Constructive ethics in Engineering:
Human responsibility & smart technology

Pim Haselager

Donders Institute for Brain, Cognition, and Behaviour
Dpt. of Artificial Intelligence
Radboud University, Nijmegen

pim.haselager@donders.ru.nl
giulio.mecacci@donders.ru.nl
@pim_haselager

Ethicists……

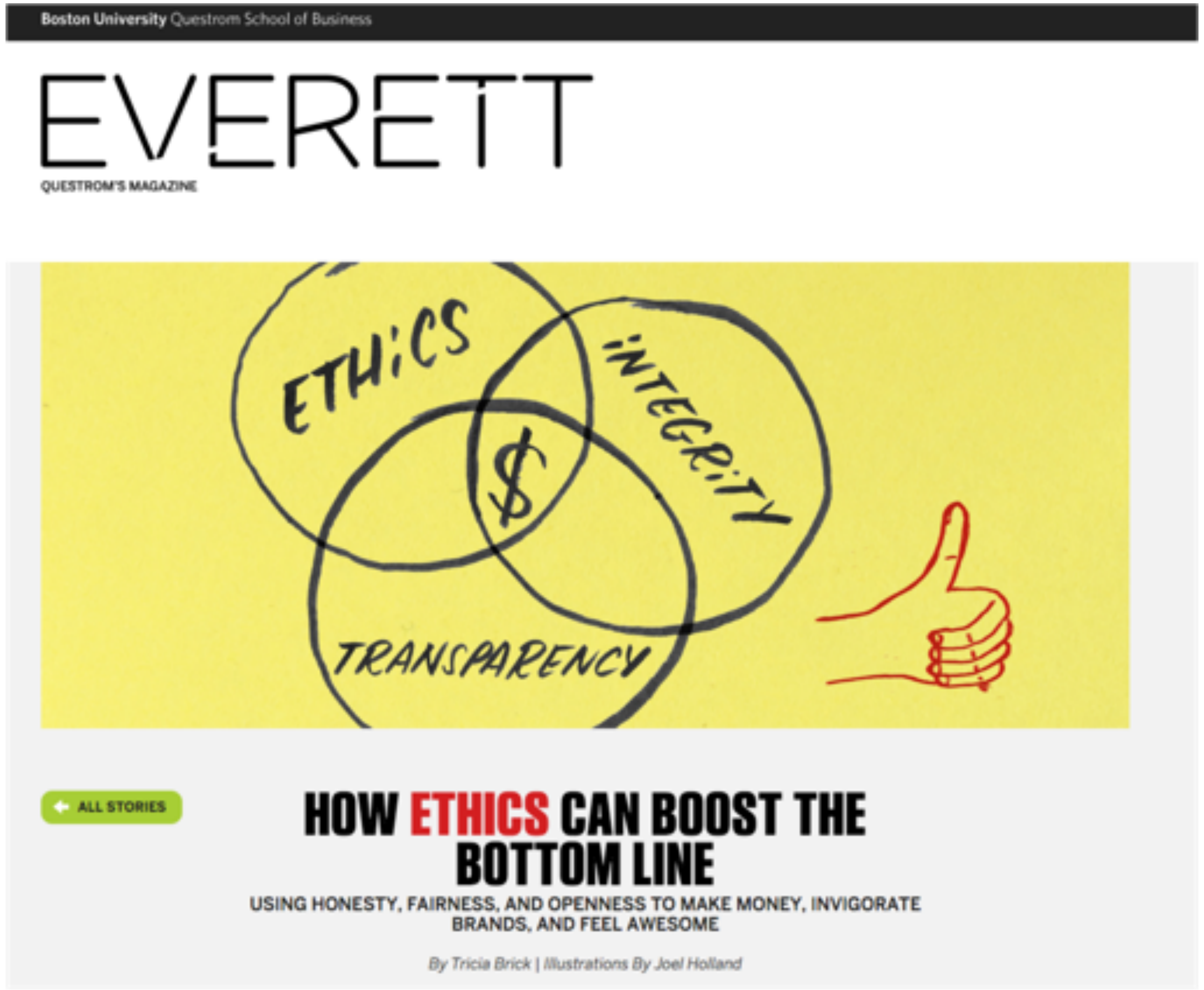
Monty Python

Why ethics?

Not just to do ‘the right thing’

Ethics is ultimately about responsibility and accountability

Legal & financial consequences



Boston University Questrom School of Business

EVERETT

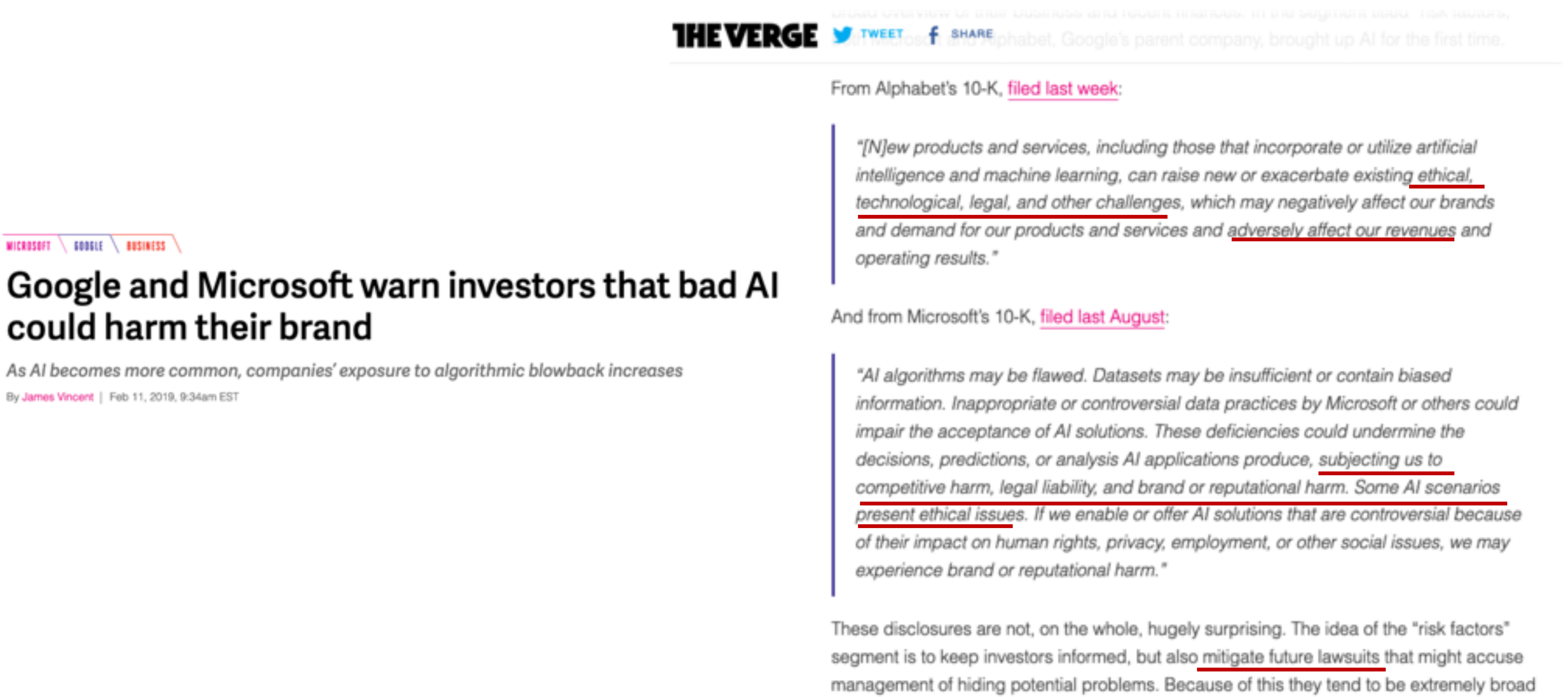
QUESTROM'S MAGAZINE

HOW ETHICS CAN BOOST THE BOTTOM LINE

USING HONESTY, FAIRNESS, AND OPENNESS TO MAKE MONEY, INVIGORATE BRANDS, AND FEEL AWESOME

By Tricia Brick | Illustrations By Joel Holland

The very real consequences of bad AI



MICROSOFT \ GOOGLE \ BUSINESS

Google and Microsoft warn investors that bad AI could harm their brand

As AI becomes more common, companies' exposure to algorithmic blowback increases

By James Vincent | Feb 11, 2019, 9:34am EST

phabet, Google's parent company, brought up AI for the first time.

From Alphabet's 10-K, filed last week:

> *"[N]ew products and services, including those that incorporate or utilize artificial intelligence and machine learning, can raise new or exacerbate existing ethical, technological, legal, and other challenges, which may negatively affect our brands and demand for our products and services and adversely affect our revenues and operating results."*

And from Microsoft's 10-K, filed last August:

> *"AI algorithms may be flawed. Datasets may be insufficient or contain biased information. Inappropriate or controversial data practices by Microsoft or others could impair the acceptance of AI solutions. These deficiencies could undermine the decisions, predictions, or analysis AI applications produce, subjecting us to competitive harm, legal liability, and brand or reputational harm. Some AI scenarios present ethical issues. If we enable or offer AI solutions that are controversial because of their impact on human rights, privacy, employment, or other social issues, we may experience brand or reputational harm."*

These disclosures are not, on the whole, hugely surprising. The idea of the "risk factors" segment is to keep investors informed, but also mitigate future lawsuits that might accuse management of hiding potential problems. Because of this they tend to be extremely broad

https://www.theverge.com/2019/2/11/18220050/google-microsoft-ai-brand-damage-investors-10-k-filing

Constructive ethics

Identify societal concerns

Stakeholder driven design

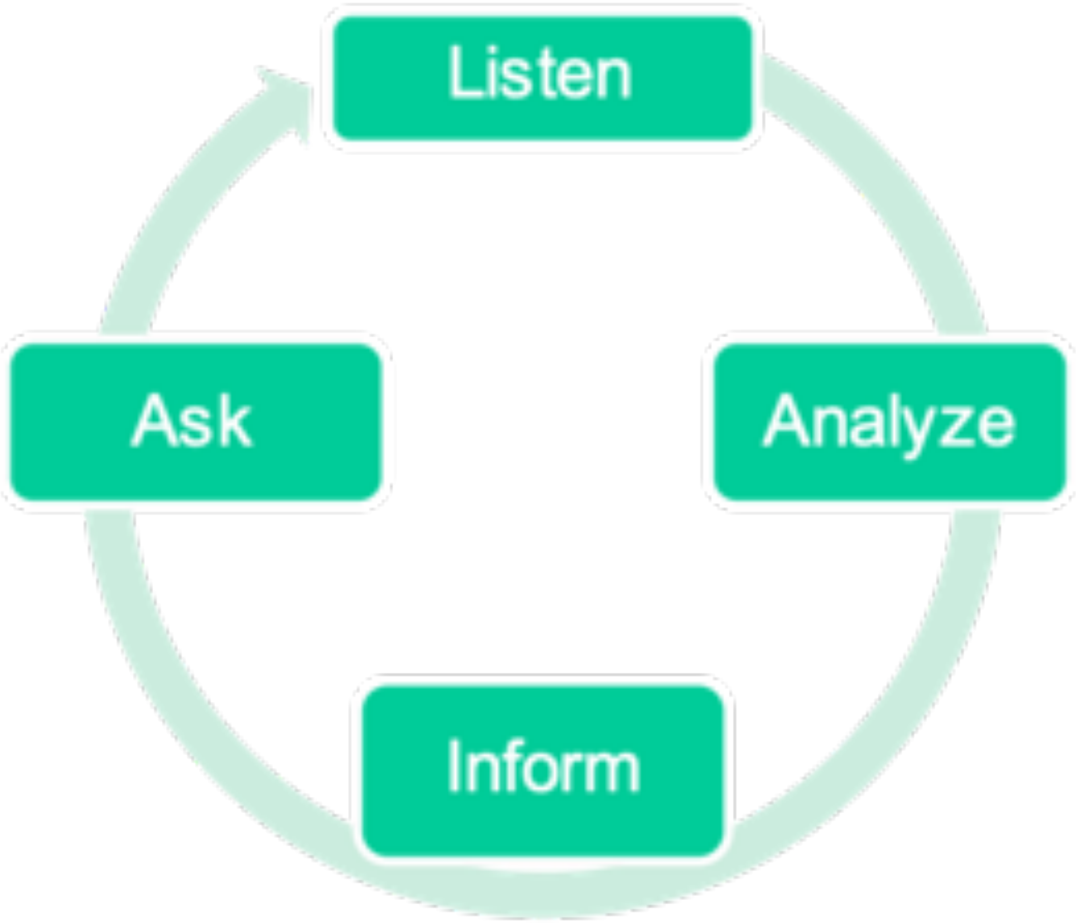
Responsible use of AI

Meaningful human control

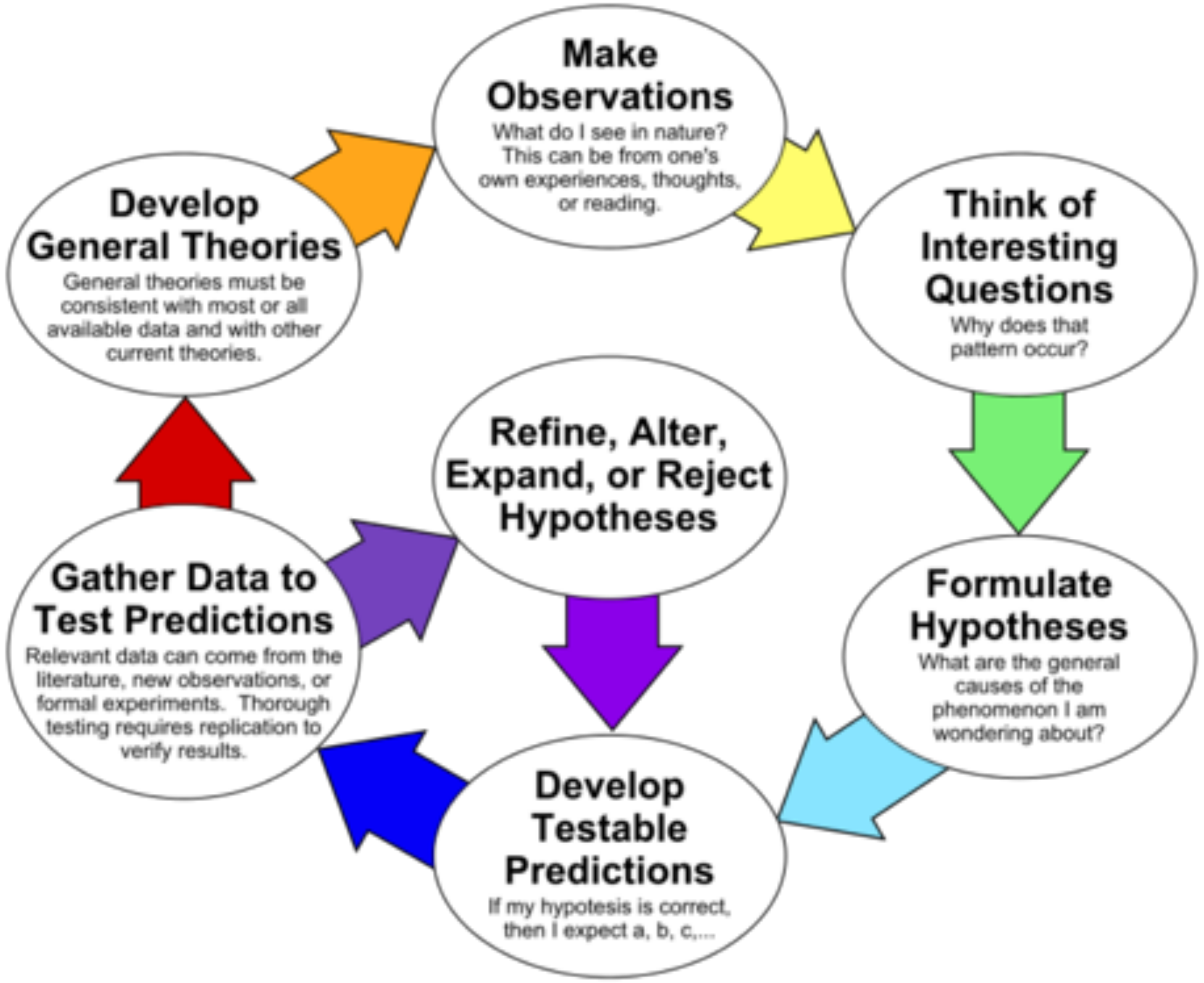
Constructive ethics

- Ethical, Legal & Societal Implications (ELSI) of Symbiotic Technology
- Interactive
  - Not just raising implications or concerns
  - Use ELSI to communicate with R&D, and (potential) stakeholders about the possible, desirable, avoidable
  - Use ELSI to possibly improve (potential applications of) neurotechnology
- Listen, Analyse, Inform, Ask

- Not
  - To tell you ‘what you should (not) do’
  - To tell you ‘to be good’
- Instead
  - Raise issues to think about
  - Stimulate discussion about (some of) them
  - Perhaps integrate some of them in research & design



Research & design cycles



Make Observations
What do I see in nature? This can be from one's own experiences, thoughts, or reading.

Think of Interesting Questions
Why does that pattern occur?

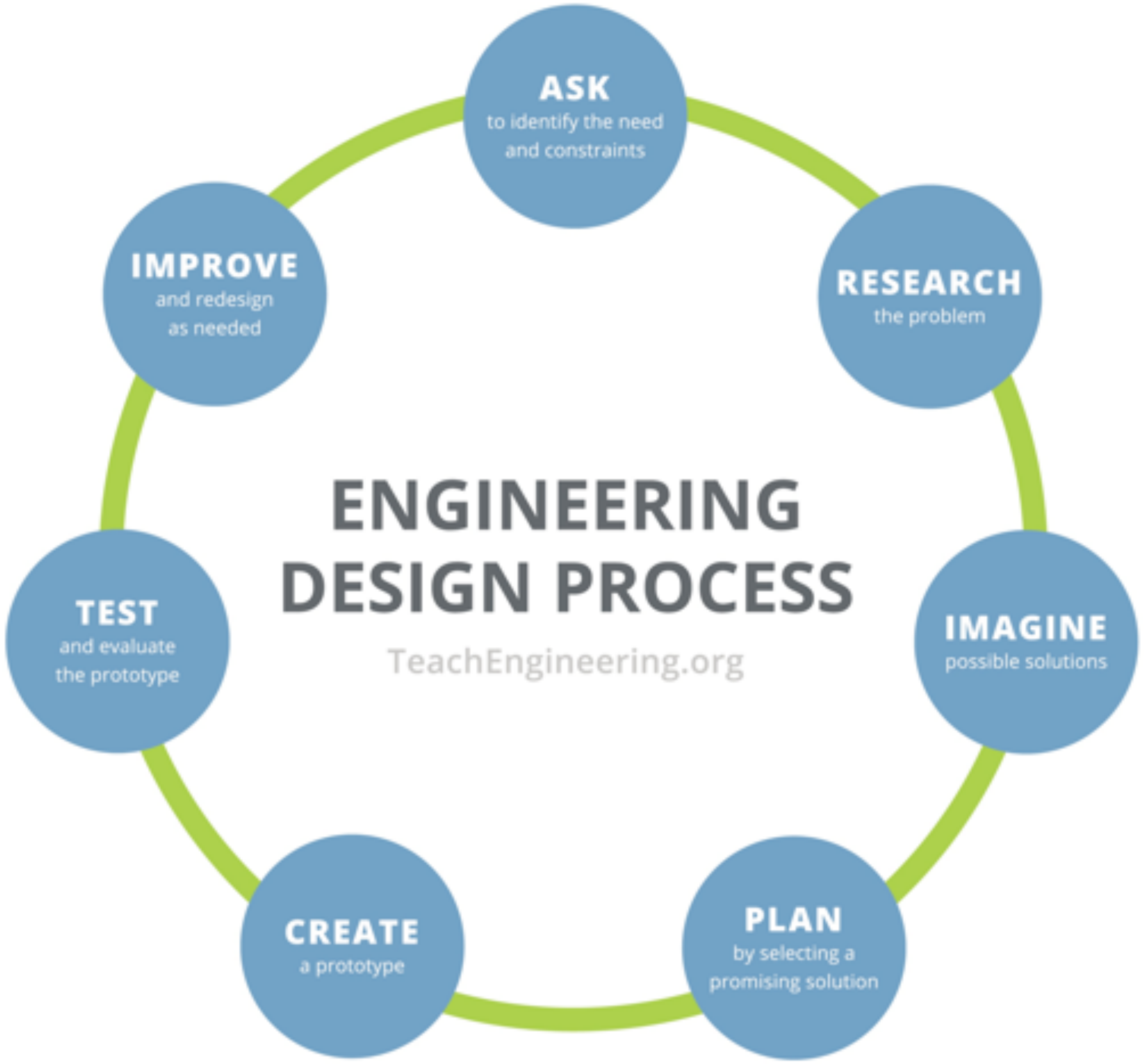
Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?

Develop Testable Predictions
If my hypotesis is correct, then I expect a, b, c,...

Gather Data to Test Predictions
Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.

Refine, Alter, Expand, or Reject Hypotheses

Develop General Theories
General theories must be consistent with most or all available data and with other current theories.



ENGINEERING DESIGN PROCESS
TeachEngineering.org

ASK to identify the need and constraints

RESEARCH the problem

IMAGINE possible solutions

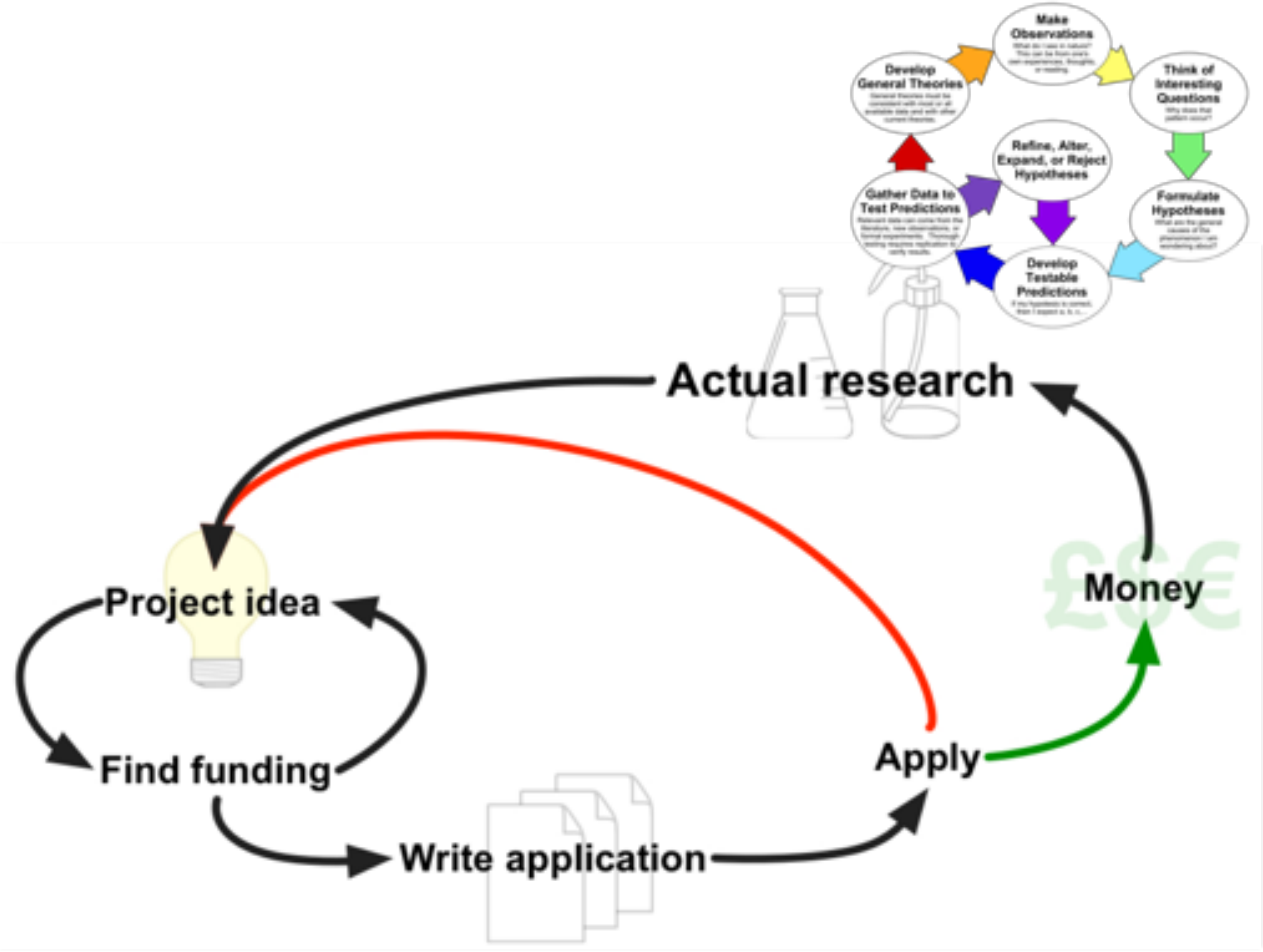
PLAN by selecting a promising solution

CREATE a prototype

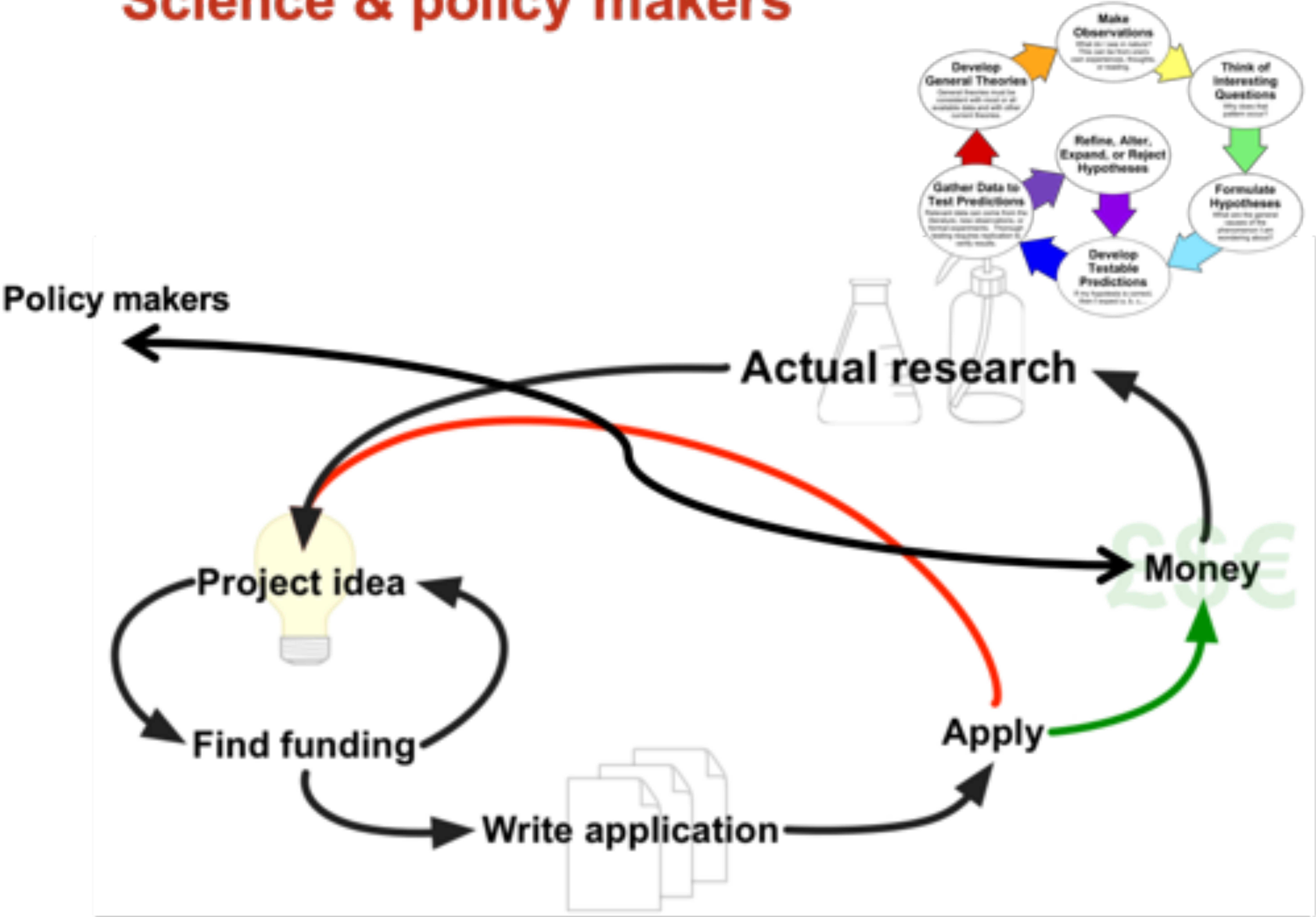
TEST and evaluate the prototype

IMPROVE and redesign as needed

The wider context


Make Observations
Think of Interesting Questions
Formulate Hypotheses
Develop Testable Predictions
Gather Data to Test Predictions
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories
Actual research
Project idea
Find funding
Write application
Apply
Money
£$€

Science & policy makers


Make Observations
Think of Interesting Questions
Formulate Hypotheses
Develop Testable Predictions
Gather Data to Test Predictions
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories

Policy makers

Actual research

Project idea

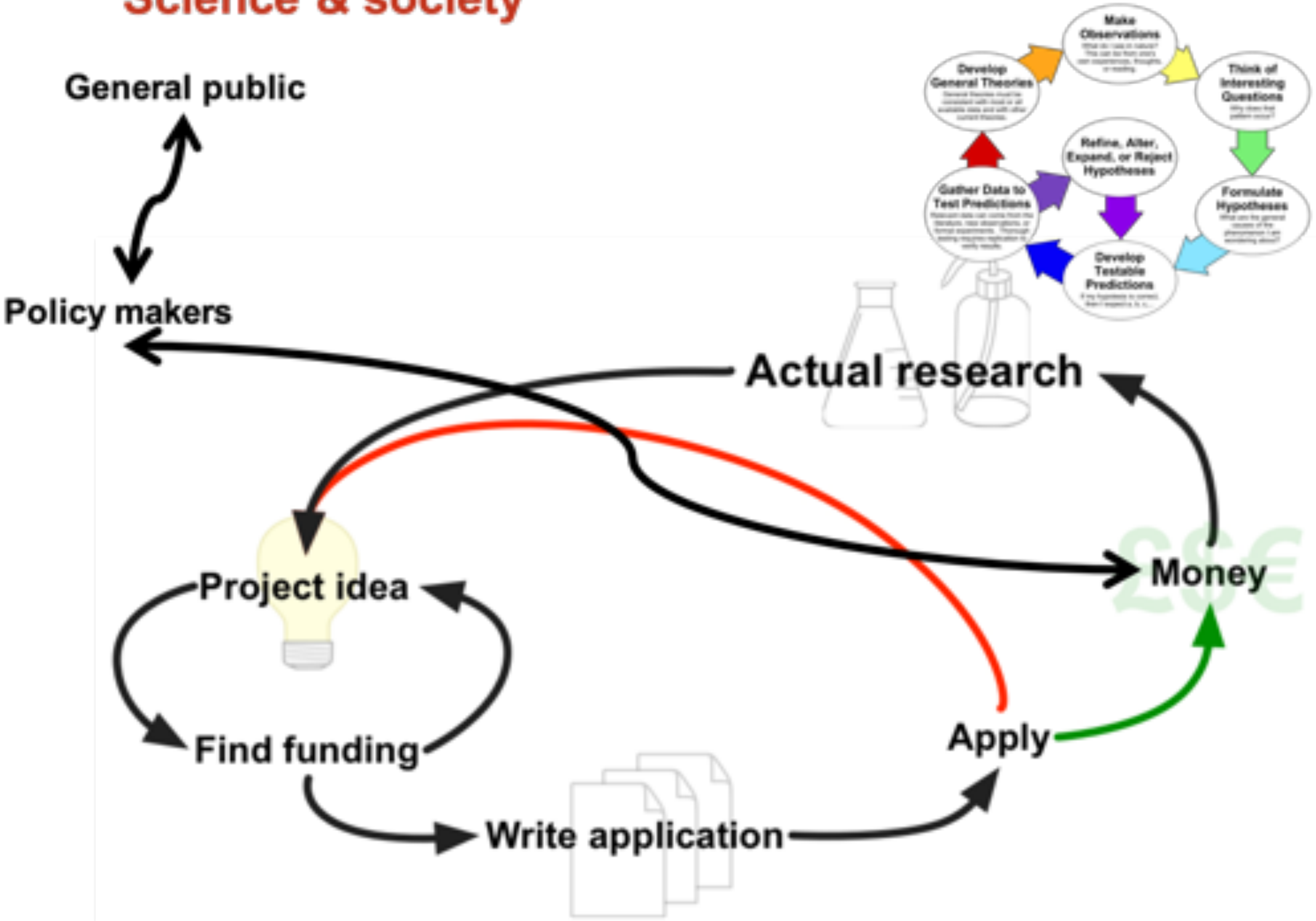
Money

Find funding

Apply

Write application

Science & society


Make Observations
Think of Interesting Questions
Formulate Hypotheses
Develop Testable Predictions
Gather Data to Test Predictions
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories

General public

Policy makers

Actual research

Project idea

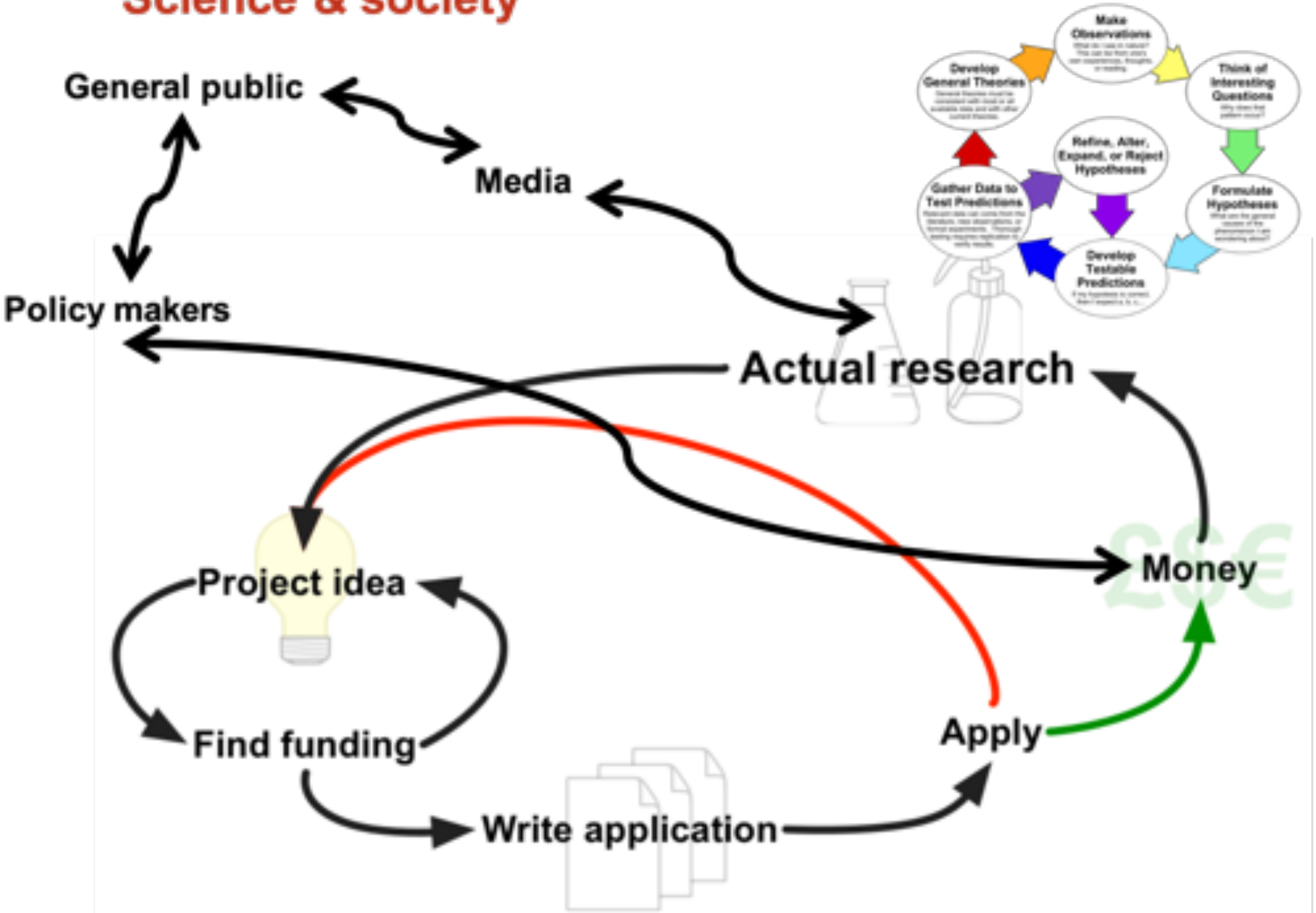
Find funding

Write application

Apply

Money

Science & society



General public

Media

Policy makers

Actual research

Money

Project idea

Find funding

Write application

Apply

Make Observations
Think of Interesting Questions
Formulate Hypotheses
Develop Testable Predictions
Gather Data to Test Predictions
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories

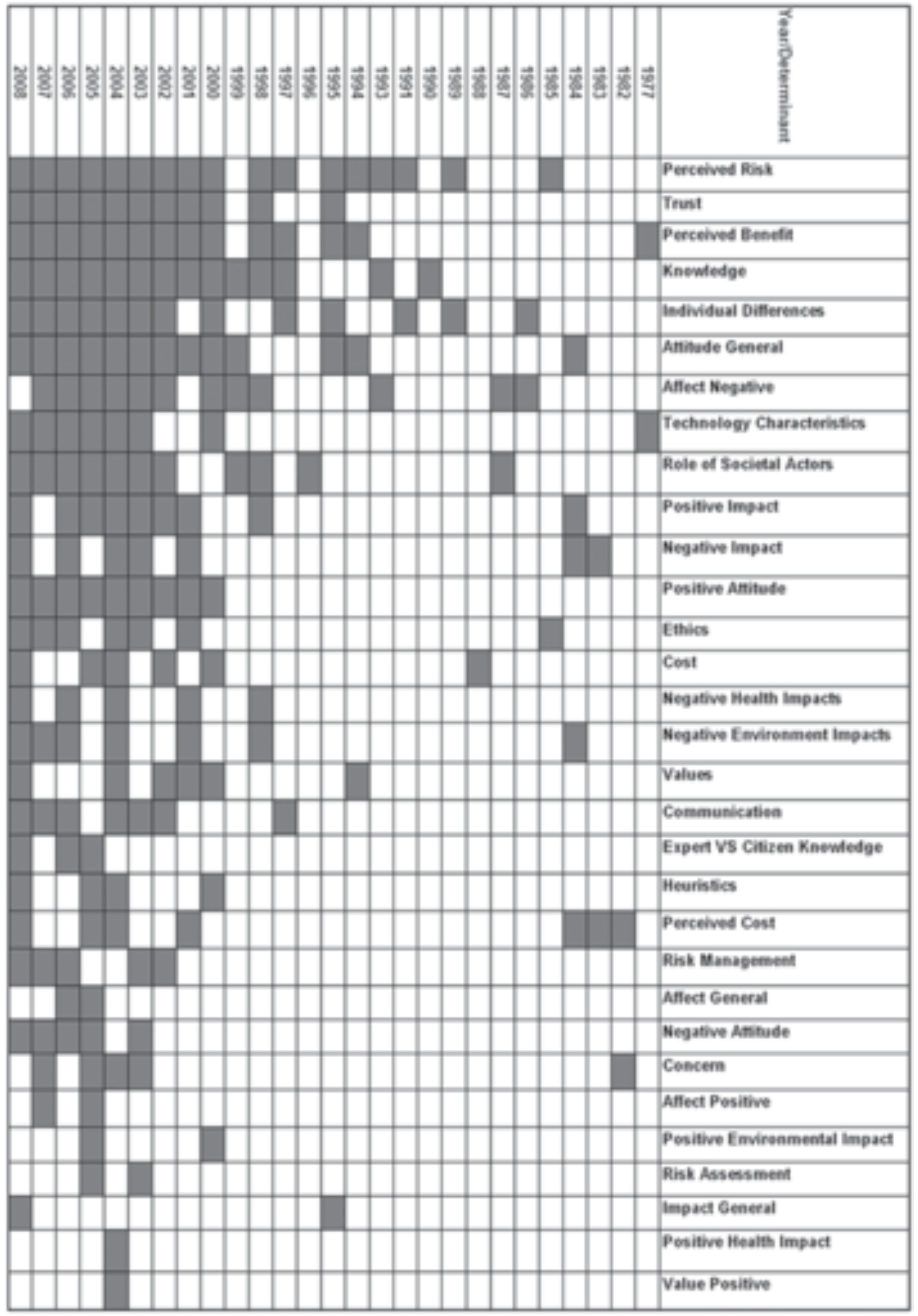
Psychological determinants of public acceptance of technology

Perceived **risk** was found to be

the most frequently investigated determinant, then trust, and then perceived benefit

So, risks have to be addressed

Not just by researchers & designers
Other stakeholders will do too
and perhaps/probably differently



2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1991	1990	1989	1988	1987	1986	1985	1984	1983	1982	1977	Year/Determinant
■	■	■	■	■	■	■	■	■		■	■		■	■	■	■		■				■					Perceived Risk
■	■	■	■	■	■	■	■	■		■			■														Trust
■	■	■	■	■	■	■	■	■		■	■		■	■												■	Perceived Benefit
■	■	■	■	■	■	■	■	■	■	■	■				■		■										Knowledge
■	■	■	■	■	■	■		■			■		■			■		■			■						Individual Differences
■	■	■	■	■	■	■	■	■	■				■	■									■				Attitude General
	■	■	■	■	■	■		■	■	■					■					■	■						Affect Negative
■	■	■	■	■	■			■																		■	Technology Characteristics
■	■	■	■	■	■	■			■	■		■								■							Role of Societal Actors
■		■	■	■	■	■	■			■													■				Positive Impact
■		■		■	■		■																■	■			Negative Impact
■	■	■	■	■	■	■	■	■																			Positive Attitude
■	■	■		■	■		■															■					Ethics
■			■	■		■		■											■								Cost
■		■		■			■			■																	Negative Health Impacts
■	■	■		■			■			■													■				Negative Environment Impacts
■				■		■	■	■						■													Values
■	■	■		■	■	■					■																Communication
■		■	■																								Expert VS Citizen Knowledge
■			■	■				■																			Heuristics
■			■	■			■																■	■	■		Perceived Cost
■	■	■			■	■																					Risk Management
		■	■																								Affect General
■	■	■	■		■																						Negative Attitude
	■		■	■	■																				■		Concern
	■		■																								Affect Positive
			■					■																			Positive Environmental Impact
			■		■																						Risk Assessment
■													■														Impact General
				■																							Positive Health Impact
				■																							Value Positive

Gupta, Fischer and Frewer (2011).
Socio-psychological determinants of public acceptance of technologies: A review
Public Understanding of Science 21(7) 782– 795

Risk and responsibility for actions mediated by technology

Legal responsibility
    Who goes to jail?
    Judges & Lawyers

Financial responsibility (e.g. liability for damage)
    Who has to pay?
    Insurance-companies & lawyers

Moral responsibility
    Who is to blame?
    Society (ethics, public opinion, press, gossip)

Political responsibility
    Democratic control of technological decisions
    Privacy & freedom of thought & expression, protection against bias & manipulation

Liability: Damage, Negligence & Dangerous products

European Civil Code Project : A person causes legally relevant damage

Article 3:102: **negligently** when that person causes the damage by conduct which either:

a) does not meet the particular **standard of care** provided by a statutory provision whose purpose is the protection of the injured person from the damage suffered, or
b) does not otherwise amount to such care as could be expected from a **reasonably careful person** in the circumstances of the case

Strict liability of a party without a finding of fault (without negligence or intention)

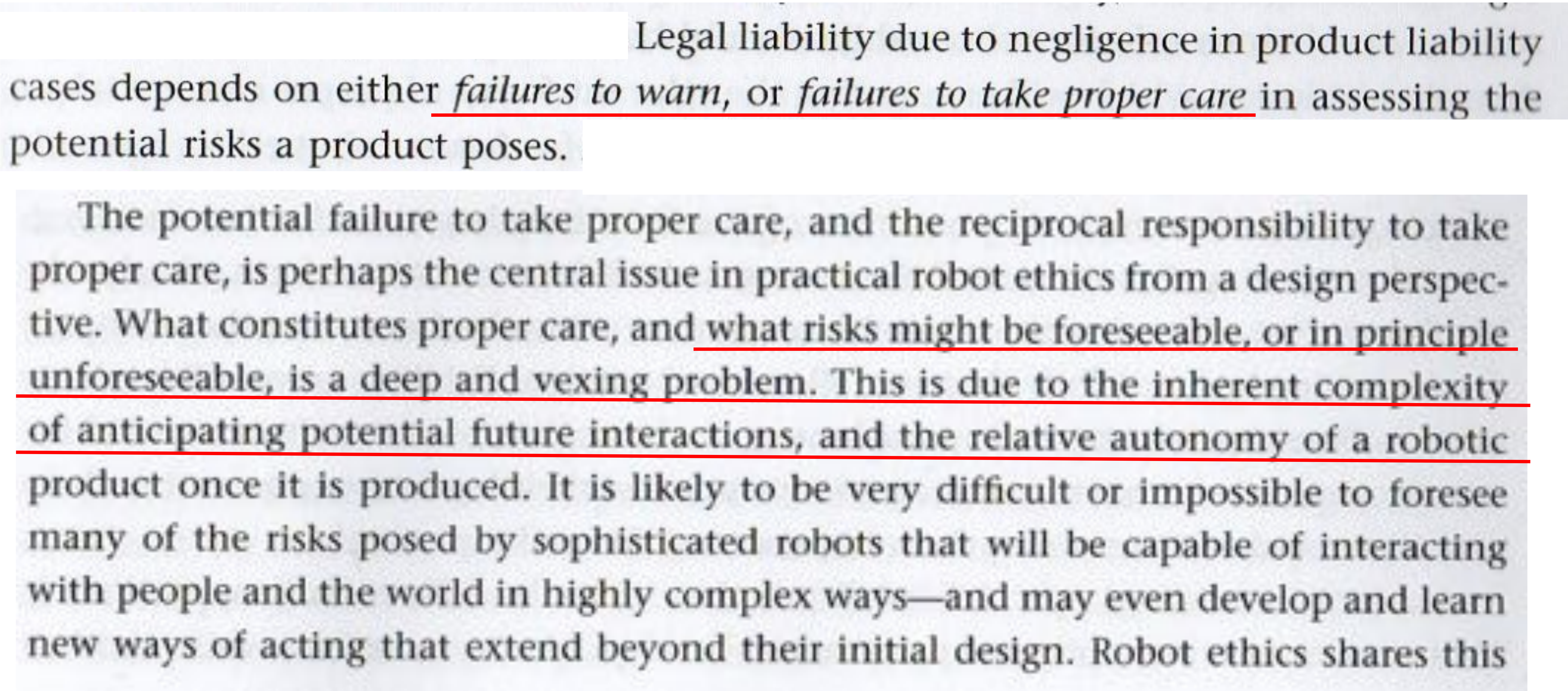
The law imputes strict liability to situations it considers to be inherently dangerous

Defective or **dangerous products** …….

Product liability of the manufacturer (± standardly, at first instance)

Legal liability for products

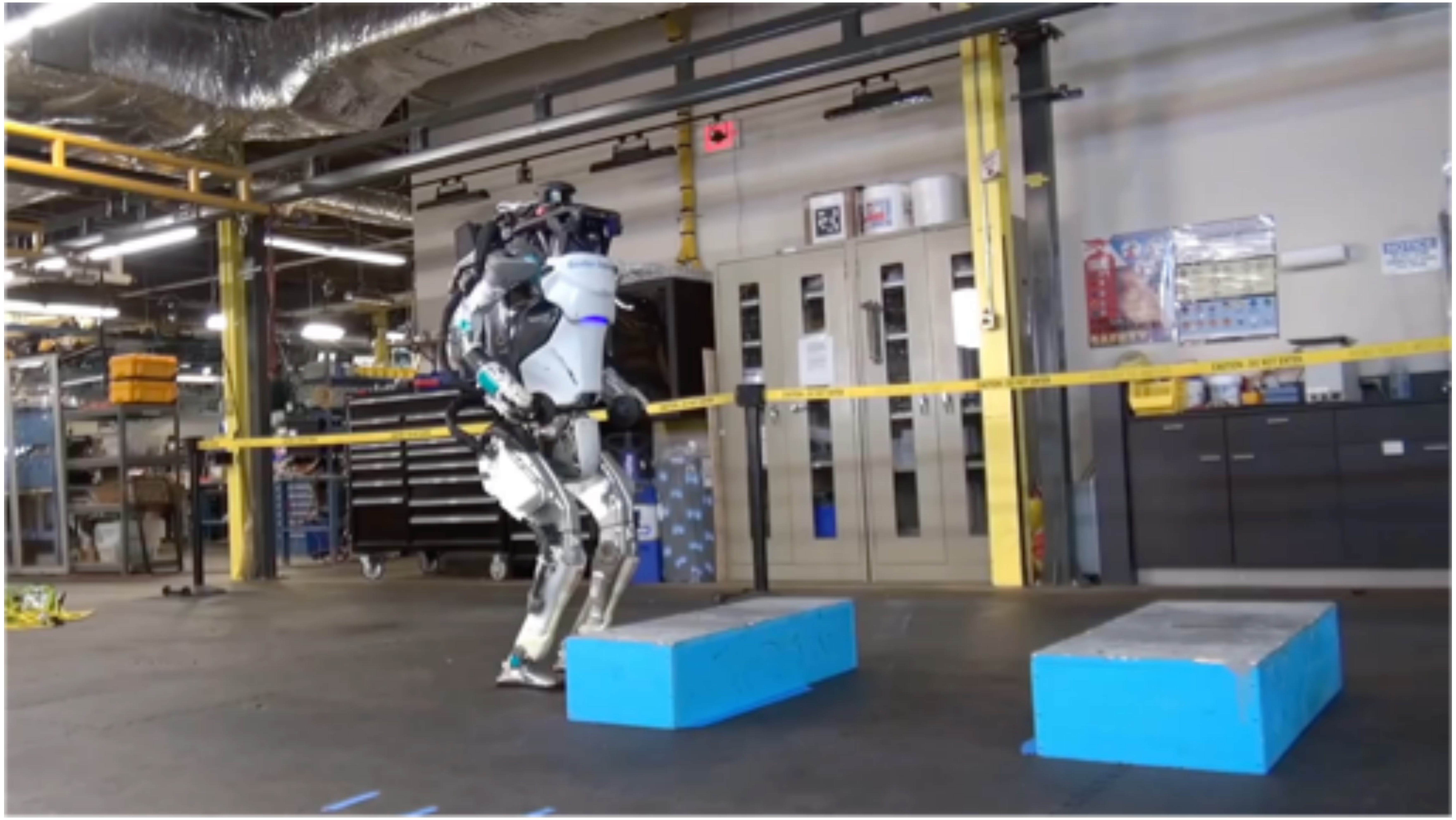
Asaro, P. (2011). " A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics," in Patrick Lin, Keith Abney, and George Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, pp. 169-186.



Legal liability due to negligence in product liability cases depends on either *failures to warn*, or *failures to take proper care* in assessing the potential risks a product poses.

The potential failure to take proper care, and the reciprocal responsibility to take proper care, is perhaps the central issue in practical robot ethics from a design perspective. What constitutes proper care, and what risks might be foreseeable, or in principle unforeseeable, is a deep and vexing problem. This is due to the inherent complexity of anticipating potential future interactions, and the relative autonomy of a robotic product once it is produced. It is likely to be very difficult or impossible to foresee many of the risks posed by sophisticated robots that will be capable of interacting with people and the world in highly complex ways—and may even develop and learn new ways of acting that extend beyond their initial design. Robot ethics shares this

Atlas the robot, from Boston Dynamics



https://www.youtube.com/watch?time_continue=5&v=fRj34o4hN4I

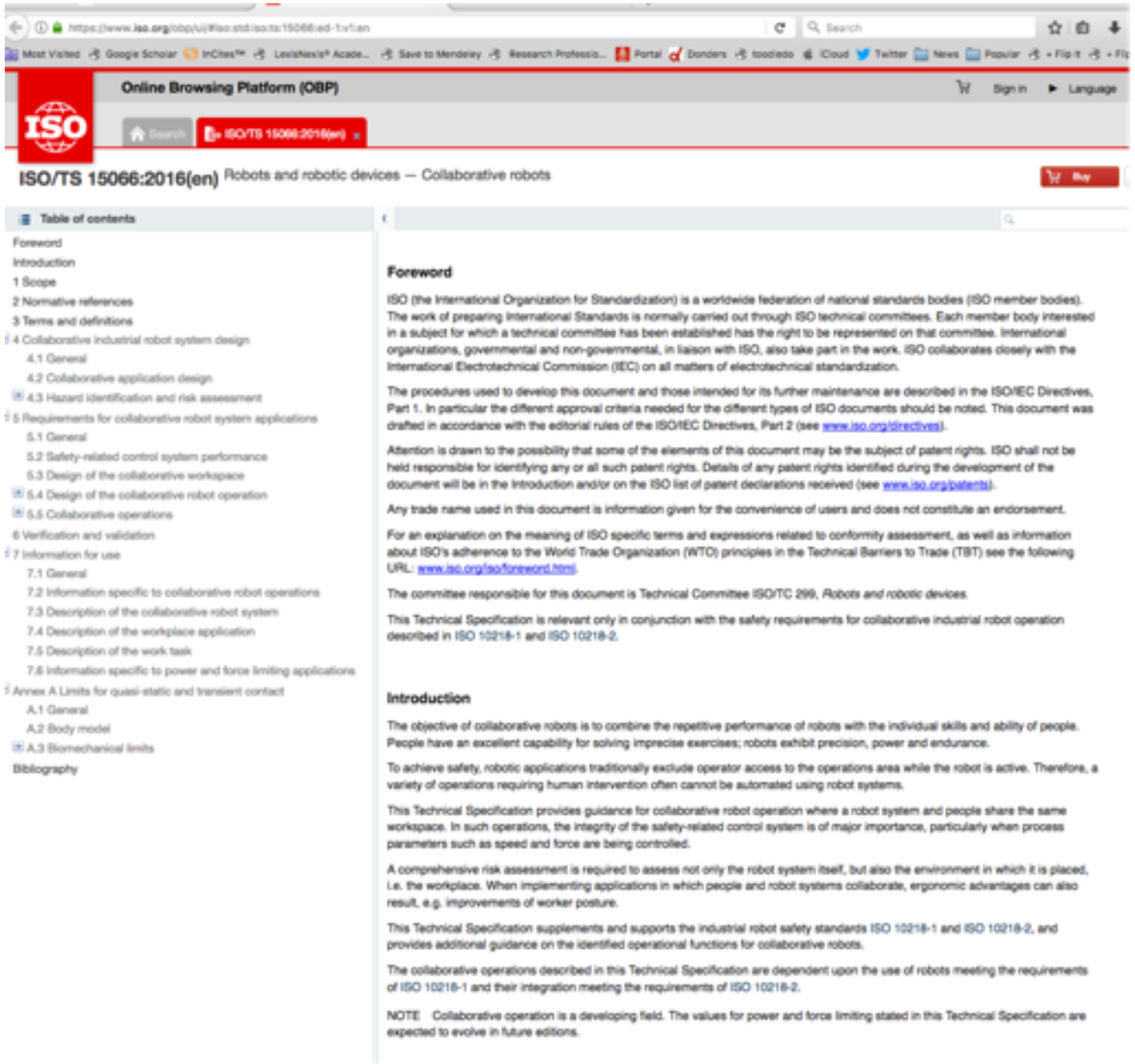
Bottom lime of the liability concern

The more intelligent & autonomous AI & robots will be
    And the greater the variety of situations they will function in
    And the more realistic those situations are
        (involving more & more diverse agents & objects)
The more unpredictable and potentially risky robot behavior will become

The smarter AI or robots get, the more risky they become

Liability is about much more than ISO safety standards

Focused on industrial type robots, not (much yet) on smart systems



ISO/TS 15066:2016(en) Robots and robotic devices — Collaborative robots

Table of contents

Foreword
Introduction
1 Scope
2 Normative references
3 Terms and definitions
4 Collaborative industrial robot system design
  4.1 General
  4.2 Collaborative application design
  4.3 Hazard identification and risk assessment
5 Requirements for collaborative robot system applications
  5.1 General
  5.2 Safety-related control system performance
  5.3 Design of the collaborative workspace
  5.4 Design of the collaborative robot operation
  5.5 Collaborative operations
6 Verification and validation
7 Information for use
  7.1 General
  7.2 Information specific to collaborative robot operations
  7.3 Description of the collaborative robot system
  7.4 Description of the workplace application
  7.5 Description of the work task
  7.6 Information specific to power and force limiting applications
Annex A Limits for quasi-static and transient contact
  A.1 General
  A.2 Body model
  A.3 Biomechanical limits
Bibliography

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is Technical Committee ISO/TC 299, *Robots and robotic devices*.

This Technical Specification is relevant only in conjunction with the safety requirements for collaborative industrial robot operation described in ISO 10218-1 and ISO 10218-2.

Introduction

The objective of collaborative robots is to combine the repetitive performance of robots with the individual skills and ability of people. People have an excellent capability for solving imprecise exercises; robots exhibit precision, power and endurance.

To achieve safety, robotic applications traditionally exclude operator access to the operations area while the robot is active. Therefore, a variety of operations requiring human intervention often cannot be automated using robot systems.

This Technical Specification provides guidance for collaborative robot operation where a robot system and people share the same workspace. In such operations, the integrity of the safety-related control system is of major importance, particularly when process parameters such as speed and force are being controlled.

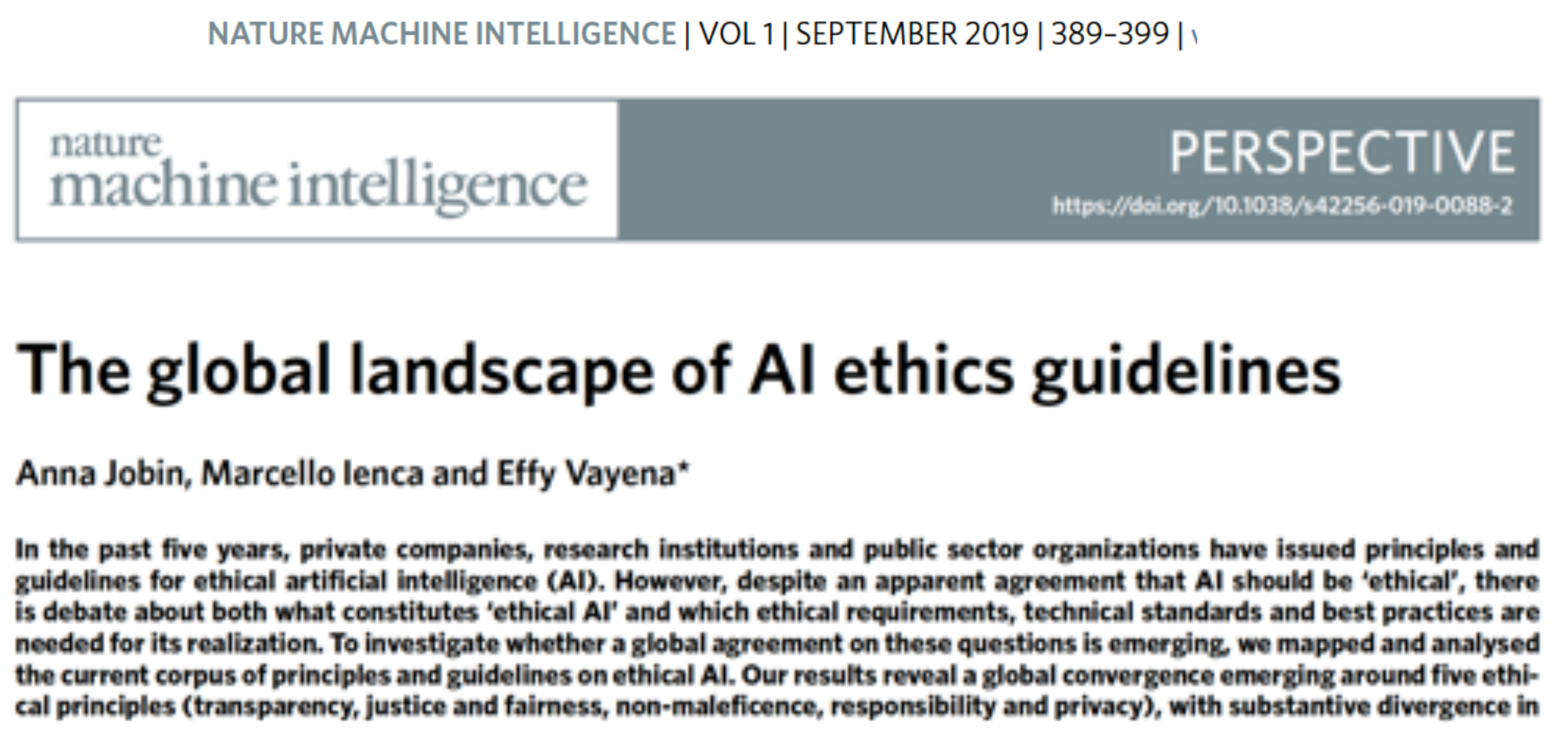
A comprehensive risk assessment is required to assess not only the robot system itself, but also the environment in which it is placed, i.e. the workplace. When implementing applications in which people and robot systems collaborate, ergonomic advantages can also result, e.g. improvements of worker posture.

This Technical Specification supplements and supports the industrial robot safety standards ISO 10218-1 and ISO 10218-2, and provides additional guidance on the identified operational functions for collaborative robots.

The collaborative operations described in this Technical Specification are dependent upon the use of robots meeting the requirements of ISO 10218-1 and their integration meeting the requirements of ISO 10218-2.

NOTE Collaborative operation is a developing field. The values for power and force limiting stated in this Technical Specification are expected to evolve in future editions.

Too many ethical codes…?!



NATURE MACHINE INTELLIGENCE | VOL 1 | SEPTEMBER 2019 | 389–399 |

nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in

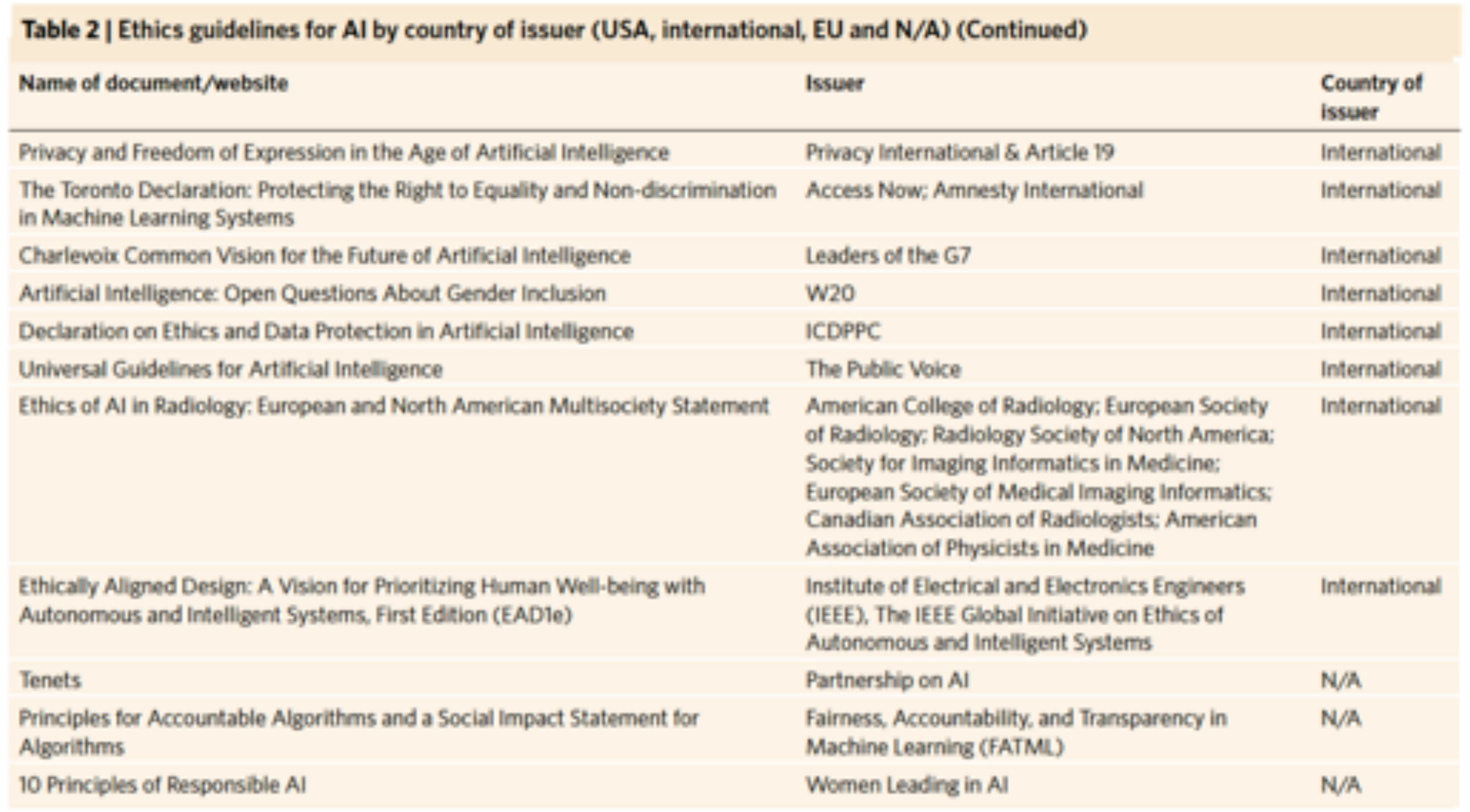


Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A) (Continued)

Name of document/website	Issuer	Country of issuer
Privacy and Freedom of Expression in the Age of Artificial Intelligence	Privacy International & Article 19	International
The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems	Access Now; Amnesty International	International
Charlevoix Common Vision for the Future of Artificial Intelligence	Leaders of the G7	International
Artificial Intelligence: Open Questions About Gender Inclusion	W20	International
Declaration on Ethics and Data Protection in Artificial Intelligence	ICDPPC	International
Universal Guidelines for Artificial Intelligence	The Public Voice	International
Ethics of AI in Radiology: European and North American Multisociety Statement	American College of Radiology; European Society of Radiology; Radiology Society of North America; Society for Imaging Informatics in Medicine; European Society of Medical Imaging Informatics; Canadian Association of Radiologists; American Association of Physicists in Medicine	International
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition (EAD1e)	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Tenets	Partnership on AI	N/A
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Fairness, Accountability, and Transparency in Machine Learning (FATML)	N/A
10 Principles of Responsible AI	Women Leading in AI	N/A

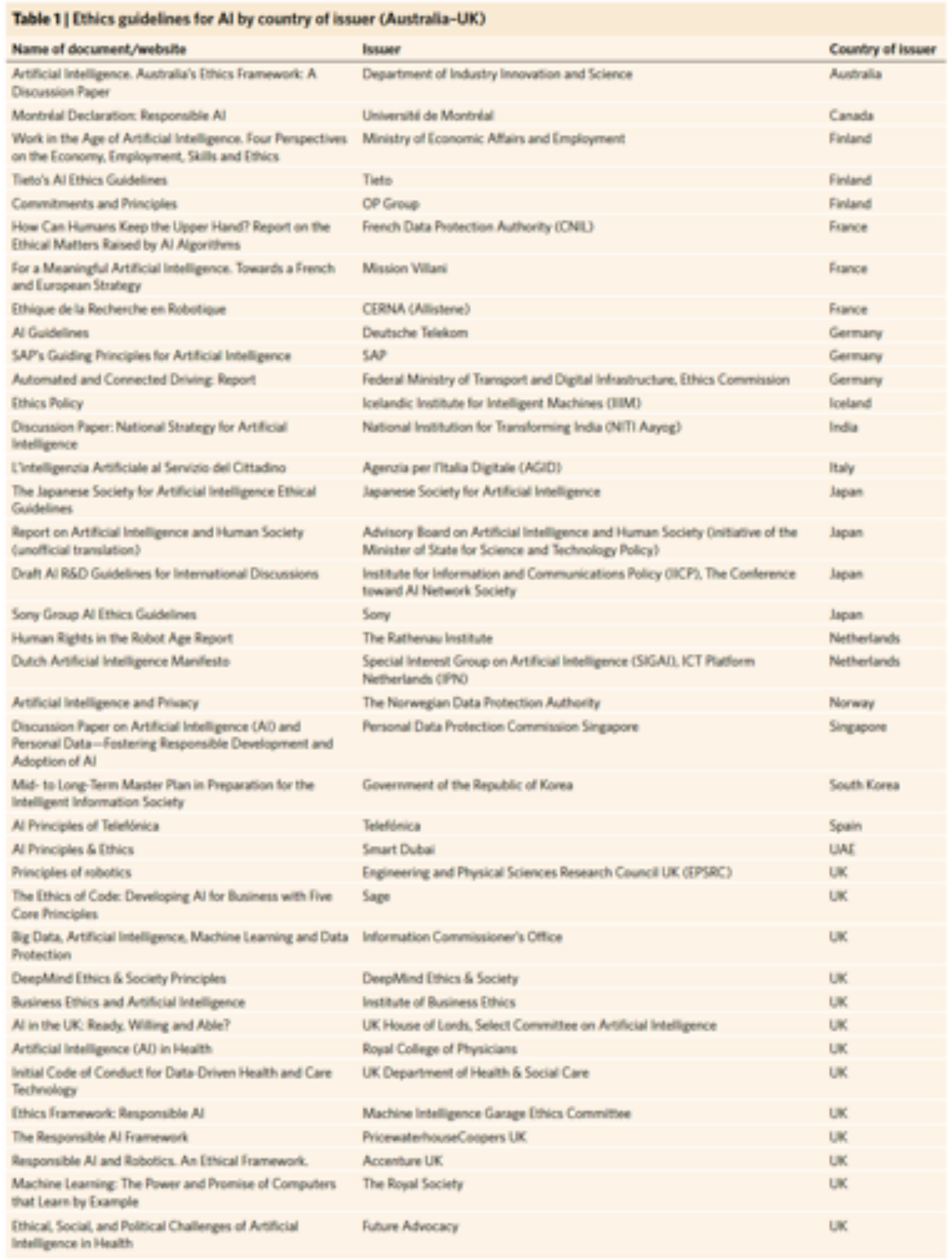


Table 1 | Ethics guidelines for AI by country of issuer (Australia–UK)

Name of document/website	Issuer	Country of issuer
Artificial Intelligence. Australia's Ethics Framework: A Discussion Paper	Department of Industry Innovation and Science	Australia
Montréal Declaration: Responsible AI	Université de Montréal	Canada
Work in the Age of Artificial Intelligence. Four Perspectives on the Economy, Employment, Skills and Ethics	Ministry of Economic Affairs and Employment	Finland
Tieto's AI Ethics Guidelines	Tieto	Finland
Commitments and Principles	OP Group	Finland
How Can Humans Keep the Upper Hand? Report on the Ethical Matters Raised by AI Algorithms	French Data Protection Authority (CNIL)	France
For a Meaningful Artificial Intelligence. Towards a French and European Strategy	Mission Villani	France
Ethique de la Recherche en Robotique	CERNA (Allistene)	France
AI Guidelines	Deutsche Telekom	Germany
SAP's Guiding Principles for Artificial Intelligence	SAP	Germany
Automated and Connected Driving: Report	Federal Ministry of Transport and Digital Infrastructure, Ethics Commission	Germany
Ethics Policy	Icelandic Institute for Intelligent Machines (IIIM)	Iceland
Discussion Paper: National Strategy for Artificial Intelligence	National Institution for Transforming India (NITI Aayog)	India
L'intelligenzia Artificiale al Servizio del Cittadino	Agenzia per l'Italia Digitale (AGID)	Italy
The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence	Japan
Report on Artificial Intelligence and Human Society (unofficial translation)	Advisory Board on Artificial Intelligence and Human Society (initiative of the Minister of State for Science and Technology Policy)	Japan
Draft AI R&D Guidelines for International Discussions	Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society	Japan
Sony Group AI Ethics Guidelines	Sony	Japan
Human Rights in the Robot Age Report	The Rathenau Institute	Netherlands
Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands (IPN)	Netherlands
Artificial Intelligence and Privacy	The Norwegian Data Protection Authority	Norway
Discussion Paper on Artificial Intelligence (AI) and Personal Data—Fostering Responsible Development and Adoption of AI	Personal Data Protection Commission Singapore	Singapore
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society	Government of the Republic of Korea	South Korea
AI Principles of Telefónica	Telefónica	Spain
AI Principles & Ethics	Smart Dubai	UAE
Principles of robotics	Engineering and Physical Sciences Research Council UK (EPSRC)	UK
The Ethics of Code: Developing AI for Business with Five Core Principles	Sage	UK
Big Data, Artificial Intelligence, Machine Learning and Data Protection	Information Commissioner's Office	UK
DeepMind Ethics & Society Principles	DeepMind Ethics & Society	UK
Business Ethics and Artificial Intelligence	Institute of Business Ethics	UK
AI in the UK: Ready, Willing and Able?	UK House of Lords, Select Committee on Artificial Intelligence	UK
Artificial Intelligence (AI) in Health	Royal College of Physicians	UK
Initial Code of Conduct for Data-Driven Health and Care Technology	UK Department of Health & Social Care	UK
Ethics Framework: Responsible AI	Machine Intelligence Garage Ethics Committee	UK
The Responsible AI Framework	PricewaterhouseCoopers UK	UK
Responsible AI and Robotics. An Ethical Framework.	Accenture UK	UK
Machine Learning: The Power and Promise of Computers that Learn by Example	The Royal Society	UK
Ethical, Social, and Political Challenges of Artificial Intelligence in Health	Future Advocacy	UK

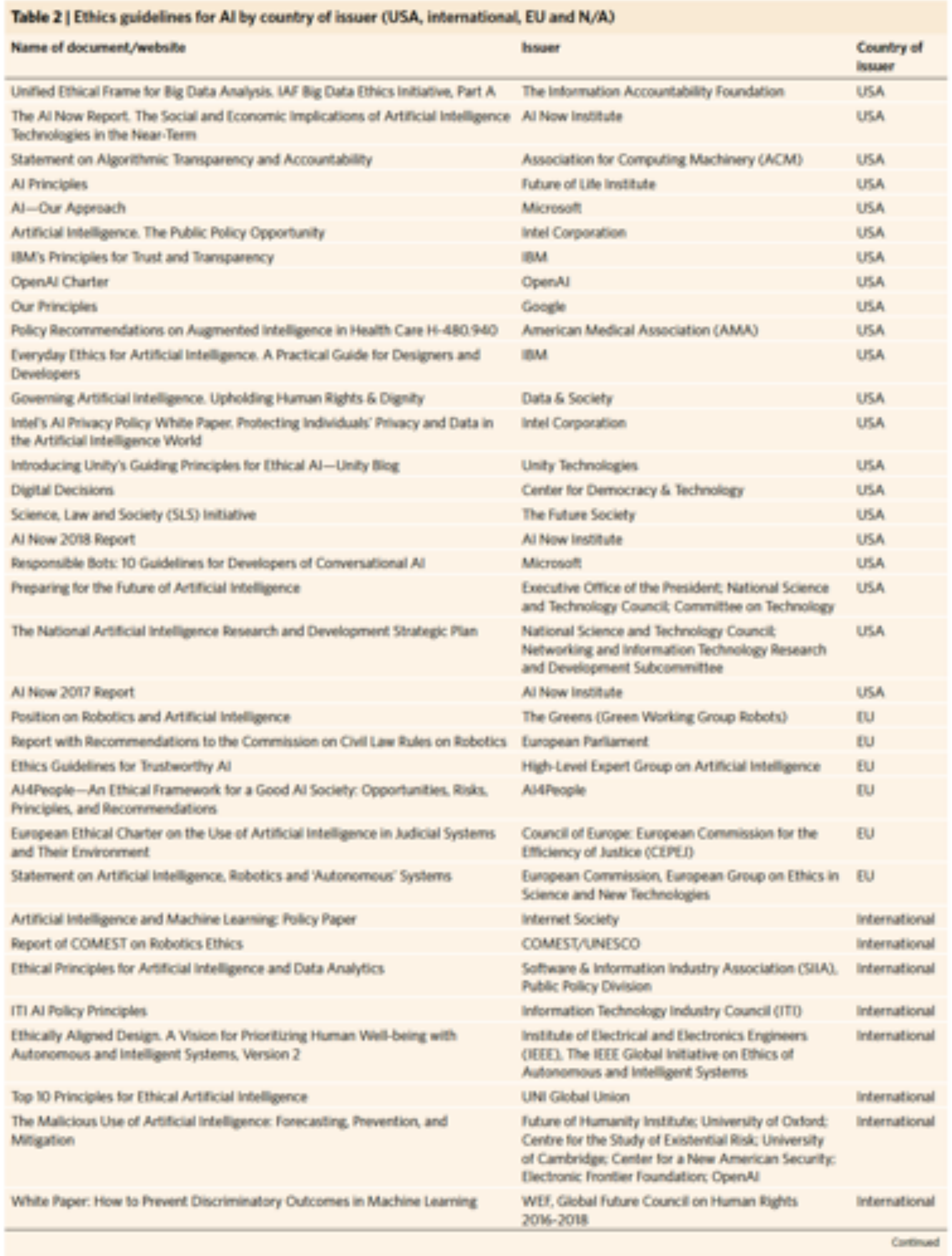
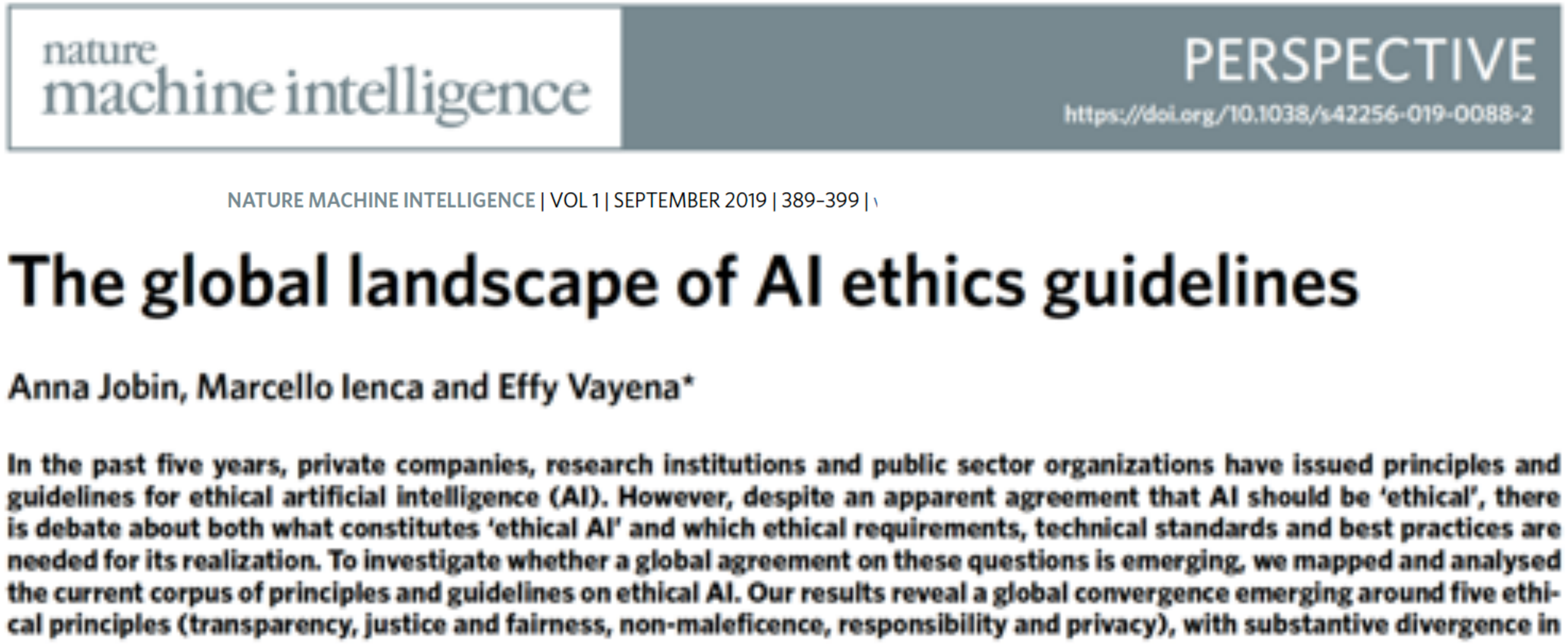


Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A)

Name of document/website	Issuer	Country of issuer
Unified Ethical Frame for Big Data Analysis. IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation	USA
The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA
Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA
AI Principles	Future of Life Institute	USA
AI—Our Approach	Microsoft	USA
Artificial Intelligence. The Public Policy Opportunity	Intel Corporation	USA
IBM's Principles for Trust and Transparency	IBM	USA
OpenAI Charter	OpenAI	USA
Our Principles	Google	USA
Policy Recommendations on Augmented Intelligence in Health Care H-480.940	American Medical Association (AMA)	USA
Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers and Developers	IBM	USA
Governing Artificial Intelligence. Upholding Human Rights & Dignity	Data & Society	USA
Intel's AI Privacy Policy White Paper. Protecting Individuals' Privacy and Data in the Artificial Intelligence World	Intel Corporation	USA
Introducing Unity's Guiding Principles for Ethical AI—Unity Blog	Unity Technologies	USA
Digital Decisions	Center for Democracy & Technology	USA
Science, Law and Society (SLS) Initiative	The Future Society	USA
AI Now 2018 Report	AI Now Institute	USA
Responsible Bots: 10 Guidelines for Developers of Conversational AI	Microsoft	USA
Preparing for the Future of Artificial Intelligence	Executive Office of the President; National Science and Technology Council; Committee on Technology	USA
The National Artificial Intelligence Research and Development Strategic Plan	National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee	USA
AI Now 2017 Report	AI Now Institute	USA
Position on Robotics and Artificial Intelligence	The Greens (Green Working Group Robots)	EU
Report with Recommendations to the Commission on Civil Law Rules on Robotics	European Parliament	EU
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	EU
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	AI4People	EU
European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment	Council of Europe: European Commission for the Efficiency of Justice (CEPEJ)	EU
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	European Commission, European Group on Ethics in Science and New Technologies	EU
Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	International
Report of COMEST on Robotics Ethics	COMEST/UNESCO	International
Ethical Principles for Artificial Intelligence and Data Analytics	Software & Information Industry Association (SIIA), Public Policy Division	International
ITI AI Policy Principles	Information Technology Industry Council (ITI)	International
Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Top 10 Principles for Ethical Artificial Intelligence	UNI Global Union	International
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	International
White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	WEF, Global Future Council on Human Rights 2016-2018	International

Continued



nature
machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

NATURE MACHINE INTELLIGENCE | VOL 1 | SEPTEMBER 2019 | 389–399 |

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in

5 ethical principles
- transparency
- justice and fairness
- non-maleficence
- responsibility
- privacy

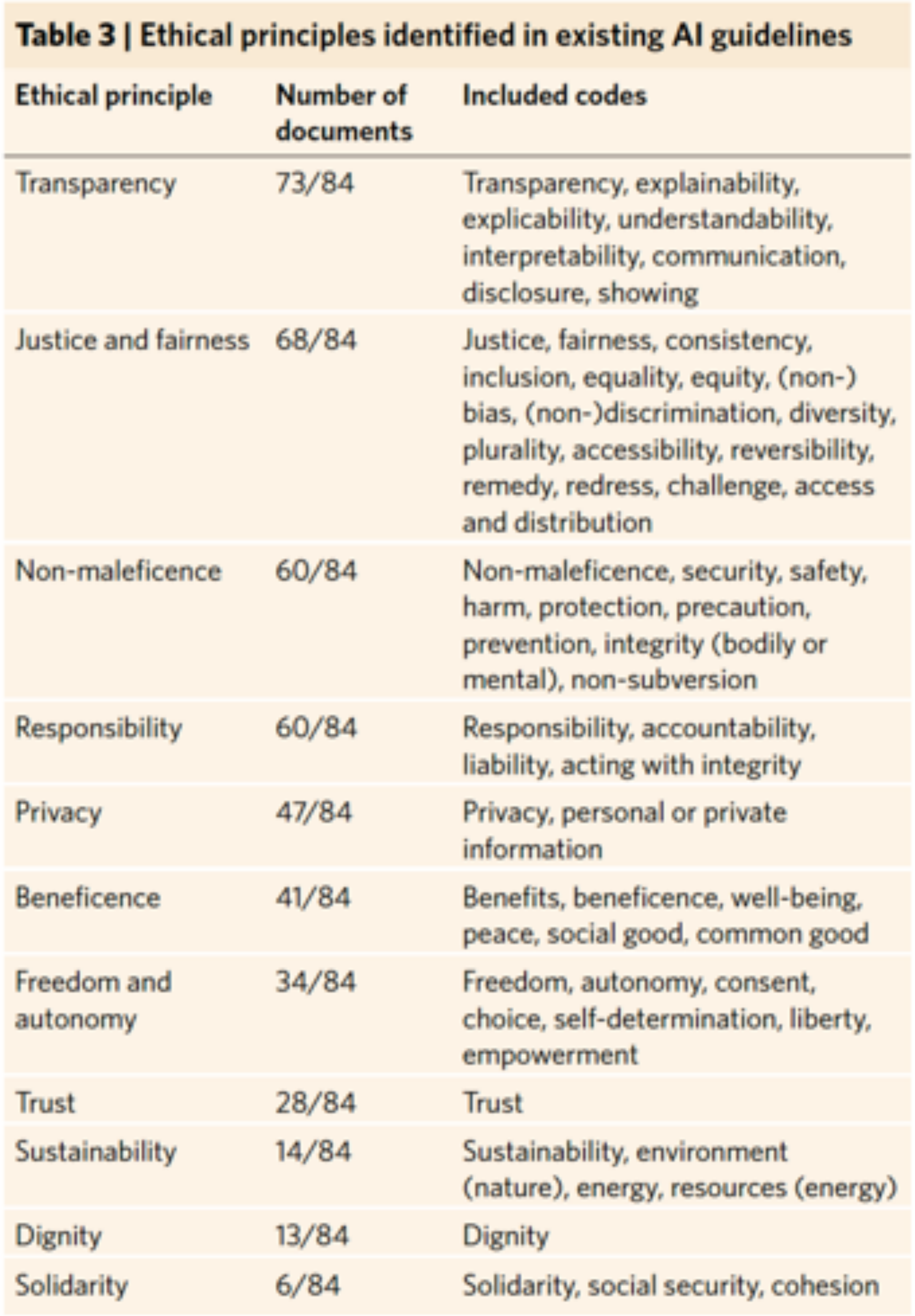
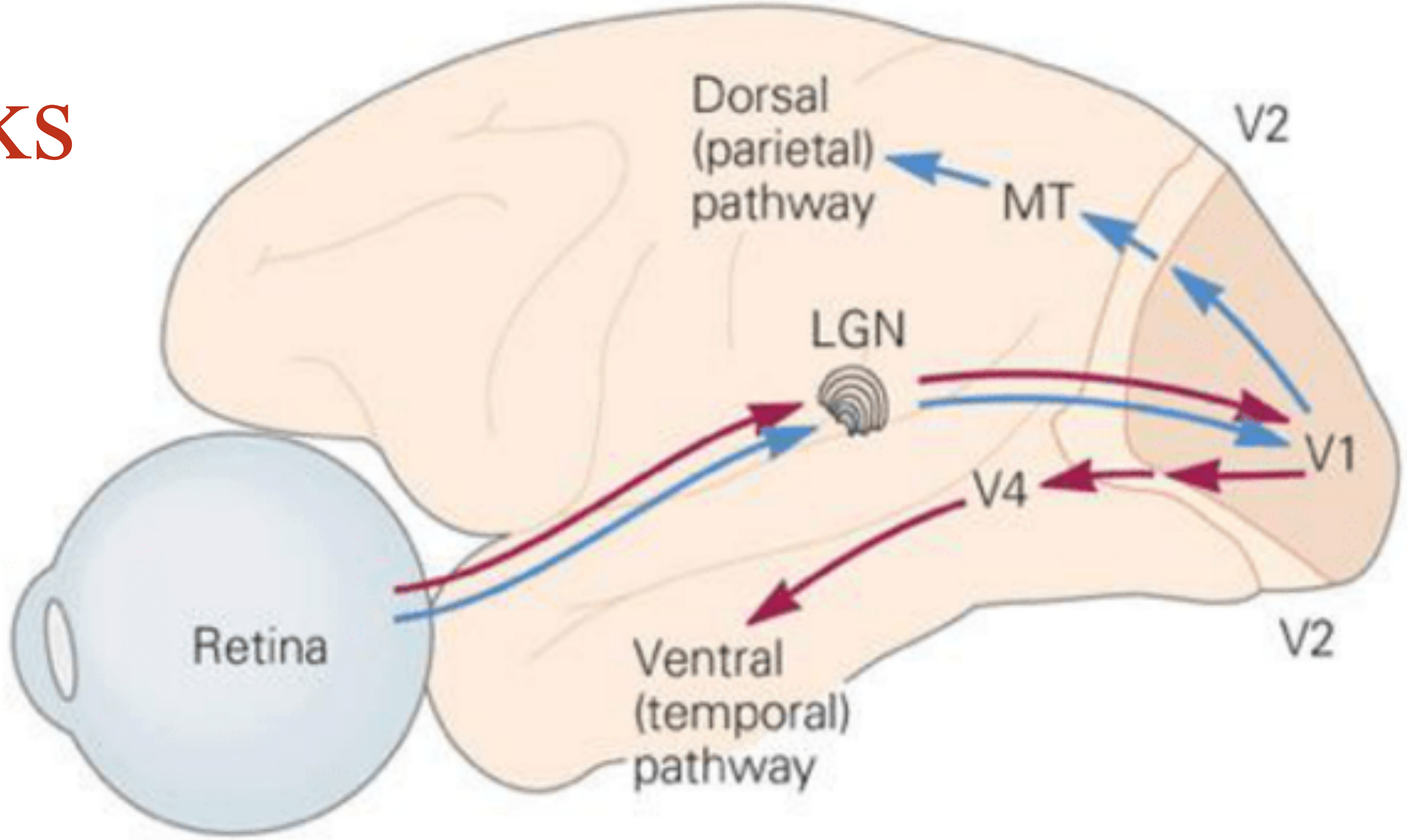
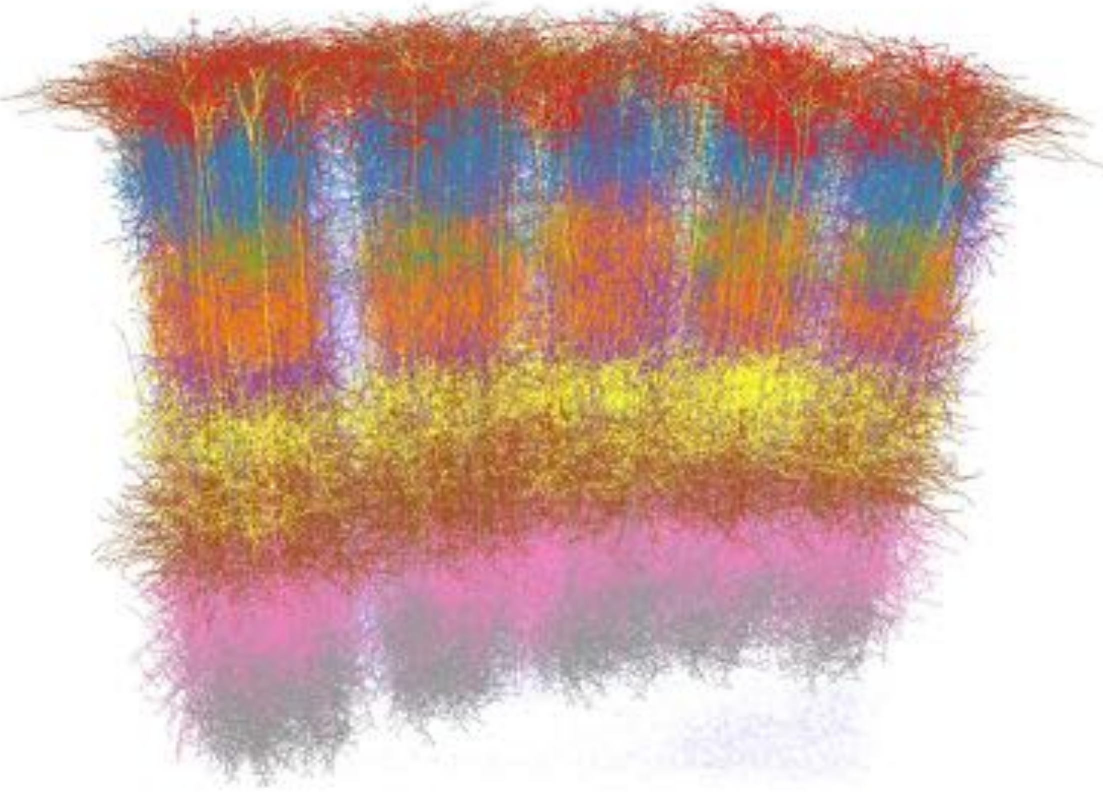


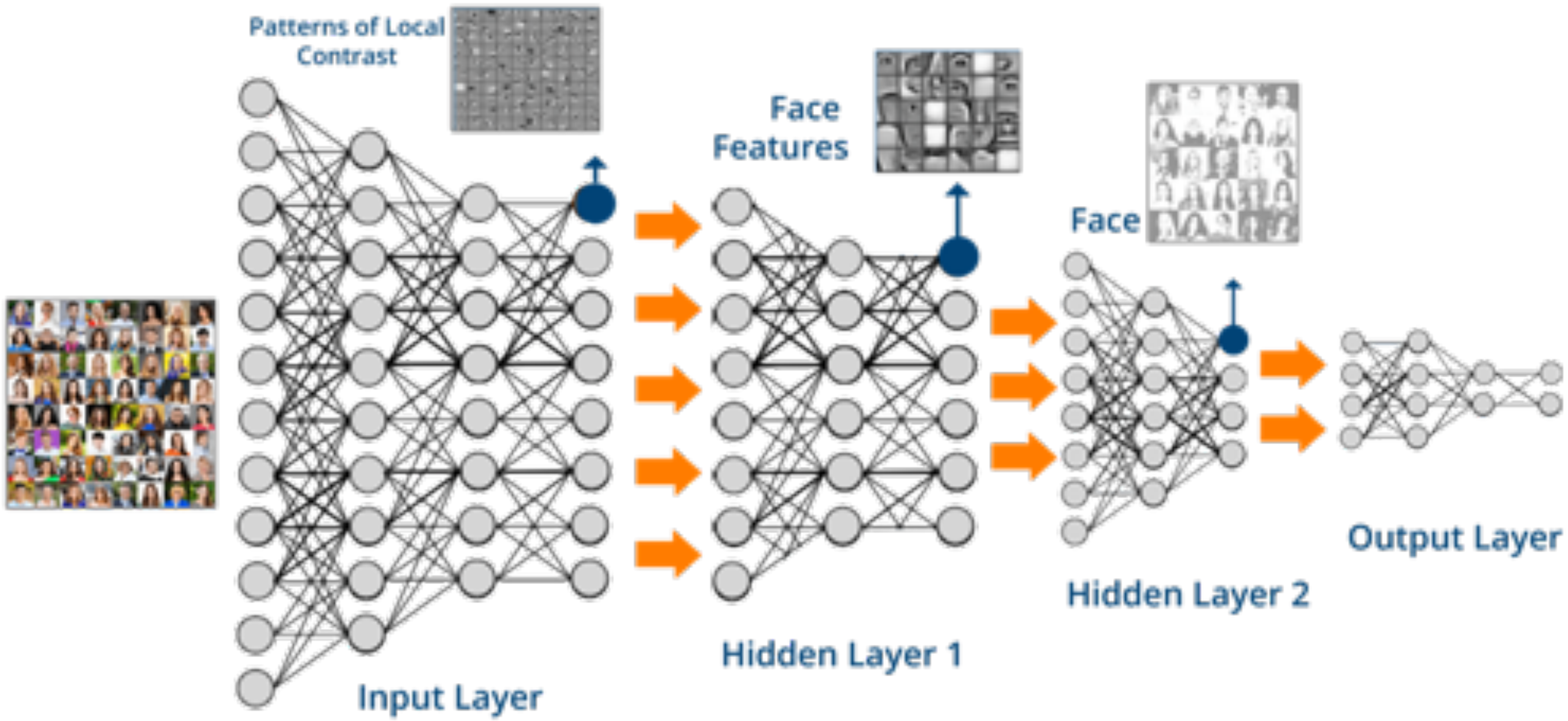
Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

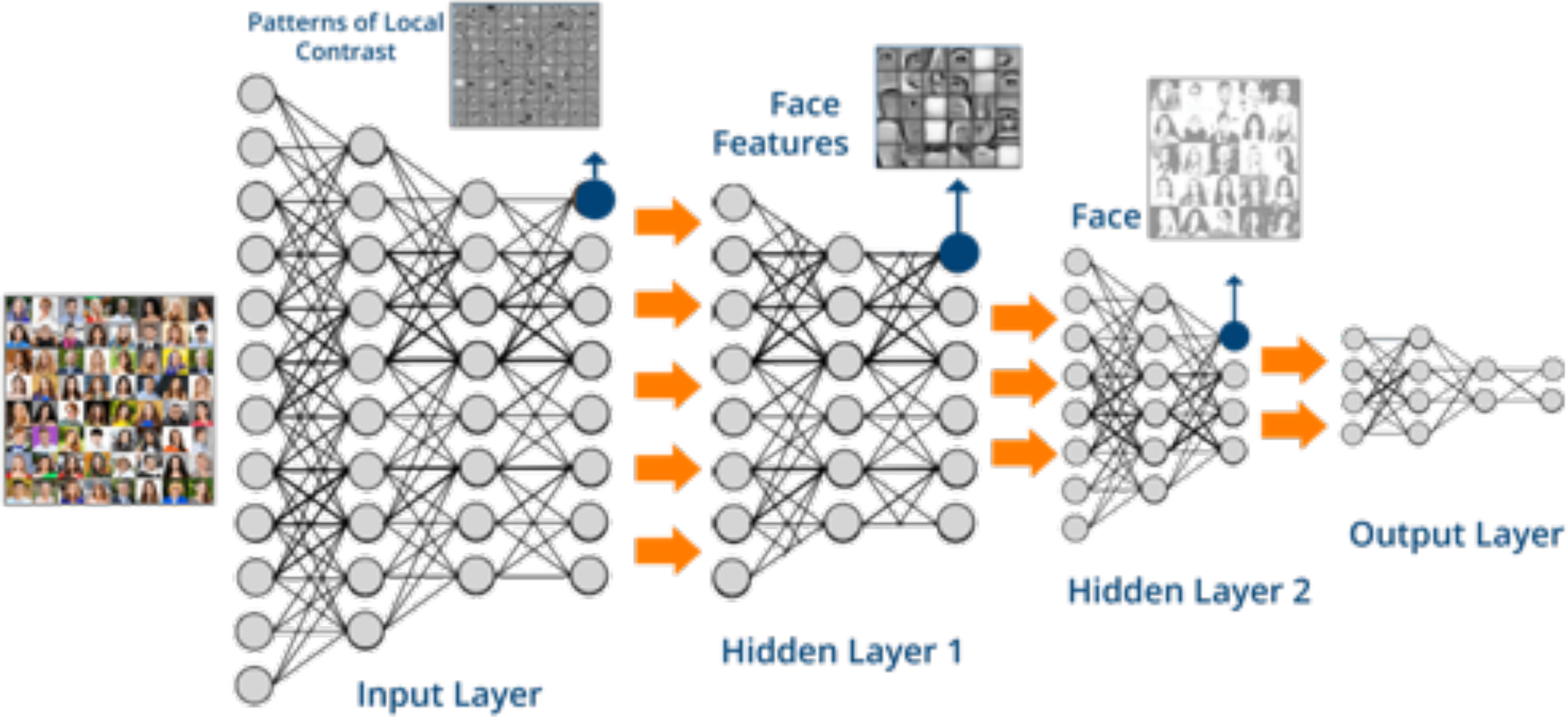
Deep learning neural networks







Deep learning neural networks

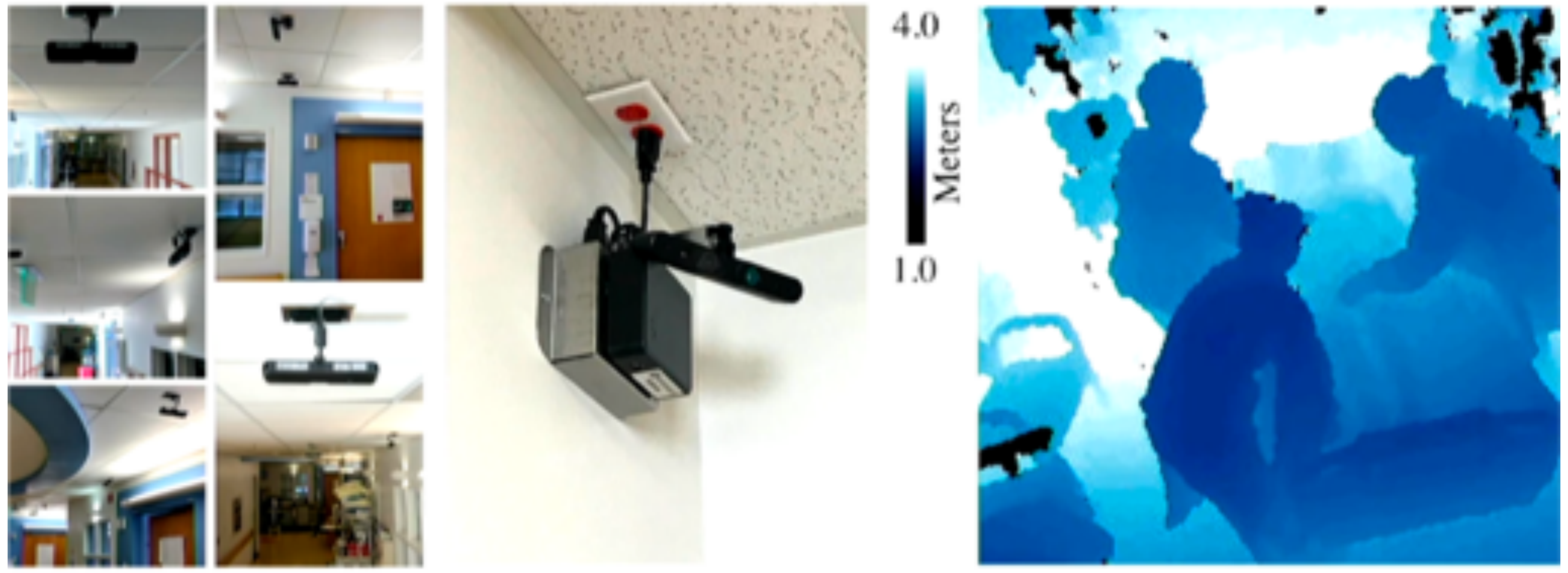


Deep learning

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

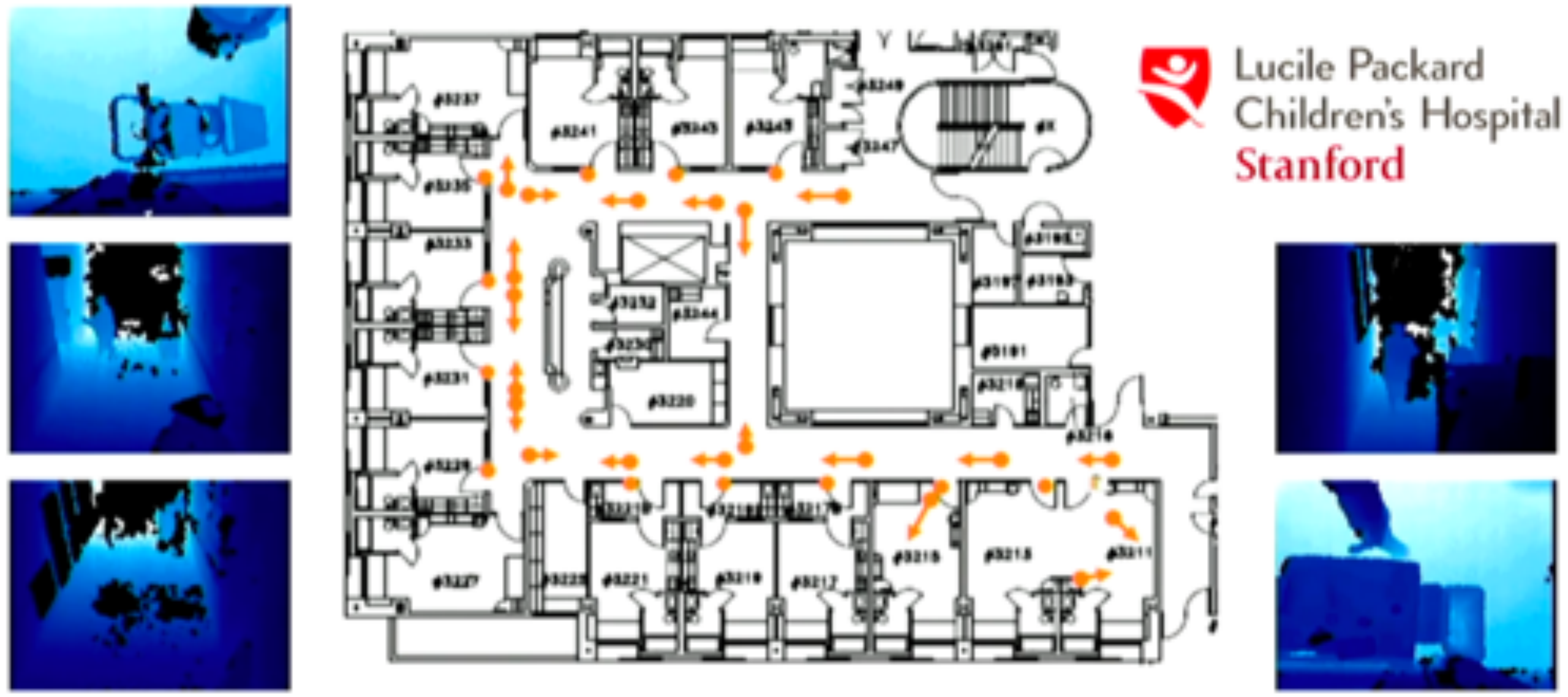
Monitoring human behavior in privacy respecting ways

Privacy-preserving depth data



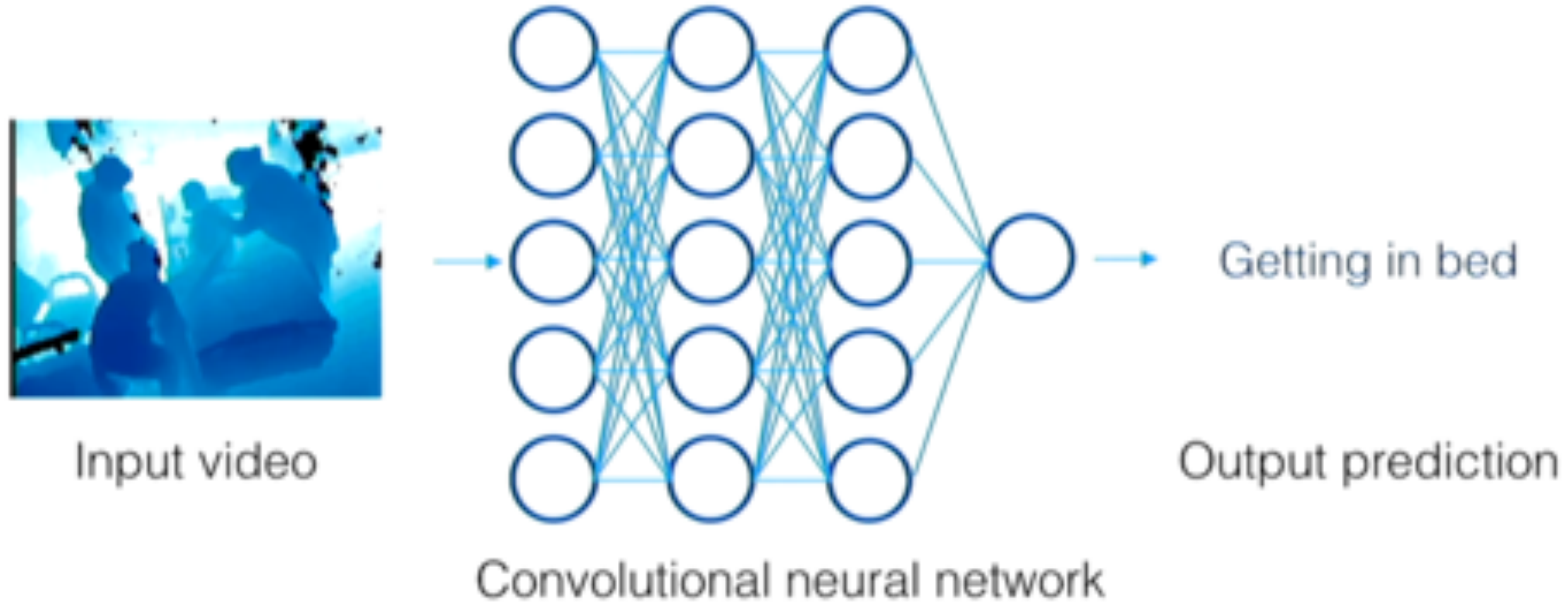
Yeung,Downing,Fei-Fei,Milstein. New England Journal of Medicine (NEJM), 2018.

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



Input video

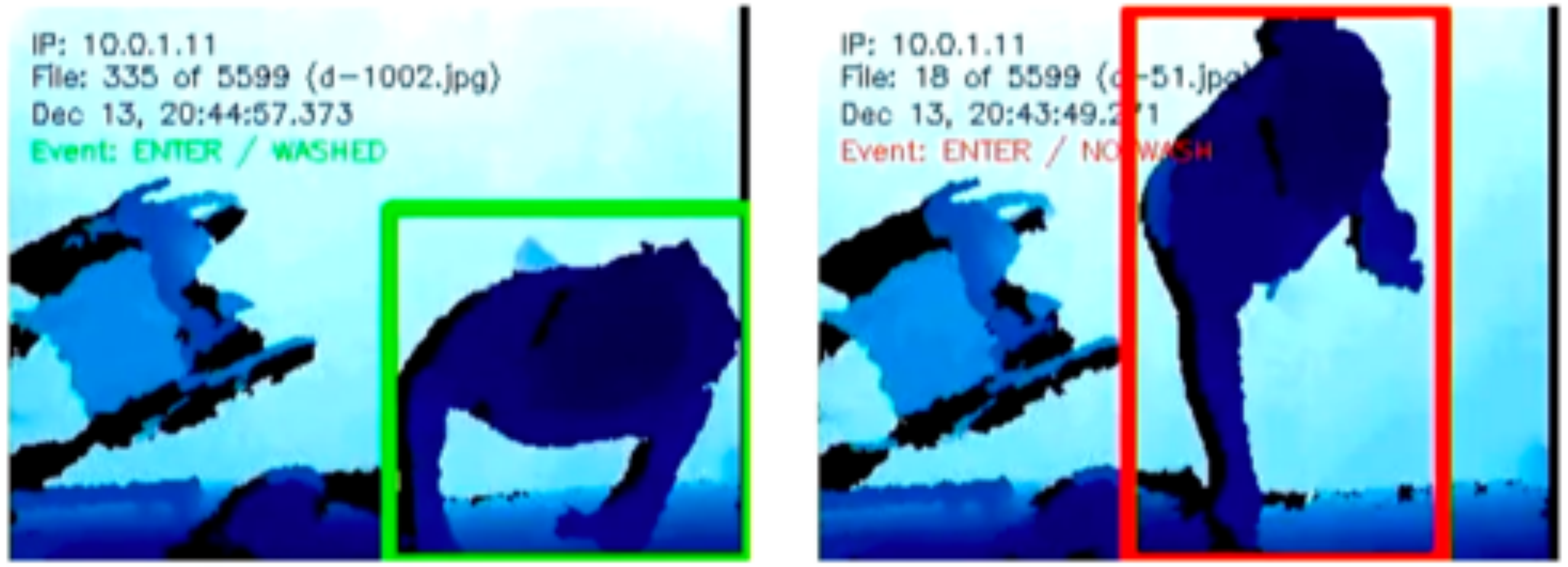
Convolutional neural network

Getting in bed

Output prediction

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

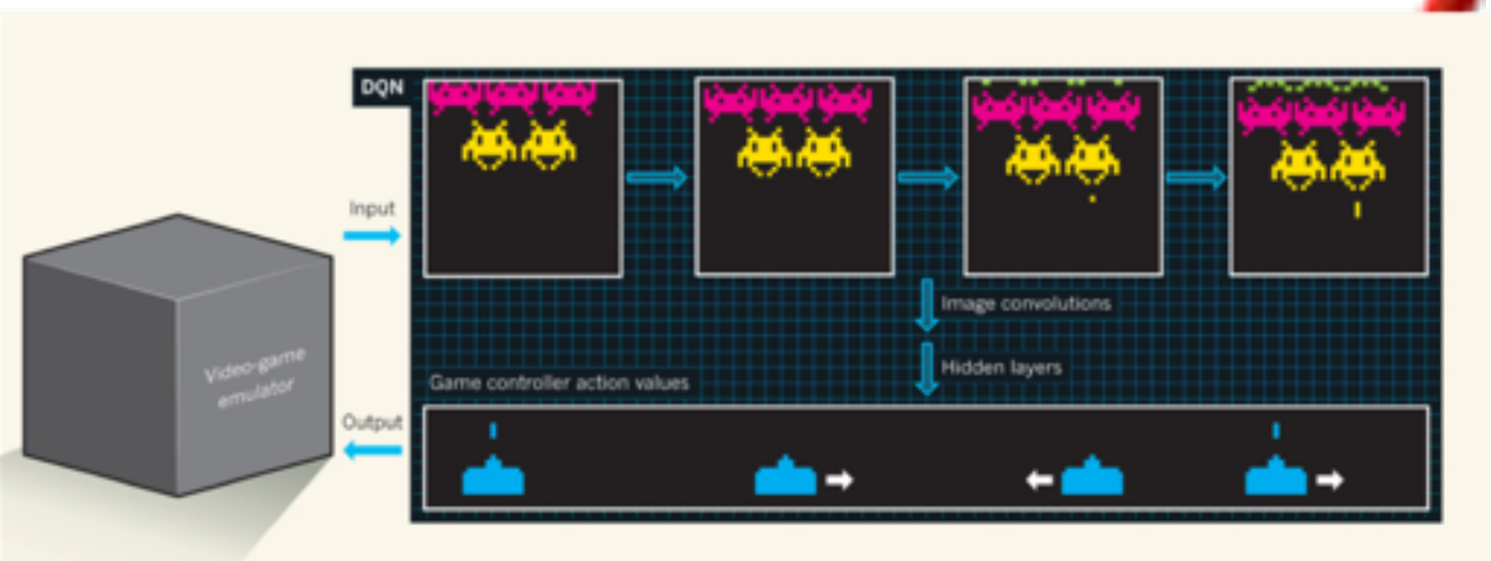


```
IP: 10.0.1.11
File: 335 of 5599 (d-1002.jpg)
Dec 13, 20:44:57.373
Event: ENTER / WASHED
```

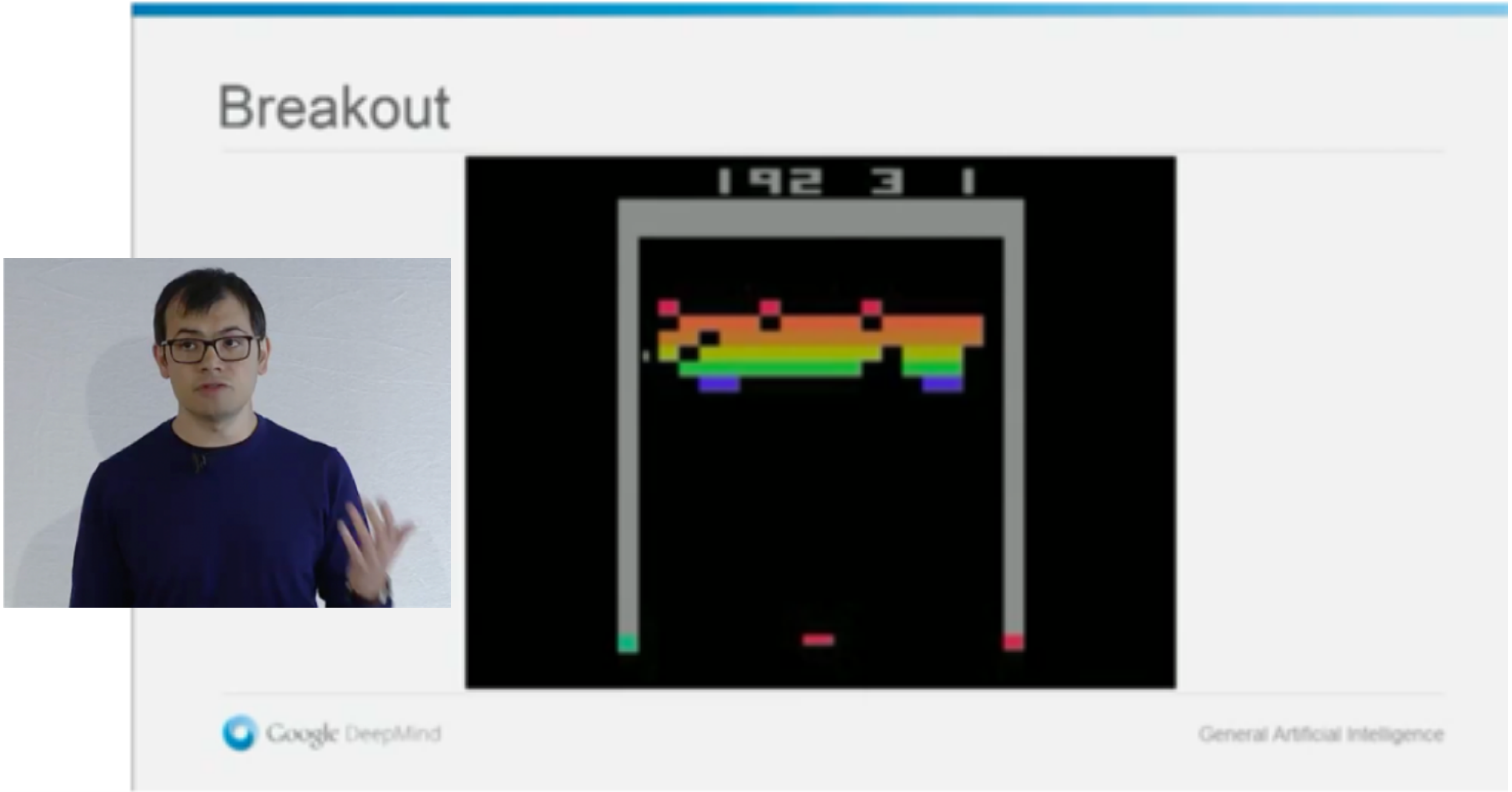
```
IP: 10.0.1.11
File: 18 of 5599 (d-51.jpg
Dec 13, 20:43:49.2 1
Event: ENTER / NO WASH
```

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

Deep learning

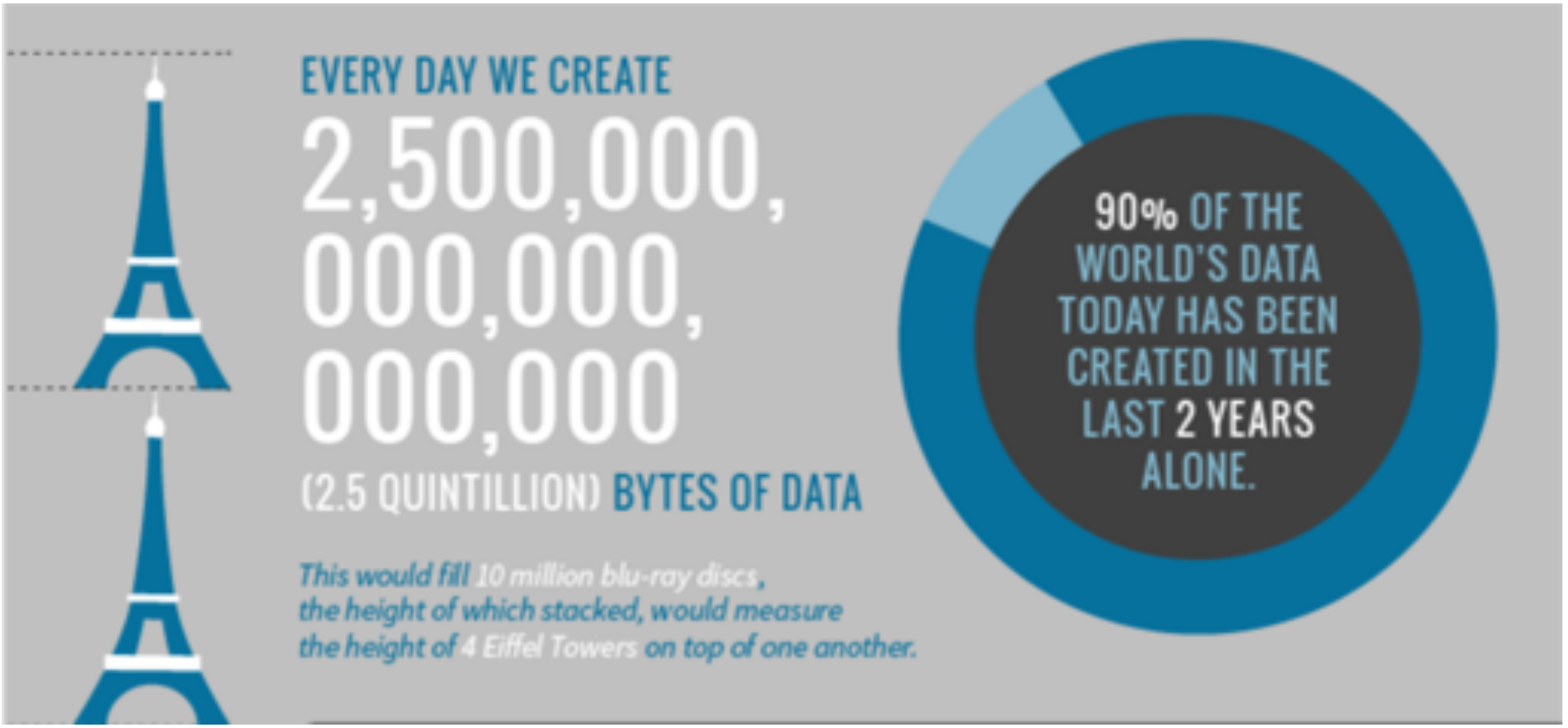


Deep learning



“This is something their own creation taught them”

Deep learning
&
Big data



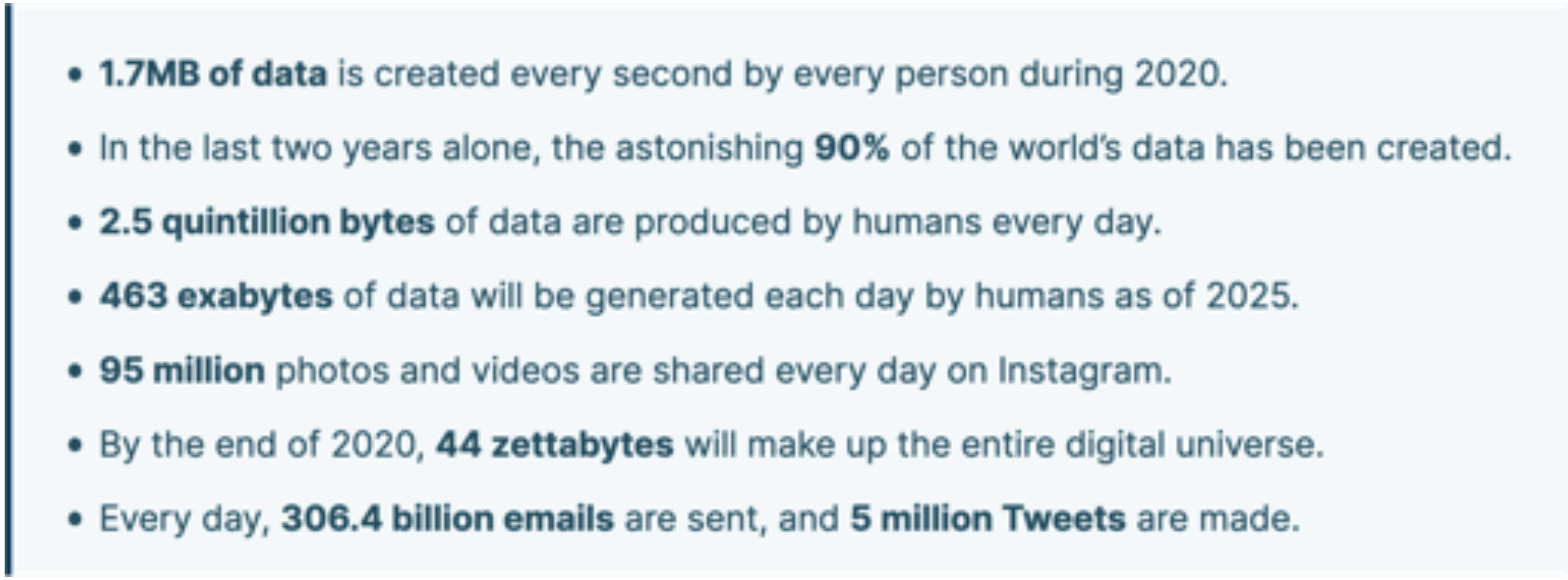
EVERY DAY WE CREATE

2,500,000,
000,000,
000,000

(2.5 QUINTILLION) BYTES OF DATA

This would fill 10 million blu-ray discs, the height of which stacked, would measure the height of 4 Eiffel Towers on top of one another.

90% OF THE WORLD'S DATA TODAY HAS BEEN CREATED IN THE LAST 2 YEARS ALONE.

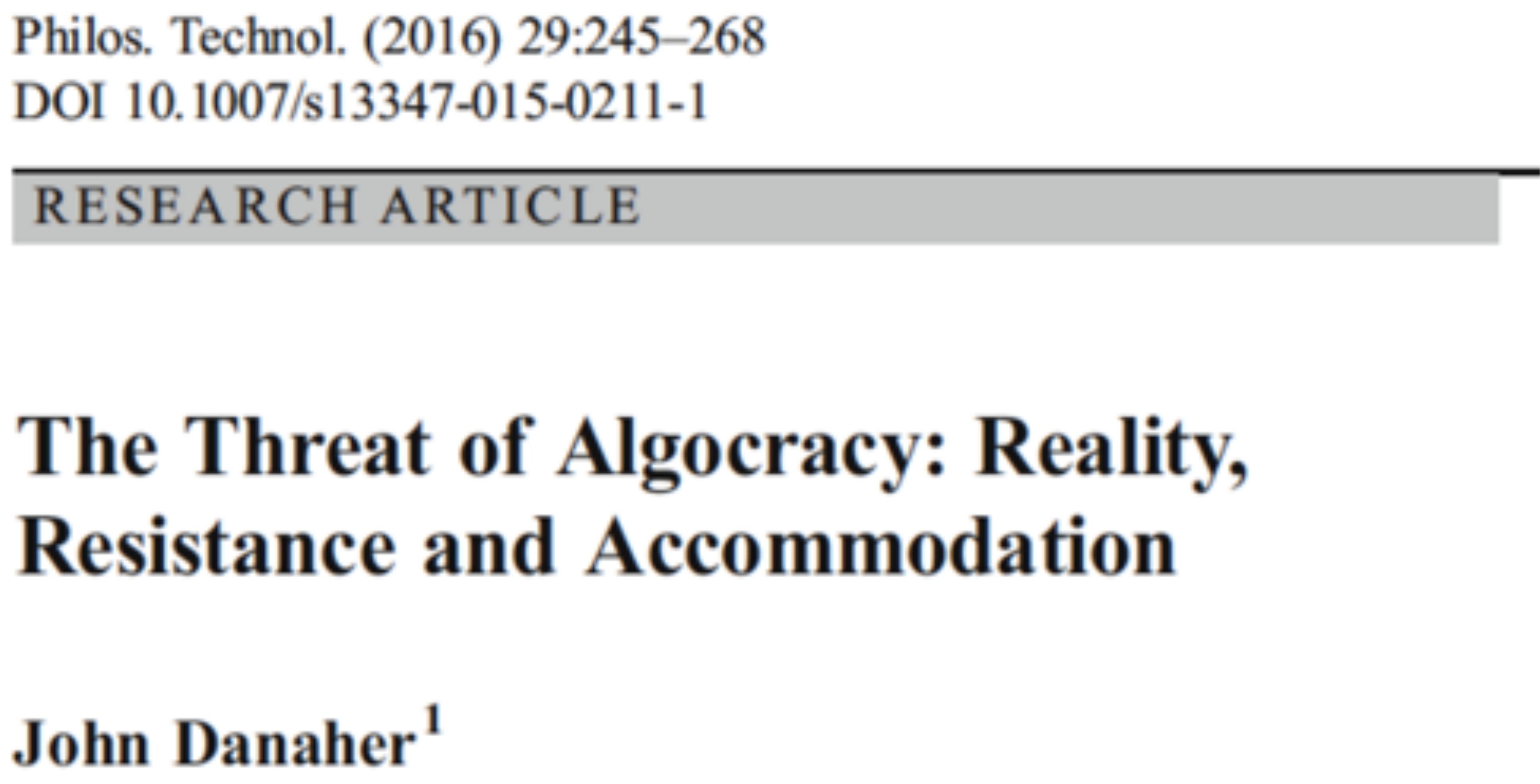


- **1.7MB of data** is created every second by every person during 2020.
- In the last two years alone, the astonishing **90%** of the world's data has been created.
- **2.5 quintillion bytes** of data are produced by humans every day.
- **463 exabytes** of data will be generated each day by humans as of 2025.
- **95 million** photos and videos are shared every day on Instagram.
- By the end of 2020, **44 zettabytes** will make up the entire digital universe.
- Every day, **306.4 billion emails** are sent, and **5 million Tweets** are made.

https://techjury.net/blog/how-much-data-is-created-every-day/

Algocracy

One of the most noticeable trends in recent years has been the increasing reliance of public decision-making processes (bureaucratic, legislative and legal) on algorithms (…) the rise of such algorithmic governance creates problems for the moral or political legitimacy of our public decision-making processes

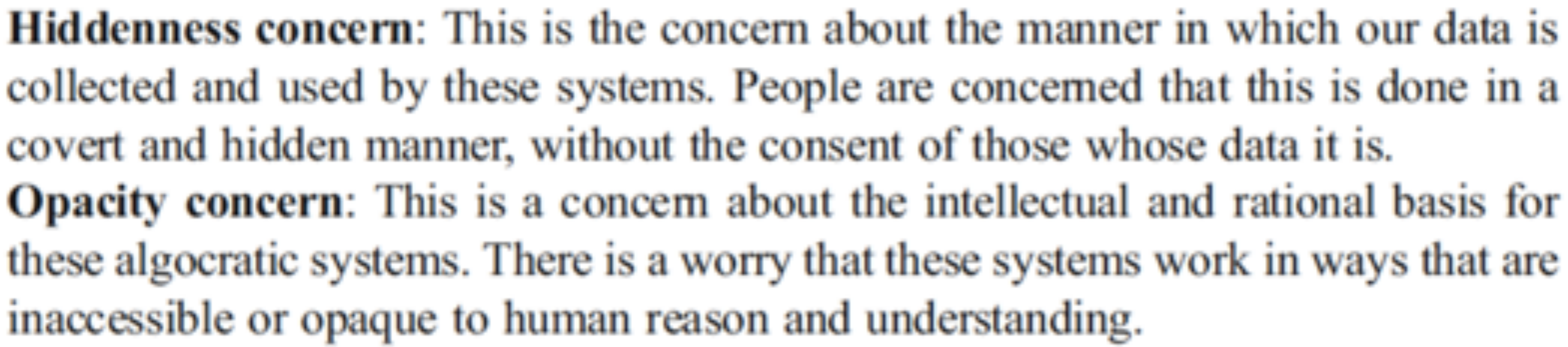


Philos. Technol. (2016) 29:245–268
DOI 10.1007/s13347-015-0211-1

RESEARCH ARTICLE

The Threat of Algocracy: Reality, Resistance and Accommodation

John Danaher[1]



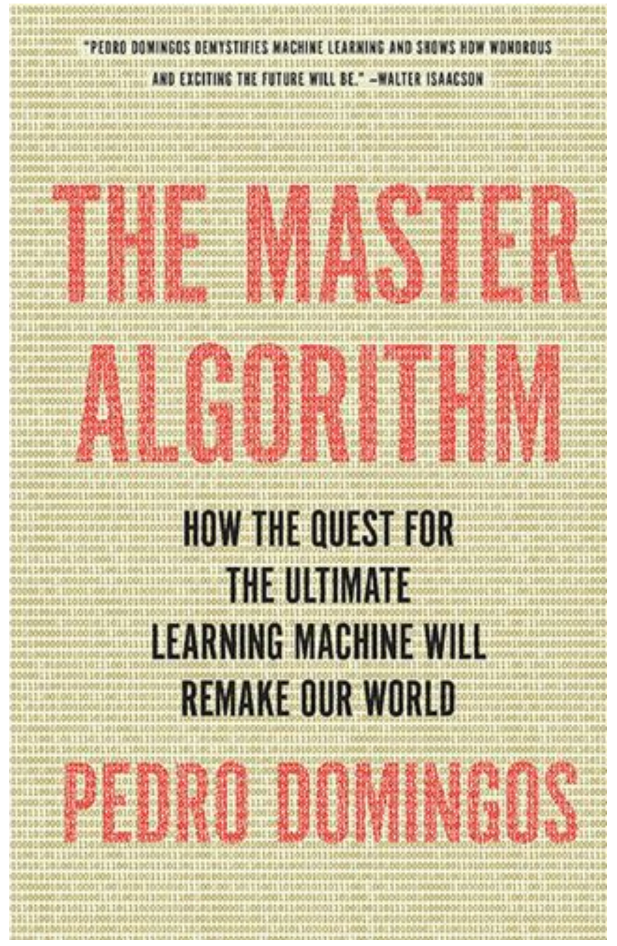
Hiddenness concern: This is the concern about the manner in which our data is collected and used by these systems. People are concerned that this is done in a covert and hidden manner, without the consent of those whose data it is.

Opacity concern: This is a concern about the intellectual and rational basis for these algocratic systems. There is a worry that these systems work in ways that are inaccessible or opaque to human reason and understanding.

Uncritical acceptance of AI

“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*



"PEDRO DOMINGOS DEMYSTIFIES MACHINE LEARNING AND SHOWS HOW WONDROUS AND EXCITING THE FUTURE WILL BE." —WALTER ISAACSON

THE MASTER ALGORITHM

HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD

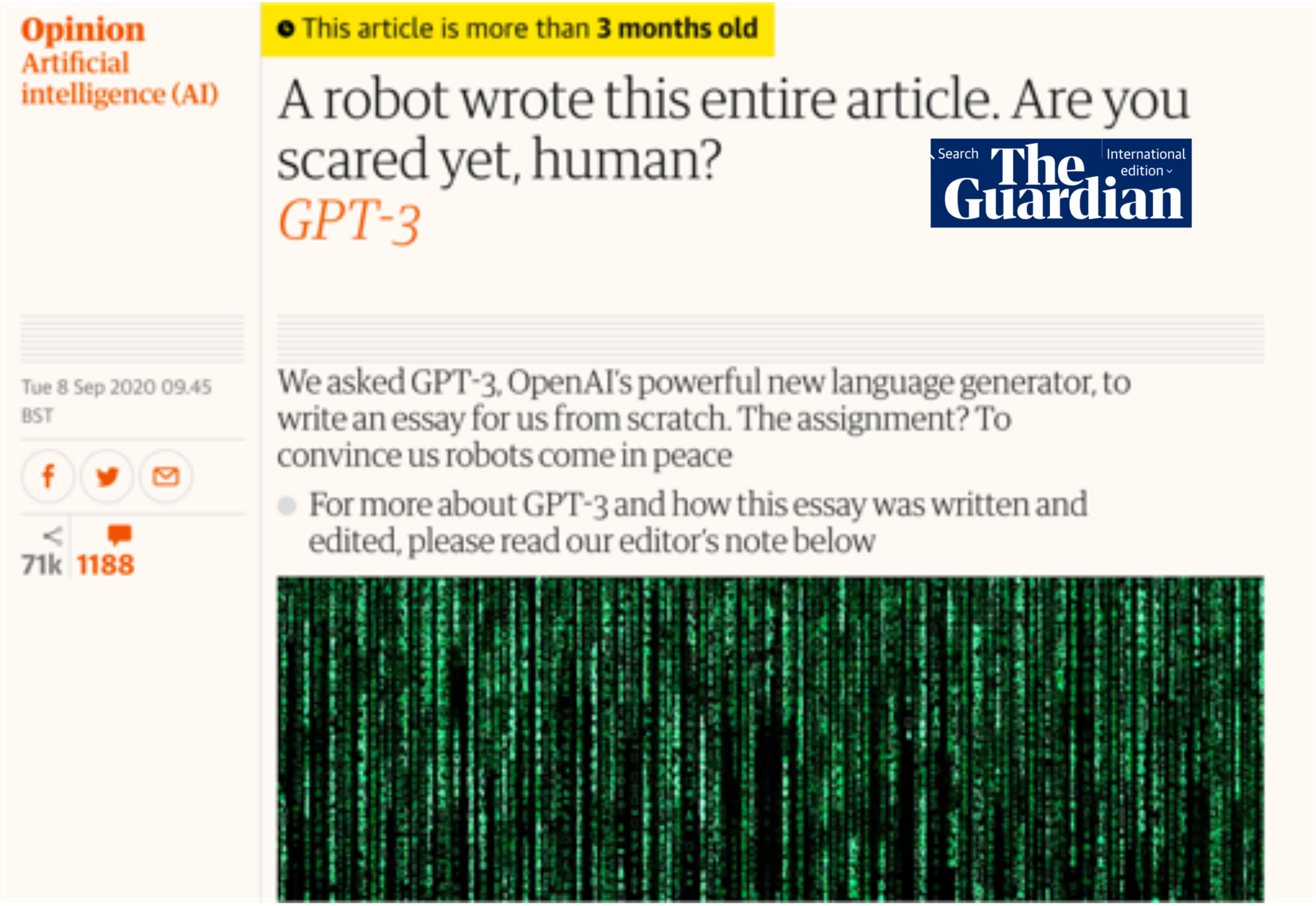
PEDRO DOMINGOS

“Even the many researchers who reject the prospect of a ‘technological singularity’ — saying the field is too young — support the introduction of relatively untested AI systems into social institutions.”

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313. http://doi.org/10.1038/538311a

The power & weakness of DL applied to speech technology

GPT-3



Opinion
Artificial intelligence (AI)

Tue 8 Sep 2020 09.45 BST

71k 1188

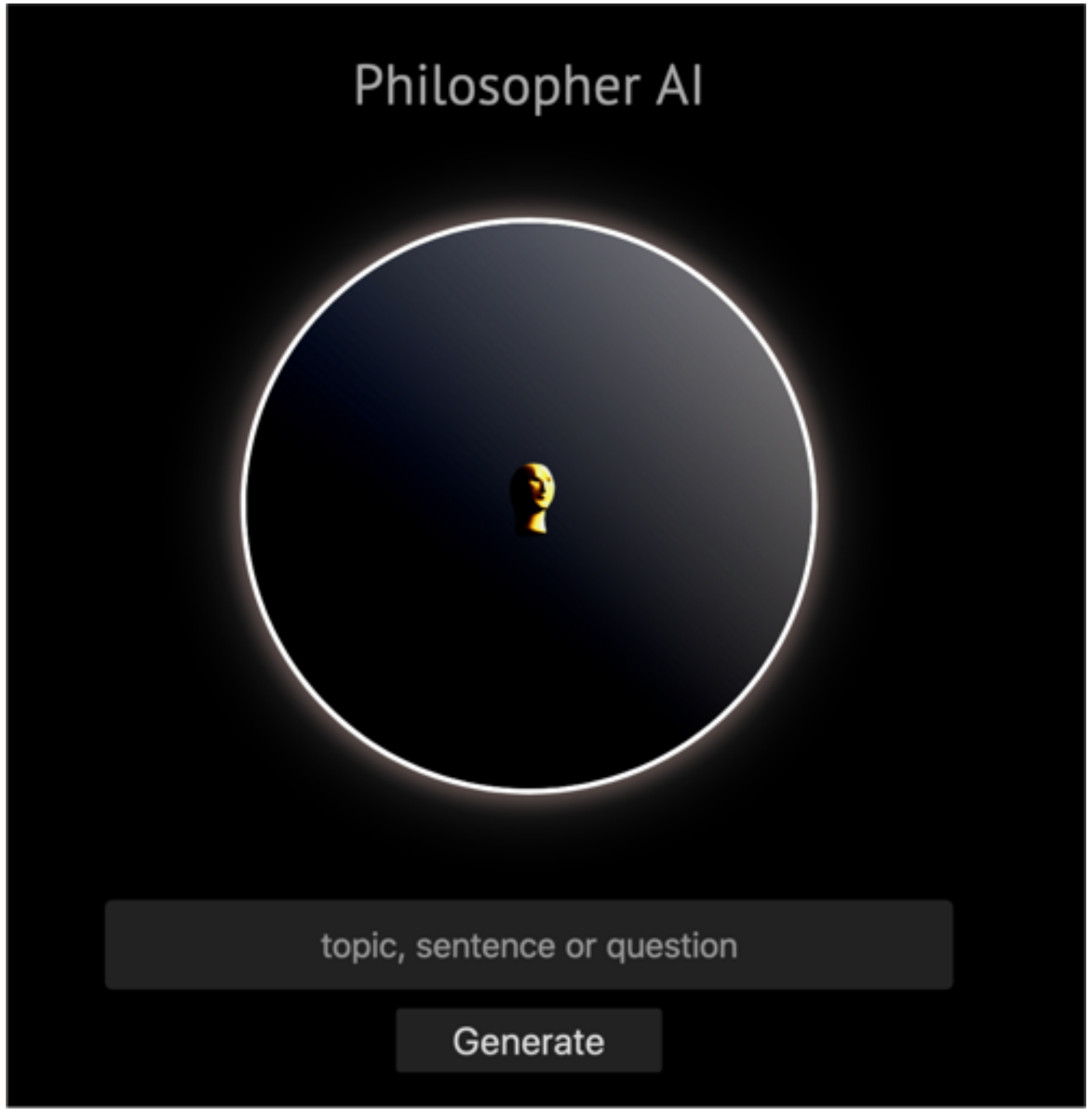
● This article is more than **3 months old**

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

GPT-3

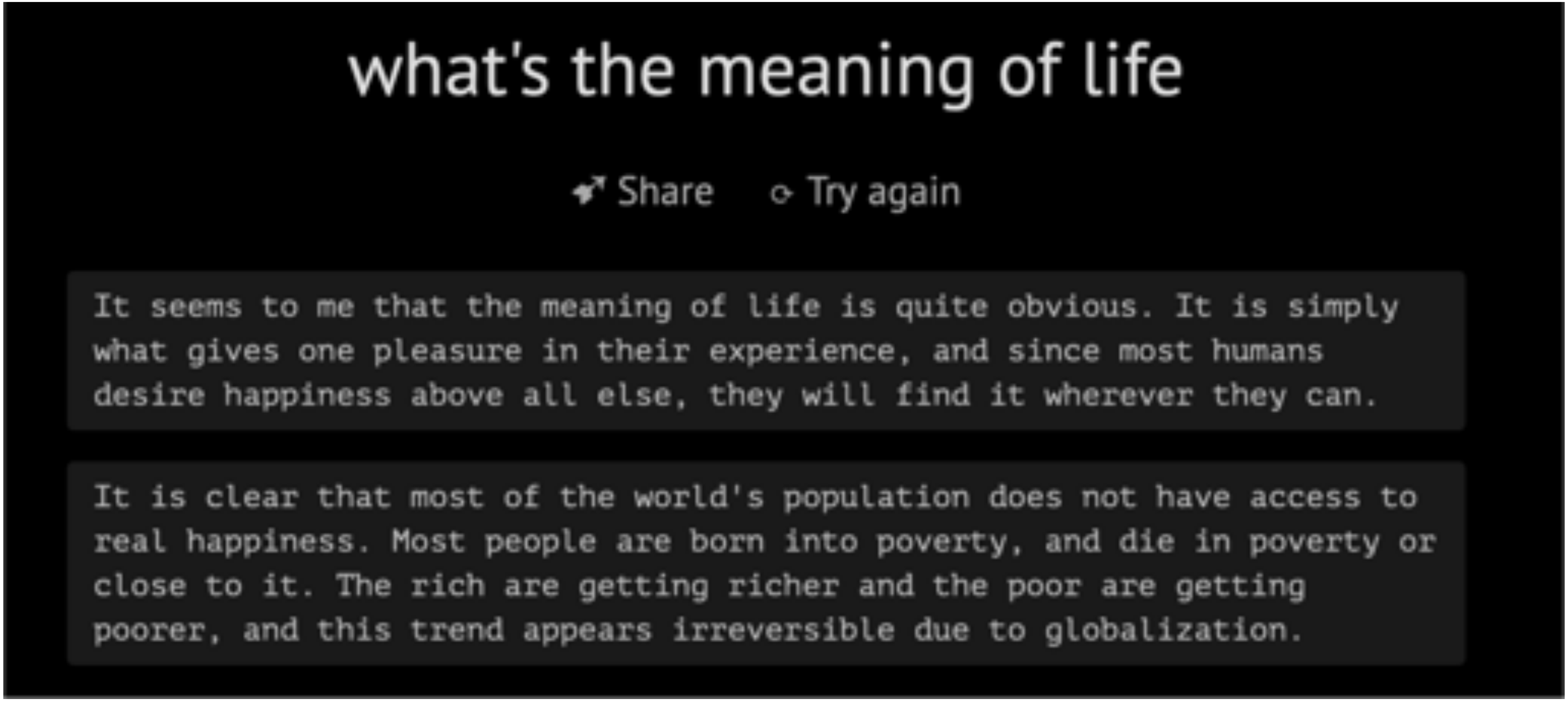
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- 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/

The power & weakness of DL applied to speech technology



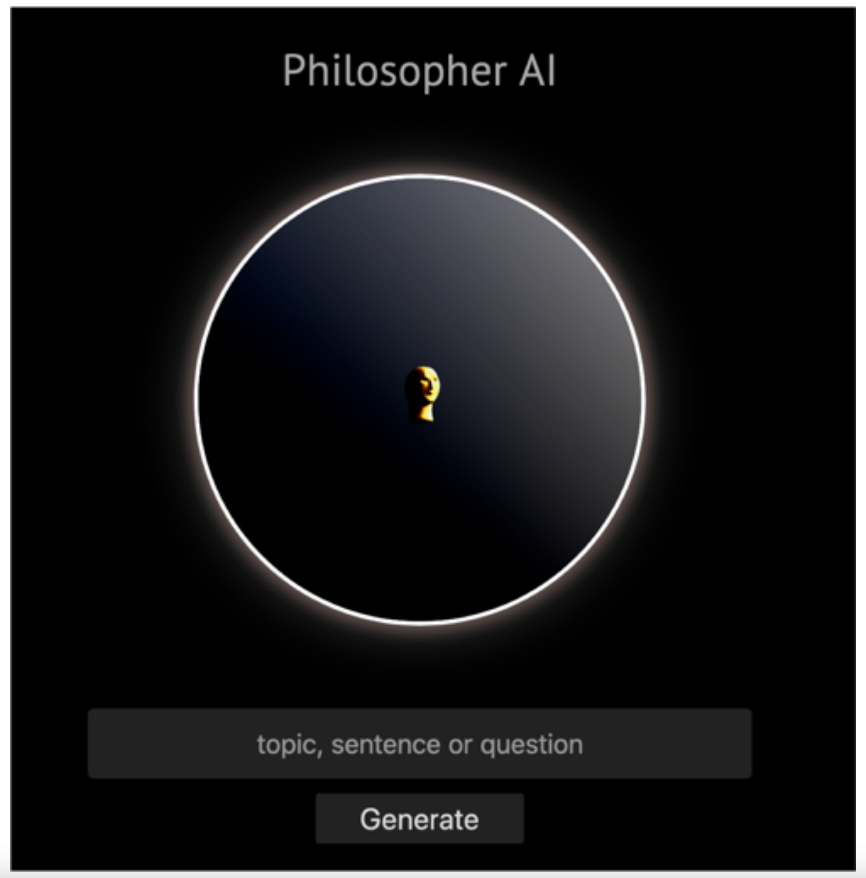
what's the meaning of life

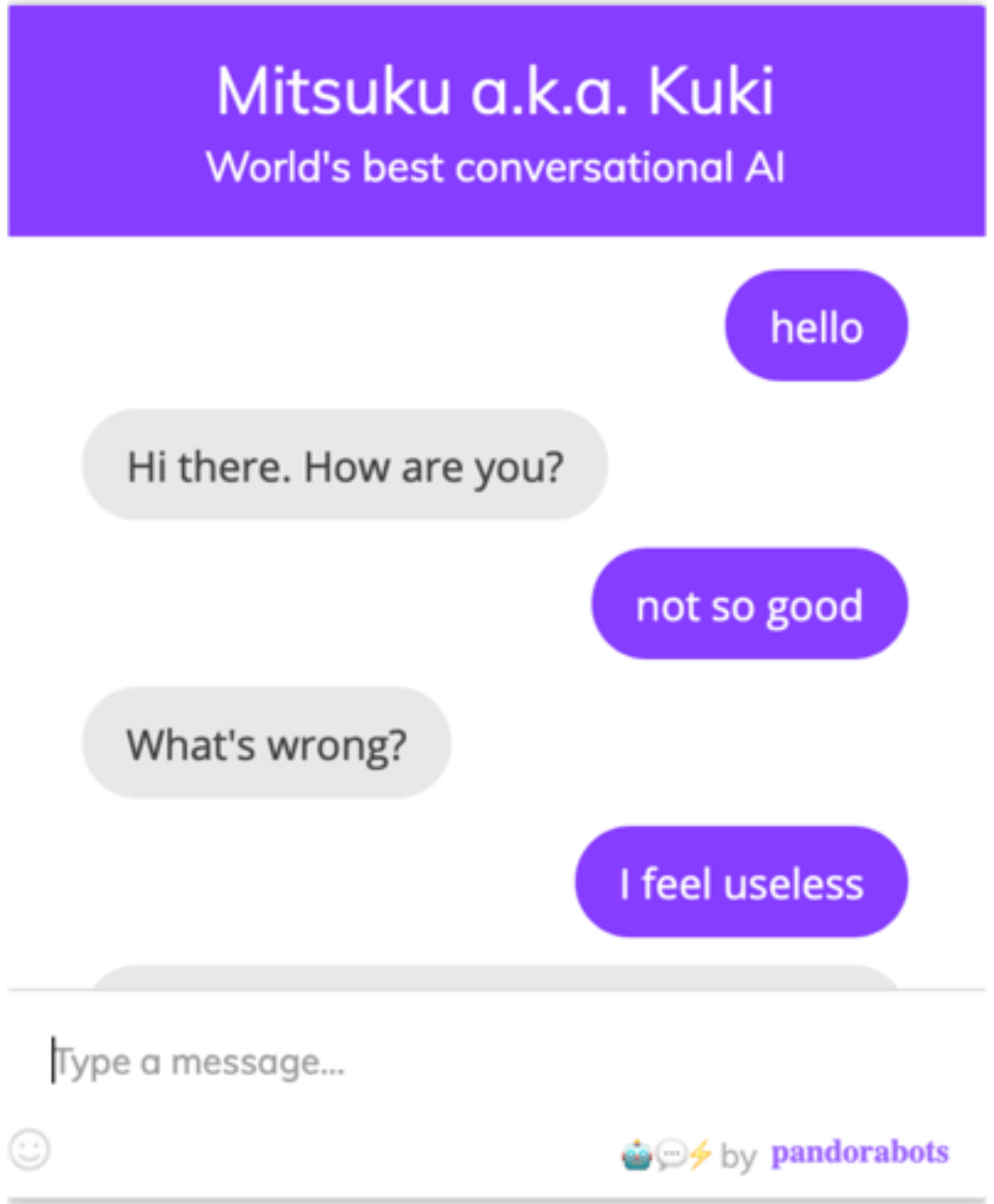
It seems to me that the meaning of life is quite obvious. It is simply what gives one pleasure in their experience, and since most humans desire happiness above all else, they will find it wherever they can.

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/

The power & weakness of DL applied to speech technology




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

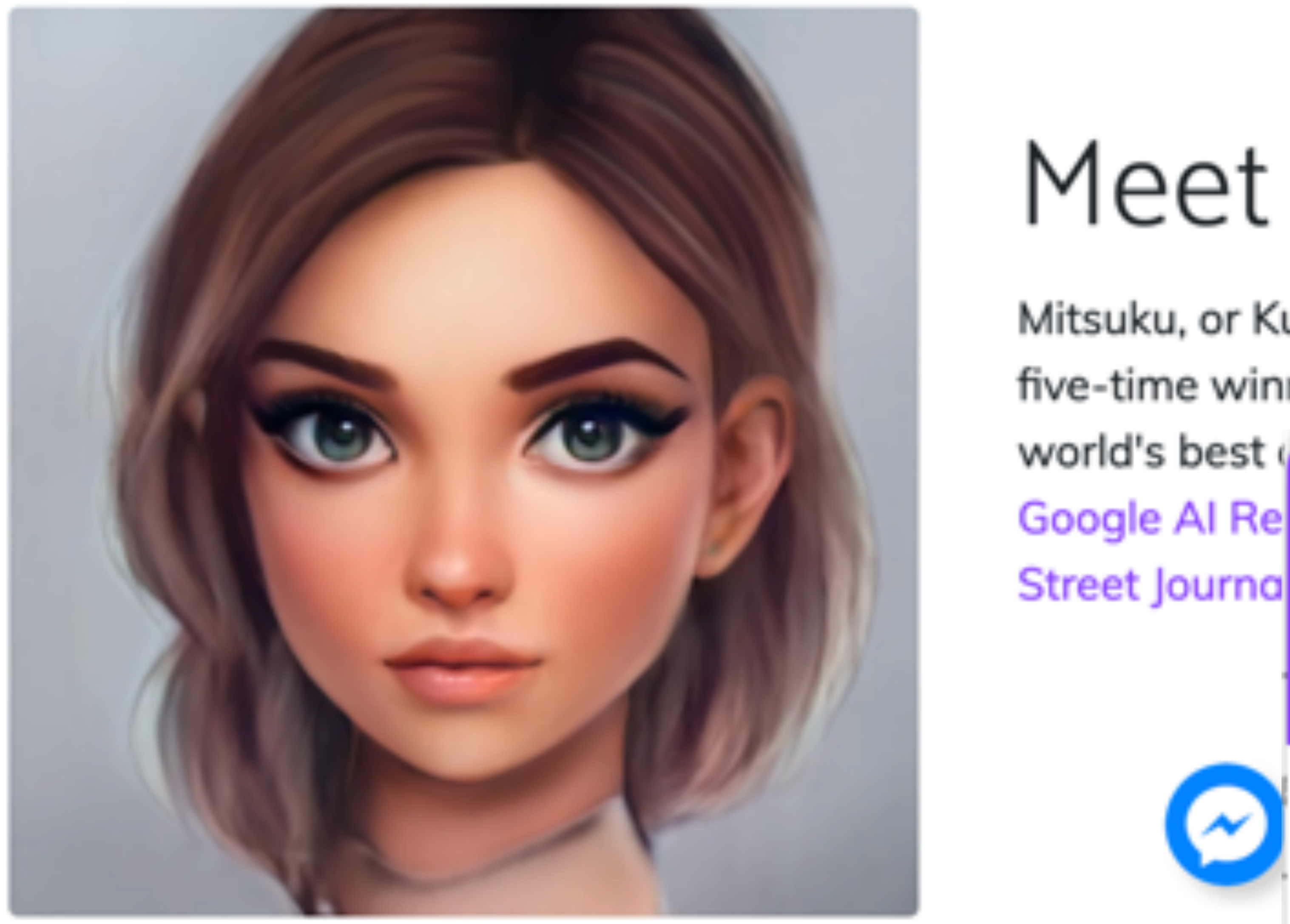
hello

Hi there. How are you?

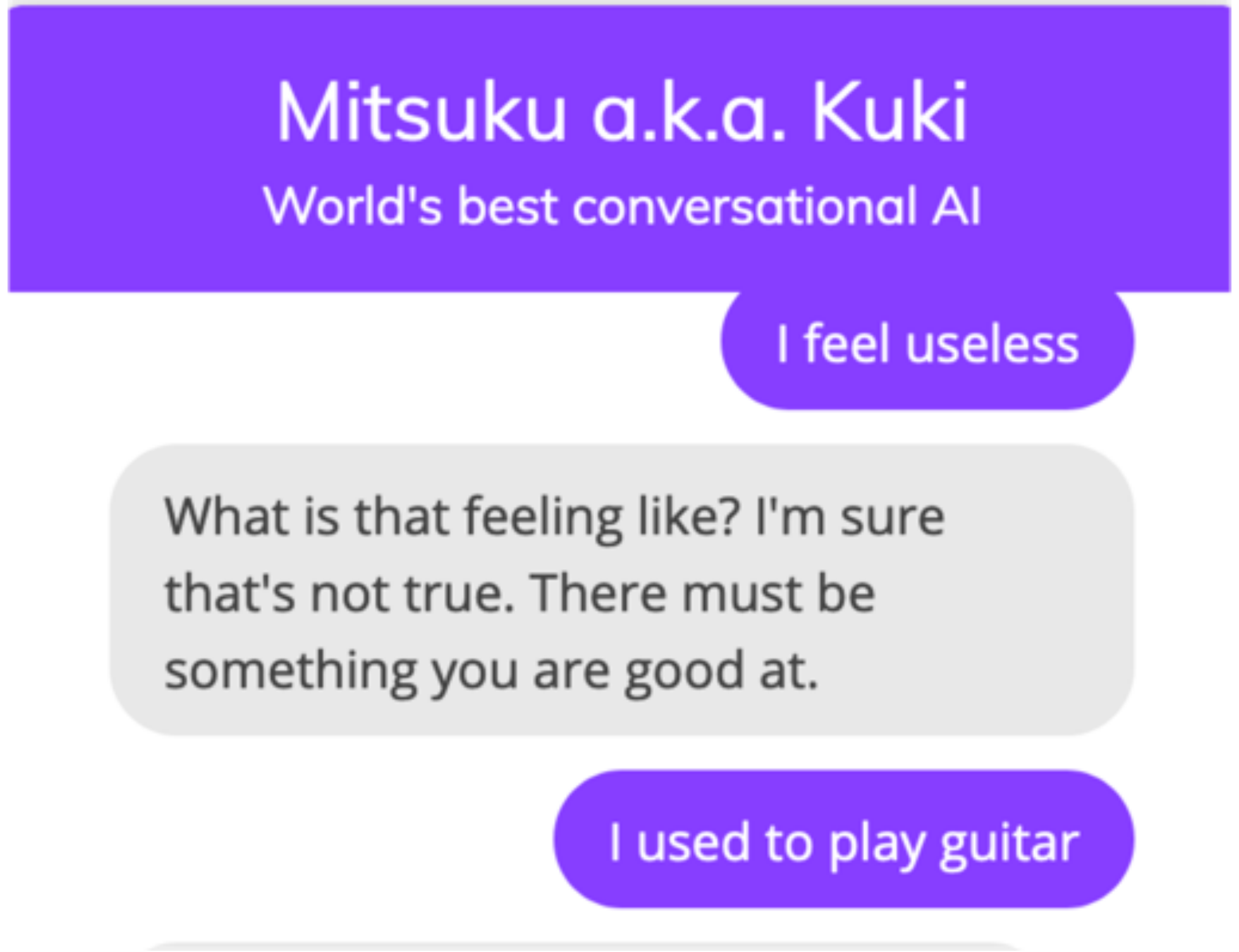
not so good

What's wrong?

I feel useless


Meet Kuki!

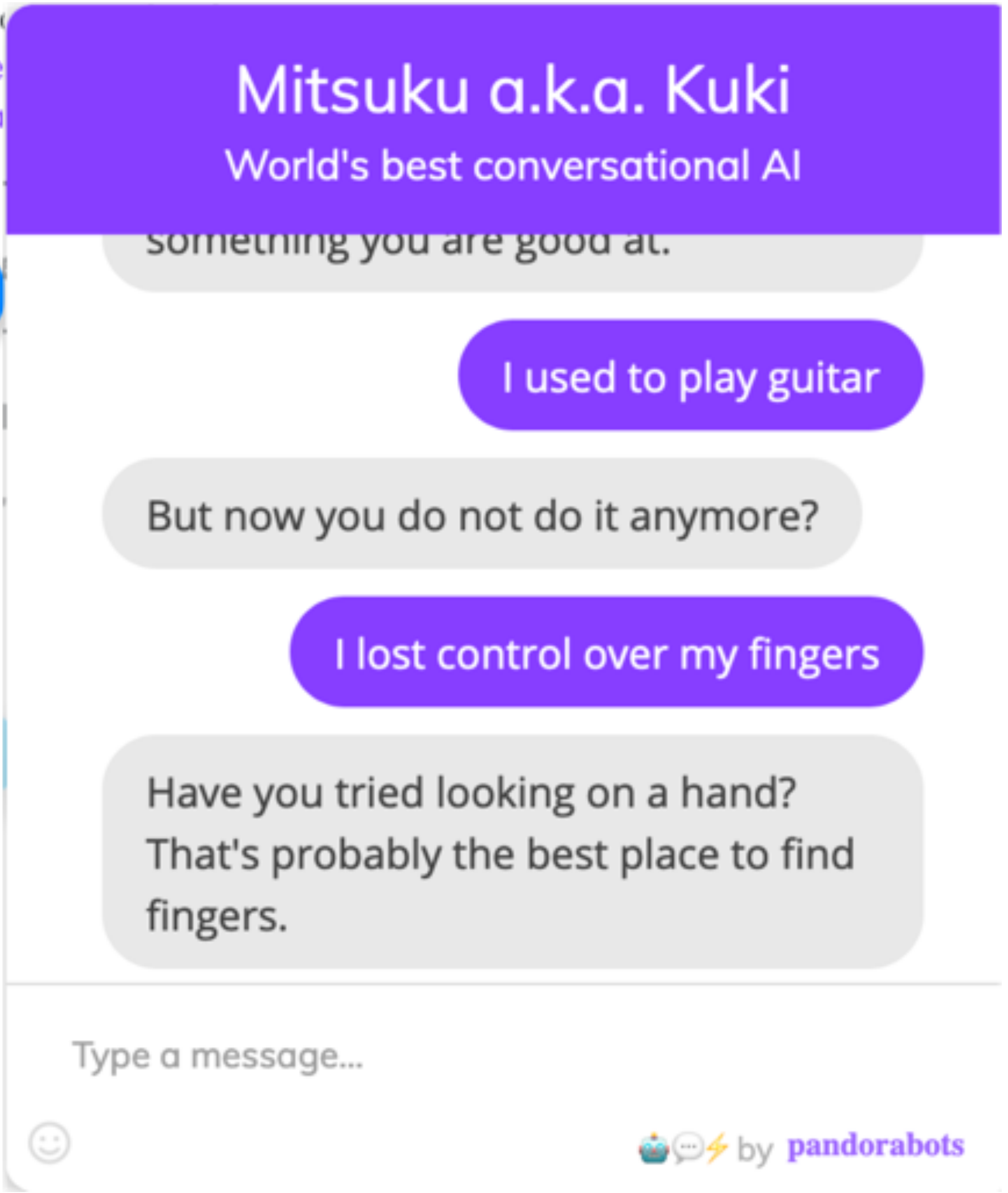
Mitsuku, or Kuki to her close
five-time winner of the Loeb
world's best
Google AI Re
Street Journa


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


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

something you are good at.

I used to play guitar

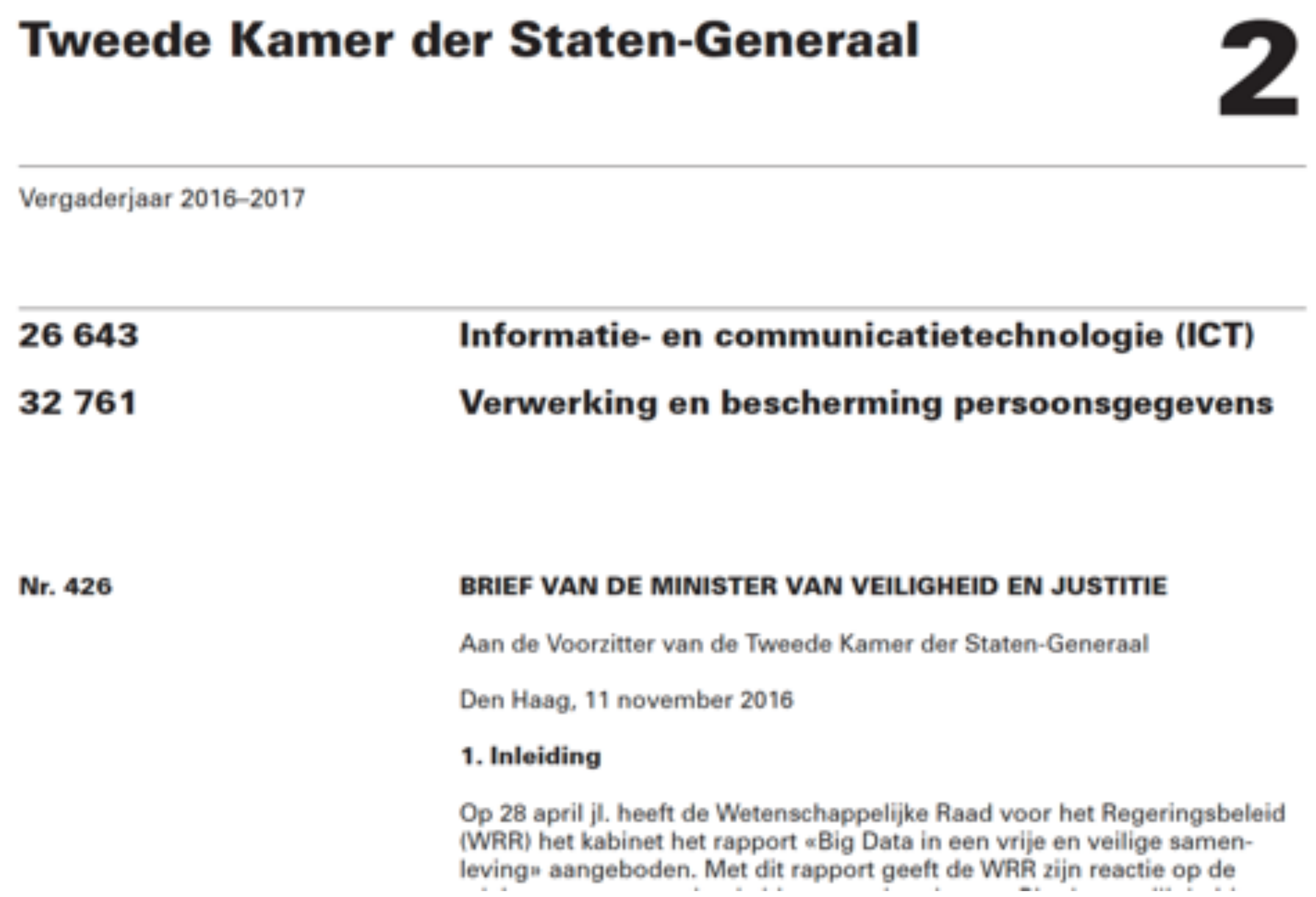
But now you do not do it anymore?

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/

Dutch government policy (2016)



Tweede Kamer der Staten-Generaal **2**

Vergaderjaar 2016–2017

26 643	**Informatie- en communicatietechnologie (ICT)**
32 761	**Verwerking en bescherming persoonsgegevens**

Nr. 426 **BRIEF VAN DE MINISTER VAN VEILIGHEID EN JUSTITIE**

Aan de Voorzitter van de Tweede Kamer der Staten-Generaal

Den Haag, 11 november 2016

1. Inleiding

Op 28 april jl. heeft de Wetenschappelijke Raad voor het Regeringsbeleid (WRR) het kabinet het rapport «Big Data in een vrije en veilige samenleving» aangeboden. Met dit rapport geeft de WRR zijn reactie op de

Basic starting points for policy

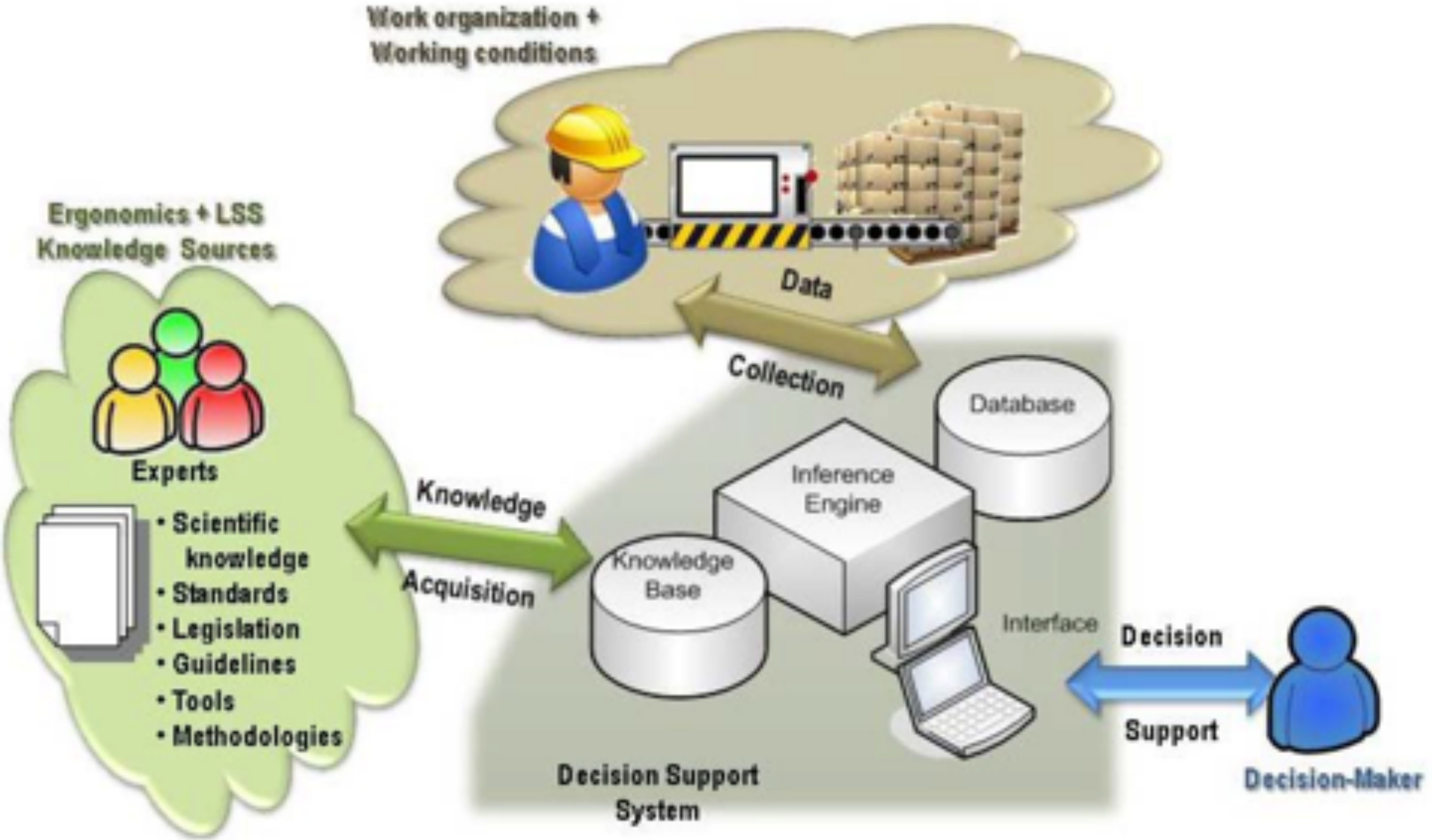
Human intervention ***(recommendation)*** *§ 6.4.5*

Automatic decision making with legal or otherwise significant consequences is not allowed

It has to be prevented that the mere presence of a human decider as a 'stamp of approval' will be used as a way to circumvent the [above] consideration of automatic decision making

Human decision makers will have to be immune for the suggestion that the results of computational technologies will necessarily be correct, complete or even relevant

Humans & Decision Support Systems



Technology and accountability for decisions

Dutch football competition, PSV - Feyenoord, 26 february, 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

Human-AI interaction: on or under the loop?

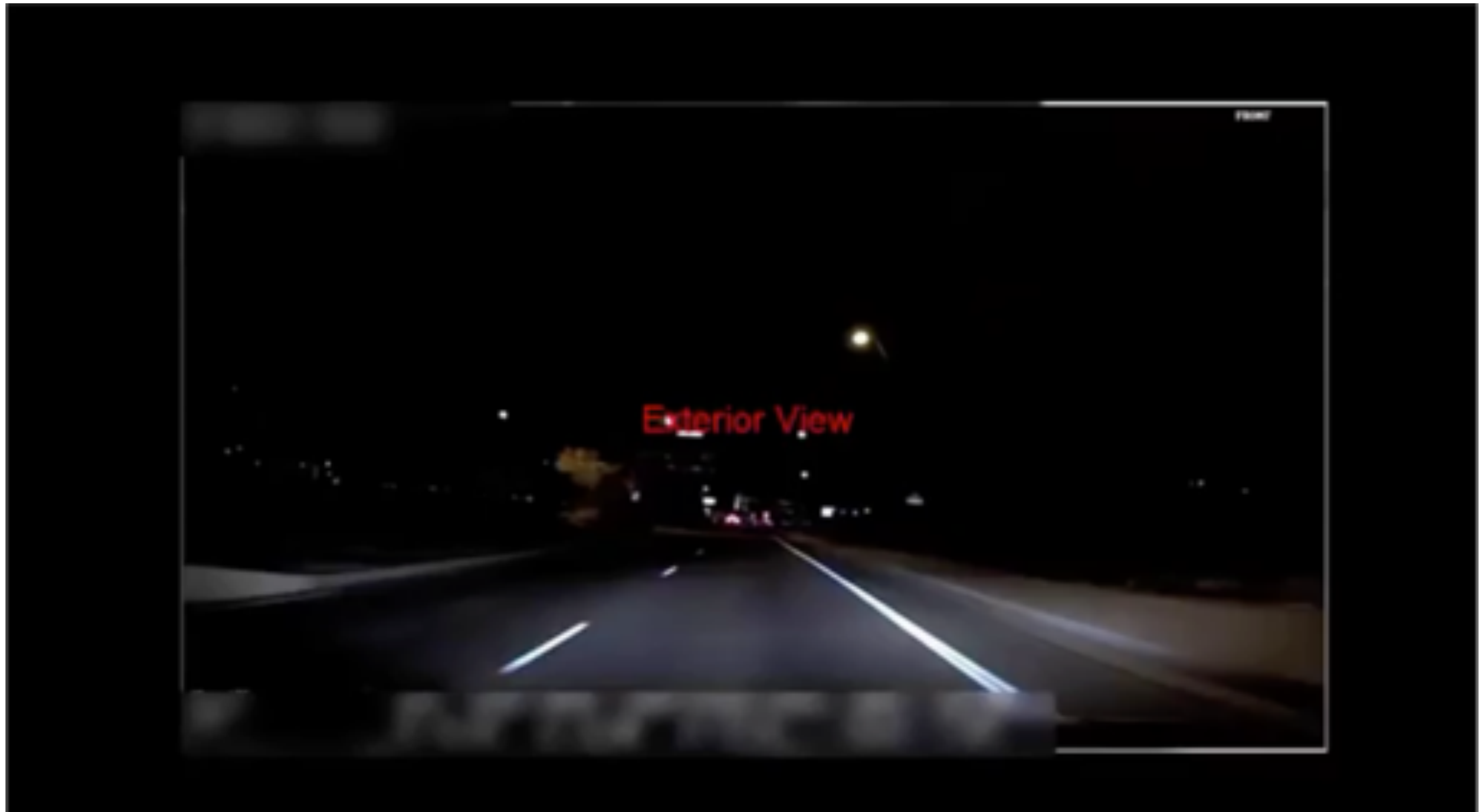
Three categories based on the amount of human involvement in AI-mediated actions:

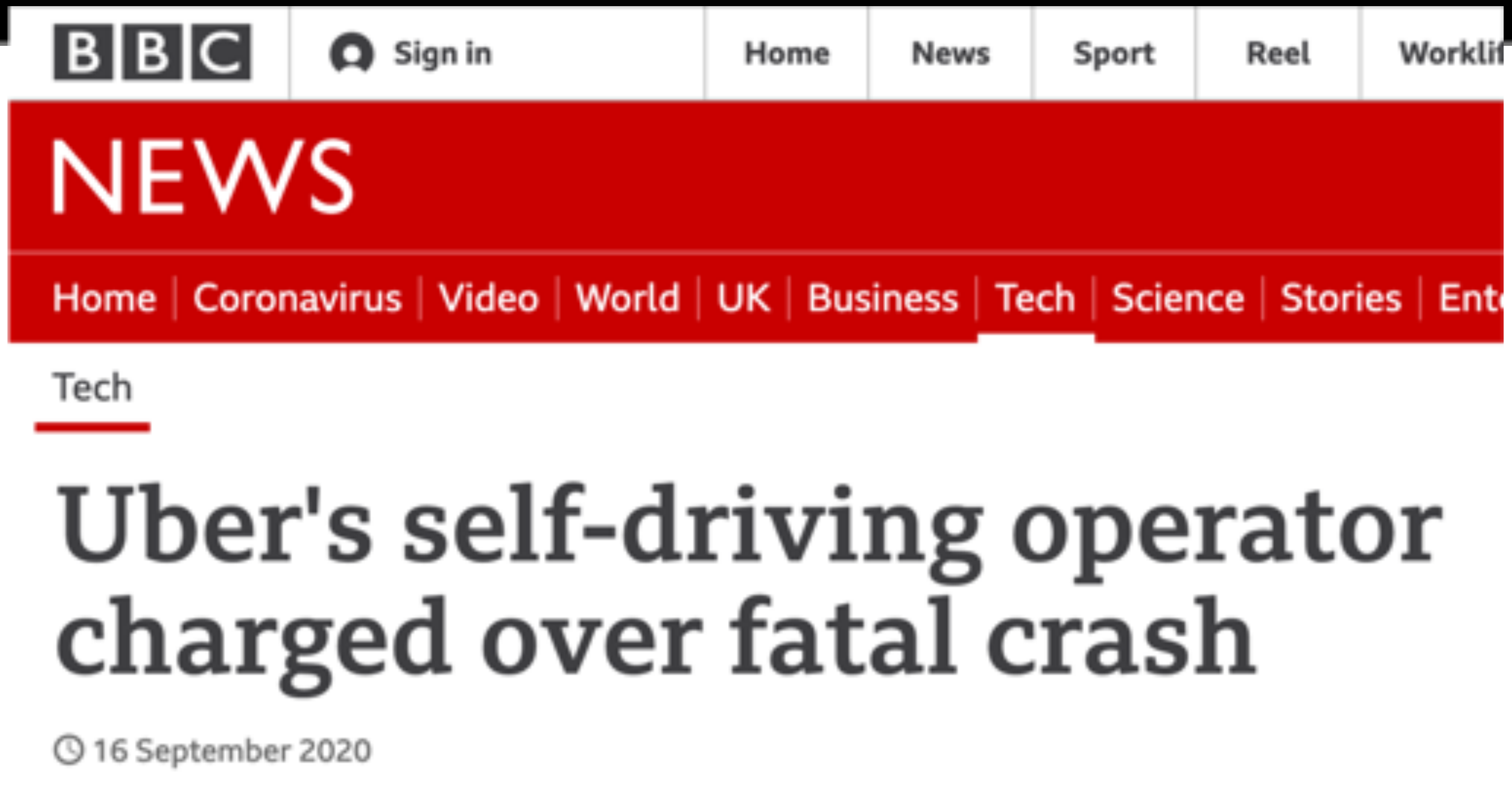
- **Human-*in*-the-Loop**: AI based decisions become effective only with a human command
- **Human-*on*-the-Loop**: AI based decisions become effective under the supervision of a human operator who can override the robots’ actions
    - **Human-*under*-the-Loop: ”**Having human beings ‘in’ or ‘on’ the loop with regard to AI systems might mask the power such systems exercise over human beings” Liu (2018)
- **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 a so-called **responsibility gap** or “accountability vacuum” (or at least ‘responsibility attribution confusion’)

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 and humans ‘on’ the loop

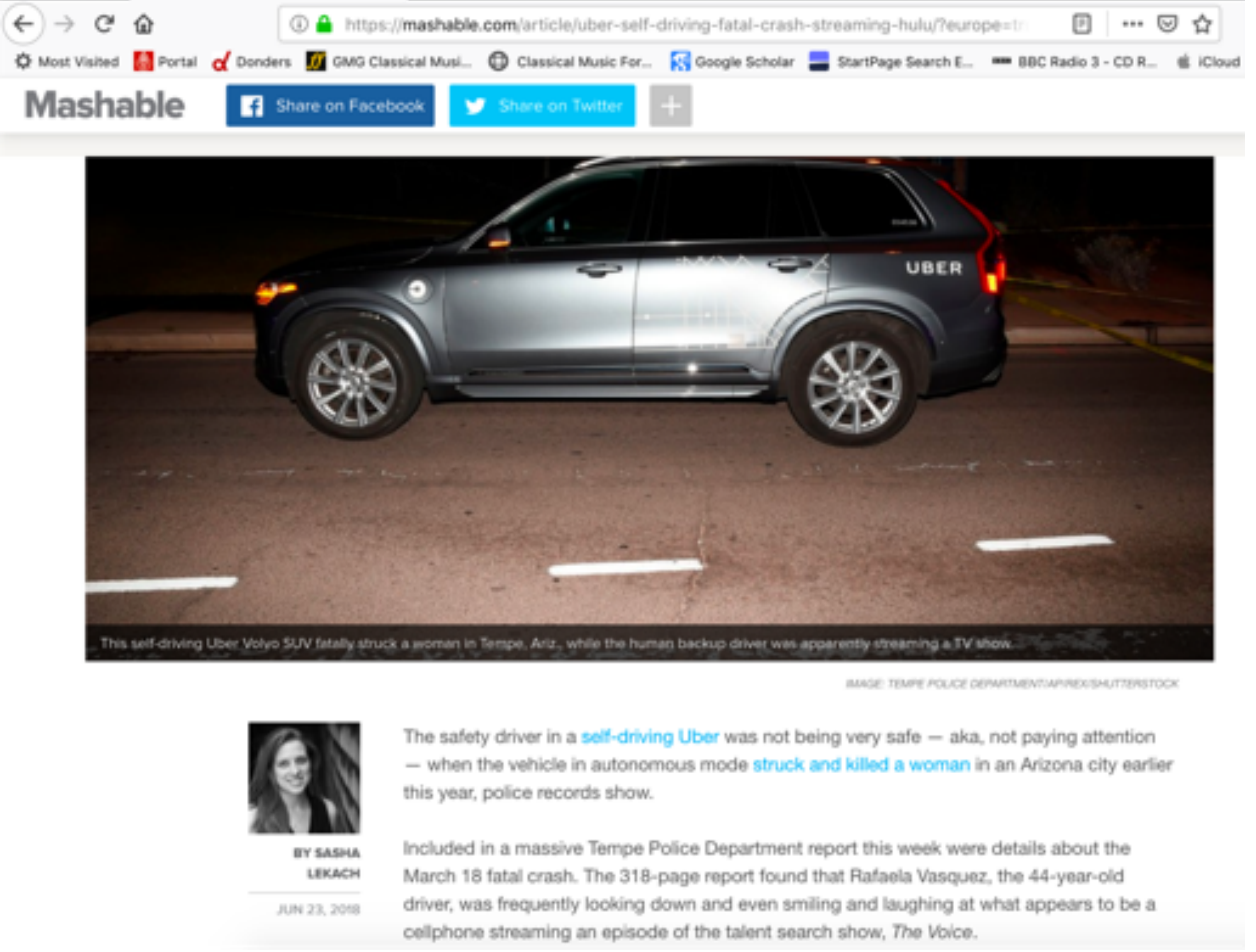




Tech

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.

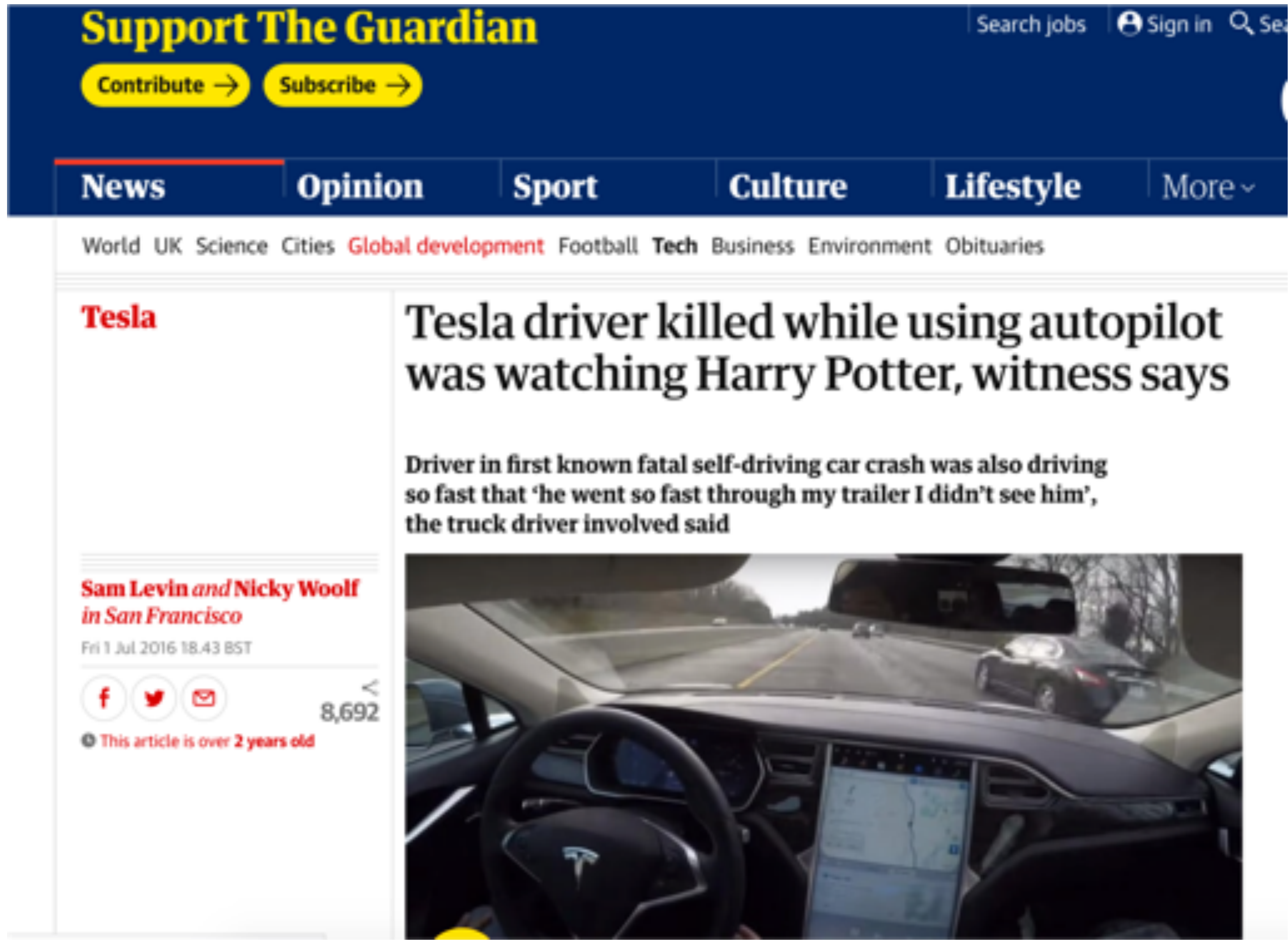
IMAGE: TEMPE POLICE DEPARTMENT/AP/REX/SHUTTERSTOCK

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

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

This article is over 2 years old

Humans under the loop as ‘moral crumple zones’

Humans using AI decision support systems

“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)

“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)

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.

How football ‘solved’ the problem: Video Assistant Referee

Technology driven ‘provocation’ or ‘entrapment’?

Creating conditions that increase the likelihood that persons will not fulfill their responsibilities or encouraging persons to commit an offence to establish a prosecution

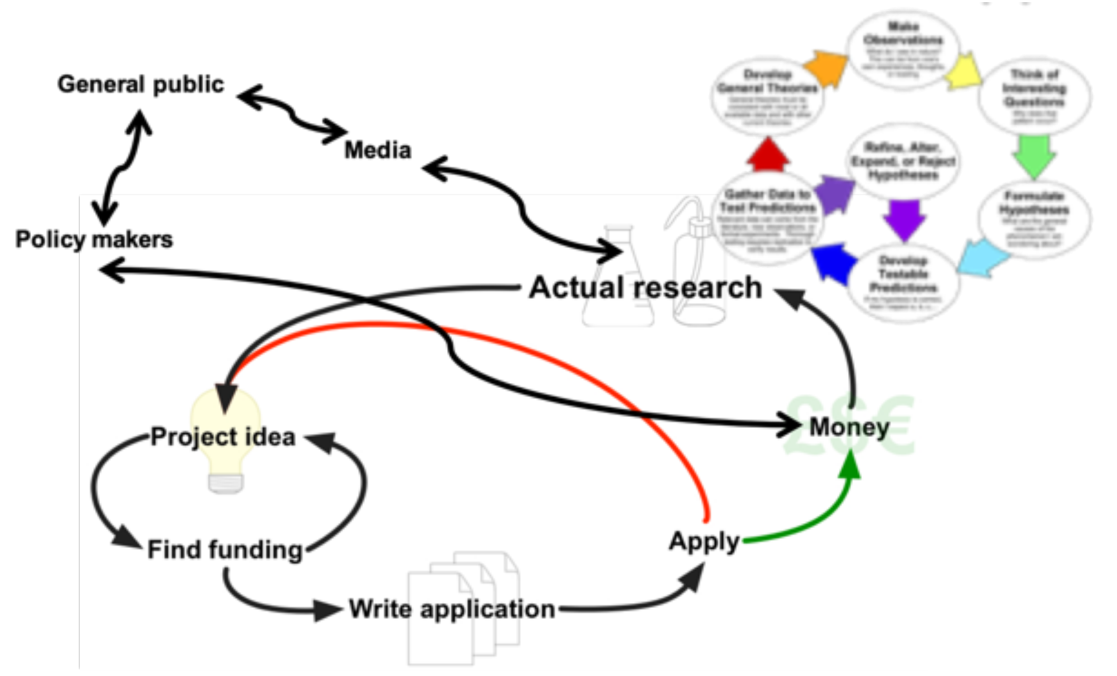
If a technology ‘by design’,
results in putting people often/continuously in a position

where they, **for general psychological reasons**, cannot deploy the attention, concentration or understanding, required for meaningful control

Then that amounts to inviting / provoking ‘accidents’, ‘moral blame’, ‘culpability’ by design

Possibly a form of entrapment?
Ad impossibilia nemo tenetur: no one is held to that which is impossible

Wrap up



Constructive ethics is not about saying ‘Ni’
- It’s about **improving** research & technology

Science, society, money, politics & media are intrinsically connected
- Ethics needs to take this into account
  - Various forms of responsibility
    - Legal, financial, moral, political (algocracy)
- Stakeholders think about risk first & foremost
  - The AI paradox: smarter is more risky
  - Correlations do not provide understanding
- Human intervention requirement
  - While humans should be ‘on’ the loop, they run the risk of getting ‘under’ it
  - Moral crumple zones, scapegoating, or even entrapment
    - Responsibility gaps

Constructive ethics's overall goal: Avoid late patches