Social and environmental value assessment of AI/ML technologies

Brigita Jurisic

International Iberian Nanotechnology Laboratory

Braga, Portugal

What will we cover?

1. INL & the FORGING project
2. Responsible Research and Innovation
3. Ethics, morals, laws and values
4. Ethics of AI & ML technologies

4.1 Privacy and data protection

4.2 Fairness and bias

4.3 Role of human judgement

5. Guidelines, policy and legal frameworks


INTERNATIONAL IBERIAN
NANOTECHNOLOGY
LABORATORY

Responsible innovation for digital and green transition

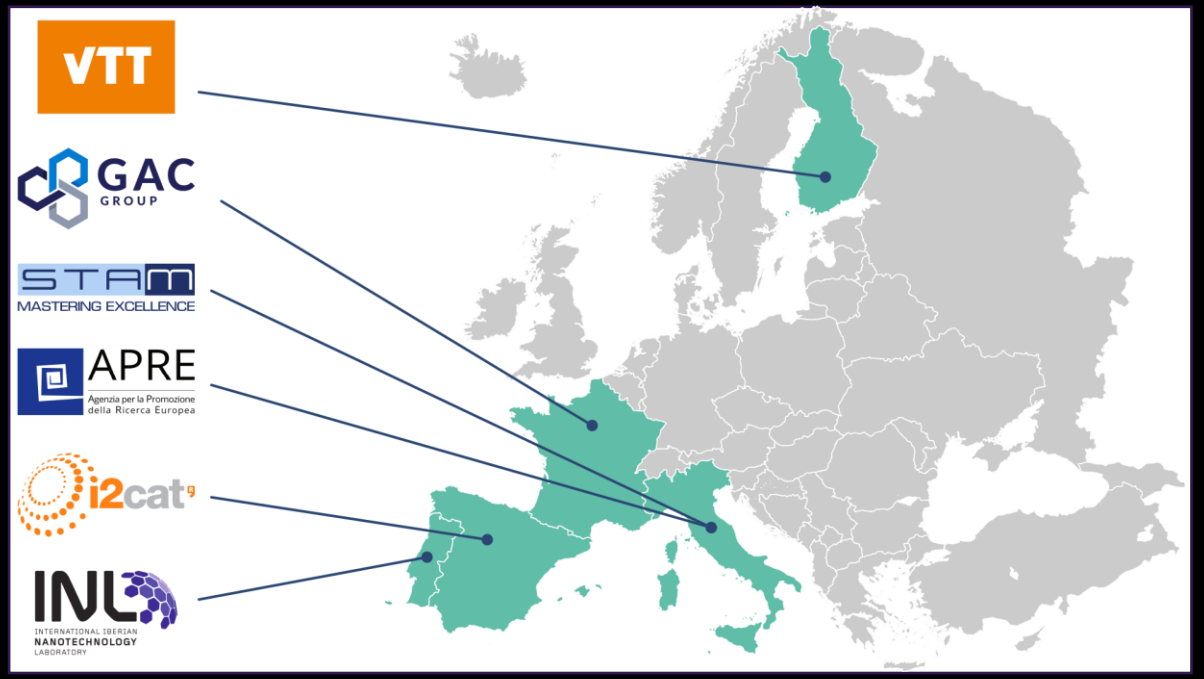
Build the bridge between research and industry

Enable cross-sectorial thinking

Use foresight as a creativity enabling tool

FORGING

INDUSTRY-ACADEMIA FORUM TO UNCOVER THE POTENTIAL OF EMERGING ENABLING TECHNOLOGIES



Society is facing many challenges today...

Health, demographic change, and wellbeing

Food, agriculture and forestry, and water

Secure, clean and efficient energy

Smart, green and integrated transport

Climate action, environment, and resources

Europe in a changing world: inclusive, innovative and reflective societies

Secure societies: freedom and security of Europe and its citizens

Responsible Research and Innovation tackles these challenges by aligning values, needs and expectations of all actors involved in Research and Innovation

And there is also a large consensus that changes are needed throughout the R&I system

Certain key issues (or policy agendas) need to be taken into account:

ETHICS
Research integrity and ethical acceptability of the R&I outcomes

GENDER EQUALITY
Human resources, decision bodies and research dimension

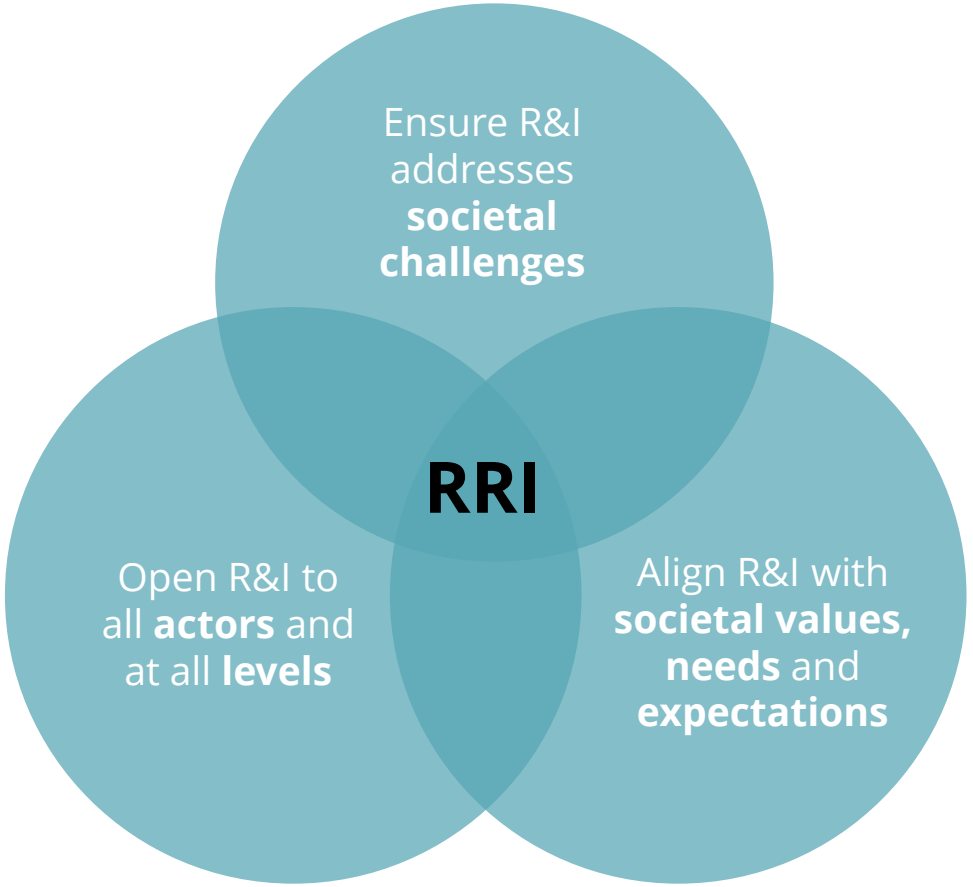
GOVERNANCE
Structural changes to include all these issues in the R&I system

OPEN ACCESS
To results from publicly funded research, privacy issues and even more: open science

PUBLIC ENGAGEMENT
Towards a more open and inclusive R&I

SCIENCE EDUCATION
Provide competences for the responsible citizens society needs

Therefore more complex and multifaceted societies demand more democracy in science and more science in democracy



Responsible Research & Innovation is a **new governance and values framework** to build a new path where these requests can blossom

The RRI Toolkit: A wealth of resources to help you implement RRI

TOOLS
Use manuals, guidelines, and 'how tos' to implement RRI.

INSPIRING PRACTICES
Find inspiration in RRI success stories across Europe.

PROJECTS
Get to know other projects on RRI and find potential partners.

LIBRARY
Learn of RRI from articles, reports, cross-analyses, and more.

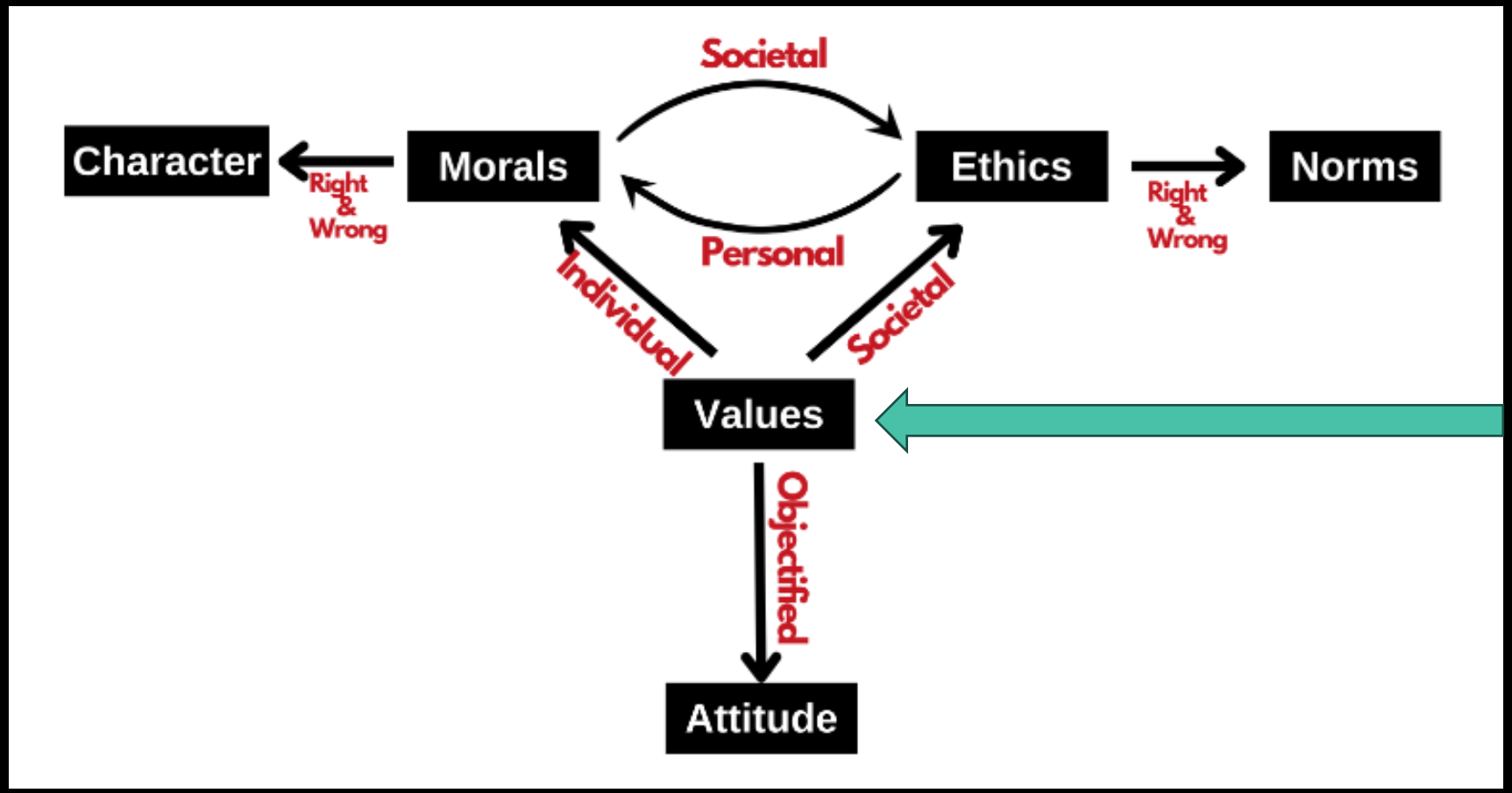
HOW TOs
Get concrete examples on how to put RRI into practice in different contexts.

SELF-REFLECTION TOOL
Reflect on how RRI your own professional practice is.

TRAINING MATERIALS
Organise trainings in RRI using showcases and presentations.

COMMUNICATION MATERIALS
Spread the word on RRI with videos and presentations.

Ethics, Morals, Laws and Values



Reason
Dignity
Responsibility
Freedom
Action
Civilization
Continuous development
Democracy
Human-technology co-evolution
Transformation
Participation
Protection of future generations

Values, Ethics, Morals and Attitude - civilspedia.com

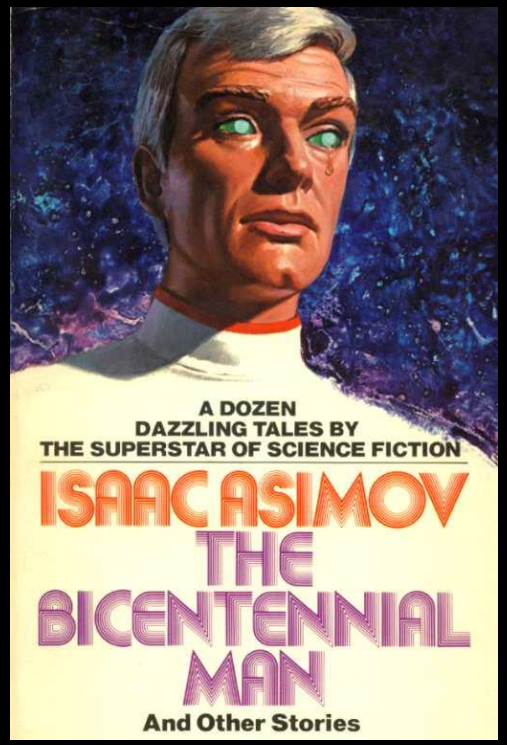
Ethics and AI/ML technologies

Three Laws of Robotics

1. May not injure a human being, or, through inaction, allow a human being to come to harm.

2. Must obey the orders given by human beings, except where such orders would conflict with the First Law.

3. Must protect their own existence, as long as such protection does not conflict with the First or Second Law.



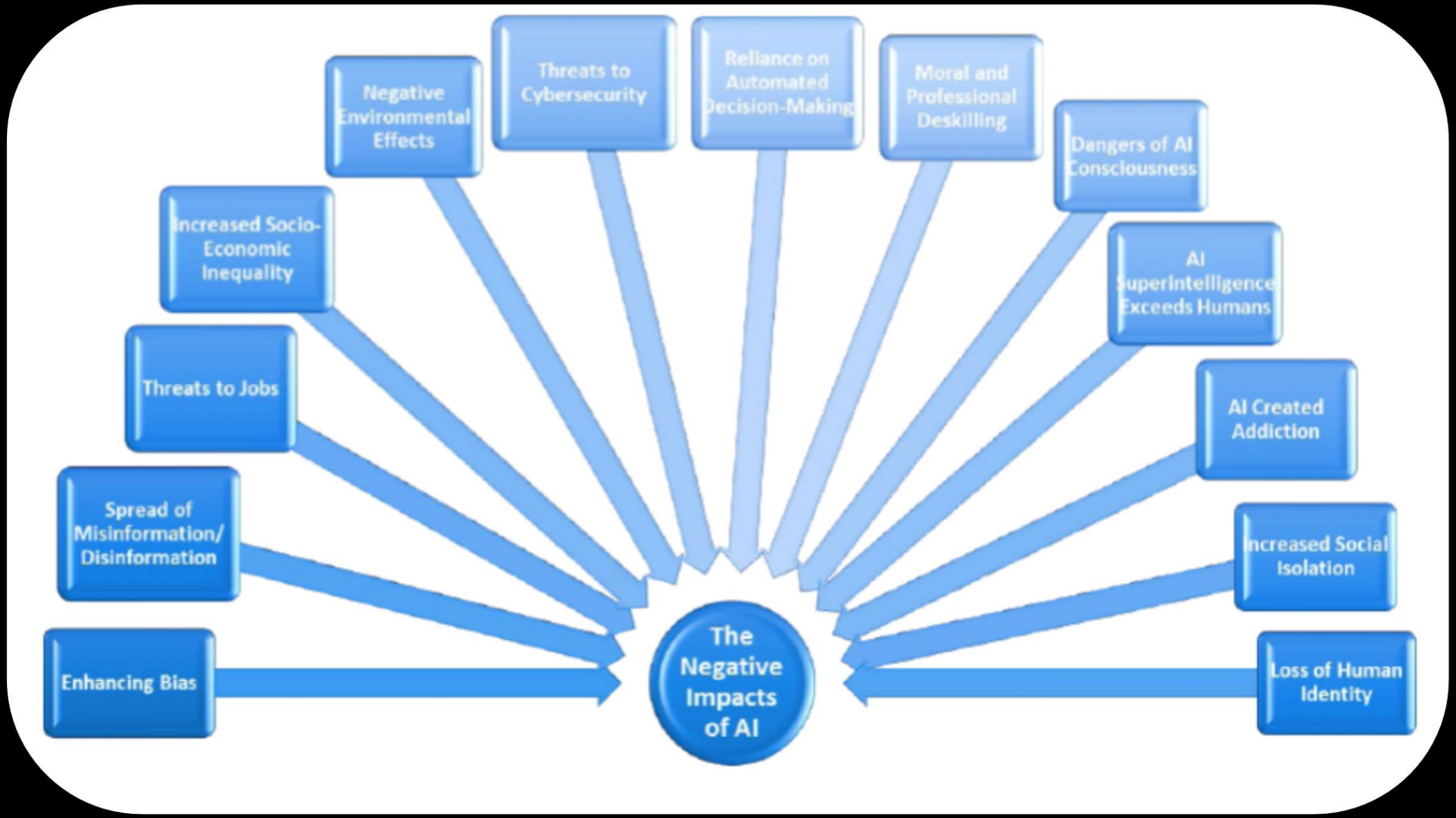
©Bronasbooks.com

Major ethical concerns of AI/ML technologies

Privacy and surveillance

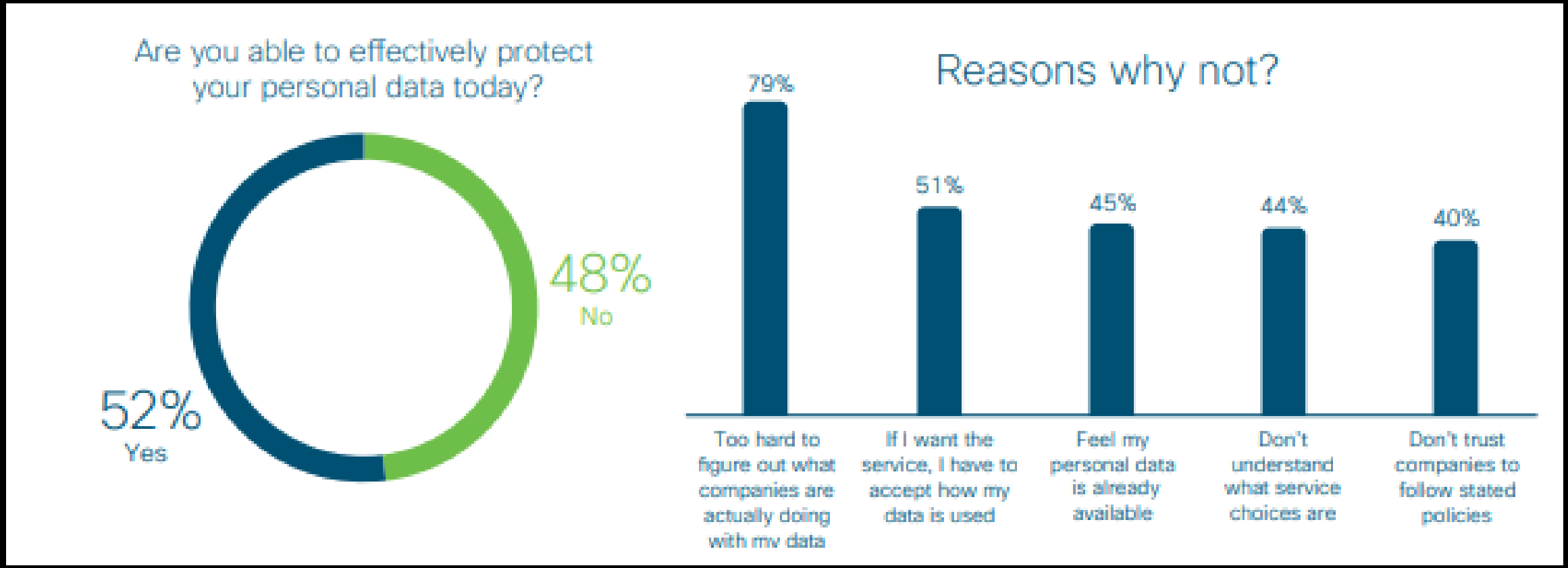
Fairness and bias

Role of human judgement



Mehan, Julie (2024). Artificial Intelligence - Ethical, social, and security impacts for the present and the future, Second edition

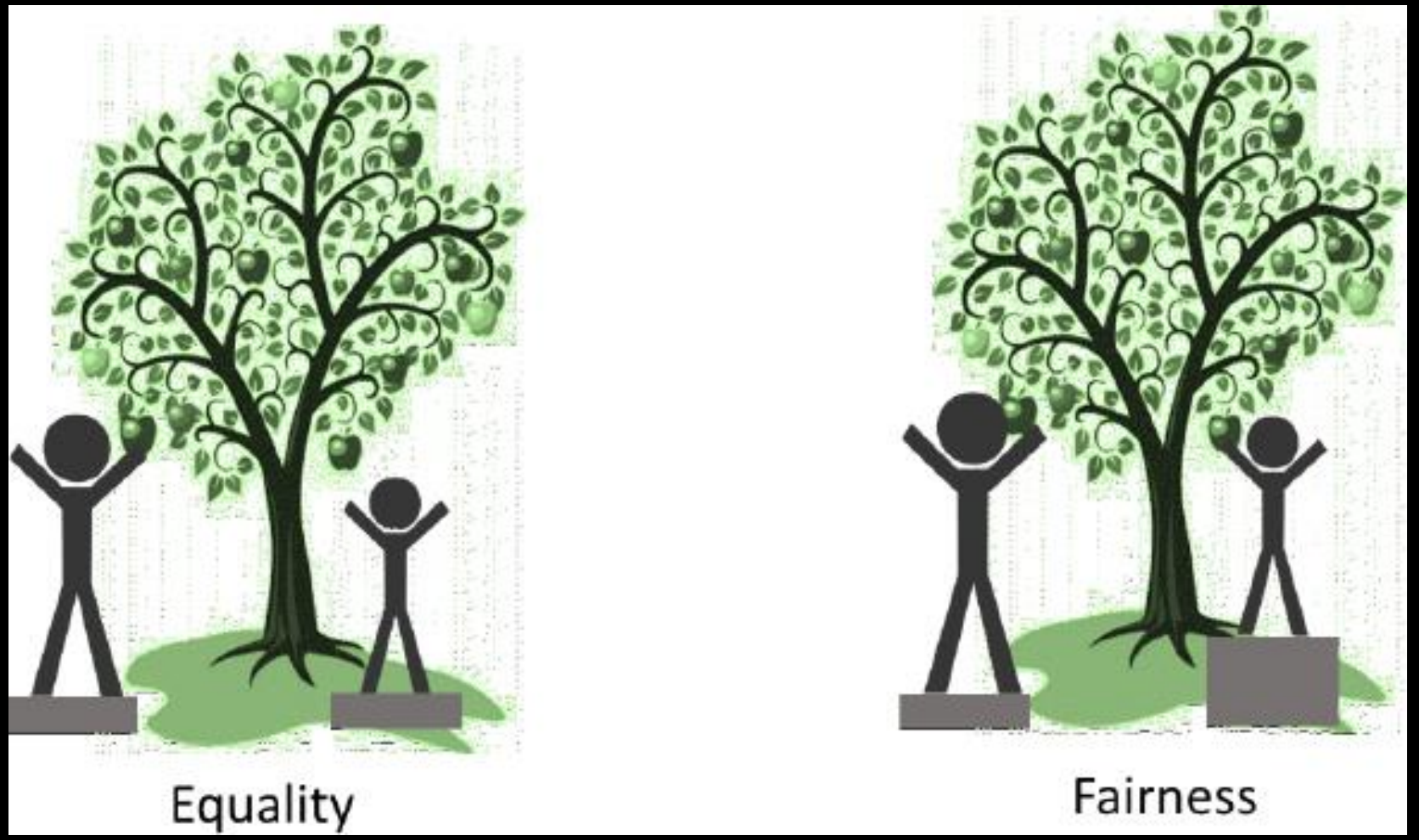
Privacy and Data Protection



Cisco, 2020. Consumer Privacy Survey: Protecting Data Privacy to Maintain Digital Trust

Data privacy - an individual's right of self-determination regarding when, how, and to what extent personal information about them is collected, shared with, or communicated to others.

Fairness and bias



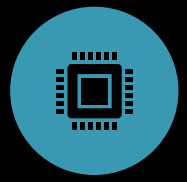
Mehan, Julie (2024). Artificial Intelligence - Ethical, social, and security impacts for the present and the future, Second edition

Fairness and bias tools

IBM’s AI Fairness 360 Toolkit: a Python toolkit focusing on technical solutions through fairness metrics and algorithms to help users examine, report, and mitigate discrimination and bias in ML models.

Facebook´s “Fairness Flow” internal tool to identify bias in ML models.

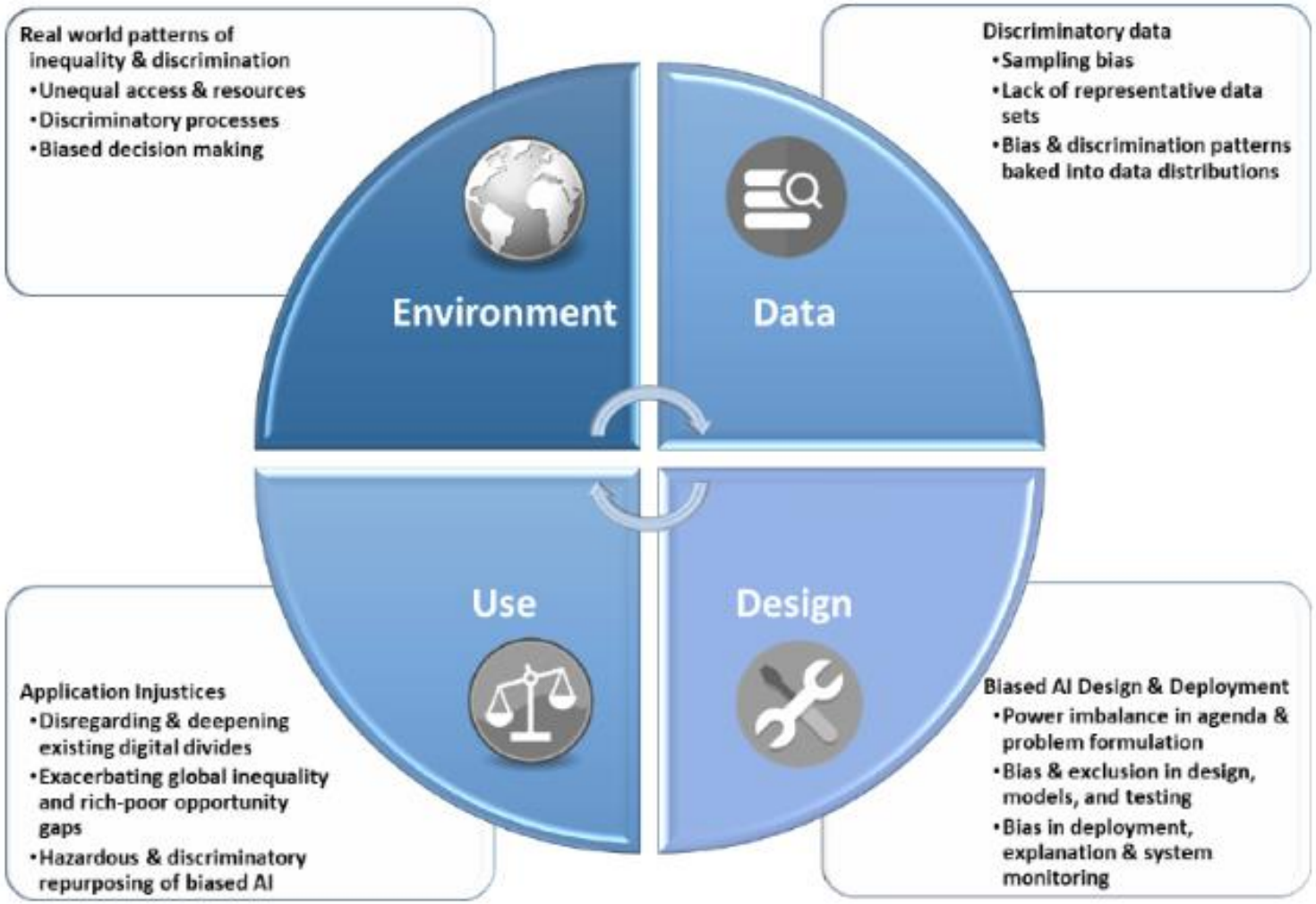
Google’s What-If Tool: a tool to explore a models’ performance on a dataset, including examining several preset definitions of fairness constraints (e.g., equality of opportunity).



Microsoft’s fairlean.py: a Python package that implements a variety of algorithms that seek to mitigate “unfairness” in supervised machine learning.

Co-designed AI checklist listing what needs to be considered at different stages of an AI system’s development and deployment life cycle (i.e. envision, define, prototype, build, launch, and evolve)

Where does bias come from?



Mehan, Julie (2024). Artificial Intelligence - Ethical, social, and security impacts for the present and

Role of human judgement

- Reliance on automated decision making and AI dependency
- Job threats and increase in socio-economic inequality
- Moral and professional deskilling
- Evolution of self-aware AI
- Accountability and responsibility
- AI as moral agent – artificial moral agents

©century.edu

Member-only story

Who Killed Elaine Herzberg?

Who is ultimately to blame in the first self-driving car fatality — the technology, the victim, the safety driver, Uber, or the American city itself?

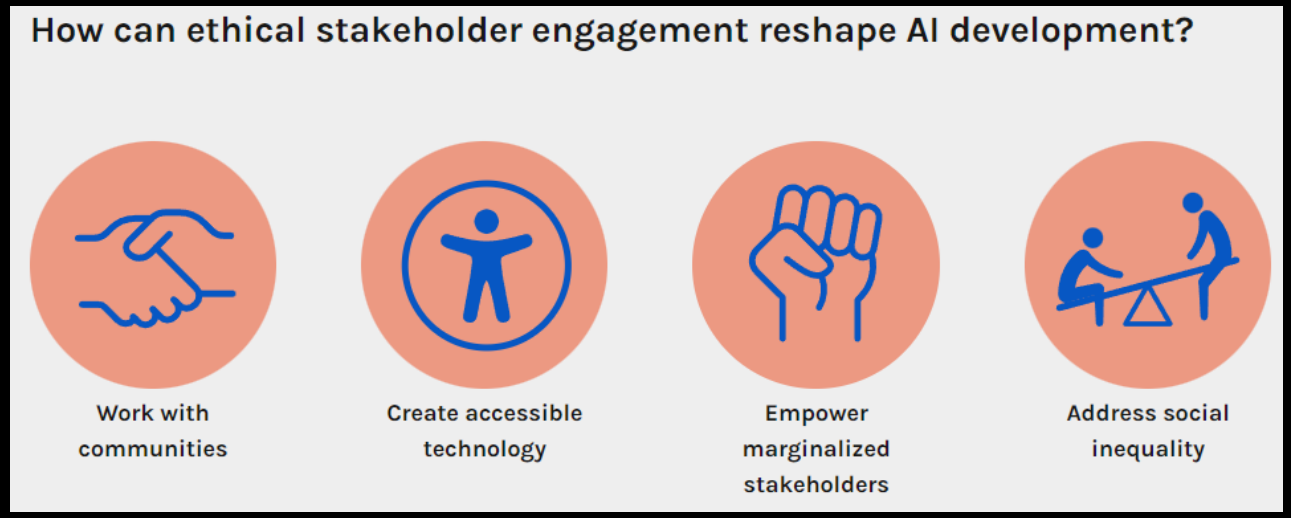
Jack Stilgoe · Follow
Published in OneZero · 8 min read · Dec 12, 2019

Ethical approaches

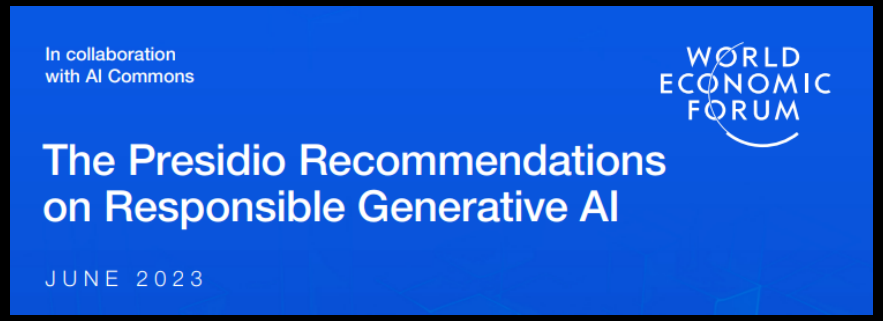
- **Utilitarian ethics** is about maximizing overall happiness, while minimizing overall suffering.
- **Kantian ethics** is about adopting a set of basic principles (“maxims”) fit to serve as universal laws, in accordance with which all are treated as ends-in-themselves and never as mere means.
- **Virtue ethics** is about cultivating and then fully realizing a set of basic virtues and excellences.
- **Confucian ethics** is similar to virtue ethics, and also places emphasis on the creation and maintenance of social harmony.
- **Ubuntu ethics,** which we also mentioned above, is about relating to each other in communal ways that allow us to fully realize our humanity.

Nyholm, Sven (2023). This is Technology Ethics: An Introduction, John Wiley & Sons

Guidelines, policy and legal frameworks



Inclusive Research & Design - Partnership on AI



Design of transparent and inclusive AI systems - AI Governance Alliance (weforum.org)

A European strategy for data | Shaping Europe's digital future (europa.eu) (2020)

European Declaration on Digital Rights and Principles | Shaping Europe's digital future (europa.eu) (2022)

The EU's Digital Services Act (europa.eu) (2022)

Data Act enters into force: what it means for you - European Commission (europa.eu) (2023)

AI Act | Shaping Europe's digital future (europa.eu) (2024)

People at the centre

Digital technologies should **protect people's rights, support democracy, and ensure that all digital players act responsibly and safely**. The EU promotes these values around the world.

Solidarity and inclusion

Technology should **unite, not divide, people**. Everyone should have access to the internet, to digital skills, to digital public services, and to fair working conditions.

Freedom of choice

People should benefit from a **fair online environment, be safe from illegal and harmful content, and be empowered** when they interact with new and evolving technologies like artificial intelligence.

Participation

Citizens should be able to **engage in the democratic process** at all levels and have **control over their own data.**

Safety and security

The digital environment should be **safe and secure**. All users, from childhood to old age, should be empowered and protected.

Sustainability

Digital devices should support **sustainability and the green transition**. People need to know about the environmental impact and energy consumption of their devices.

European Digital Rights and Principles

AI Pact

- Adopting AI governance strategy
- High-risk AI systems 'mapping
- Promoting AI literacy

OECD Declaration on a Trusted, Sustainable and Inclusive Digital Future

The Vision for the OECD for the Next Decade, which directs the OECD to support countries in harnessing the potential of digitalisation for economic growth and social inclusion to support open societies in the digital and data driven age and to advance responses to the challenges of digitalisation, including:

- **guarding against threats to democracy**,
- **digital security** and
- **digital privacy** and
- **combatting disinformation online**, as well as to seek initiatives that
- **enhance and promote data free flow with trust**.

Ethics washing

“The Trustworthy AI story is a marketing narrative invented by industry, a bedtime story for tomorrow's customers. The underlying guiding idea of a “trustworthy AI” is, first and foremost, conceptual nonsense. Machines are not trustworthy; only humans can be trustworthy (or untrustworthy).” Metzinger, 2019

What are the problems AI/ML technologies should solve?

For the curious

- Werthner, Prem, Lee & Ghezzi (2022). Perspectives on digital humanism Perspectives on Digital Humanism | SpringerLink
- Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *Zenodo*. *https://doi.org/10.5281/zenodo.3240529*.
- WEF global risks report (2023). Available at WEF_Global_Risks_Report_2023.pdf (weforum.org)
- Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (December 11, 2019). Algorithmic Decision-Making and the Control Problem. *SpringerLink*. Available at *https://link.springer.com/article/10.1007/s11023-019-09513-7*
- Smith, G. (2020). What does "fairness" mean for machine learning systems? *Center for Equity, Gender & Leadership (EGAL) at Berkeley Haas*. Available at *https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf*
- Burt, A. (December 13, 2019). The AI Transparency Paradox. *Harvard Business Review*. Available at *https://hbr.org/2019/12/the-ai-transparency-paradox*
- Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (December 11, 2019). Algorithmic Decision-Making and the Control Problem. *SpringerLink*. Available at *https://link.springer.com/article/10.1007/s11023-019-09513-7*
- NSCAI (March 1, 2021). Final Report. *National Security Commission on Artificial Intelligence.* Available at *www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf*
- Moor, J. (1985) *What is Computer Ethics?* Available at *https://web.cs.ucdavis.edu/~rogaway/classes/188/spring06/papers/moor.html*
- Nyholm, Sven (2023). This is Technology Ethics: An Introduction, John Wiley & Sons
- Anderson, M. and Anderson, S.L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28 (4): 15–26
- OECD Declaration on a Trusted, Sustainable and Inclusive Digital Future (2022). Available at Documents (oecd-events.org)
- Decision - 2022/2481 - EN - EUR-Lex (europa.eu) - Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030
- Europe's digital decade: 2030 targets | European Commission (europa.eu)
- European Commission: Directorate-General for Communications Networks, Content and Technology, Study to support the monitoring of the Declaration on Digital Rights and Principles – Final report, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2759/875696
- EU's high-level expert group's "Ethics Guidelines for Trustworthy AI" can be downloaded here. https://ec.europa.eu/futurium/en/ai-alliance-consultation. 1.html.

Awful AI

Awful AI is a curated list to track *current* scary usages of AI - hoping to raise awareness to its misuses in society

Artificial intelligence in its current state is unfair, easily susceptible to attacks and notoriously difficult to control. Often, AI systems and predictions amplify existing systematic biases even when the data is balanced. Nevertheless, more and more concerning uses of AI technology are appearing in the wild. This list aims to track *all of them*. We hope that *Awful AI* can be a platform to spur discussion for the development of possible preventive technology (to fight back!).

You can cite the list and raise more awareness through Zenodo.

DOI 10.5281/zenodo.5855972

Table of Contents