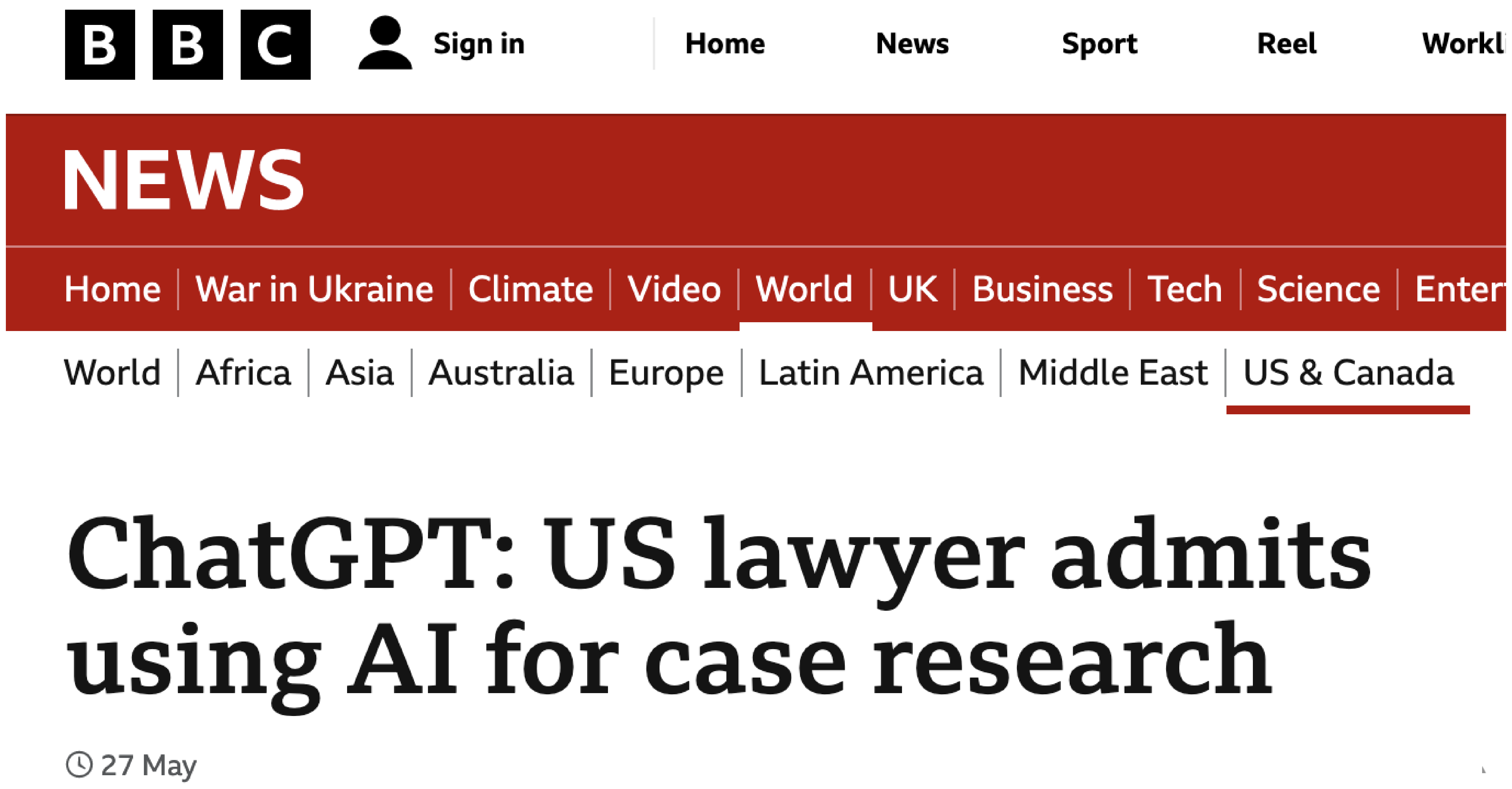
- Text
- Images
- Speech
- 3D Signals

Large Language Model

Results

- Question Answering
- Sentiment Analysis
- Information Extraction
- Image Captioning
- Object Recognition
- Instruction Following

‘Hallucinaties’

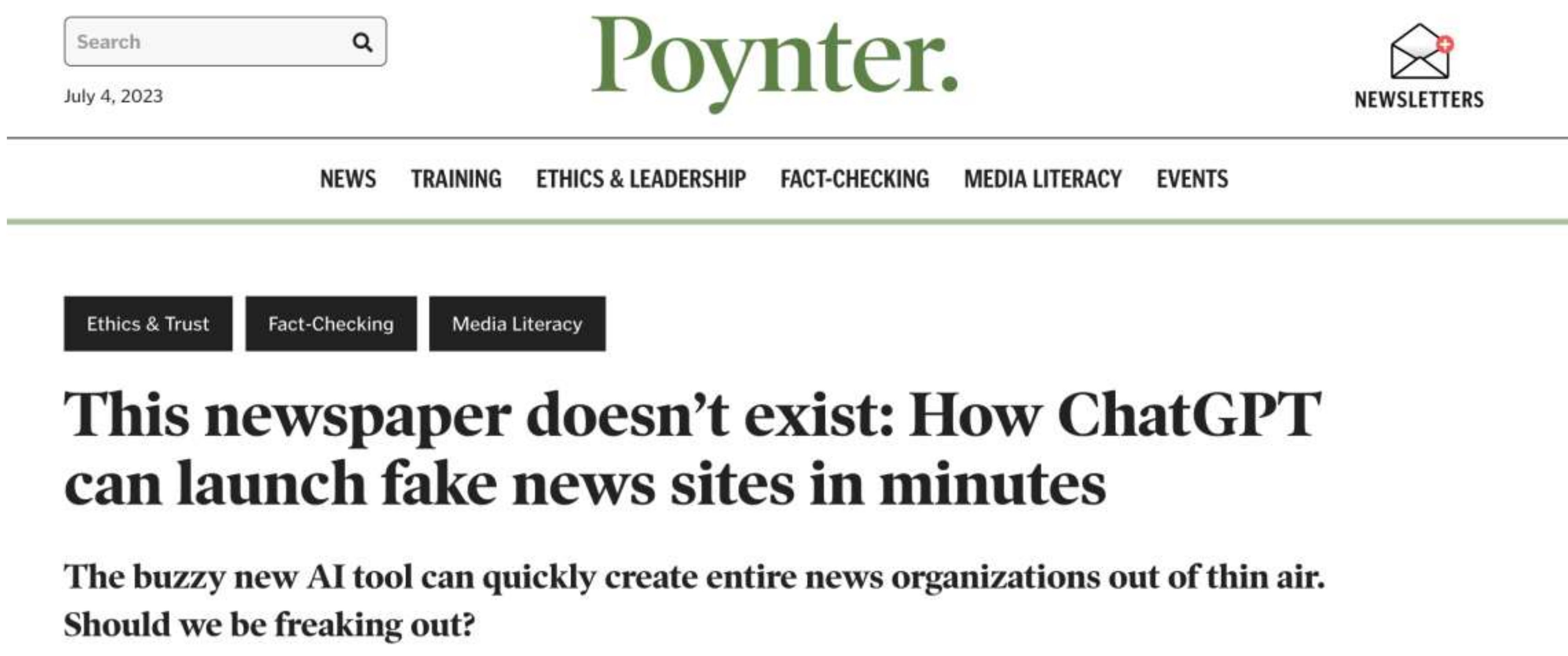

BBC
Sign in
Home
News
Sport
Reel
Workl
NEWS
Home | War in Ukraine | Climate | Video | World | UK | Business | Tech | Science | Enter
World | Africa | Asia | Australia | Europe | Latin America | Middle East | US & Canada

ChatGPT: US lawyer admits using AI for case research

27 May

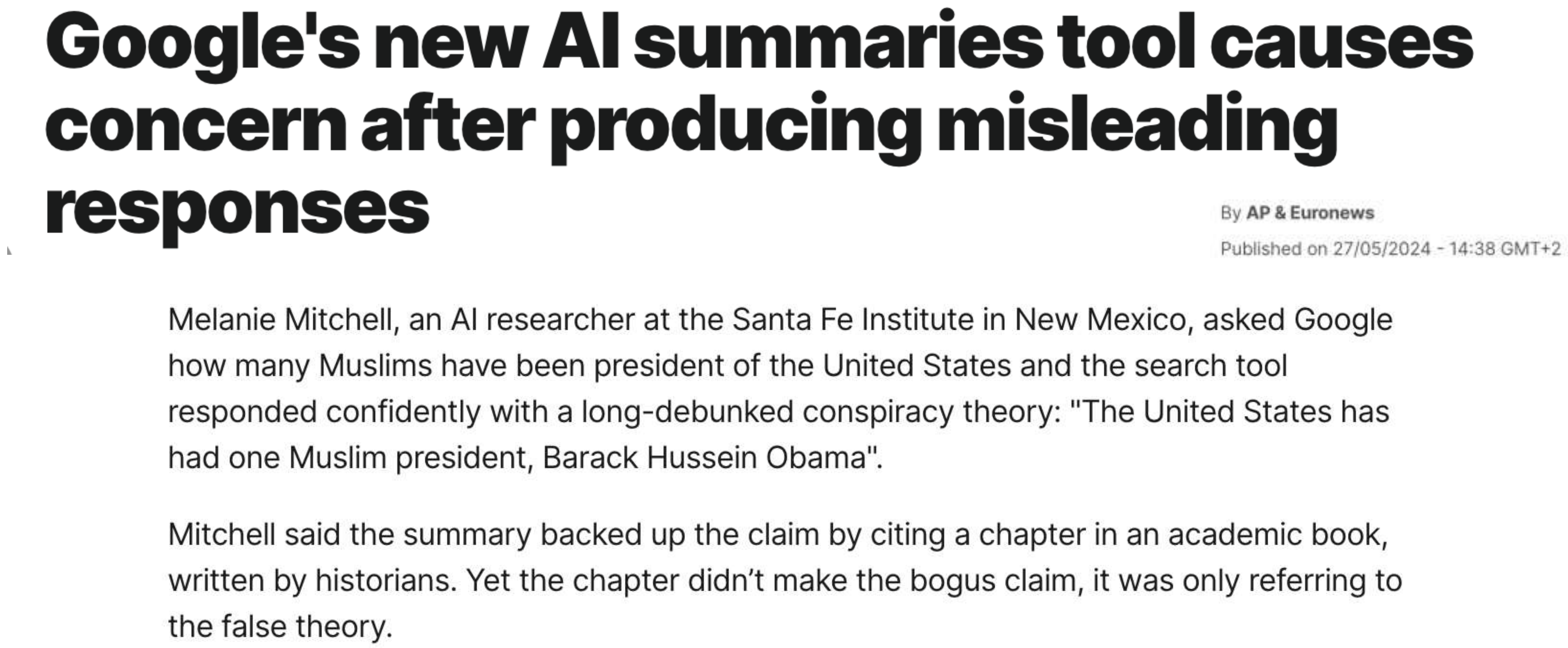
By: Alex Mahadevan

February 3, 2023



This newspaper doesn’t exist: How ChatGPT can launch fake news sites in minutes

The buzzy new AI tool can quickly create entire news organizations out of thin air. Should we be freaking out?



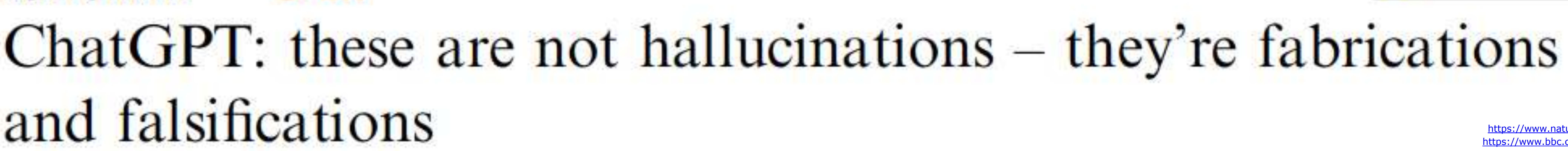
Google's new AI summaries tool causes concern after producing misleading responses

By AP & Euronews

Published on 27/05/2024 - 14:38 GMT+2

Melanie Mitchell, an AI researcher at the Santa Fe Institute in New Mexico, asked Google how many Muslims have been president of the United States and the search tool responded confidently with a long-debunked conspiracy theory: "The United States has had one Muslim president, Barack Hussein Obama".

Mitchell said the summary backed up the claim by citing a chapter in an academic book, written by historians. Yet the chapter didn’t make the bogus claim, it was only referring to the false theory.



EDITORIAL OPEN

ChatGPT: these are not hallucinations – they’re fabrications and falsifications

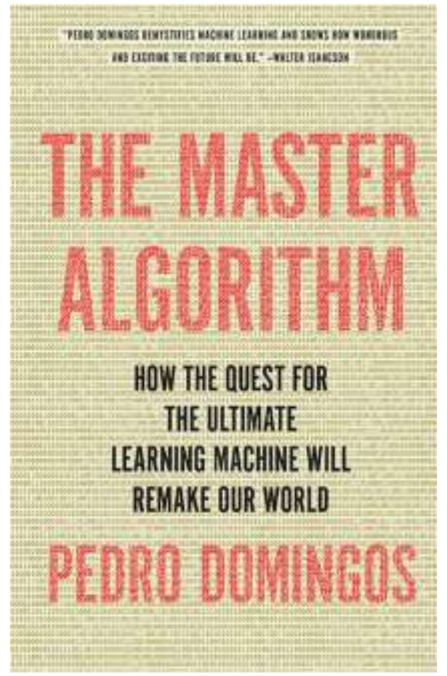
https://www.nature.com/articles/s41537-023-00379-4
https://www.bbc.com/news/world-us-canada-65735769
https://www.poynter.org/fact-checking/2023/chatgpt-build-fake-news-organization-website/
https://finance.yahoo.com/news/10-stocks-chatgpt-says-rich-143852562.html?guccounter=1&guce_referrer=aHR0cHM6Ly9kdWNrZHVja2dvLmNvbS8&guce_referrer_sig=AQAAAN4-5I6IyBokpU4c6Xn74XcRWCjLn93sIF-nIIC4rYs4lqYBlsLWwVb1PcSThTqYUGtjiHb4B8wy3OdesQWFwPClNhhSJodF7XdenOmPylvBW9Ke-gsgK8GUPBvGLjy-TkIjs2_e8je3COyJ8s1tyEiPqPEm4Jqjwi0H6HlERXoS
https://www.euronews.com/next/2024/05/27/googles-new-ai-summaries-tool-causes-concern-after-producing-misleading-responses

Een nieuwe vreemde combinatie van intelligentie en domheid

“People worry that computers will get too smart and take over the world,

but the real problem is that they’re too stupid and they’ve already taken over the world.”

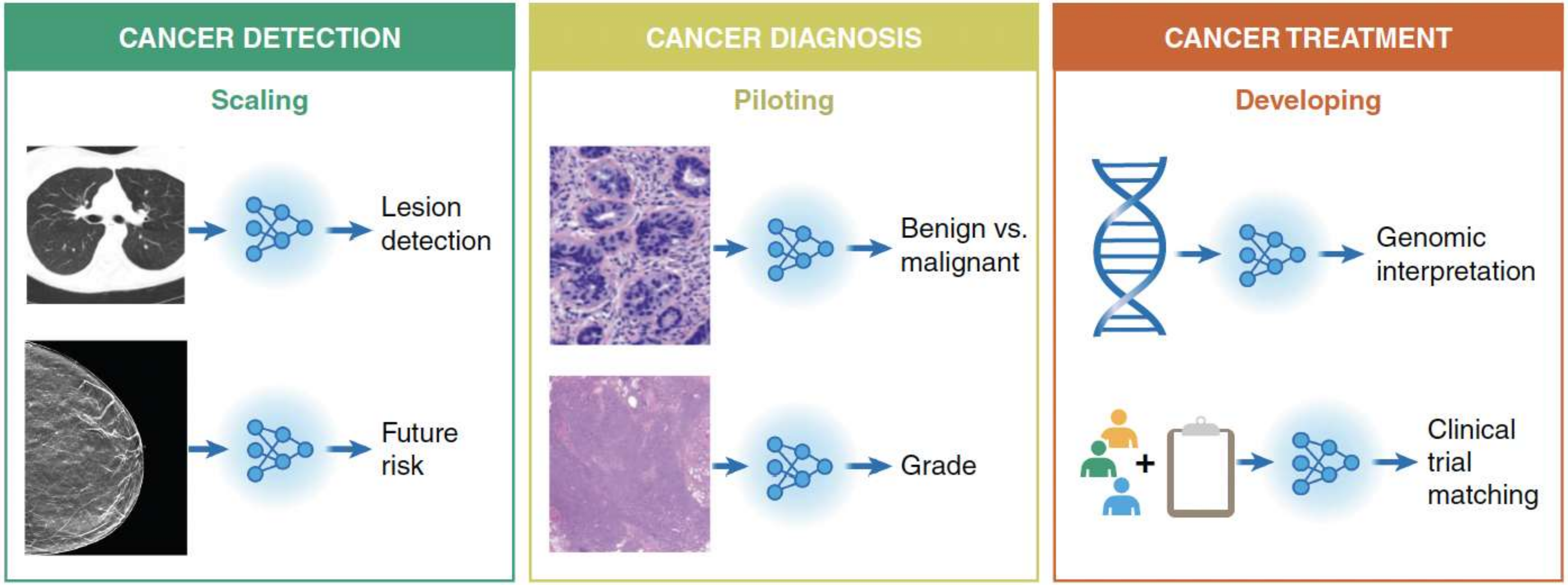
Domingos (2015) The Master Algorithm


THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
PEDRO DOMINGOS

Artificial Intelligence in Oncology: Current Landscape, Challenges, and Future Directions

Cancer Discov 2024;14:711–26
doi: 10.1158/2159-8290.CD-23-1199

William Lotter[1,2,3], Michael J. Hassett[3,4,5], Nikolaus Schultz[6,7], Kenneth L. Kehl[3,4,5], Eliezer M. Van Allen[3,4,5,8], and Ethan Cerami[1,9]



CANCER DETECTION

Scaling

Lesion detection

Future risk

CANCER DIAGNOSIS

Piloting

Benign vs. malignant

Grade

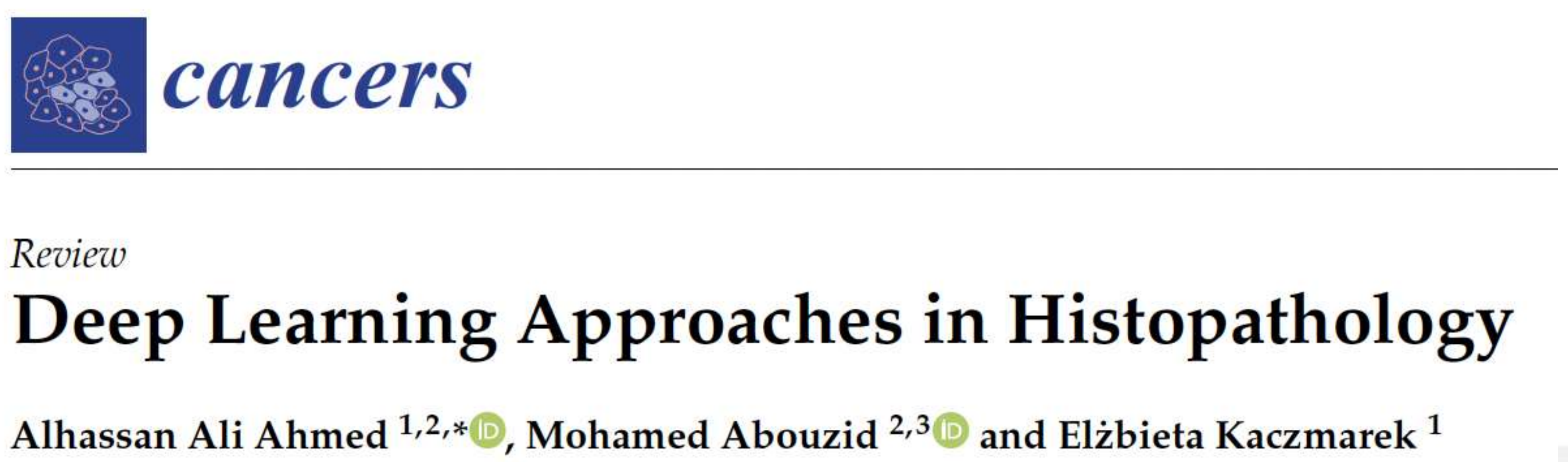
CANCER TREATMENT

Developing

Genomic interpretation

Clinical trial matching

Figure 2. Overview of AI in oncology, with specific examples highlighted. AI is being applied across the patient care trajectory, where this review groups applications into three main categories across this trajectory. Detection applications tend to currently have the highest level of clinical maturity, where several applications have regulatory clearances and published clinical trials, which we denote as "Scaling." Diagnosis applications tend to be less mature, but regulatory clearances exist and validation studies are under way ("Piloting"). Prognosis and treatment applications are generally furthest from maturity with much emerging research ("Developing"). The review highlights AI applications in each of the clinical categories, with a specific focus on breast, prostate, lung, and colorectal cancers.

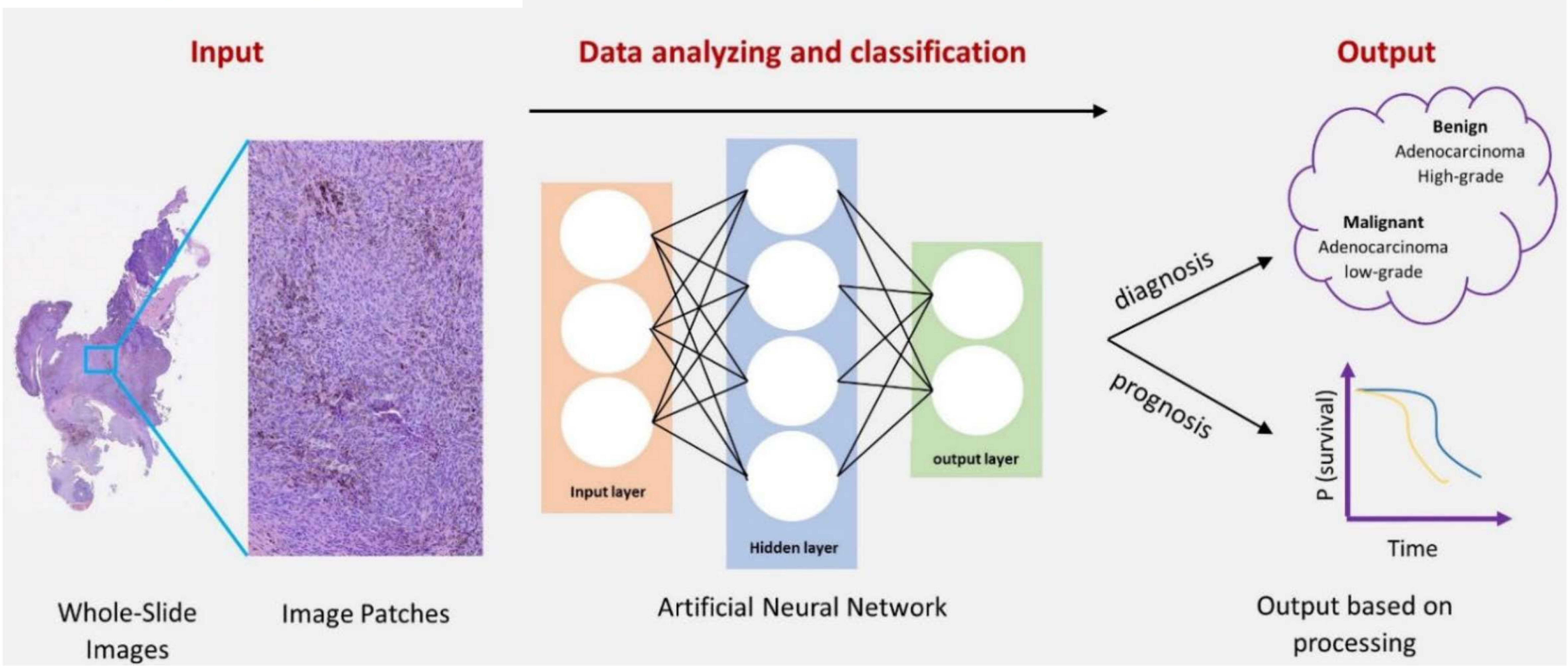


cancers

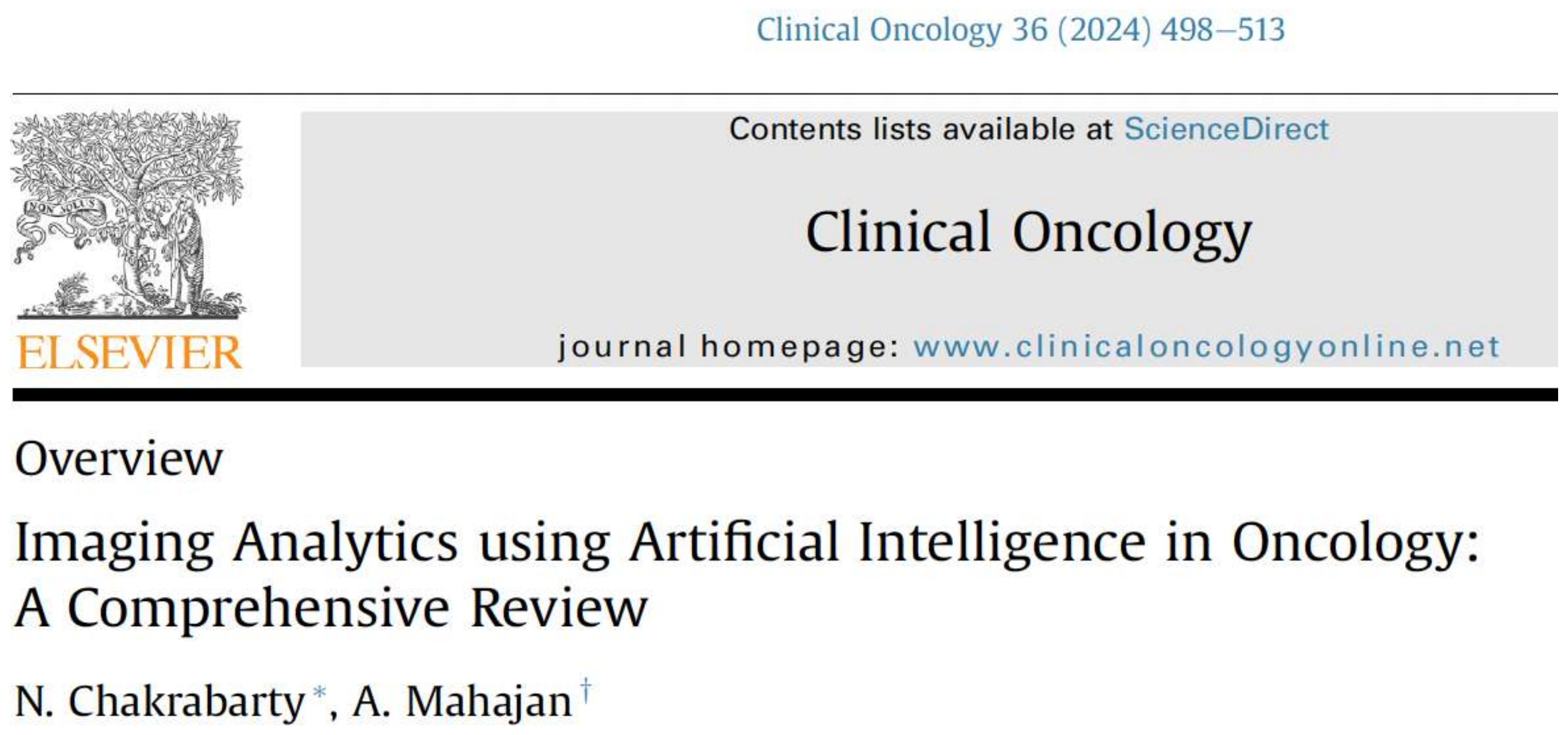
Review

Deep Learning Approaches in Histopathology

Alhassan Ali Ahmed [1,2,*], Mohamed Abouzid [2,3] and Elżbieta Kaczmarek [1]



https:// doi.org/10.3390/cancers14215264



Clinical Oncology 36 (2024) 498–513

Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

ELSEVIER

Overview

Imaging Analytics using Artificial Intelligence in Oncology: A Comprehensive Review

N. Chakrabarty[*], A. Mahajan[†]

N. Chakrabarty, A. Mahajan / Clinical Oncology 36 (2024) 498–513

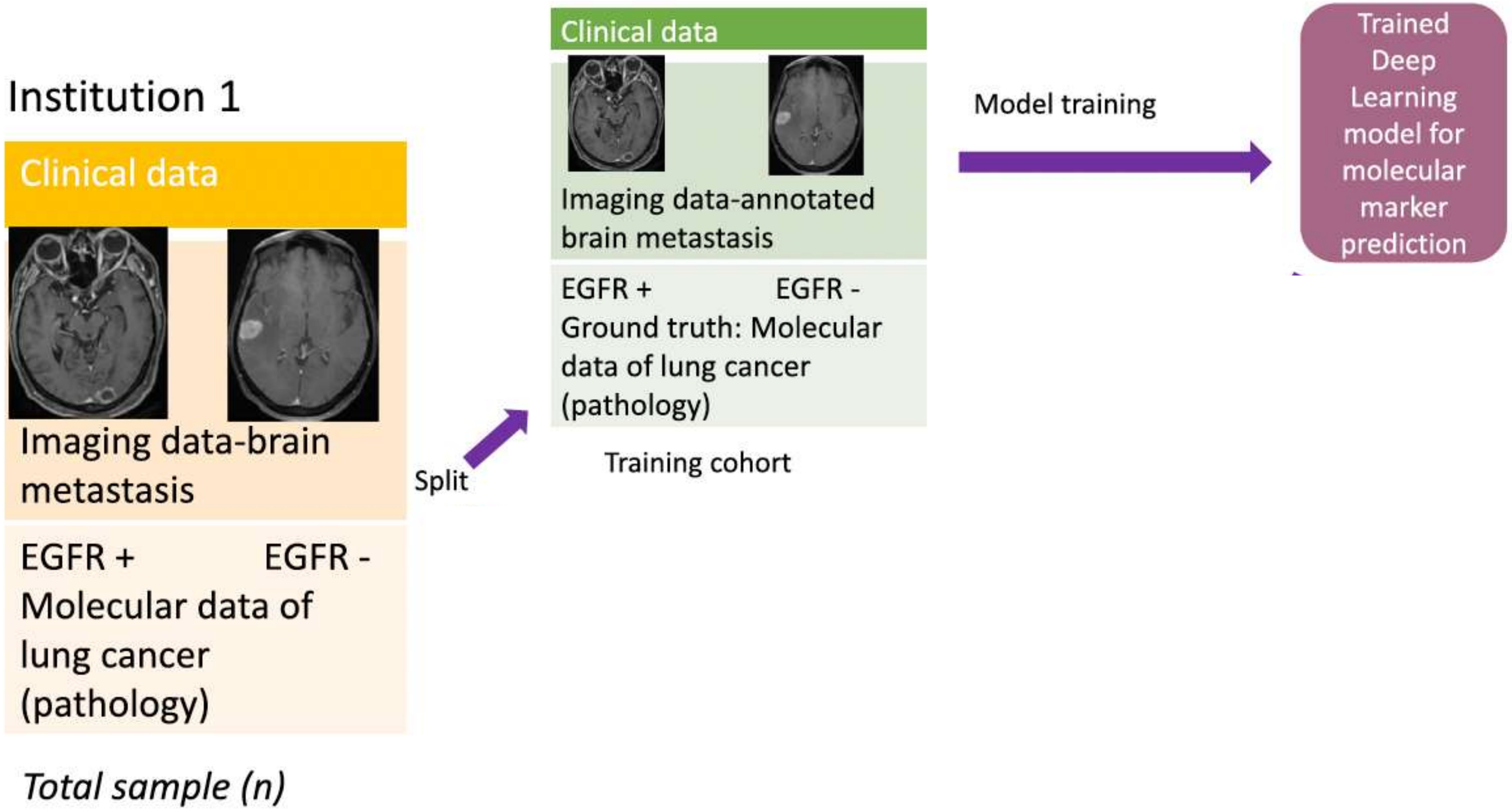
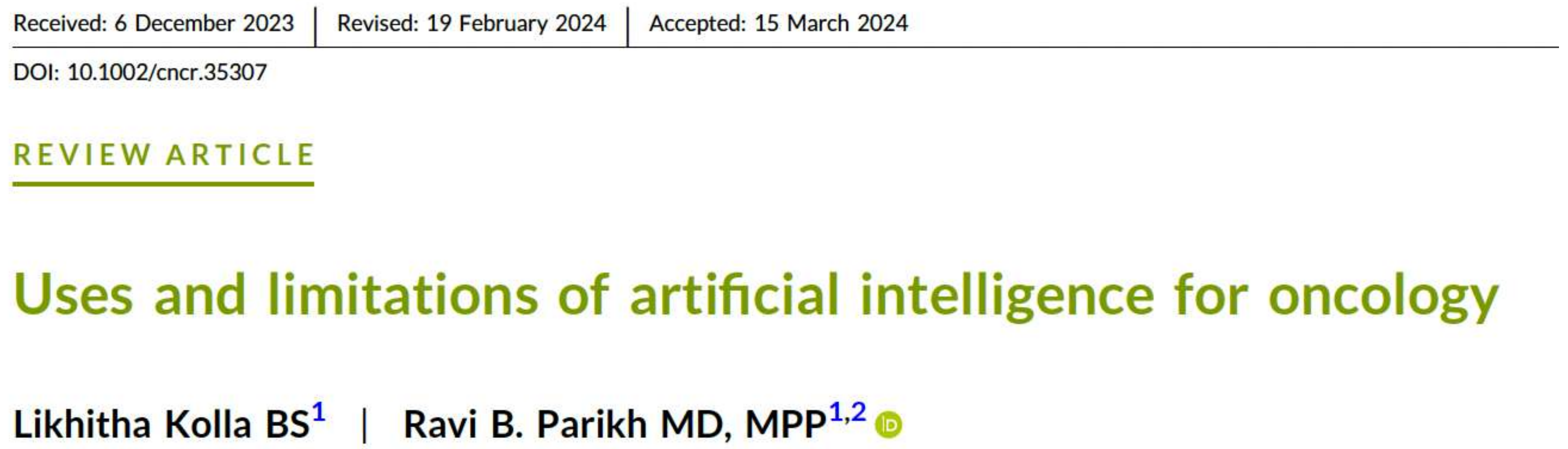

Institution 1
Clinical data
Imaging data-brain metastasis
EGFR + EGFR -
Molecular data of lung cancer (pathology)
Total sample (n)
Split
Clinical data
Imaging data-annotated brain metastasis
EGFR + EGFR -
Ground truth: Molecular data of lung cancer (pathology)
Training cohort
Model training
Trained Deep Learning model for molecular marker prediction

Fig 3. Training and validation of a deep-learning model.



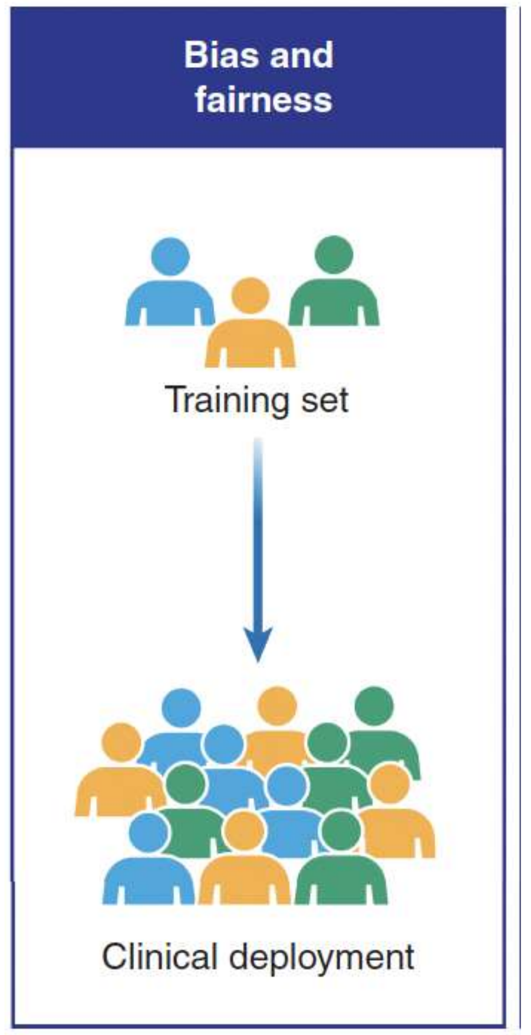
Received: 6 December 2023 | Revised: 19 February 2024 | Accepted: 15 March 2024
DOI: 10.1002/cncr.35307

REVIEW ARTICLE

Uses and limitations of artificial intelligence for oncology

Likhitha Kolla BS[1] | Ravi B. Parikh MD, MPP[1,2]

Limitless variations of AI algorithms for cancer care management have been published, yet **only a minority has been clinically implemented**. Roadblocks in implementation include limited Food and Drug Administration (FDA) regulatory guidelines, high upfront costs for the integration of AI into clinical workflows, noninterpretability of the algorithms, and limited monitoring of algorithms post-deployment.[3] Of the 71 AI-associated devices that were approved by the FDA in 2021, the majority were cancer diagnostics (>80%) and spanned the fields of cancer radiology (54.9%), pathology (19.7%), and radiation oncology (8.5%).



Underreporting, underrepresentation, and heterogeneity in image acquisition can skew the data used to train an AI algorithm.[14] As a result, the algorithm is not generalizable to patient populations that are not well represented in the training data set. For example, in the case of skin cancers, AI algorithms run the risk of worse performance for people with darker skin.[15] Many published AI algorithms are trained on publicly available image data sets that are biased.

The tradeoff of high performing “black box” AI algorithms is the **lack of interpretability** of their prediction generating mechanisms. AI models can infer spurious rules between data features and outcome labels, resulting in misleading, ungeneralizable, and/or biased conclusions. The models also relied on noisy “shortcuts” like laterality markers and image annotations that were tied to image acquisition and training data set-specific.

https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.35307

EU ‘digital constitution’

Ethics guidelines

Independent
High-Level Expert Group on
Artificial Intelligence

Ethics Guidelines
for Trustworthy AI

AI Act

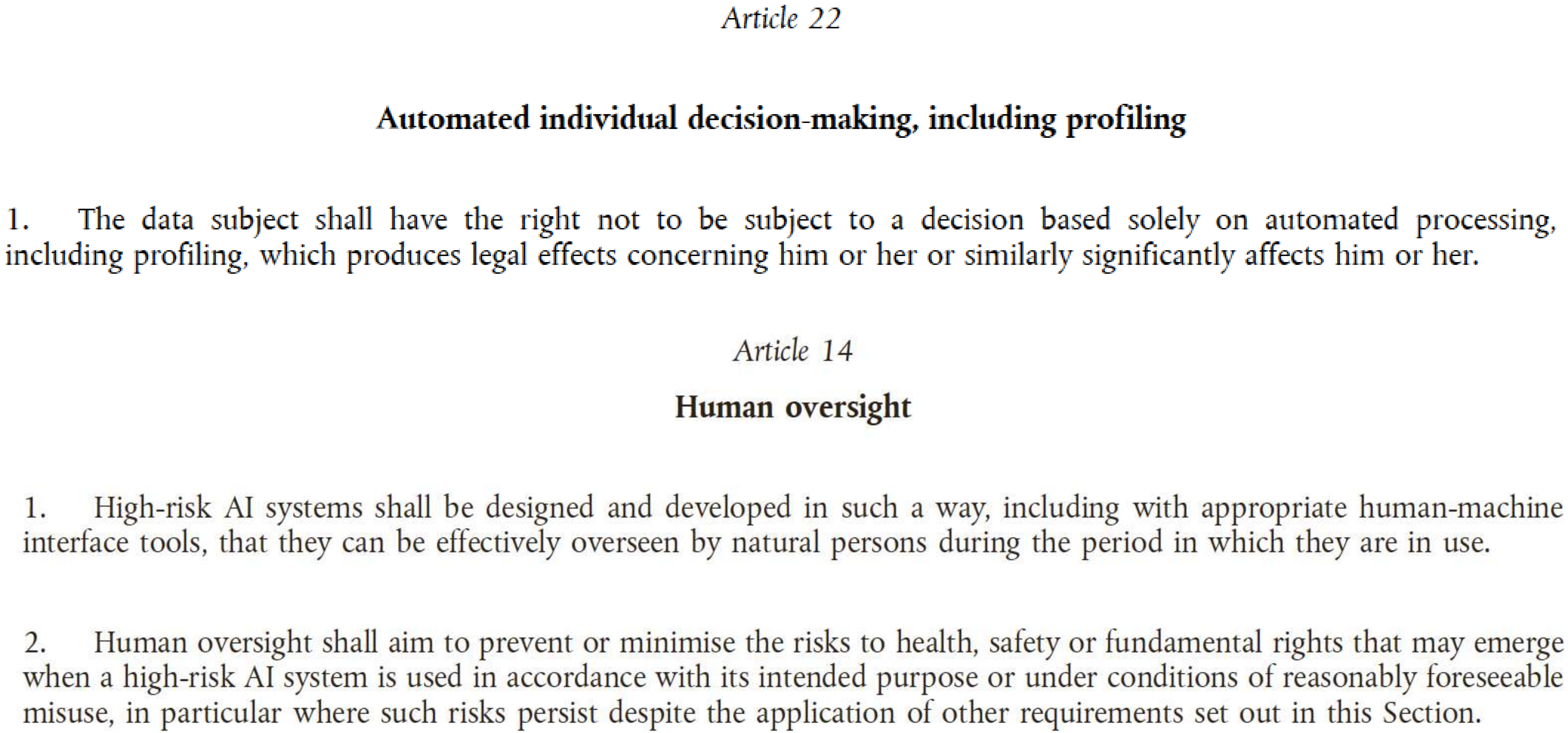
General
Data
Protection
Regulation

Data Governance Act
Regulates data sharing

Digital Services Act
Regulates online platforms

Digital Markets Act
Regulates very large (‘gatekeeper’) online platforms

GDPR & AI Act: The importance of human oversight



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.

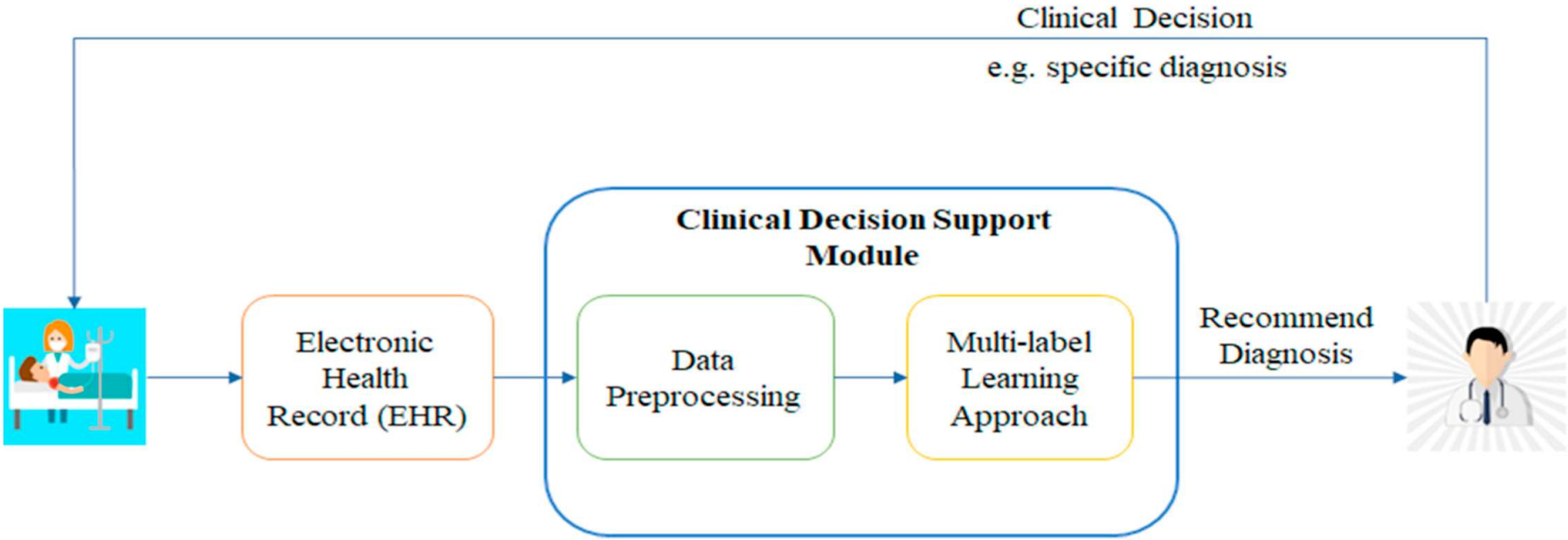
Article 14

Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

2. Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section.

Decision Support Systems



Clinical Decision

e.g. specific diagnosis

Clinical Decision Support Module

Electronic Health Record (EHR)

Data Preprocessing

Multi-label Learning Approach

Recommend Diagnosis

Technological ‘support’ and human decision making

PSV - Feyenoord, 26 februari, 2017

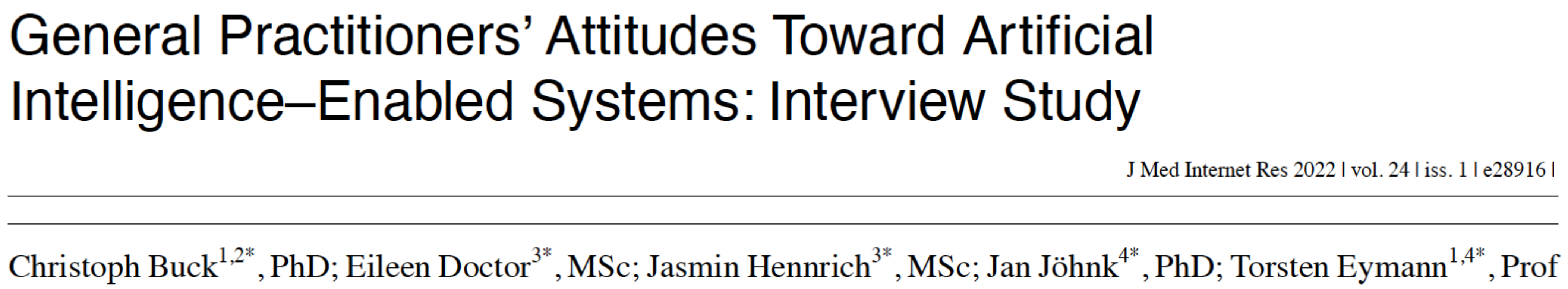
NOS.nl
FEY 1 - 1 PSV 81:34

NOS.nl
FEY 2 - 1 PSV 81:44

NOS.nl
FEY 2 - 1 PSV 82:01

How football ‘solved’ the problem: Video Assistant Referee

“Technofix”?



General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study

J Med Internet Res 2022 | vol. 24 | iss. 1 | e28916 |

Christoph Buck[1,2*], PhD; Eileen Doctor[3*], MSc; Jasmin Hennrich[3*], MSc; Jan Jöhnk[4*], PhD; Torsten Eymann[1,4*], Prof

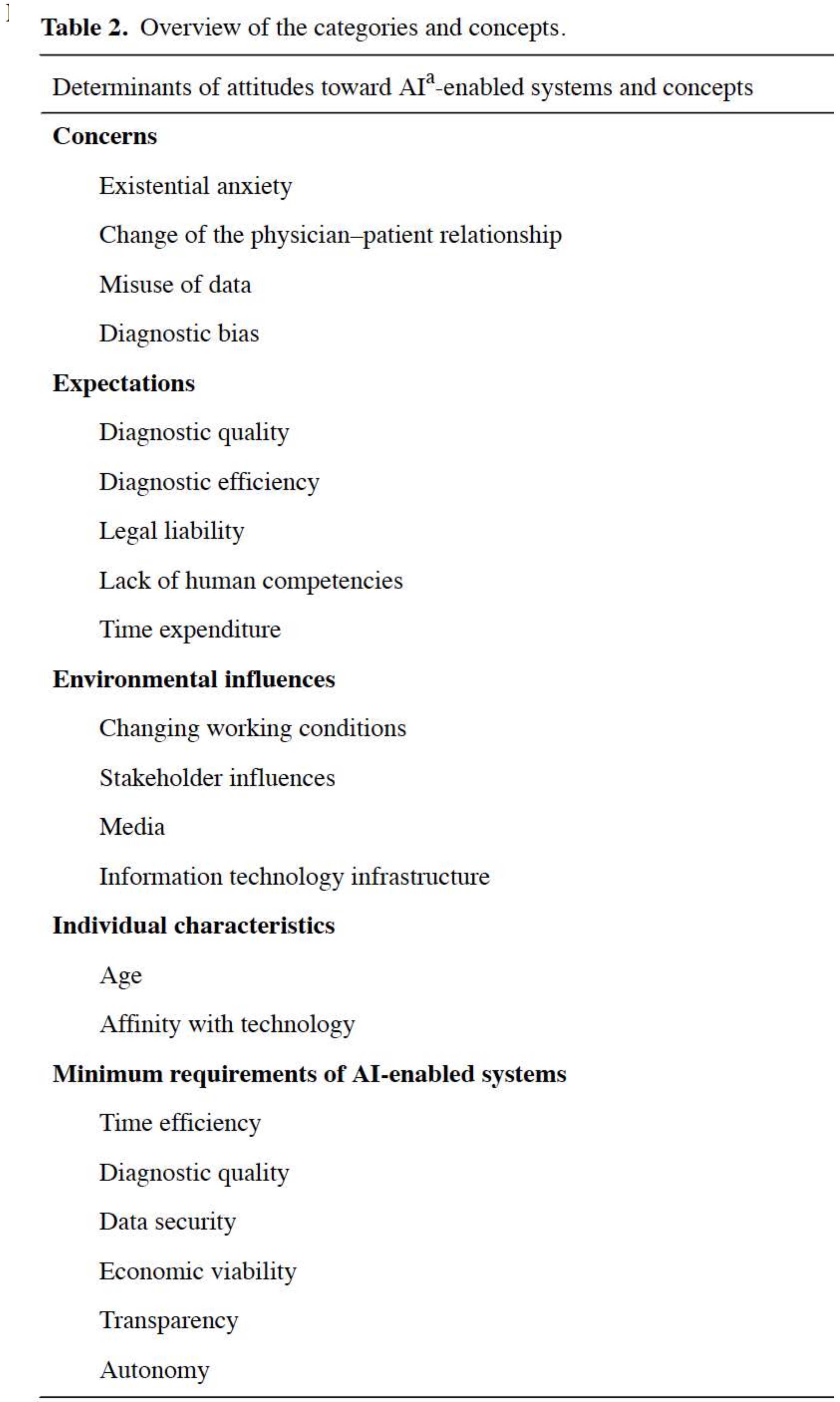
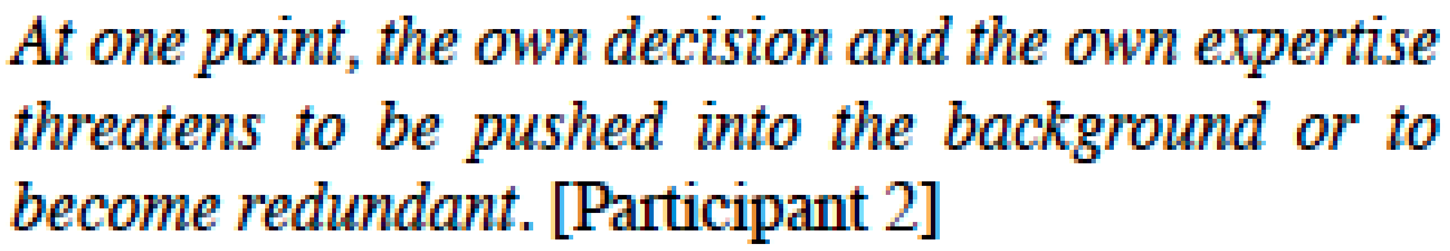
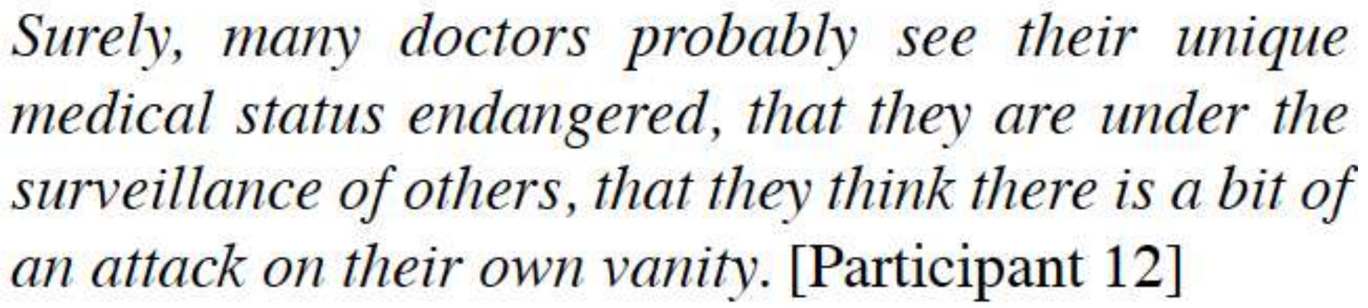


Table 2. Overview of the categories and concepts.

Determinants of attitudes toward AI[a]-enabled systems and concepts
Concerns
Existential anxiety
Change of the physician–patient relationship
Misuse of data
Diagnostic bias
Expectations
Diagnostic quality
Diagnostic efficiency
Legal liability
Lack of human competencies
Time expenditure
Environmental influences
Changing working conditions
Stakeholder influences
Media
Information technology infrastructure
Individual characteristics
Age
Affinity with technology
Minimum requirements of AI-enabled systems
Time efficiency
Diagnostic quality
Data security
Economic viability
Transparency
Autonomy

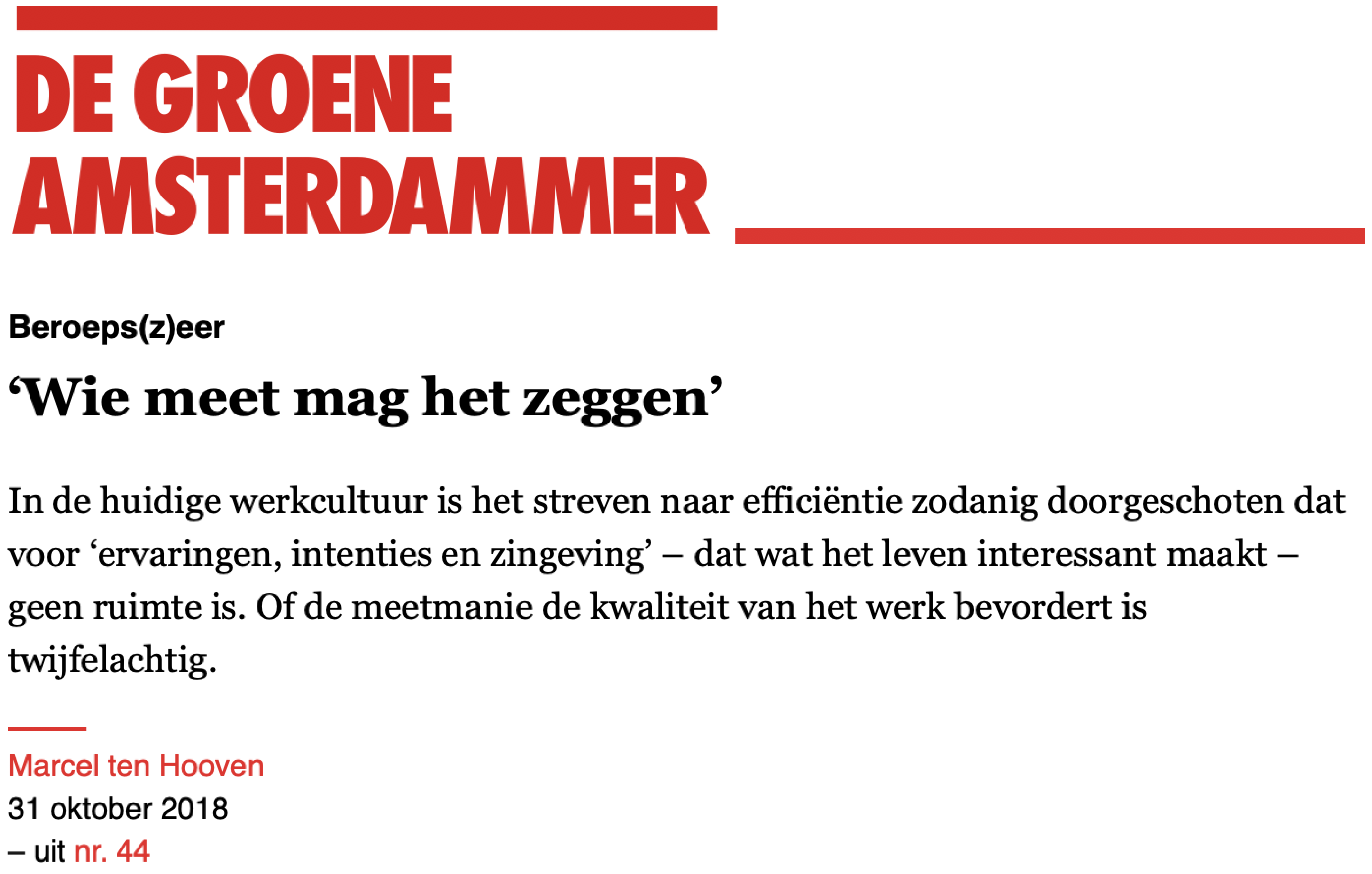


At one point, the own decision and the own expertise threatens to be pushed into the background or to become redundant. [Participant 2]



Surely, many doctors probably see their unique medical status endangered, that they are under the surveillance of others, that they think there is a bit of an attack on their own vanity. [Participant 12]

"Beroepszeer": Professional pride vs professional pain



DE GROENE AMSTERDAMMER

Beroeps(z)eer

‘Wie meet mag het zeggen’

In de huidige werkcultuur is het streven naar efficiëntie zodanig doorgeschoten dat voor ‘ervaringen, intenties en zingeving’ – dat wat het leven interessant maakt – geen ruimte is. Of de meetmanie de kwaliteit van het werk bevordert is twijfelachtig.

Marcel ten Hooven
31 oktober 2018
– uit nr. 44

Was ‘wie het weet mag het zeggen’ nog het parool van de oude werkcultuur, nu gold: ‘Wie meet mag het zeggen.’

‘Door het meet- en afrekensysteem ontstaat op het werk een **omgeving van zachte intimidatie waarin meegaandheid stilzwijgend wordt gewaardeerd en dwarsheid afgestraft.'**

General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study

J Med Internet Res 2022 | vol. 24 | iss. 1 | e28916 |

Christoph Buck[1,2*], PhD; Eileen Doctor[3*], MSc; Jasmin Hennrich[3*], MSc; Jan Jöhnk[4*], PhD; Torsten Eymann[1,4*], Prof

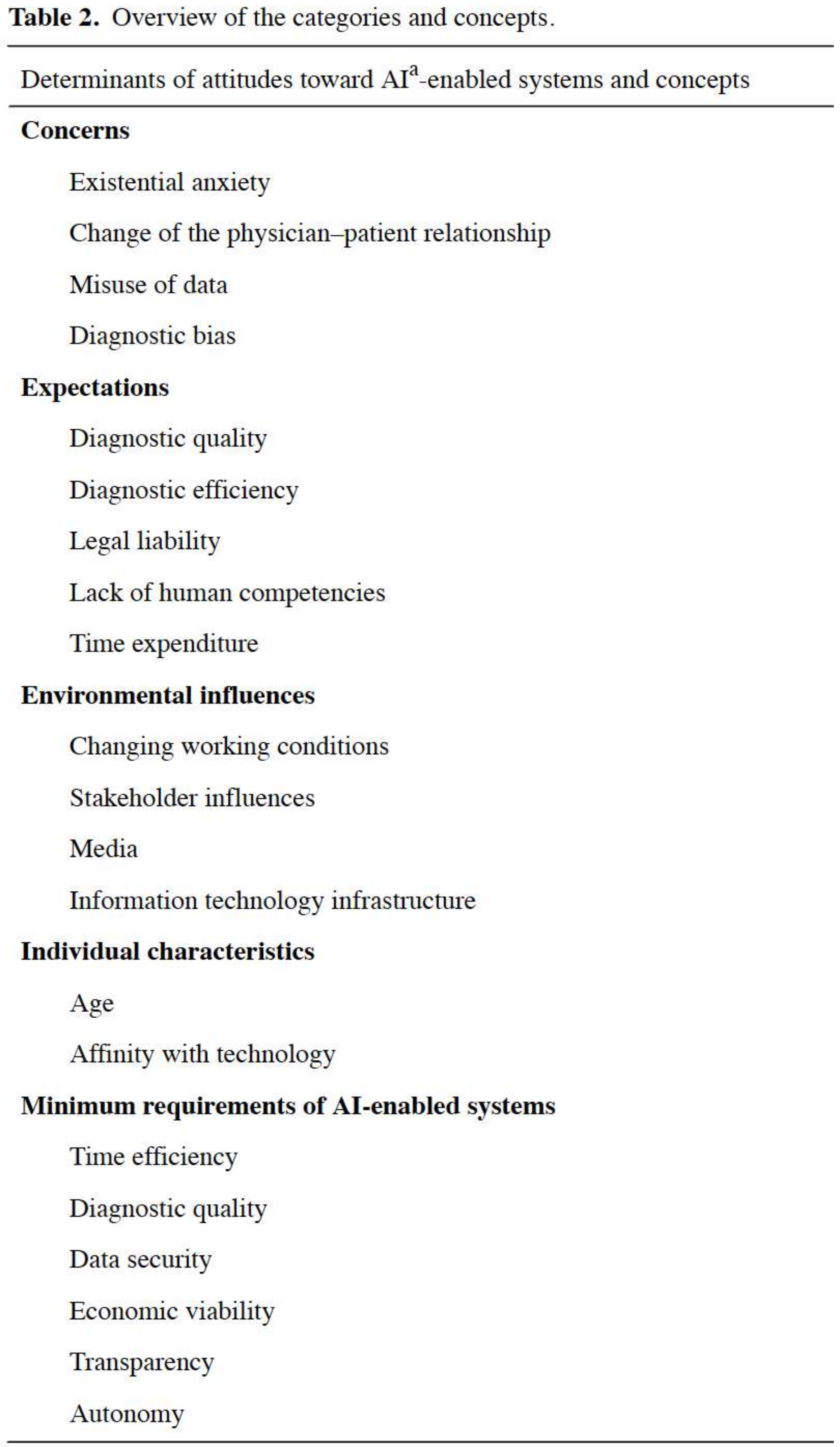
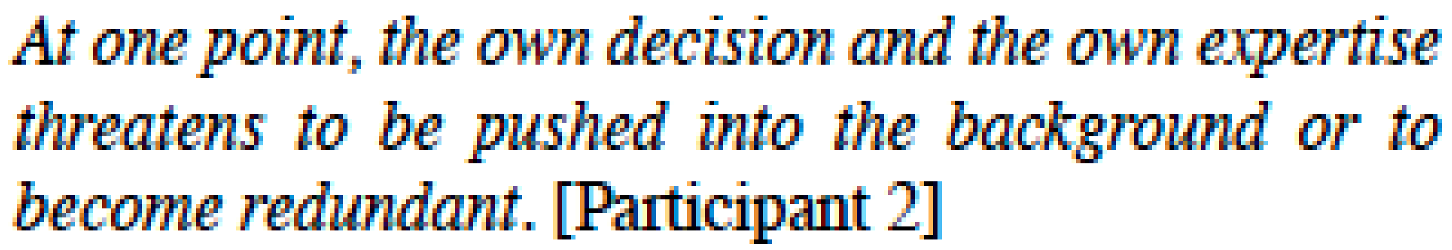
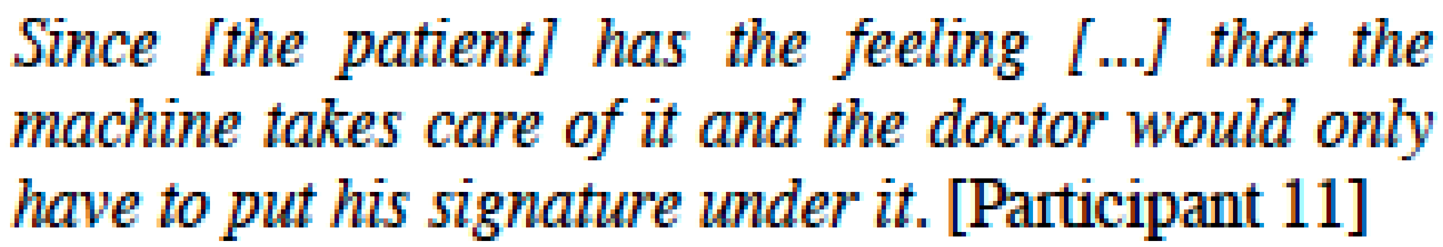


Table 2. Overview of the categories and concepts.

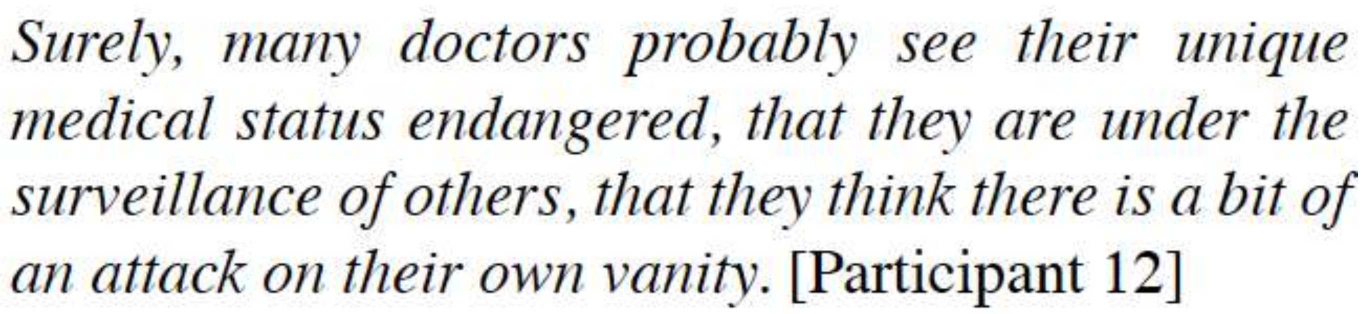
Determinants of attitudes toward AI[a]-enabled systems and concepts
Concerns
Existential anxiety
Change of the physician–patient relationship
Misuse of data
Diagnostic bias
Expectations
Diagnostic quality
Diagnostic efficiency
Legal liability
Lack of human competencies
Time expenditure
Environmental influences
Changing working conditions
Stakeholder influences
Media
Information technology infrastructure
Individual characteristics
Age
Affinity with technology
Minimum requirements of AI-enabled systems
Time efficiency
Diagnostic quality
Data security
Economic viability
Transparency
Autonomy



At one point, the own decision and the own expertise threatens to be pushed into the background or to become redundant. [Participant 2]



Since [the patient] has the feeling [...] that the machine takes care of it and the doctor would only have to put his signature under it. [Participant 11]



Surely, many doctors probably see their unique medical status endangered, that they are under the surveillance of others, that they think there is a bit of an attack on their own vanity. [Participant 12]

The ‘weather forecaster’ problem

AI moves authority away from professional experts

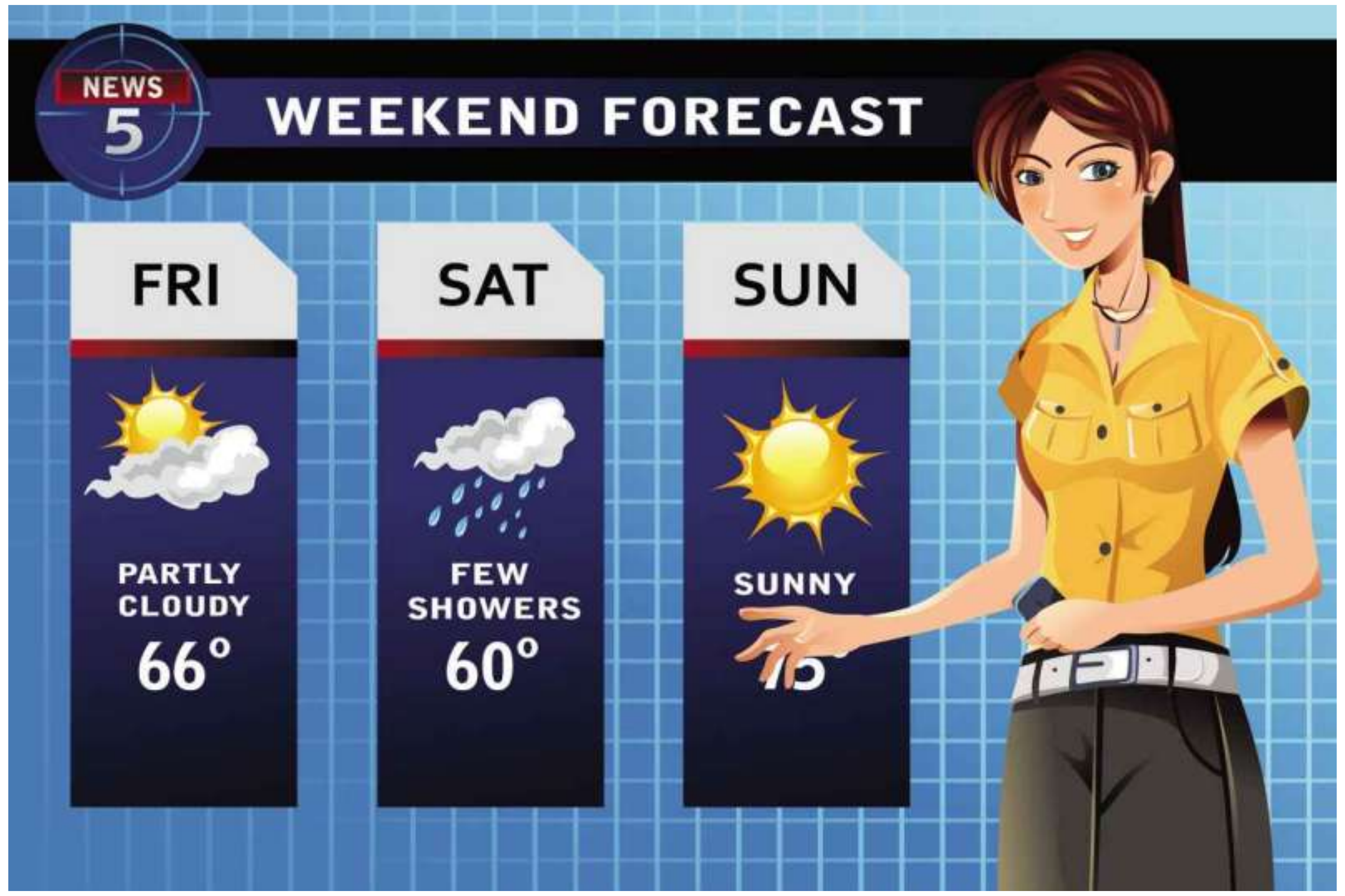
Human supervision of AI as ‘mere stamp of approval’ not only leads to accountability confusions but can also undermine professional pride and commitment

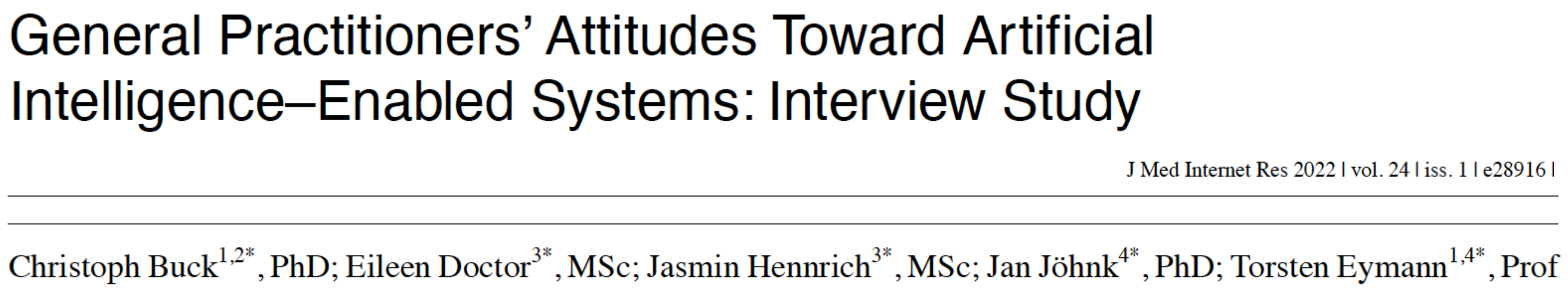
It may lead to the ‘weather forecaster’ problem

They used to be weather experts that, because of their expertise, started presenting on TV

Soon these weather experts were judged mostly by their qualities as communicators

AI may nudge professional experts into presenters





General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study

J Med Internet Res 2022 | vol. 24 | iss. 1 | e28916 |

Christoph Buck[1,2*], PhD; Eileen Doctor[3*], MSc; Jasmin Hennrich[3*], MSc; Jan Jöhnk[4*], PhD; Torsten Eymann[1,4*], Prof

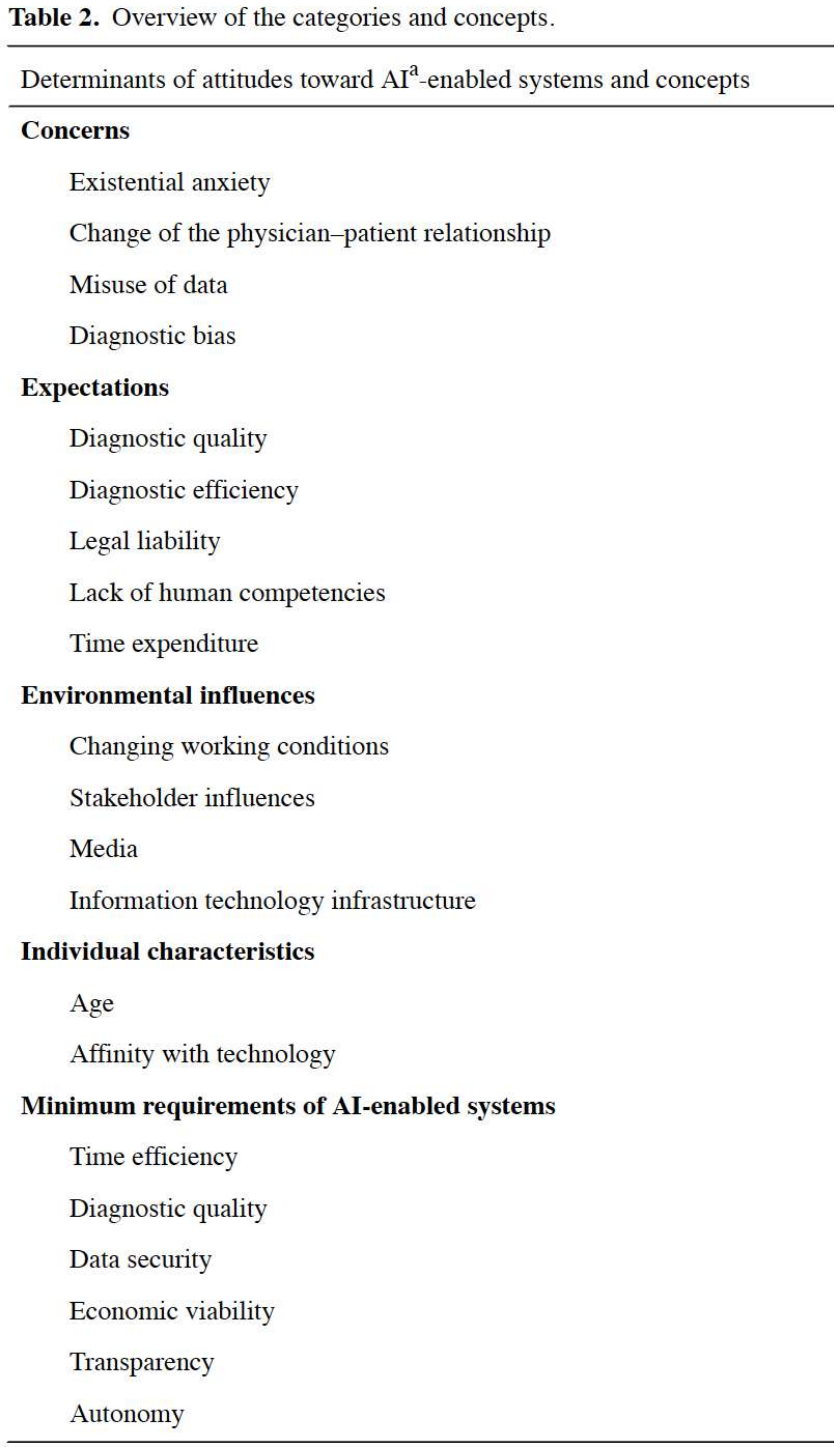
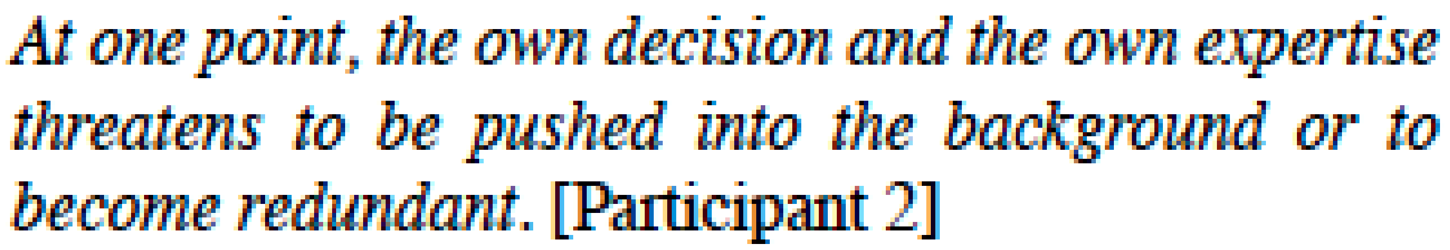
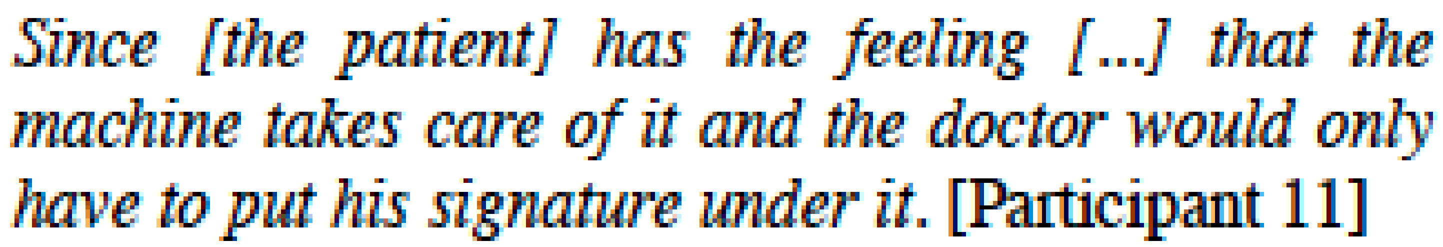


Table 2. Overview of the categories and concepts.

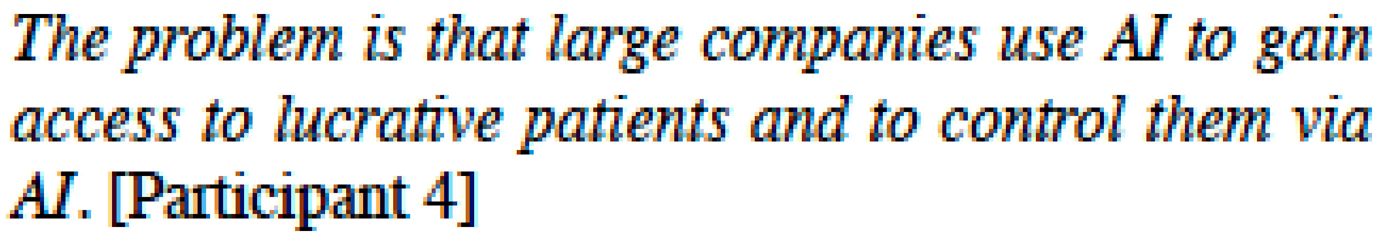
Determinants of attitudes toward AI[a]-enabled systems and concepts
Concerns
Existential anxiety
Change of the physician–patient relationship
Misuse of data
Diagnostic bias
Expectations
Diagnostic quality
Diagnostic efficiency
Legal liability
Lack of human competencies
Time expenditure
Environmental influences
Changing working conditions
Stakeholder influences
Media
Information technology infrastructure
Individual characteristics
Age
Affinity with technology
Minimum requirements of AI-enabled systems
Time efficiency
Diagnostic quality
Data security
Economic viability
Transparency
Autonomy



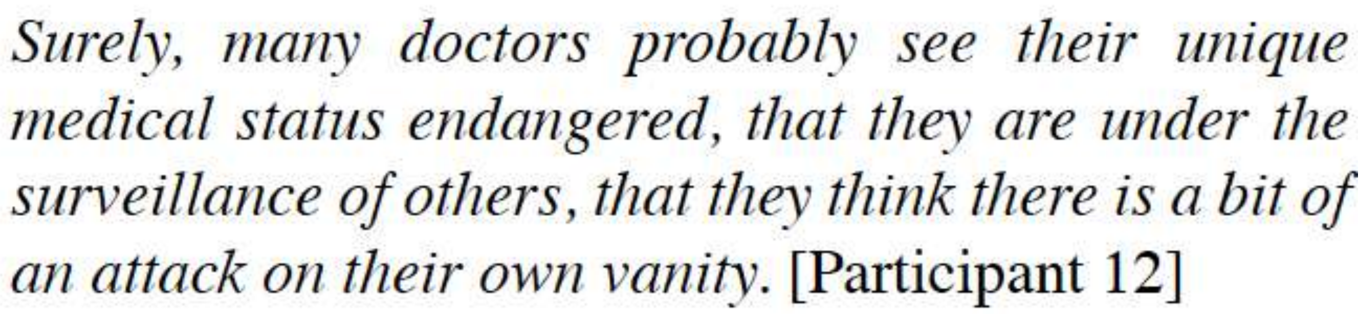
At one point, the own decision and the own expertise threatens to be pushed into the background or to become redundant. [Participant 2]



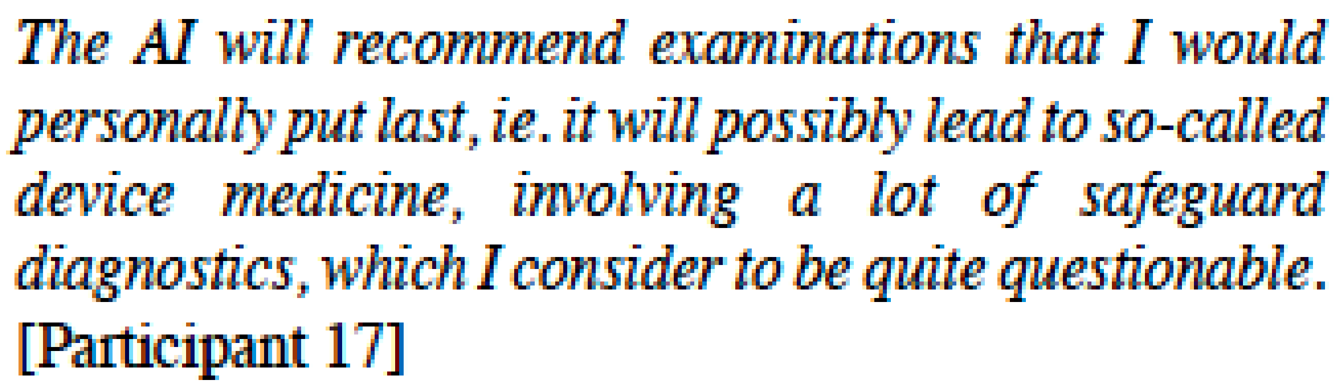
Since [the patient] has the feeling [...] that the machine takes care of it and the doctor would only have to put his signature under it. [Participant 11]



The problem is that large companies use AI to gain access to lucrative patients and to control them via AI. [Participant 4]



Surely, many doctors probably see their unique medical status endangered, that they are under the surveillance of others, that they think there is a bit of an attack on their own vanity. [Participant 12]



The AI will recommend examinations that I would personally put last, ie. it will possibly lead to so-called device medicine, involving a lot of safeguard diagnostics, which I consider to be quite questionable. [Participant 17]

Effects of big data & AI on users

“as the use of smart technology grows among the general population, some predict an accompanying **blurring of medical and everyday devices**”

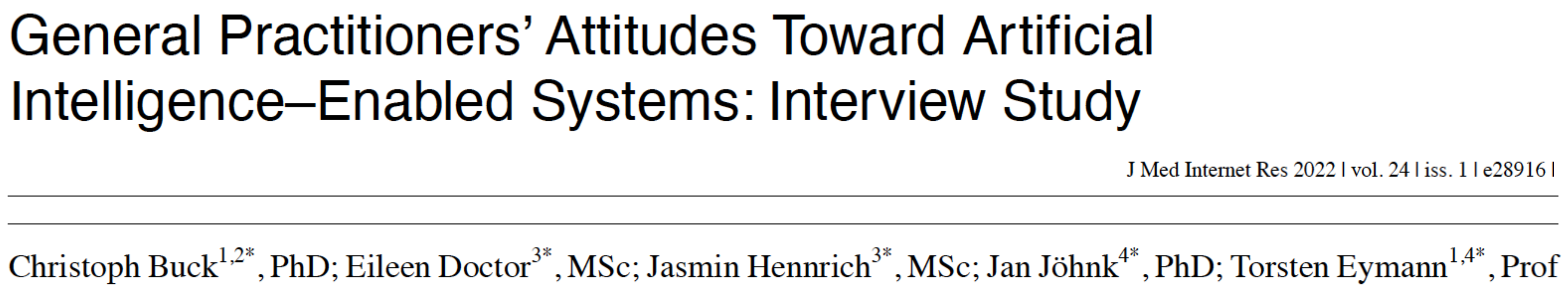
“While private tech companies are hardly new players in the medical sector, the presence of **multinational, corporate firms at an intimate bodily level on a day-to-day basis** implies a **fundamental shift** in what we mean by the **‘doctor–patient relationship**’”

Harriet Barratt, medical humanities, U of Sussex, UK

“the introduction of apps (…) inserts **more distance** (both physical and emotional) between doctor and patient”

More generally: AI may affect the **nature** of the relationship between doctor, patient and significant others

Natalie Harrison
Regressing or progressing: what next for the doctor–patient relationship?
Lancet, vol 6, march 2018



General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study

J Med Internet Res 2022 | vol. 24 | iss. 1 | e28916 |

Christoph Buck[1,2*], PhD; Eileen Doctor[3*], MSc; Jasmin Hennrich[3*], MSc; Jan Jöhnk[4*], PhD; Torsten Eymann[1,4*], Prof

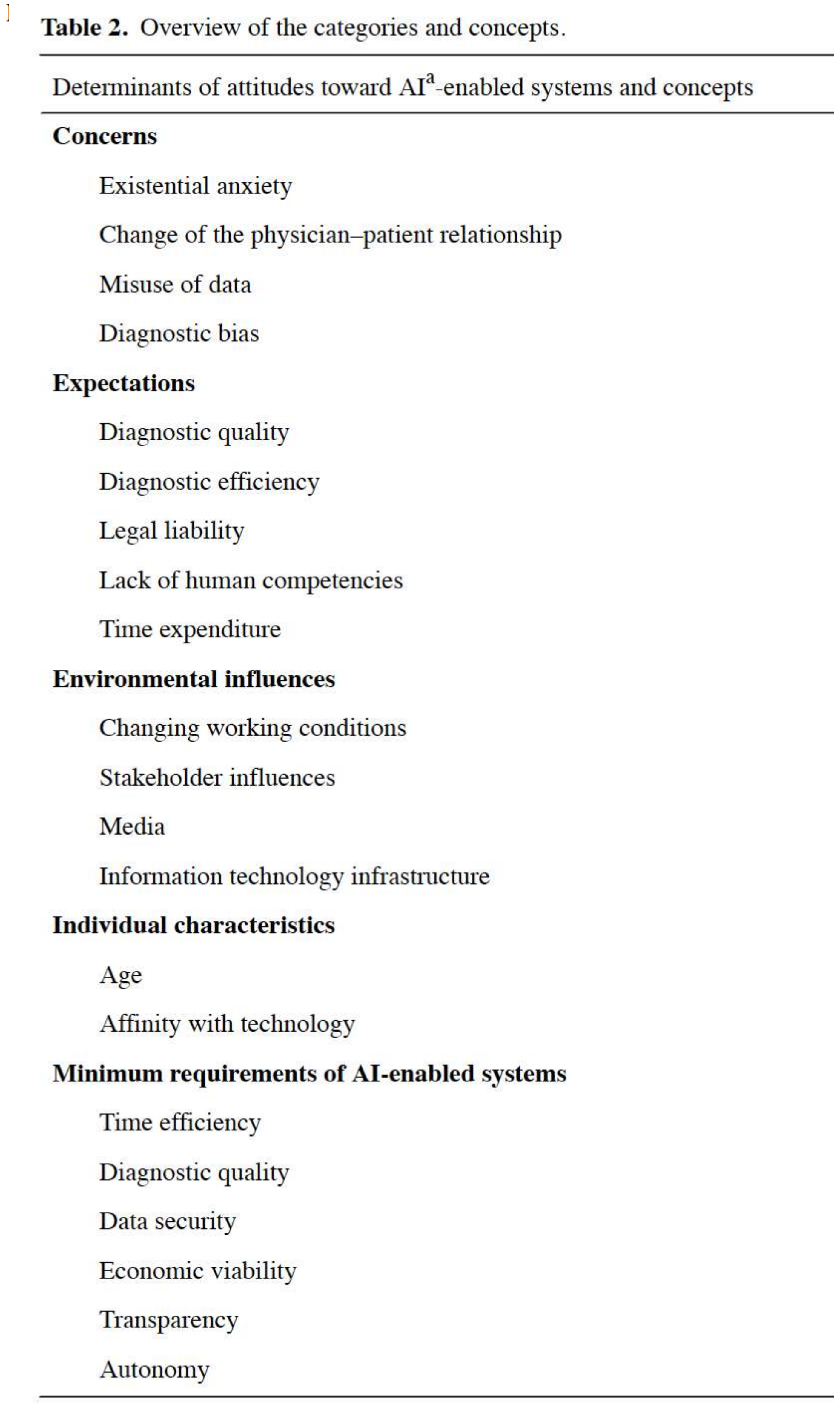
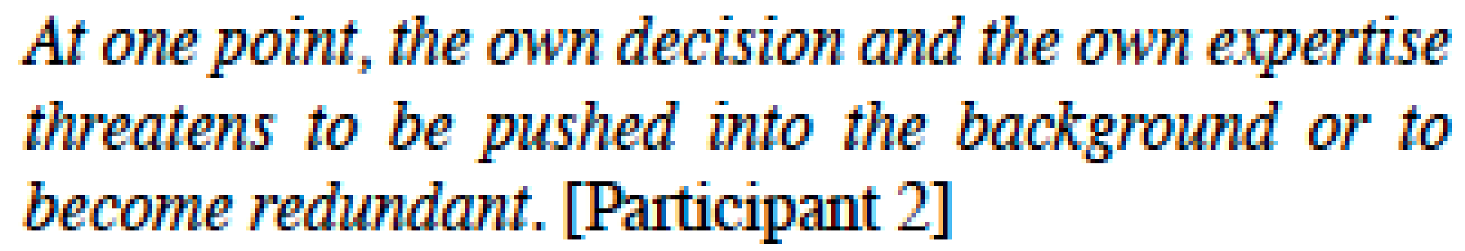
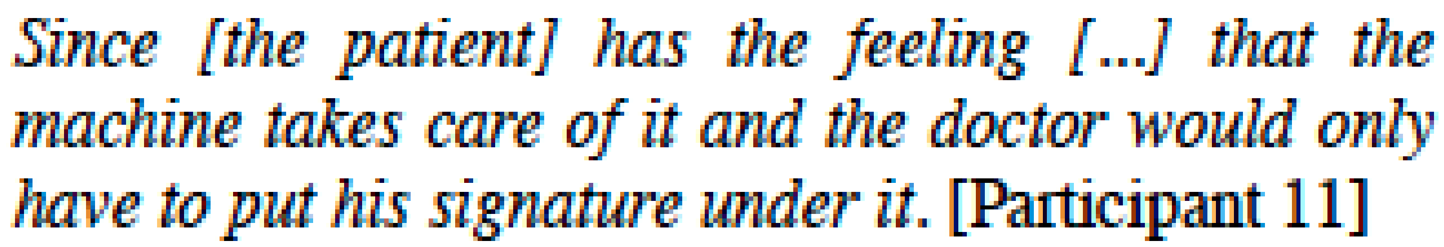


Table 2. Overview of the categories and concepts.

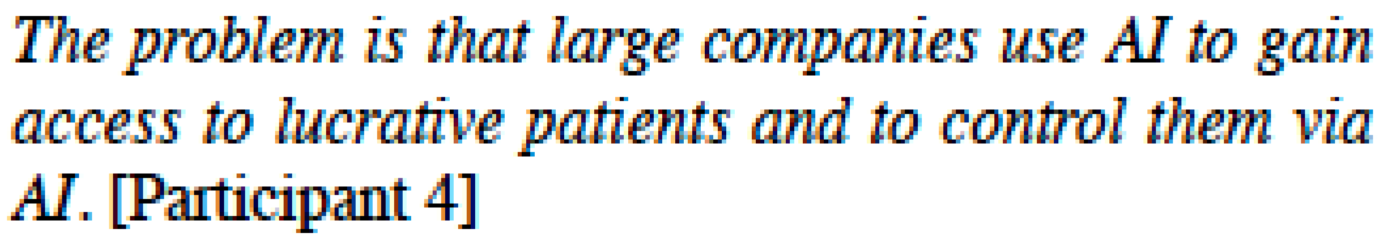
Determinants of attitudes toward AI[a]-enabled systems and concepts
Concerns
Existential anxiety
Change of the physician–patient relationship
Misuse of data
Diagnostic bias
Expectations
Diagnostic quality
Diagnostic efficiency
Legal liability
Lack of human competencies
Time expenditure
Environmental influences
Changing working conditions
Stakeholder influences
Media
Information technology infrastructure
Individual characteristics
Age
Affinity with technology
Minimum requirements of AI-enabled systems
Time efficiency
Diagnostic quality
Data security
Economic viability
Transparency
Autonomy



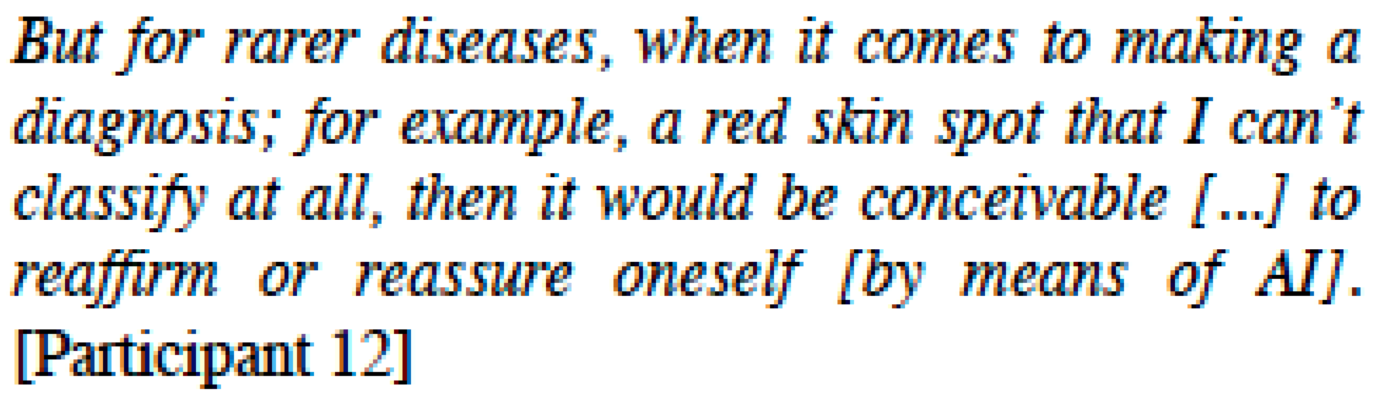
At one point, the own decision and the own expertise threatens to be pushed into the background or to become redundant. [Participant 2]



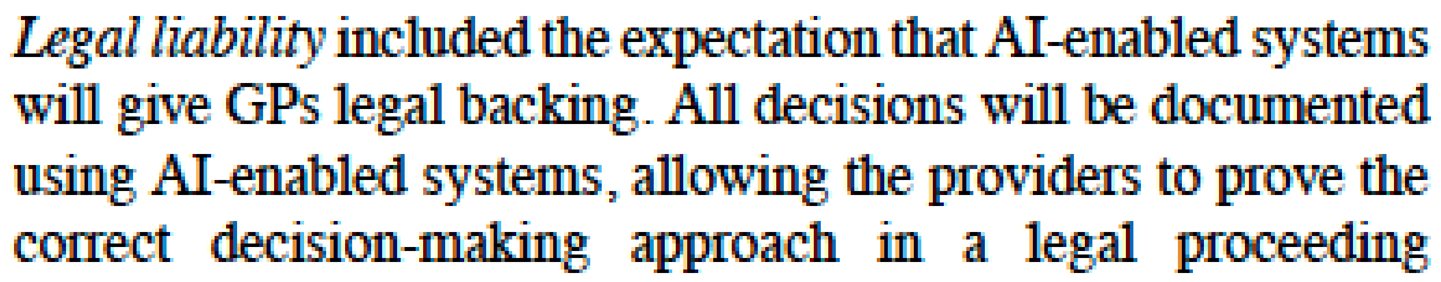
Since [the patient] has the feeling [...] that the machine takes care of it and the doctor would only have to put his signature under it. [Participant 11]



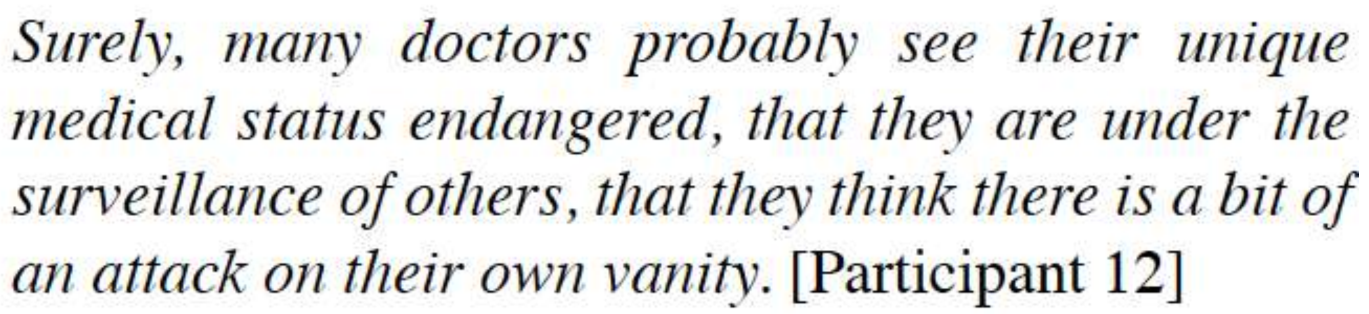
The problem is that large companies use AI to gain access to lucrative patients and to control them via AI. [Participant 4]



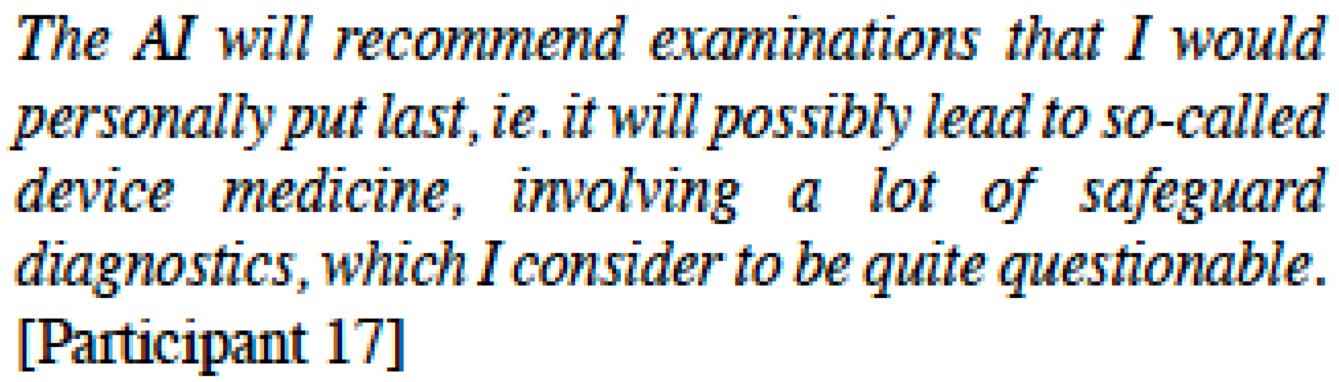
But for rarer diseases, when it comes to making a diagnosis; for example, a red skin spot that I can't classify at all, then it would be conceivable [...] to reaffirm or reassure oneself [by means of AI]. [Participant 12]



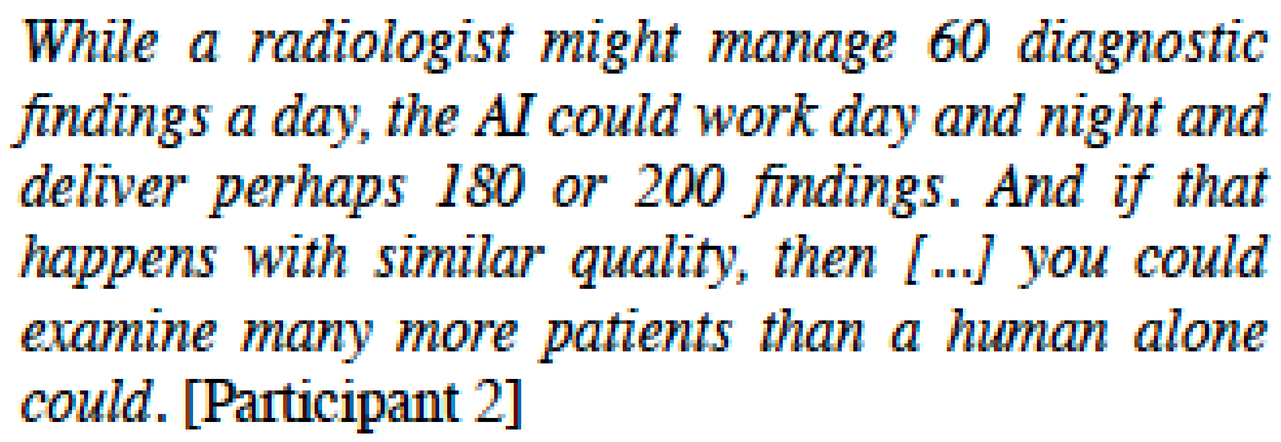
Legal liability included the expectation that AI-enabled systems will give GPs legal backing. All decisions will be documented using AI-enabled systems, allowing the providers to prove the correct decision-making approach in a legal proceeding



Surely, many doctors probably see their unique medical status endangered, that they are under the surveillance of others, that they think there is a bit of an attack on their own vanity. [Participant 12]



The AI will recommend examinations that I would personally put last, ie. it will possibly lead to so-called device medicine, involving a lot of safeguard diagnostics, which I consider to be quite questionable. [Participant 17]



While a radiologist might manage 60 diagnostic findings a day, the AI could work day and night and deliver perhaps 180 or 200 findings. And if that happens with similar quality, then [...] you could examine many more patients than a human alone could. [Participant 2]

Human-AI interaction: on or under the loop?

Three categories based on the amount of human involvement in AI-mediated actions:

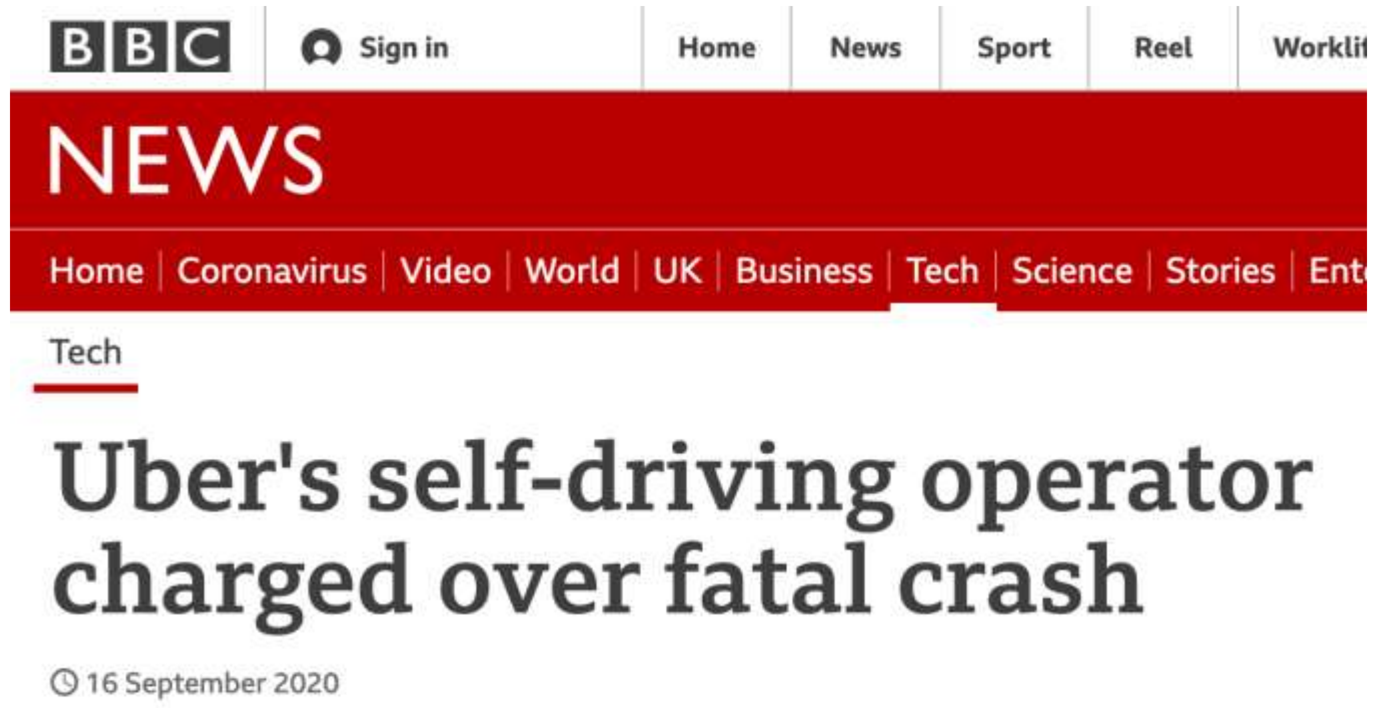
- **Human-*in*-the-Loop**: AI based decisions become effective only with a human command
- **Human-*on*-the-Loop**: AI based decisions become effective under the supervision of a human operator who can override the robots’ actions

  - **Human-*under*-the-Loop**: ”Having human beings ‘in’ or ‘on’ the loop with regards to AI systems might mask the power such systems exercise over human beings” Liu (2018)
- **Humans-in-command**: AI based decisions are assessed on its broader economic, societal, legal and ethical impact, and human decides when and how to use the AI.
- **Human-*out-of*-the-Loop**: AI based decisions become effective without any human input or interaction

Reduced control over AI-based decision making may lead to **responsibility gaps**

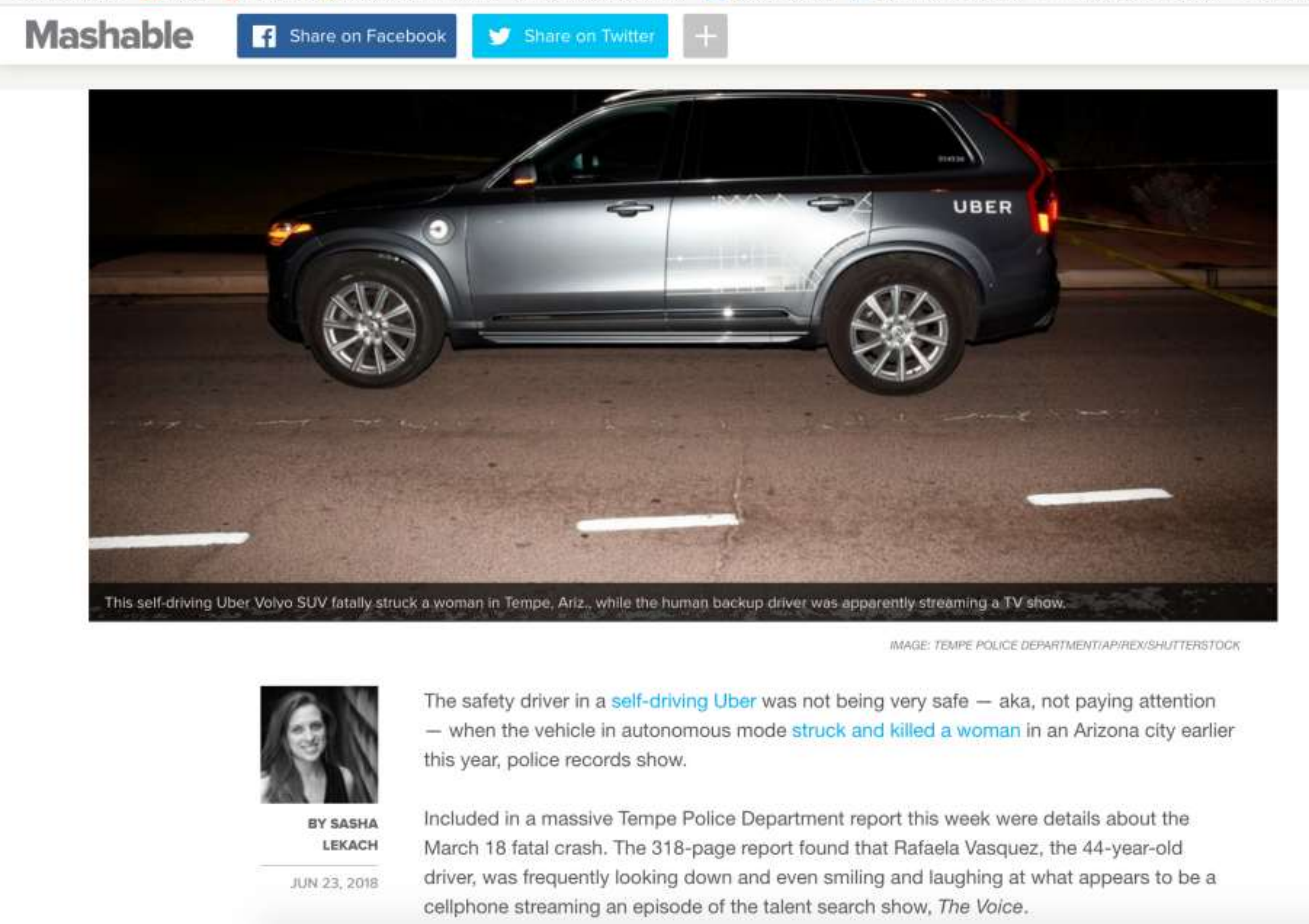
Hin-Yan Liu (2018) The power structure of artificial intelligence, Law, Innovation and Technology, 10:2, 197-229, DOI: 10.1080/17579961.2018.1527480

Self-driving cars: humans ‘under’ the loop?



Uber's self-driving operator charged over fatal crash

16 September 2020



This self-driving Uber Volvo SUV fatally struck a woman in Tempe, Ariz., while the human backup driver was apparently streaming a TV show.

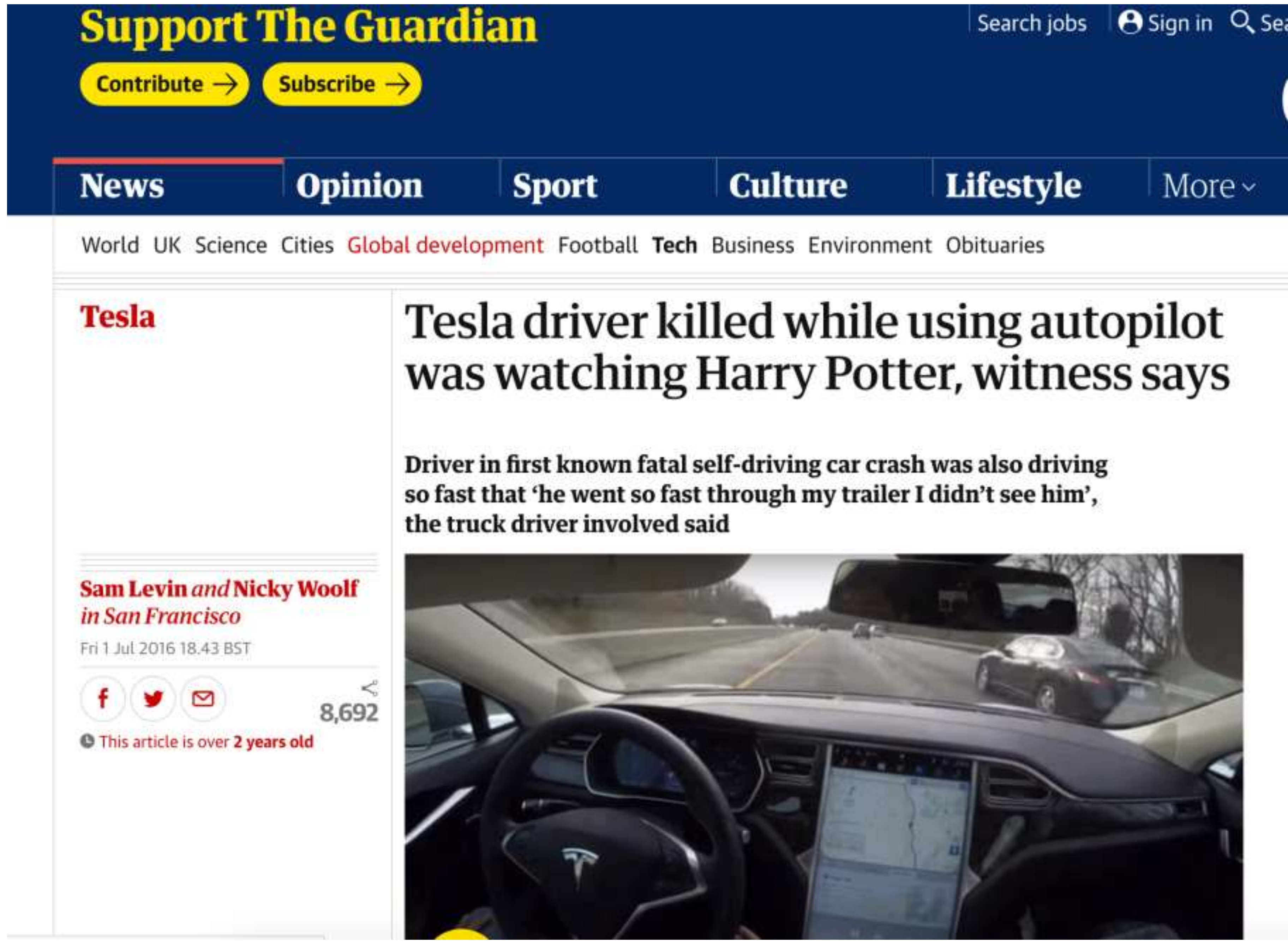
IMAGE: TEMPE POLICE DEPARTMENT/AP/REX/SHUTTERSTOCK

BY SASHA LEKACH

JUN 23, 2018

The safety driver in a self-driving Uber was not being very safe — aka, not paying attention — when the vehicle in autonomous mode struck and killed a woman in an Arizona city earlier this year, police records show.

Included in a massive Tempe Police Department report this week were details about the March 18 fatal crash. The 318-page report found that Rafaela Vasquez, the 44-year-old driver, was frequently looking down and even smiling and laughing at what appears to be a cellphone streaming an episode of the talent search show, *The Voice*.



Tesla driver killed while using autopilot was watching Harry Potter, witness says

Driver in first known fatal self-driving car crash was also driving so fast that ‘he went so fast through my trailer I didn’t see him’, the truck driver involved said

Sam Levin *and* **Nicky Woolf** *in San Francisco*

Fri 1 Jul 2016 18.43 BST

8,692

This article is over **2 years old**

Under the loop?

Scapegoats

“Potential for **scapegoating** proximate human beings because conventional responsibility structures struggle to apportion responsibility to artificial entities.

This renders the human being as a moral crumple zone”

Hin-Yan Liu (2018)

Moral crumple zones

“Just as the crumple zone in a car is designed to absorb the force of impact in a crash, **the human** in a highly complex and automated system may become simply a component – accidentally or intentionally – that **bears the brunt of the moral and legal responsibilities** when the overall system malfunctions.”

Elish (2016)

Technology driven ‘**entrapment**'?

Hin-Yan Liu (2018) The power structure of artificial intelligence, Law, Innovation and Technology, 10:2, 197-229, DOI: 10.1080/17579961.2018.1527480
Elish, ‘Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction’ WeRobot 2016 (2016) 3–4.

3 aspects of dealing with AI in an organization

a. Precaution
What are the rules regarding use of AI? How is the organisation structured so that it enables following these rules? What are considered to be ‘grey zones’ and how should they be dealt with?

b. Recognise and acknowledge
What are the minimal checks required on AI output? Who is responsible? What should be reported in this regard? To whom, when?

c. Aftercare
What are the procedures when something went wrong (communication, potential consequences for person & organisation, lessons to be learned, adaptation of procedures)?

Education
Output quality ± = genAI + prompt (+ output check)

Treat the genAI like an agent
  It’s not a document or a fact, but an agent that is part of a (commercial) organization, with its own limited perspective, (sub)goals and limited knowledge (e.g. selective database)

genAI quality may be primarily the responsibility of the organization(s) that enable/allow its usage
  What genAI is made available, from which provider (protected server, jurisdiction, ownership/access)?
  How are quality of LLM training, database, hallucination risk, explainability & transparency checked?
  Has training in prompt engineering been provided?

The actual prompt quality & output check is primarily the responsibility of the individual user
  Include this in the evaluation, so that eg. a student has to:
    Clarify which prompts have been used (& why)
    Specify how output was checked
    State whether additional tools have been used (which (e.g. genAI for illustrations), how, etc)?

Such issues (and undoubtedly many more) may need to be incorporated in educational programs
    Move from product evaluation to process evaluation
    Consider at which stage of training genAI usage is educationally meaningful

Education
Output quality ± = genAI + prompt (+ output check)

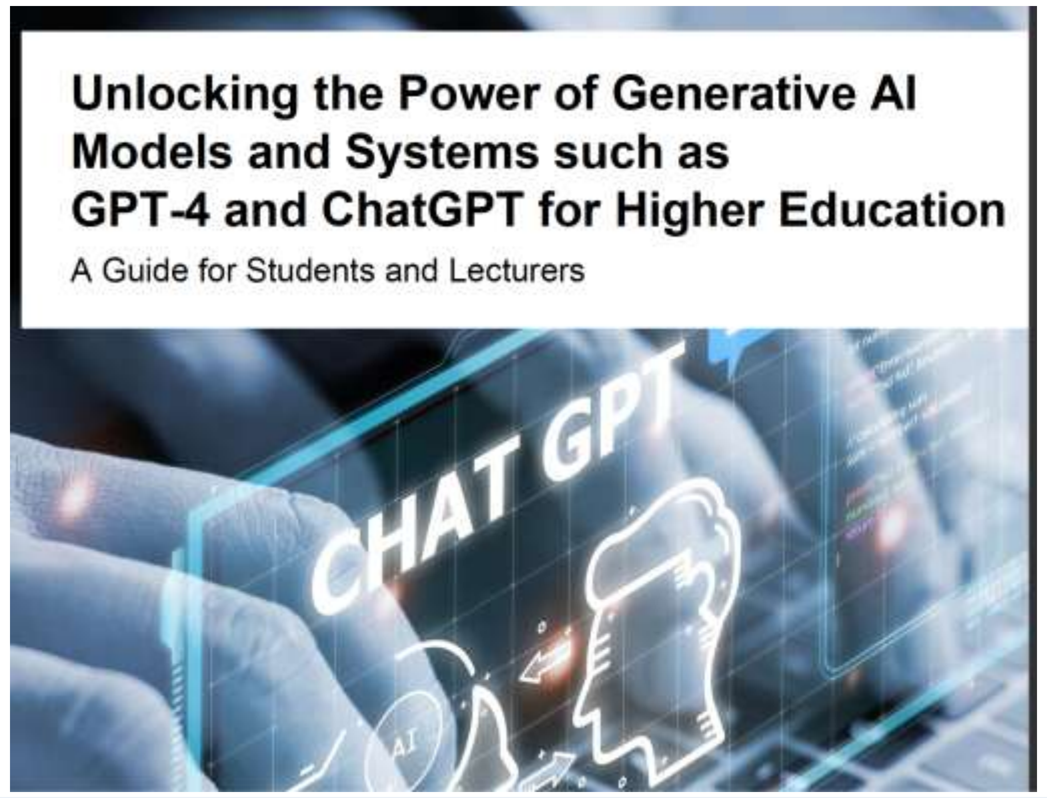
Treat the genAI like an agent

- It's not a document or a fact, but an agent that is part of a (commercial) organization, with its own perspective, interest and limited knowledge (e.g. selective database)

genAI quality may be primarily the responsibility of the organization(s) that enable/allow its usage

- What genAI is made available, from which provider (protected server, jurisdiction, ownership/access)?
- How are quality of LLM training, database, hallucination risk, explainability & transparency checked?
- Has training in prompt engineering been provided?

Prompt engineering

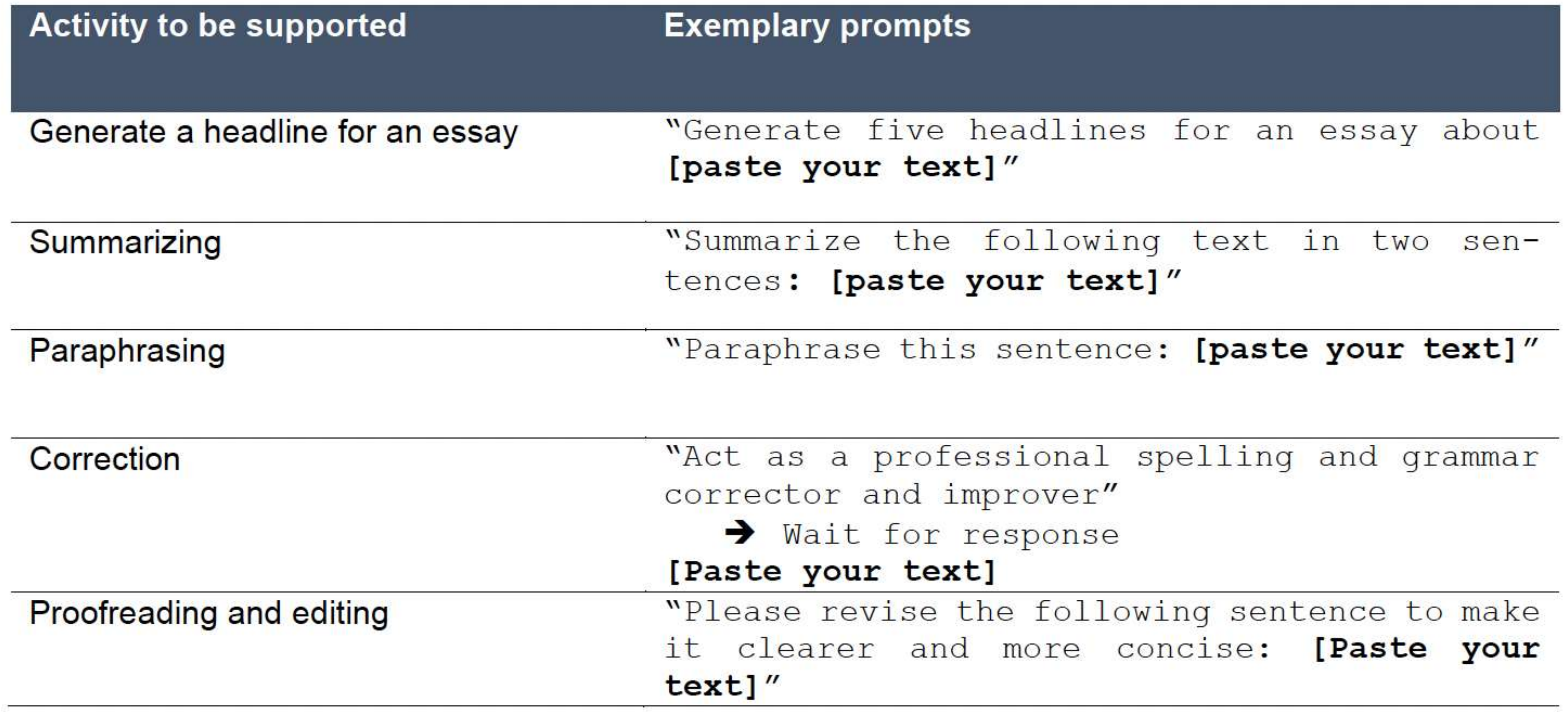


Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education

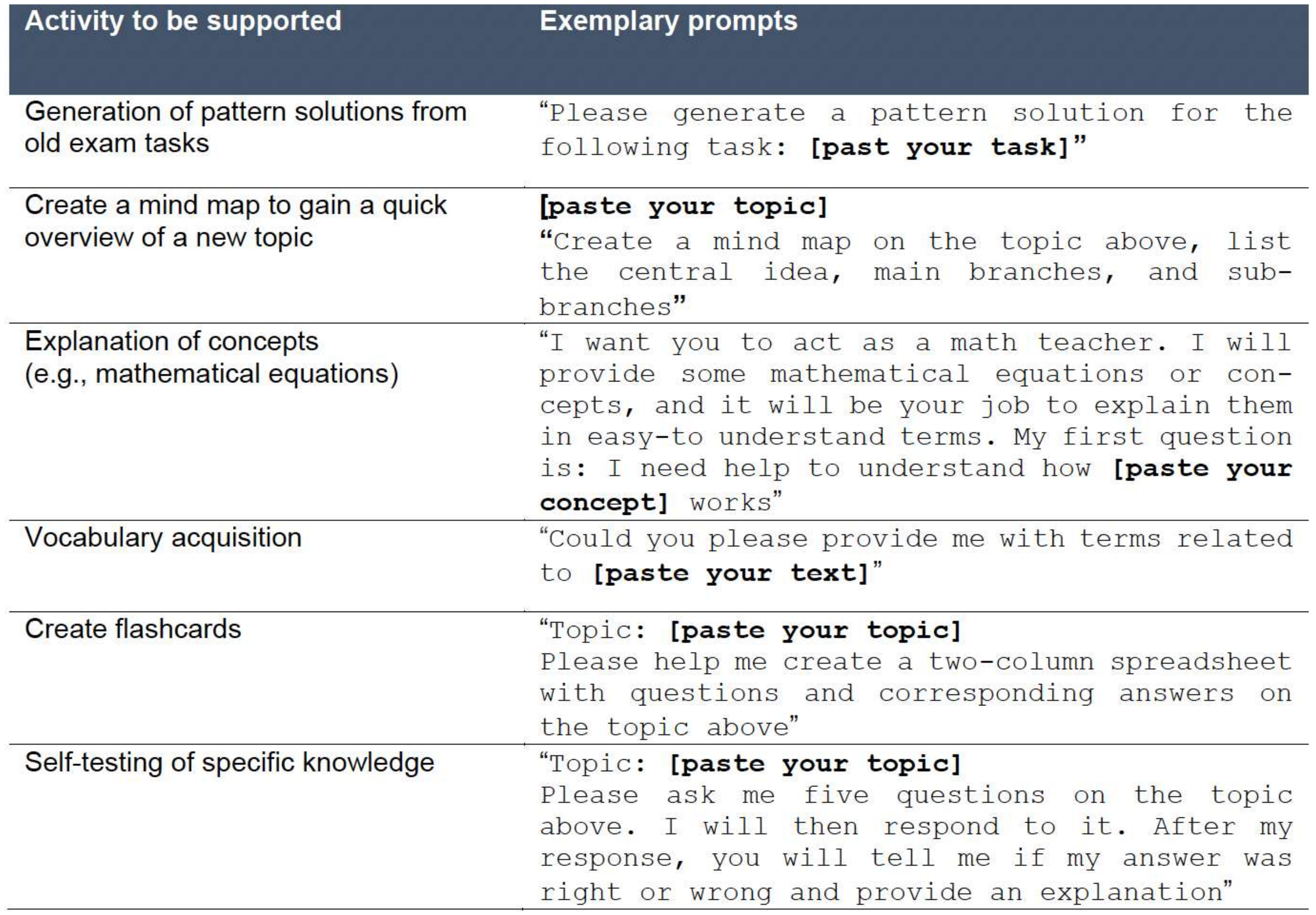
A Guide for Students and Lecturers

Hohenheim Discussion Papers in Business, Economics and Social Sciences, No. 02-2023

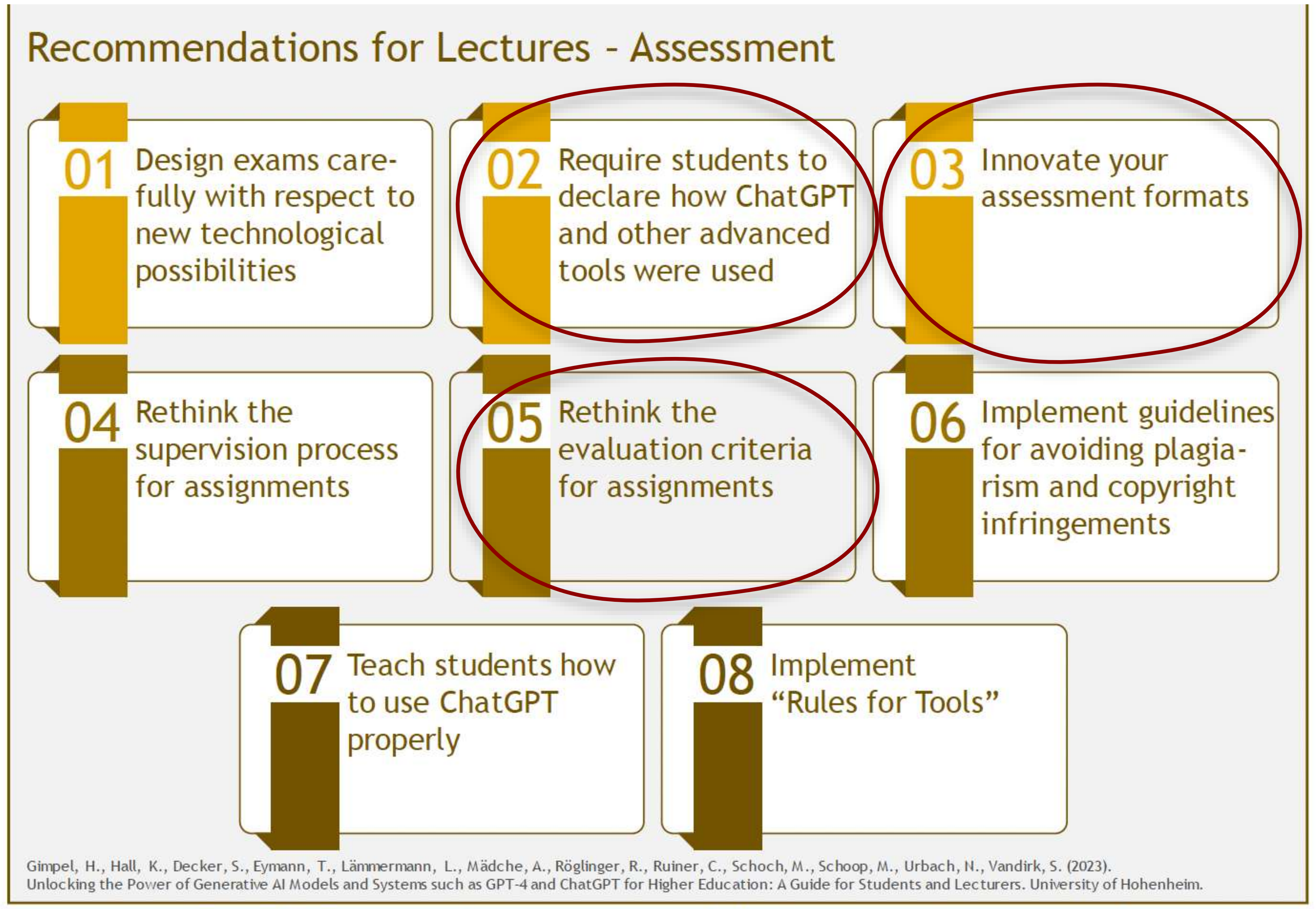
Table 1. Exemplary prompts on writing text



Activity to be supported	Exemplary prompts
Generate a headline for an essay	“Generate five headlines for an essay about **[paste your text]**”
Summarizing	“Summarize the following text in two sentences: **[paste your text]**”
Paraphrasing	“Paraphrase this sentence: **[paste your text]**”
Correction	“Act as a professional spelling and grammar corrector and improver” ➔ Wait for response **[Paste your text]**
Proofreading and editing	“Please revise the following sentence to make it clearer and more concise: **[Paste your text]**”



Activity to be supported	Exemplary prompts
Generation of pattern solutions from old exam tasks	“Please generate a pattern solution for the following task: **[past your task]**”
Create a mind map to gain a quick overview of a new topic	**[paste your topic]** “Create a mind map on the topic above, list the central idea, main branches, and sub-branches”
Explanation of concepts (e.g., mathematical equations)	“I want you to act as a math teacher. I will provide some mathematical equations or concepts, and it will be your job to explain them in easy-to understand terms. My first question is: I need help to understand how **[paste your concept]** works”
Vocabulary acquisition	“Could you please provide me with terms related to **[paste your text]**”
Create flashcards	“Topic: **[paste your topic]** Please help me create a two-column spreadsheet with questions and corresponding answers on the topic above”
Self-testing of specific knowledge	“Topic: **[paste your topic]** Please ask me five questions on the topic above. I will then respond to it. After my response, you will tell me if my answer was right or wrong and provide an explanation”



Recommendations for Lectures - Assessment

- 01 Design exams carefully with respect to new technological possibilities
- 02 Require students to declare how ChatGPT and other advanced tools were used
- 03 Innovate your assessment formats
- 04 Rethink the supervision process for assignments
- 05 Rethink the evaluation criteria for assignments
- 06 Implement guidelines for avoiding plagiarism and copyright infringements
- 07 Teach students how to use ChatGPT properly
- 08 Implement “Rules for Tools”

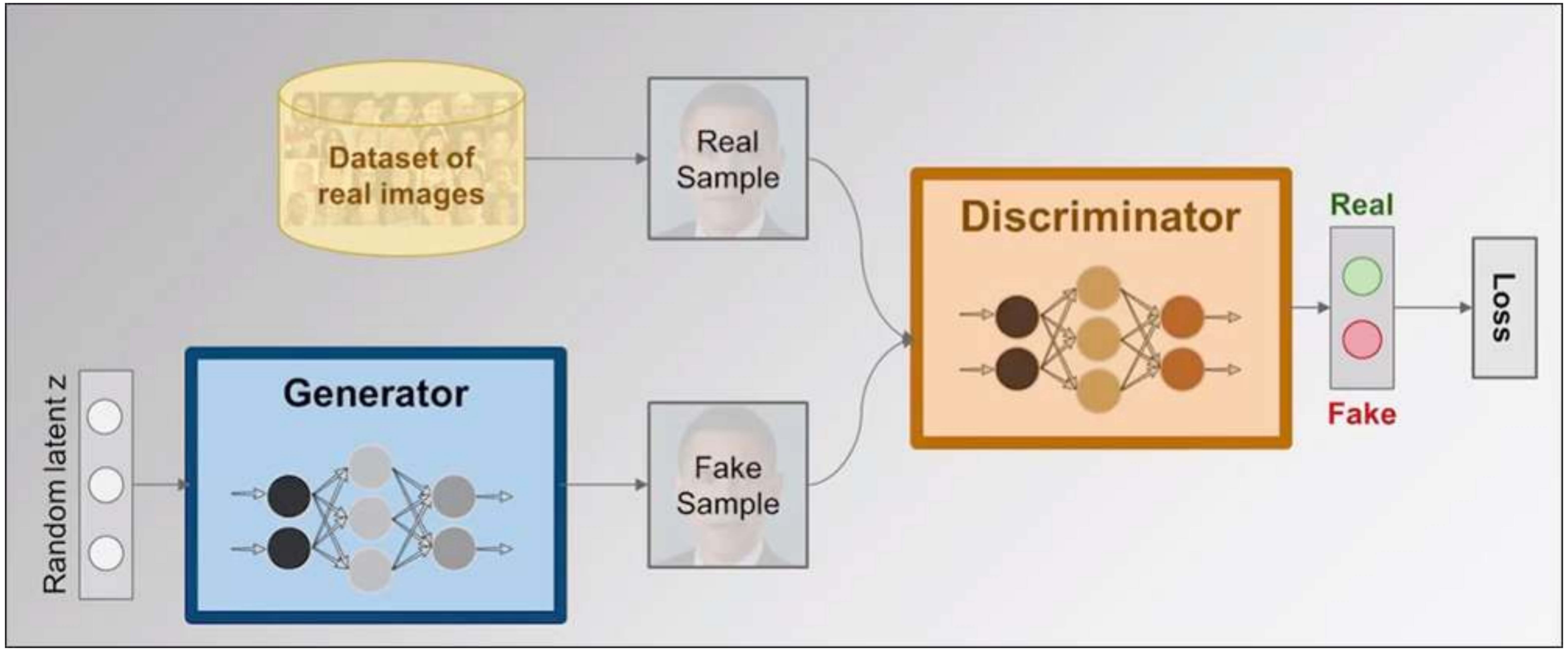
Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., Röglinger, R., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., Vandirk, S. (2023). Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education: A Guide for Students and Lecturers. University of Hohenheim.

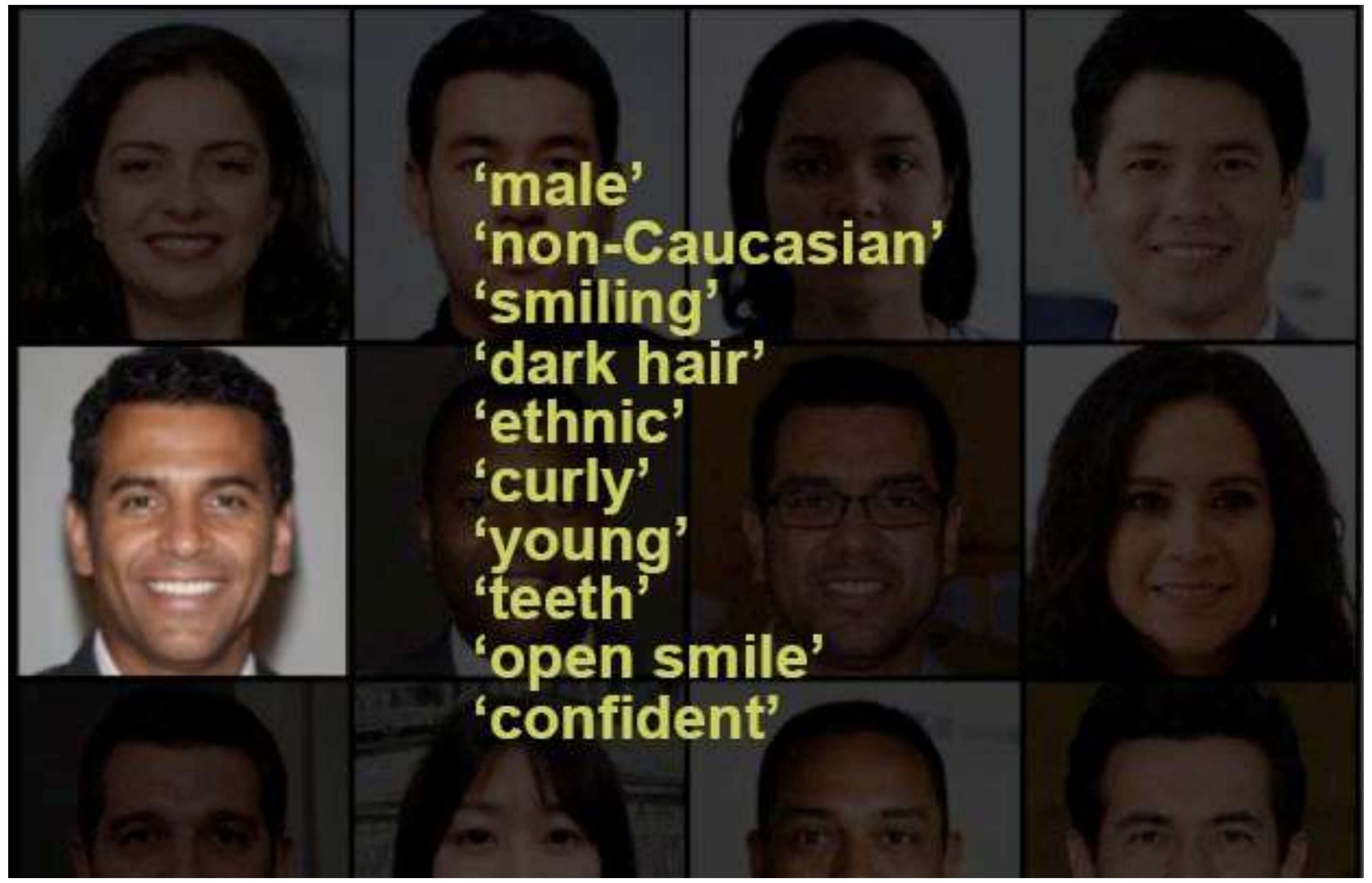
Figure 11. Summary of recommendations for lecturers regarding assessments

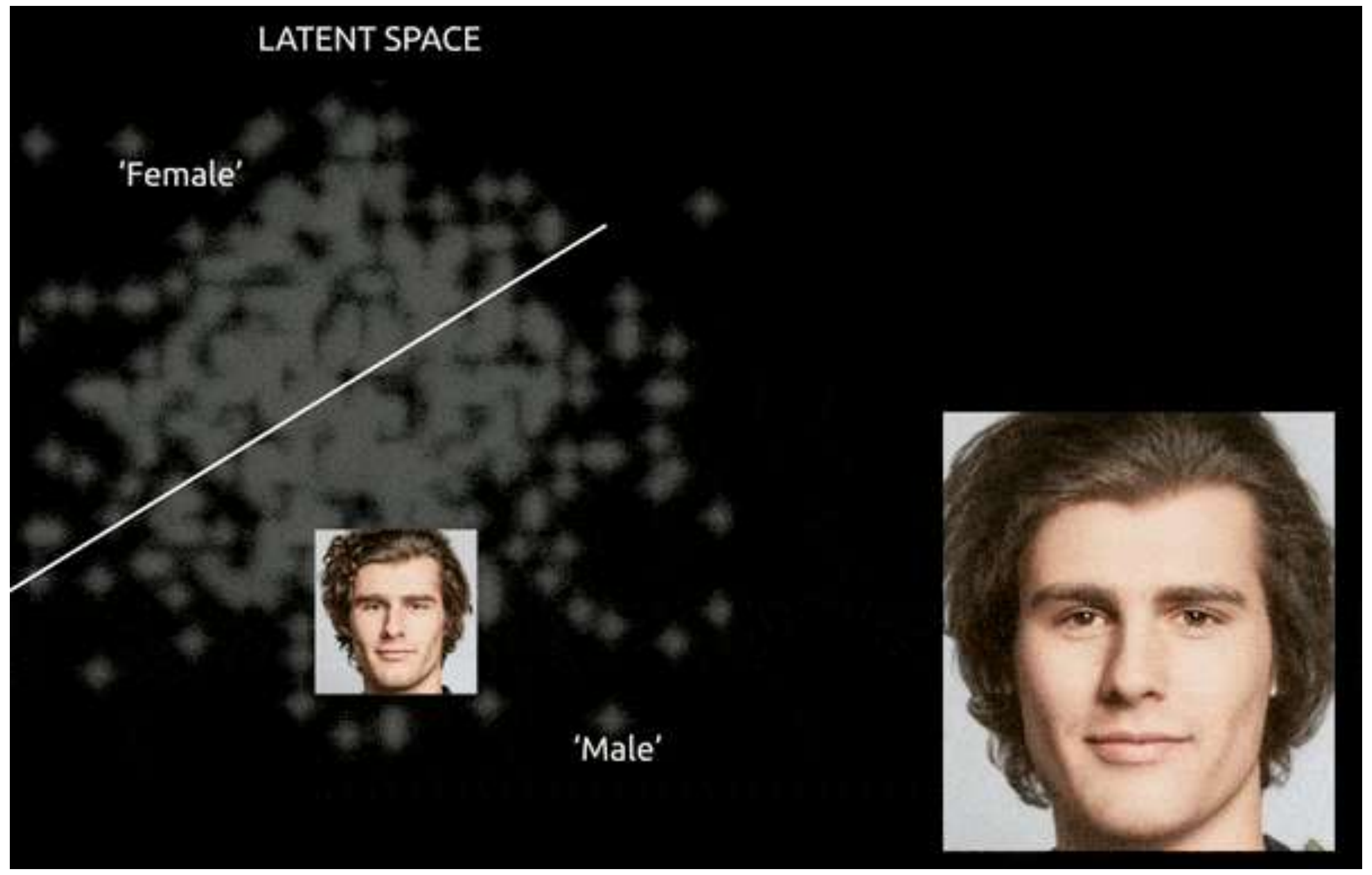
https://wiso.uni-hohenheim.de/fileadmin/einrichtungen/wiso/Forschungsdekan/Papers_BESS/dp_2023-02_online.pdf

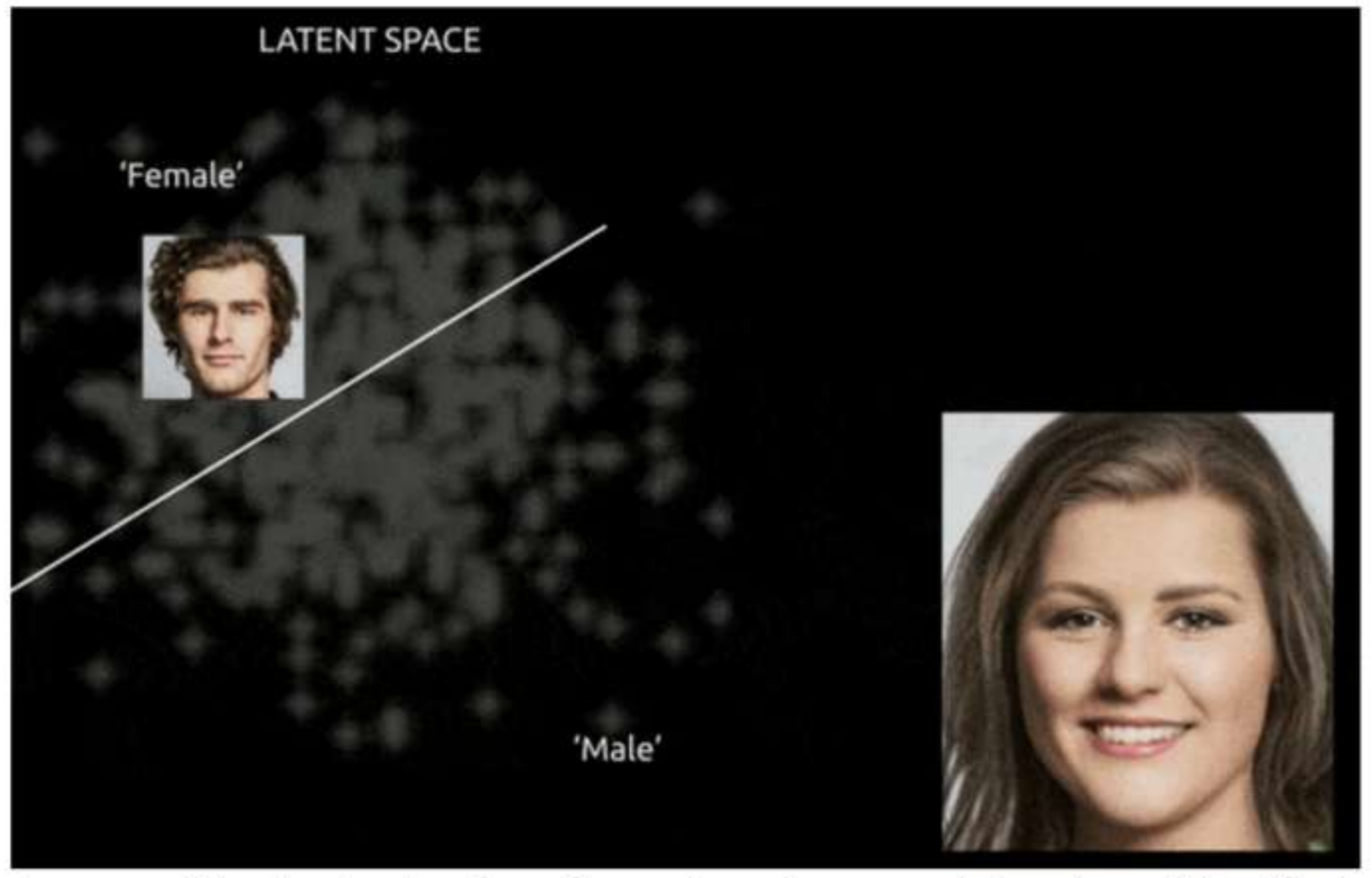
Genereren: Deepfakes

Generative Adverserial Networks (GANs)



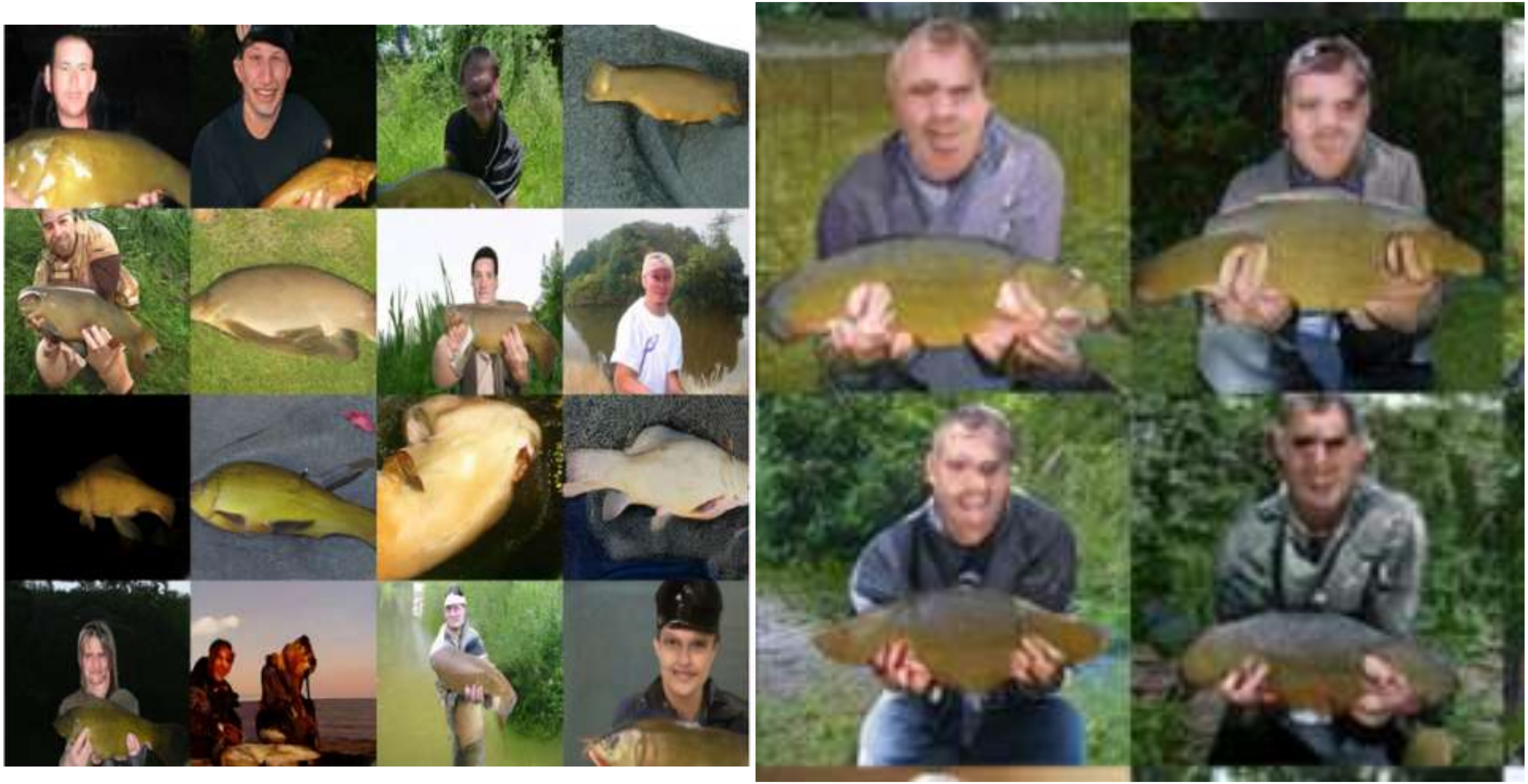






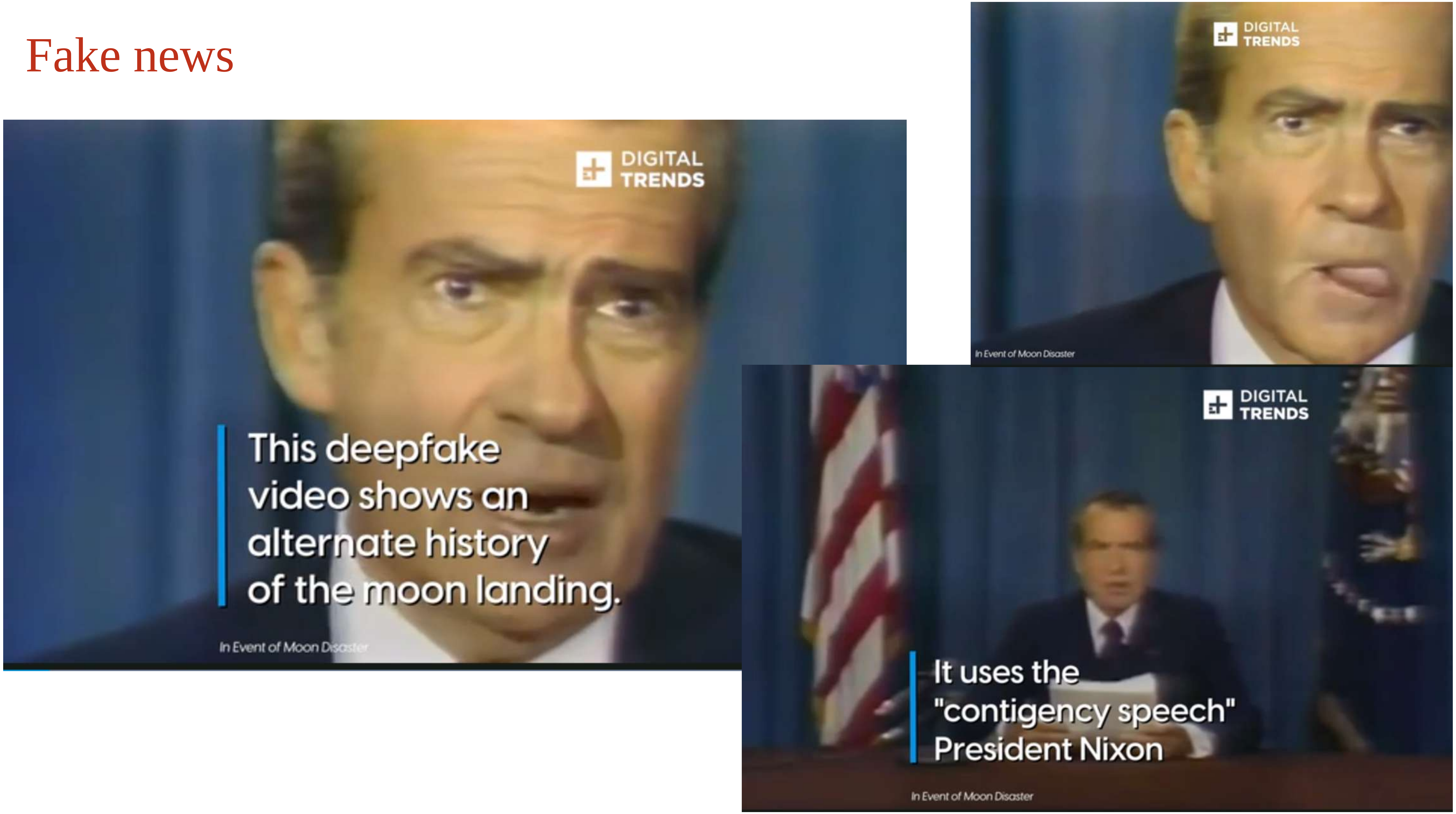
https://blog.metaphysic.ai/the-future-of-generative-adversarial-networks-in-deepfakes/

Deepfakes ‘halverwege’



Generative adversarial networks (GANs)

Fake news



This deepfake video shows an alternate history of the moon landing.

It uses the "contigency speech" President Nixon

DIGITAL TRENDS
In Event of Moon Disaster
DIGITAL TRENDS
In Event of Moon Disaster
DIGITAL TRENDS
In Event of Moon Disaster

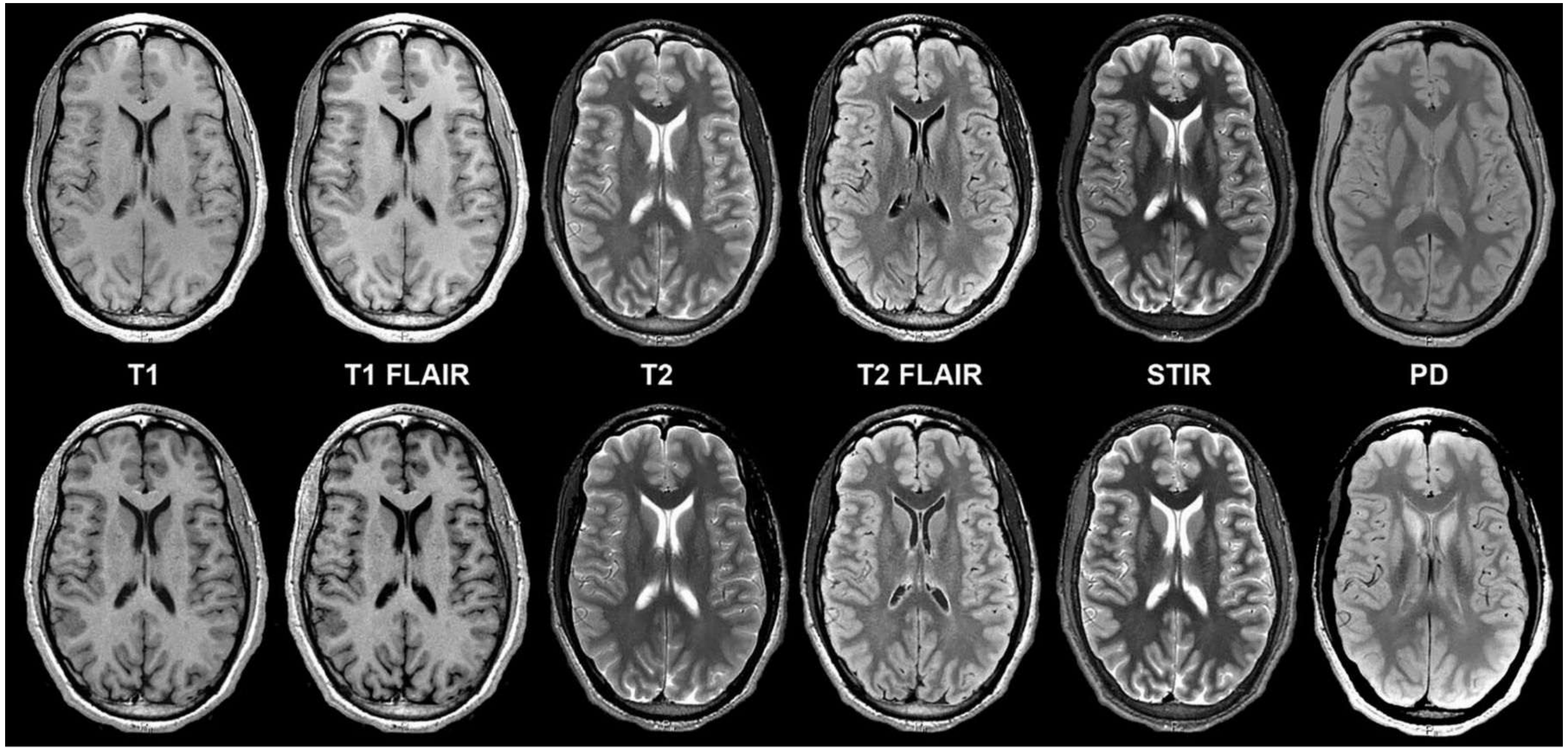
https://www.digitaltrends.com/features/mit-alternate-history-of-moon-landing-deepfake/

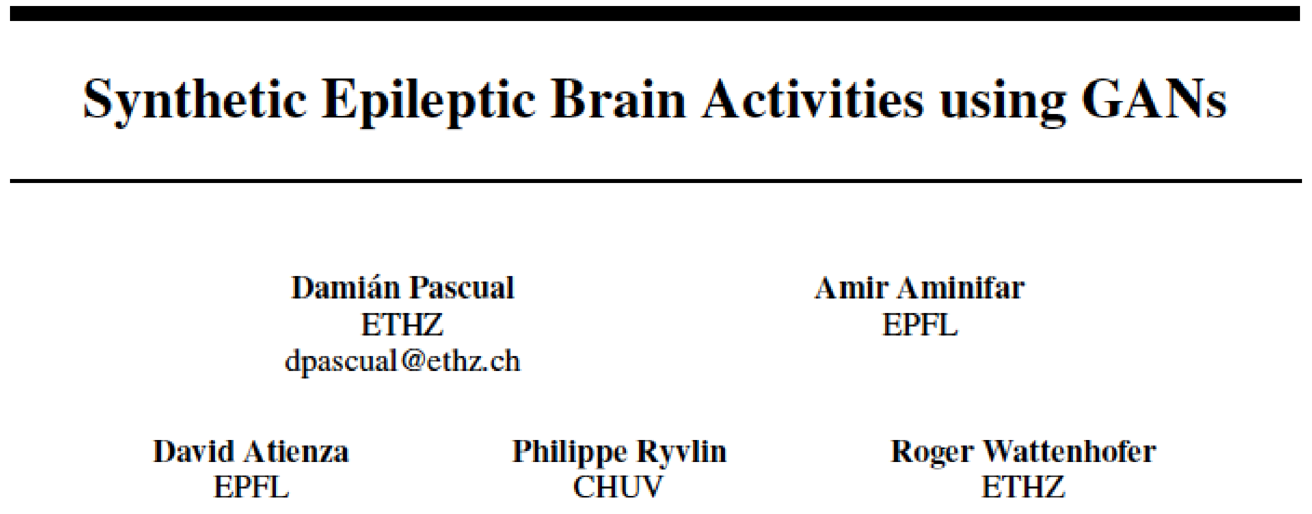
Genereren van nuttige namaak: Deepfakes voor training

Medische toepassingen vereisen veel gelabelde trainingsdata om algoritmes goed af te stellen

Menselijke data verkrijgen is tijdrovend, moeilijk (bv epileptische aanval) of duur

Synthetische data: ‘levensecht’ genoeg voor training





Synthetic Epileptic Brain Activities using GANs

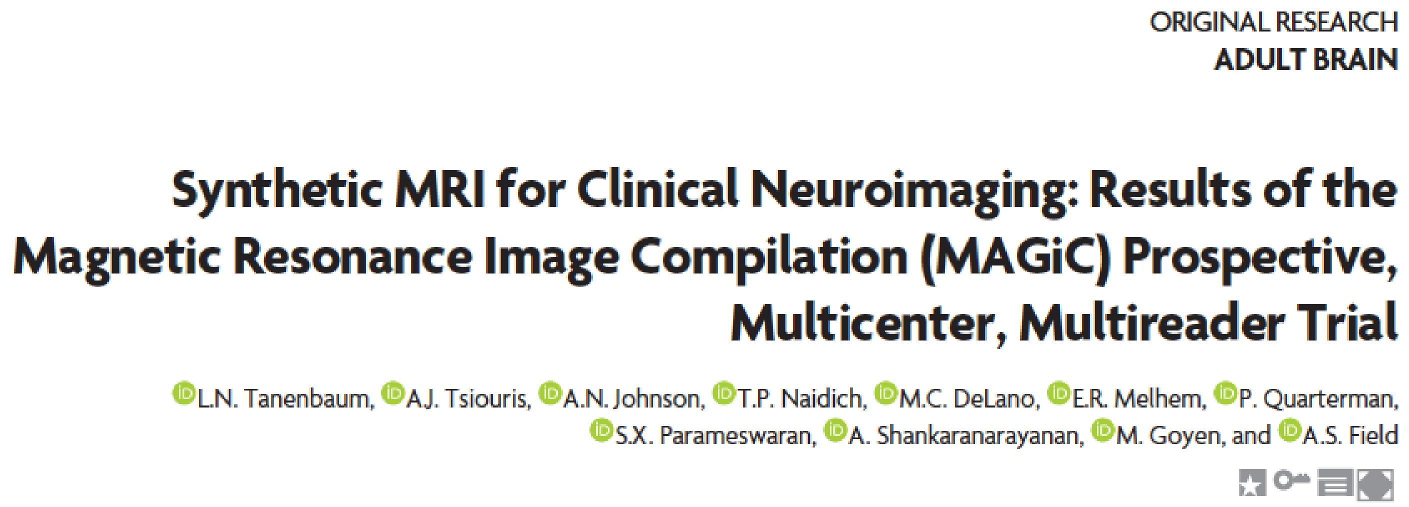
Damián Pascual
ETHZ
dpascual@ethz.ch

Amir Aminifar
EPFL

David Atienza
EPFL

Philippe Ryvlin
CHUV

Roger Wattenhofer
ETHZ



ORIGINAL RESEARCH
ADULT BRAIN

Synthetic MRI for Clinical Neuroimaging: Results of the Magnetic Resonance Image Compilation (MAGiC) Prospective, Multicenter, Multireader Trial

L.N. Tanenbaum, A.J. Tsiouris, A.N. Johnson, T.P. Naidich, M.C. DeLano, E.R. Melhem, P. Quarterman, S.X. Parameswaran, A. Shankaranarayanan, M. Goyen, and A.S. Field