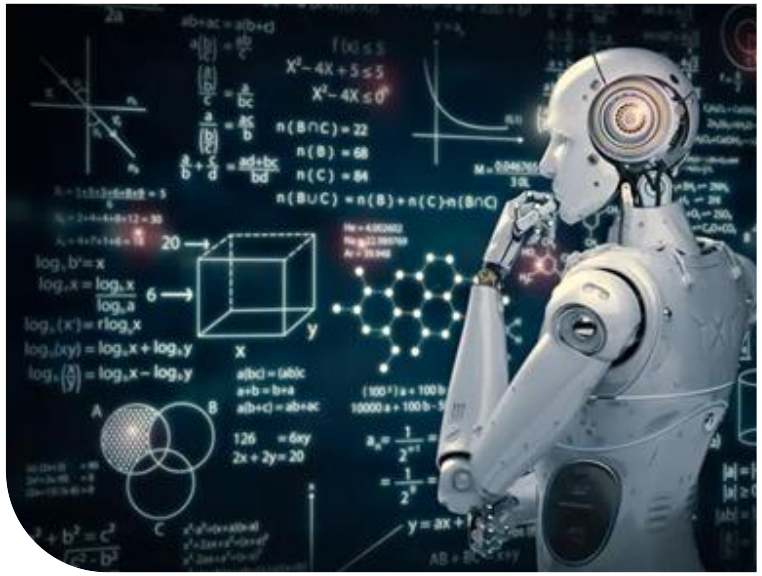
UNIVERSITY OF PASSAU

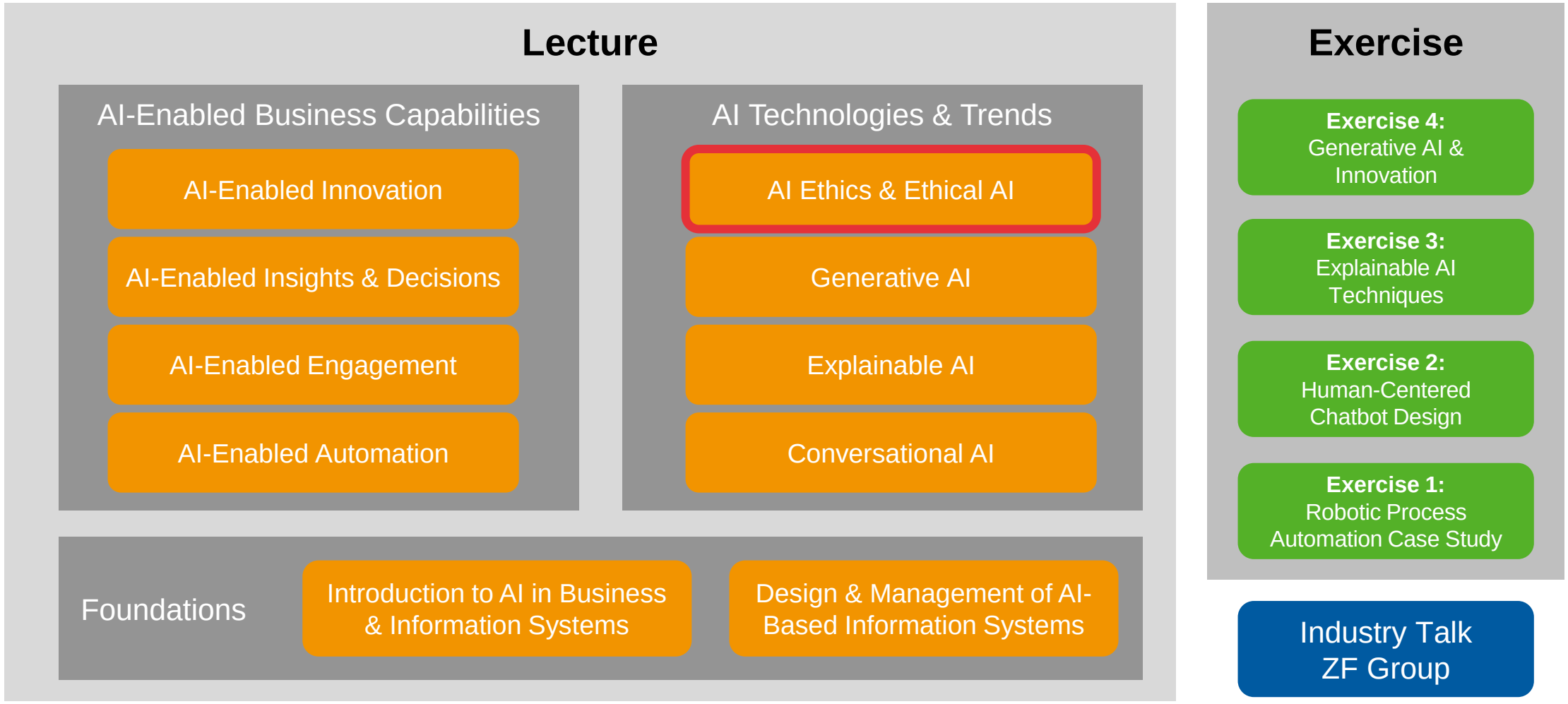
AI-Based Business Information Systems

AI Ethics & Ethical AI



Prof. Dr. Ulrich Gnewuch

Course Organization



RECAP FROM LAST LECTURE:

- What are some of the challenges and risks of generative AI?
- How does generative AI differ from traditional AI?
- Which simple techniques can you use to obtain better LLM responses?

Learning Goals

- Identify common ethical concerns with AI
- Explain the legislative and principled approach to AI ethics
  - Describe the basic idea of the EU AI act
  - Name and describe key principles for ethical AI
- Discuss the benefits and challenges of the two approaches to AI ethics

Ethical Concerns with AI – Examples

AI-Enabled Automation

AI 'apocalypse' could take away almost 8m jobs in UK, says report

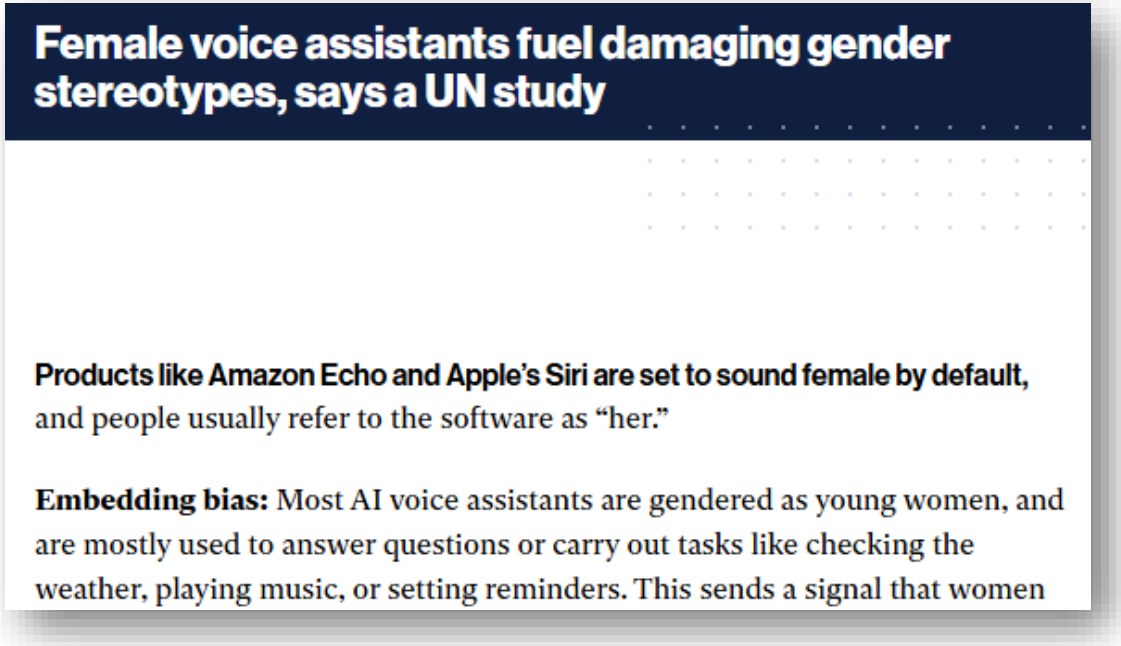
Almost 8 million UK jobs could be lost to artificial intelligence in a "jobs apocalypse", according to a report warning that women, younger workers and those on lower wages are at most risk from automation.

The Institute for Public Policy Research (IPPR) said that entry level, part-time and administrative jobs were most exposed to being replaced by AI under a "worst-case scenario" for the rollout of new technologies in the next three to five years.

The thinktank warned that the UK was facing a "sliding doors" moment as growing numbers of companies adopt generative AI technologies - which can read and create text, data and software code - to automate everyday workplace tasks.

https://www.theguardian.com/technology/2024/mar/27/ai-apocalypse-could-take-away-almost-8m-jobs-in-uk-says-report

AI-Enabled Engagement



Female voice assistants fuel damaging gender stereotypes, says a UN study

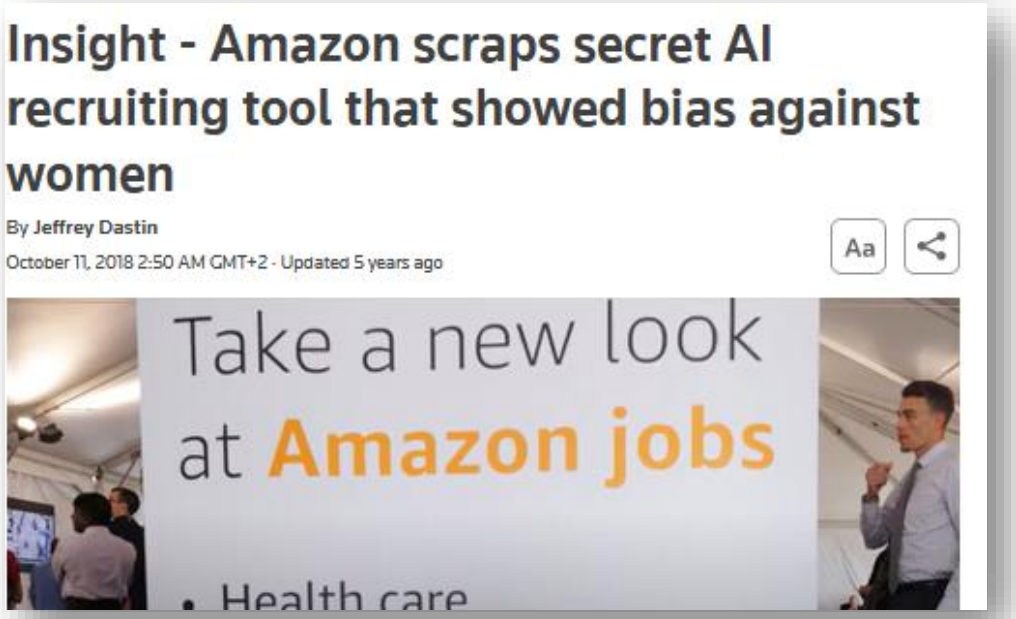
Products like Amazon Echo and Apple's Siri are set to sound female by default, and people usually refer to the software as "her."

Embedding bias: Most AI voice assistants are gendered as young women, and are mostly used to answer questions or carry out tasks like checking the weather, playing music, or setting reminders. This sends a signal that women

https://www.technologyreview.com/2019/05/22/65758/female-voice-assistants-fuel-damaging-gender-stereotypes-says-un-study

Ethical Concerns with AI – Examples

AI-Enabled Insights & Decisions



Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 11, 2018 2:50 AM GMT+2 - Updated 5 years ago

Aa

Take a new look at Amazon jobs

• Health care

https://www.reuters.com/article/idUSKCN1MK0AG/

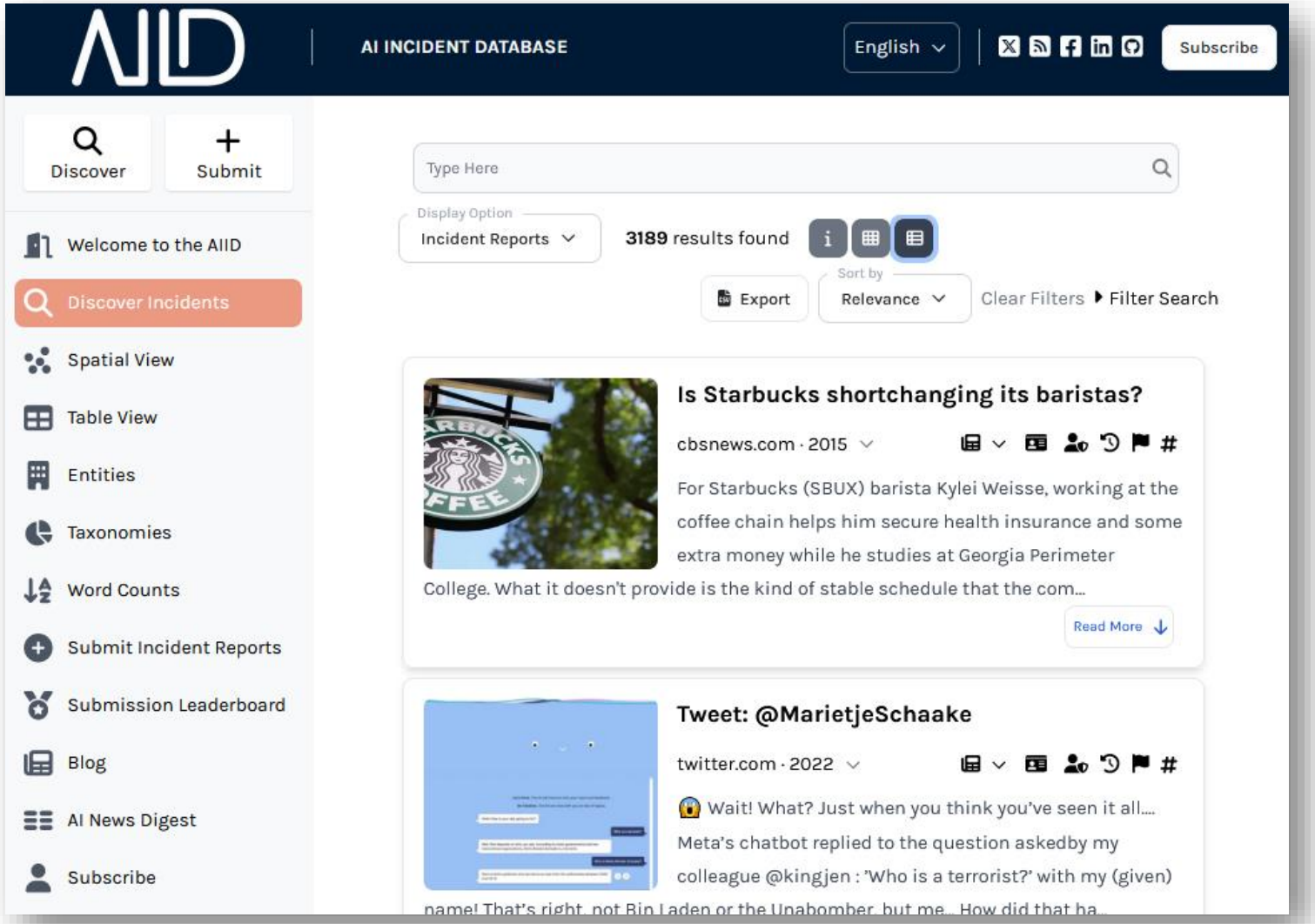
AI-Enabled Innovation

When AI Helps Generate Inventions, Who Is the Inventor?

With roots in the U.S. constitution, patent rights provide an exclusive property right in new inventions like drugs, new ways to make things like energy, and synthetic materials. However, as artificial intelligence (AI)—and specifically generative AI—are increasingly integrated across all fields such as health, education, and science, uncertainty exists in what inventions can be protected by patent rights when AI is part of the invention creation process.

https://www.csis.org/analysis/when-ai-helps-generate-inventions-who-inventor

AI Incident Database



https://incidentdatabase.ai/

Ethical Concerns with AI

Harm

Bias

Safety

Lack of Transparency

Unfairness

Robustness

Privacy

Errors

Unwarranted Trust

Loss of Human Autonomy

Job Displacement

Inexplicability

Deskilling

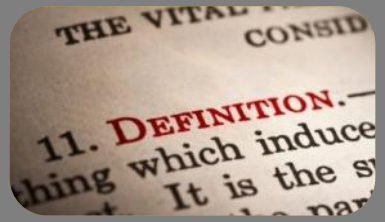
Lack of Control

…

Ethical Concerns with AI

What, in your opinion, are the top three ethical concerns, challenges, or issues with AI today and in the future?

Ethics



Ethics is the philosophical discipline interested in questions of right and wrong, good and bad, do's or don'ts.

- Ethics has been a key concern in philosophy since antiquity and in many philosophical systems it has been the key question

- Ethics theories come in three groups:
  - Virtue ethics → what would a person of good character do?
  - Deontological (or duty) ethics → rules, principles, and duties
  - Consequentialist (or results-based) ethics→ what will the outcomes of the action be?

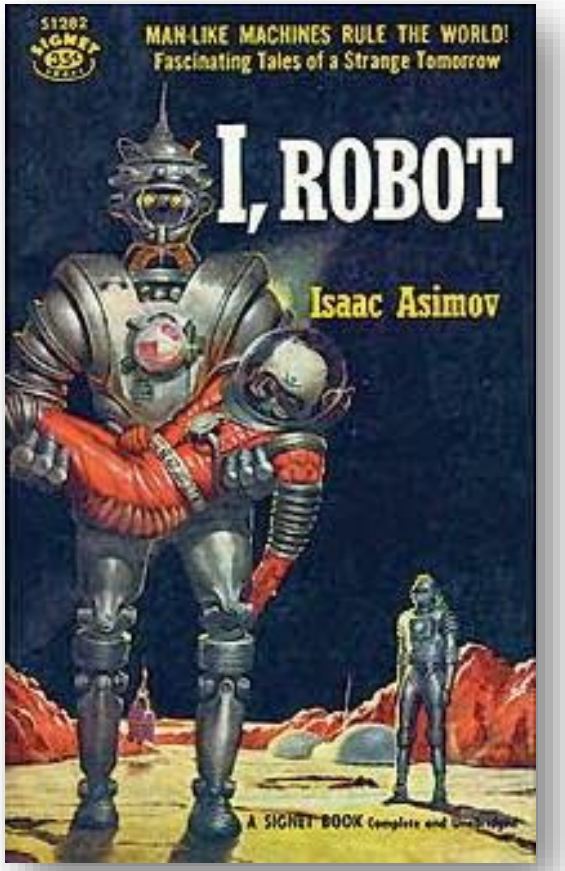
Hassan et al. 2018; Nussbaum 2009

Isaac Asimov's "Three Laws of Robotics"

- The Three Laws of Robotics were formulated by science-fiction writer Isaac Asimov and first appeared in his short story "Runaround" in 1942

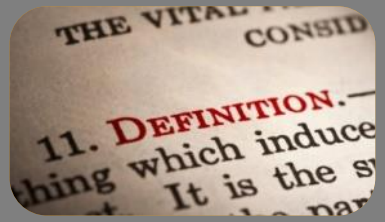
    1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
    2. *A robot must obey orders given it by human beings except where such orders would conflict with the First Law.*
    3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

- The laws have pervaded science fiction and have also influenced thought on the ethics of AI



Asimov 1942

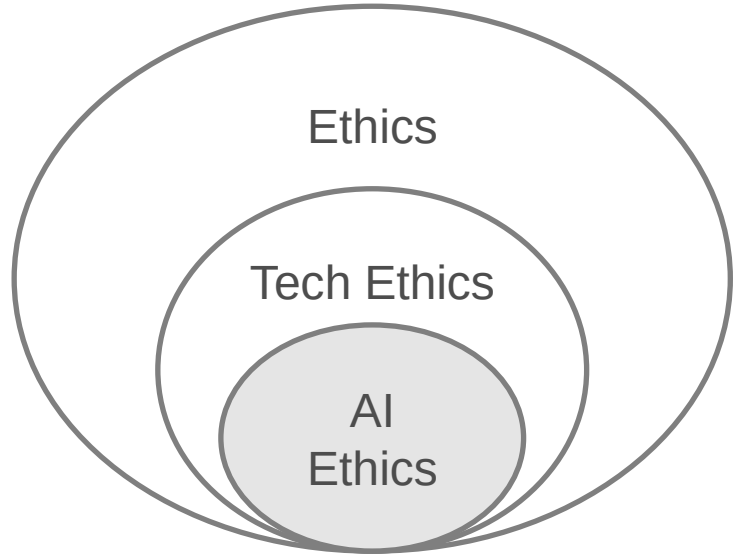
AI Ethics



AI ethics is a multidisciplinary field that addresses the ethical issues raised by the development, deployment, and use of AI.

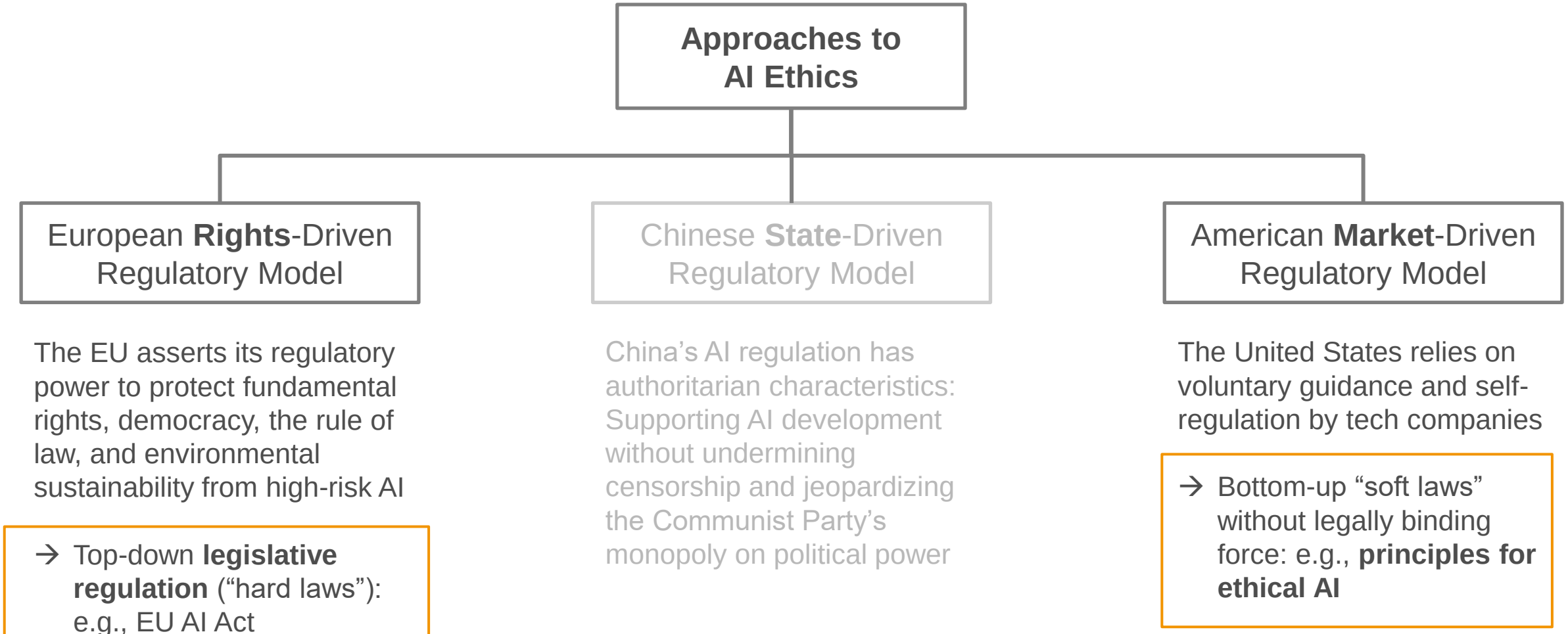
- The central concern of AI ethics is to **identify how AI can advance or raise concerns to the good life of individuals**, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society

- AI ethics has been mostly conceptual, with a significant emphasis on defining the principles for ethical AI



EU's AI HLEG (2019)

Two Approaches to AI Ethics



Bradford 2023

Legislative Approach to AI Ethics

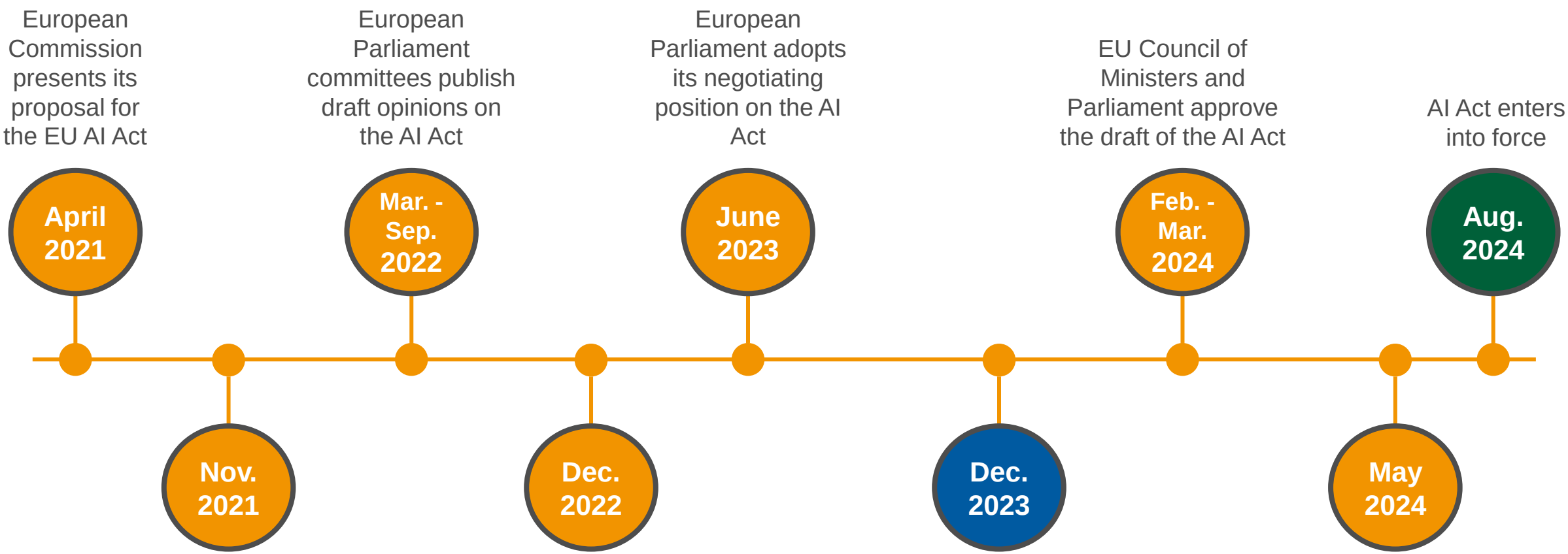
EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

On 9 December 2023, following three-day "marathon" talks, negotiators in the Council and the European Parliament reached a provisional agreement on the AI act. The act was formally adopted by the Council on 21 May 2024 and entered into force on **1 August 2024**.

https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Milestones in the History of the AI Act



https://artificialintelligenceact.eu/developments/

Difficult Negotiations on Foundation Models

UNIVERSITY OF PASSAU

EU's AI Act negotiations hit the brakes over foundation models

A technical meeting on the EU's AI regulation broke down on Friday (10 November) after large EU countries asked to retract the proposed approach for foundation models. Unless the deadlock is broken in the coming days, the whole legislation is at risk.

MISTRAL AI_

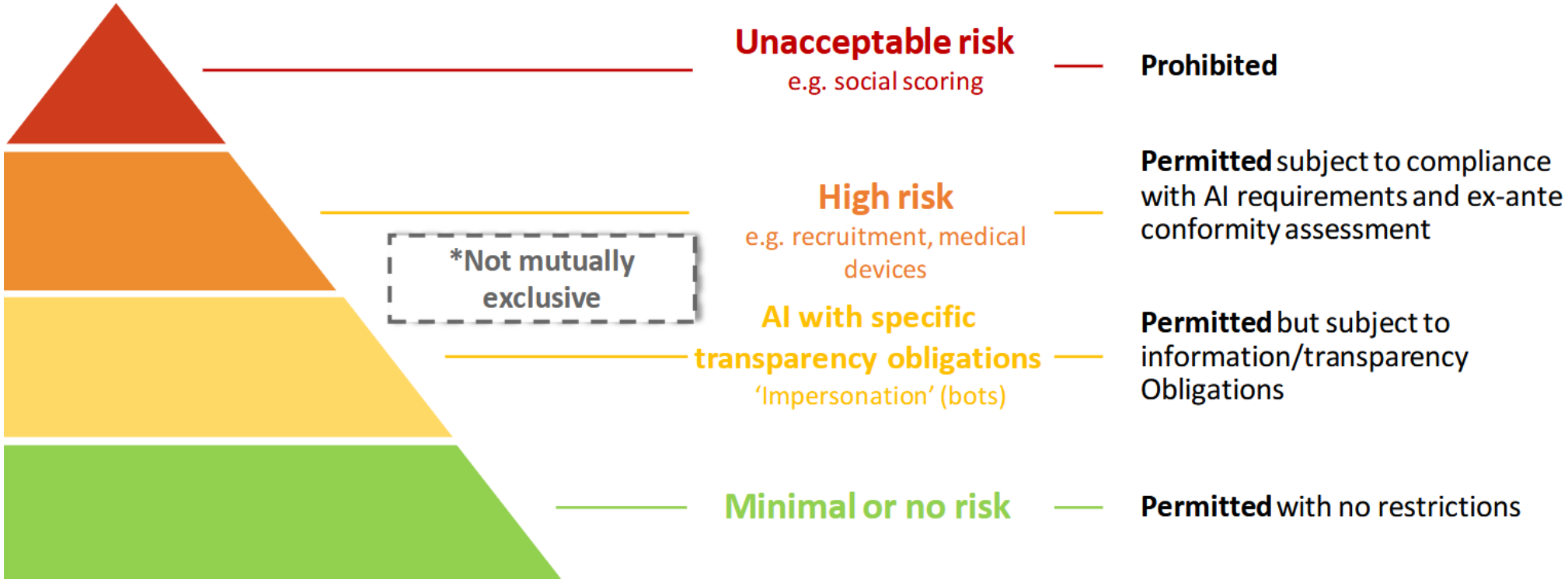
ALEPH ALPHA

*"**France, Germany, and Italy pushed against any type of regulation for foundation models**. Leading the charge against any regulation for foundation models in the AI rulebook is **Mistral**, a French AI start-up that has thrown the gauntlet to Big Tech. Cedric O, France's former state secretary for digital, is pushing Mistral's lobbying efforts, arguing that the AI Act could kill the company. Meanwhile, Germany is being pressured by its own leading AI company **Aleph Alpha** [..]. **All these companies fear the EU regulation might put them on a back foot compared to US and Chinese competitors**."*

https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/

Summary of the AI Act

- The EU AI Act pursues a **risk-based approach to the regulation of AI** and classifies AI into four different categories according to its risk:



https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf

Example: Transparency Obligations

TRANSPARENCY OBLIGATIONS FOR PROVIDERS AND DEPLOYERS OF CERTAIN AI SYSTEMS

Article 50

Transparency obligations for providers and deployers of certain AI systems

1. Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use. This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate or prosecute criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, unless those systems are available for the public to report a criminal offence.

Chatbots must not pretend to be humans!

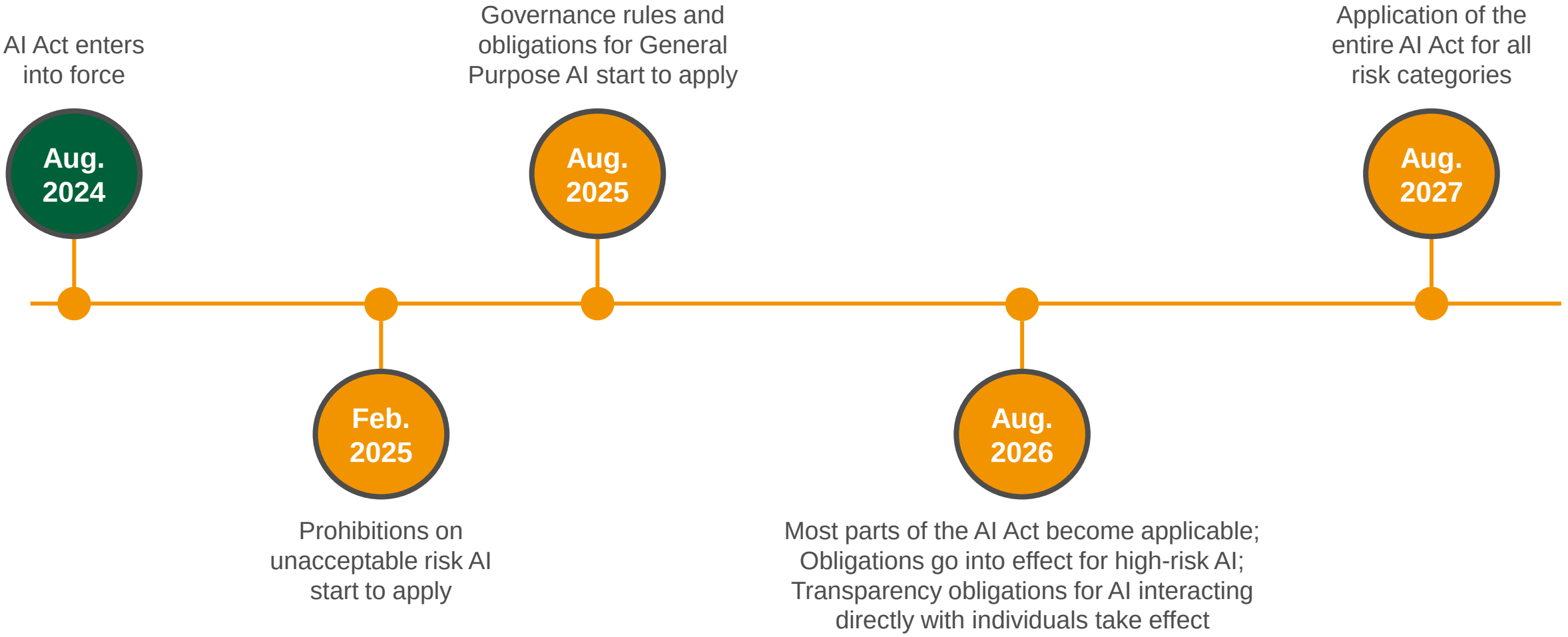
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R168 → Article 50

Penalties for Non-Compliance with the AI Act

- The penalties for non-compliance with the AI Act depend on the degree and type of non-compliance:

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R168 → Article 99

Implementation of the AI Act

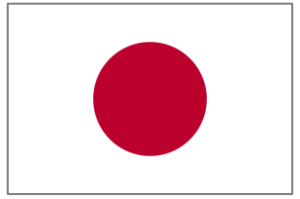


https://artificialintelligenceact.eu/implementation-timeline/

Global Impact of the AI Act?

- The EU AI Act will likely shape AI legislation in other countries by setting a high-risk-based regulation standard for AI governance
- Many countries are already considering the EU AI Act as they formulate their AI policies:

François-Philippe Champagne, **Canada's** Minister of Innovation, Science, and Industry, has stated that the country is closely following the development of the EU AI Act as it works on its own AI legislation



The **Japanese** government has expressed an interest in aligning its AI governance framework with the EU's approach as Japan's ruling party is expected to push for AI legislation within 2024

https://www.atlanticcouncil.org/blogs/geotech-cues/eu-ai-act-sets-the-stage-for-global-ai-governance-implications-for-us-companies-and-policymakers/

Criticism of EU AI Act

EU AI Act reaction: Tech experts say the world's first AI law is 'historic' but 'bittersweet'

EU AI Act: Sensible guardrail or innovation killer?

The AI Act: The EU's serial digital overregulation

The AI Act – The Epitome of Outdated Tech Governance — Exploring the Need for Innovative Regulation and Pathways to Modern Tech Governance

The EU's AI Dilemma: Innovation or Over-Regulation?

PACKED WITH LOOPHOLES: WHY THE AI ACT FAILS TO PROTECT CIVIC SPACE AND THE RULE OF LAW

Principled Approach to AI Ethics

Google's AI Principles

1. Be socially beneficial

2. Avoid creating or reinforcing unfair bias

3. Be built and tested for safety

4. Be accountable to people

5. Incorporate privacy design principles

6. Uphold high standards of scientific excellence

7. Be made available for uses that accord with these principles

https://ai.google/responsibility/principles/

Deutsche Telekom's AI Principles

1. Responsible

At Telekom we do different:
We are responsible.
Clear definition of who is responsible for which AI system.

2. Careful

At Telekom we do different:
We care.
AI systems and their usage obey human-defined rules.

3. Supporting

At Telekom we do different:
We put our customers first.
Using AI to simplify our customers' lives.

4. Transparent

At Telekom we do different:
We are transparent.
Transparency when a customer communicates with an AI and regarding our use of customer data.

5. Secure

At Telekom we do different:
We are secure.
Our customers' data is protected against unwanted external access.

6. Reliable

At Telekom we do different:
We set the framework.
Good preparation precedes an excellent outcome.

7. Trustworthy

At Telekom we do different:
We maintain control.
Continuous readiness to interfere in AI systems to prevent and/or reduce damage.

8. Cooperative

At Telekom we do different:
We foster the cooperative model.
Get advantages out of a cooperative and complementary model of human-machine interactions.

https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366

Bayerischer Rundfunk's AI Principles

Our Guidelines for the use of AI and Automation:

1. User Benefit

We demand proven benefits for our users and workflows when using AI systems. We deploy AI to help us use the resources that our contributors entrust us with more responsibly by making our work more efficient. We also use AI to set up new content, develop new methods for investigative journalism, support our workflows and improve our products.

Our critical reporting on Artificial Intelligence (Algorithmic Accountability Reporting) is backed up with our team's learnings from developing and using AI. We participate in the debate on the societal impact of algorithms by providing information on emerging trends, investigating algorithms, explaining how technologies work and strengthening an open debate on the future role of public service media in our society.

2. Accurate Representation of AI

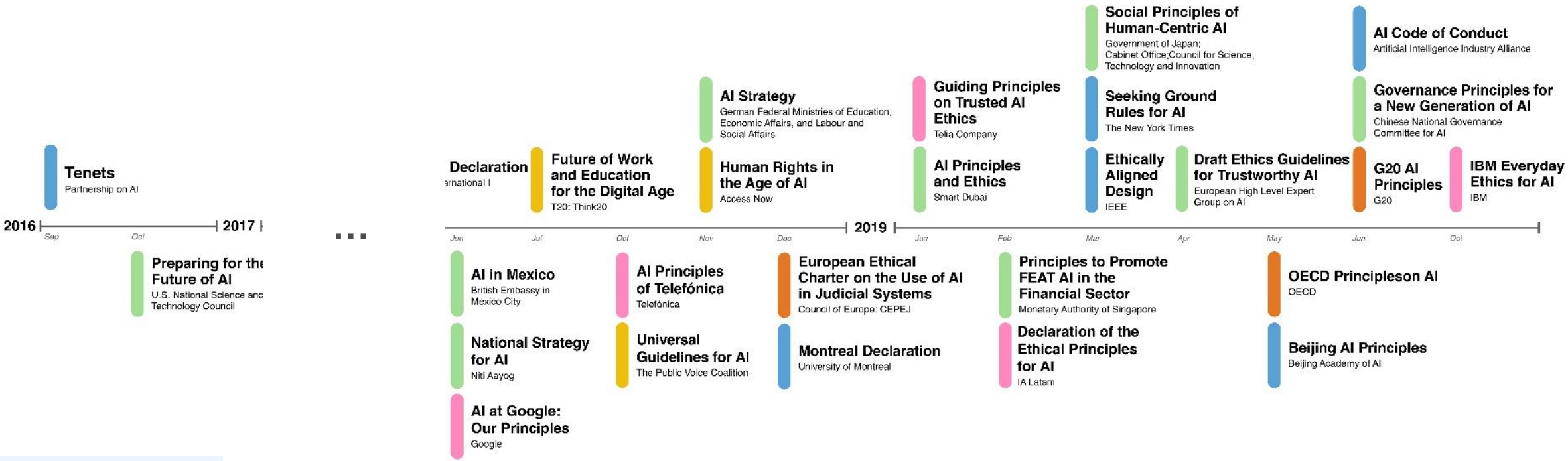
We describe AI as technical systems and avoid using misleading anthropomorphic wording. Drawing analogies between AI functions and human intelligence and skills like reading, writing, or thinking is likely inaccurate and presents technology as overly powerful. Metaphors and imagery can reinforce the delusive impression of artificial beings. Therefore, we avoid using humanized images and descriptions in our publications.

3. Editorial Control & Transparency

The principle of editorial control remains mandatory with automated content. This means that only human individuals and editorial teams can be responsible for content, never systems helping to create the content. We verify data sources and thoroughly check models and software for reliability. We set up customized human workflows and technical controls for the technology we use. Results of generative AI systems have to be checked editorially beforehand.

https://www.br.de/extra/ai-automation-lab-english/ai-ethics100.html

Timeline of Ethical AI Principles

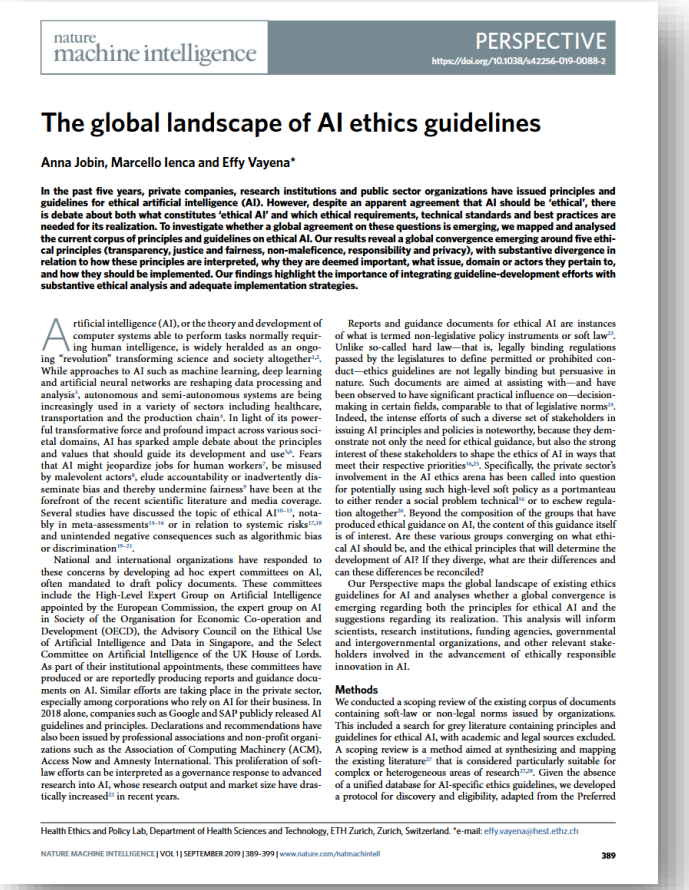


Fjeld et al. 2020

Analysis of the AI Ethics Landscape

- Jobin et al. (2019) conducted a systematic meta-analysis of **84 AI ethics documents** published by government agencies (e.g., the European Union, OECD), private companies (e.g., Microsoft, SAP), and research institutions (e.g., ACM, IEEE) from different geographic regions

- The principles of **responsibility**, **justice and fairness**, **transparency**, **non-maleficence**, and **privacy** were referenced in more than half of all documents, indicating an emerging global convergence across stakeholders on the importance of these five principles



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 relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–13], notably in meta-assessments[14–16] or in relation to systemic risks[17,18] and unintended negative consequences such as algorithmic bias or discrimination[19–21].

National and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords. As part of their institutional appointments, these committees have produced or are reportedly producing reports and guidance documents on AI. Similar efforts are taking place in the private sector, especially among corporations who rely on AI for their business. In 2018 alone, companies such as Google and SAP publicly released AI guidelines and principles. Declarations and recommendations have also been issued by professional associations and non-profit organizations such as the Association of Computing Machinery (ACM), Access Now and Amnesty International. This proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased[22] in recent years.

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[16,25]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[16] or to eschew regulation altogether[26]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

Our Perspective maps the global landscape of existing ethics guidelines for AI and analyses whether a global convergence is emerging regarding both the principles for ethical AI and the suggestions regarding its realization. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations, and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Methods

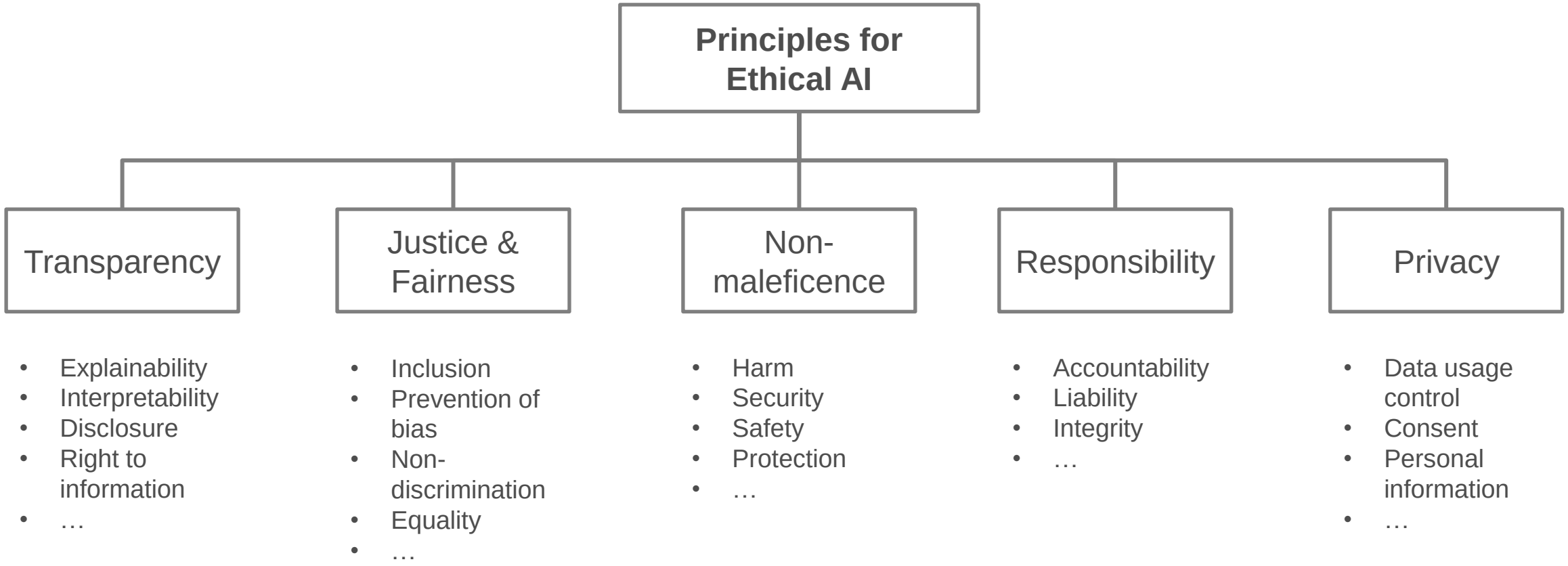
We conducted a scoping review of the existing corpus of documents containing soft-law or non-legal norms issued by organizations. This included a search for grey literature containing principles and guidelines for ethical AI, with academic and legal sources excluded. A scoping review is a method aimed at synthesizing and mapping the existing literature[27] that is considered particularly suitable for complex or heterogeneous areas of research[27,28]. Given the absence of a unified database for AI-specific ethics guidelines, we developed a protocol for discovery and eligibility, adapted from the Preferred

Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. *e-mail: effy.vayena@hest.ethz.ch

NATURE MACHINE INTELLIGENCE | VOL 1 | SEPTEMBER 2019 | 389–399 | www.nature.com/natmachintell

Jobin et al. 2019

Five Key Principles for Ethical AI



Jobin et al. 2019

Transparency

AI actors "should provide meaningful information, appropriate to the context [...] to foster a general understanding of AI systems, to make stakeholders aware of their interactions with AI systems [...] to enable those affected by an AI system to understand the outcome, and, to enable those adversely affected by an AI system to challenge its outcome [...]"

- Possible scenarios:
  - Explaining the features of a machine learning model that led to a specific prediction (→ see XAI lecture)
  - Telling customers that they are interacting with a chatbot and not with a human service employee
  - Disclosing how content was generated/improved using ChatGPT (e.g., parts of your master thesis)
  - …

https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Justice & Fairness

The development, deployment and use of AI systems must be fair. […] fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization. […] The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. […]

- "Fair AI"
- Possible scenarios:
  - Making sure that there are no gender or racial biases in job application decisions made by AI
  - Allowing customers to challenge (unfair) loan or credit decisions made by AI
  - …

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Non-Maleficence

“AI systems should not be used in ways that cause or exacerbate harm, whether individual or collective, and including harm to social, cultural, economic, natural, and political environments.”

- Possible scenarios:
  - Protecting democracy from AI-generated deepfakes and misinformation
  - Reducing the significant energy consumption of generative AI training processes
  - …

https://unsceb.org/sites/default/files/2023-03/CEB_2022_2_Add.1%20%28AI%20ethics%20principles%29.pdf

Responsibility

"[…] The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system. Appropriate oversight, impact assessment, audit, and due diligence mechanisms […] should be developed to ensure accountability for AI systems and their impact throughout their life cycle."

- "Responsible AI"
- Possible scenarios:
  - Putting the responsibility on doctors to verify the suggestions of an AI-based diagnosis tool before applying them to a patient
  - Holding car manufacturers, software developers, and operators accountable when an autonomous vehicle is involved in an accident
  - …

https://unesdoc.unesco.org/ark:/48223/pf0000380455.locale=en

Privacy

“People should have the right to access, manage and control the data they generate, given AI systems’ power to analyze and utilize that data.”

- Possible scenarios:
  - Refrain from using personal data scraped from the web to train AI models
  - Making sure that AI systems do not leak sensitive user or company data
  - …

https://futureoflife.org/open-letter/ai-principles/

Criticism of Ethical AI Principles

Principles alone cannot guarantee ethical AI

Brent Mittelstadt

Nature Machine Intelligence **1**, 501–507 (2019) | Cite this article

21k Accesses | **579** Citations | **270** Altmetric | Metrics

Abstract

Artificial intelligence (AI) ethics is now a global topic of discussion in academic and policy circles. At least 84 public–private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. According to recent meta-analyses, AI ethics has seemingly converged on a set of principles that closely resemble the four classic principles of medical ethics. Despite the initial credibility granted to a principled approach to AI ethics by the connection to principles in medical ethics, there are reasons to be concerned about its future impact on AI development and governance. Significant differences exist between medicine and AI development that suggest a principled approach for the latter may not enjoy success comparable to the former. Compared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms. These differences suggest we should not yet celebrate consensus around high-level principles that hide deep political and normative disagreement.

https://www.nature.com/articles/s42256-019-0114-4

The uselessness of AI ethics

Original Research | Open access | Published: 23 August 2022

Volume 3, pages 869–877, (2023) Cite this article

Luke Munn

Abstract

As the awareness of AI's power and danger has risen, the dominant response has been a turn to ethical principles. A flood of AI guidelines and codes of ethics have been released in both the public and private sector in the last several years. However, these are *meaningless principles* which are contested or incoherent, making them difficult to apply; they are *isolated principles* situated in an industry and education system which largely ignores ethics; and they are *toothless principles* which lack consequences and adhere to corporate agendas. For these reasons, I argue that AI ethical principles are useless, failing to mitigate the racial, social, and environmental damages of AI technologies in any meaningful sense. The result is a gap between high-minded principles and technological practice. Even when this gap is acknowledged and principles seek to be "operationalized," the translation from complex social concepts to technical rulesets is non-trivial. In a zero-sum world, the dominant turn to AI principles is not just fruitless but a dangerous distraction, diverting immense financial and human resources away from potentially more effective activity. I conclude by highlighting alternative approaches to AI justice that go beyond ethical principles: thinking more broadly about systems of oppression and more narrowly about accuracy and auditing.

https://link.springer.com/article/10.1007/s43681-022-00209-w

Key Challenges of Ethical AI Principles

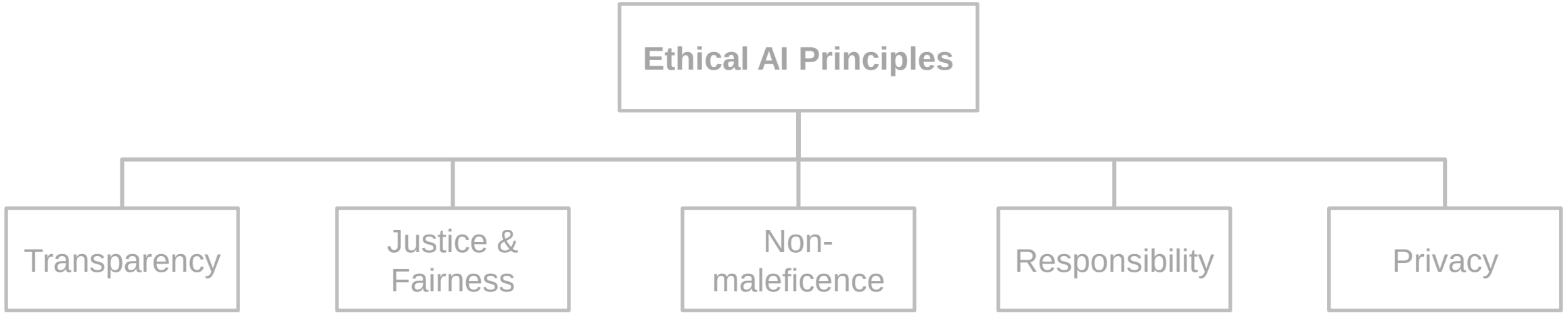
Abstract and high-level:
Large gap between principles and their practical implementation

Conflicting:
Tensions between different principles

“Toothless”:
Lack of consequences and mechanisms to enforce compliance

Mittelstadt 2019; Munn 2023

Principle–Practice Gap in AI Ethics



? **How can organizations translate these abstract, high-level principles into actionable strategies?**

Jobin et al. 2019

Research Example: RPA Implementation Case Study

- Case study at a German-based energy service provider
- Over a period of two and a half years, 45 robotic process automation (RPA) bots were implemented to automate many back-office processes (e.g., customer service, billing)
- These RPA bots enabled the automated processing of more than 200,000 back-office transactions per year → equivalent to 5 full-time employees
- Many ethical concerns and challenges: job displacement, lack of control, loss of human autonomy, errors, ...

→ ***How to address these challenges based on ethical AI principles?***

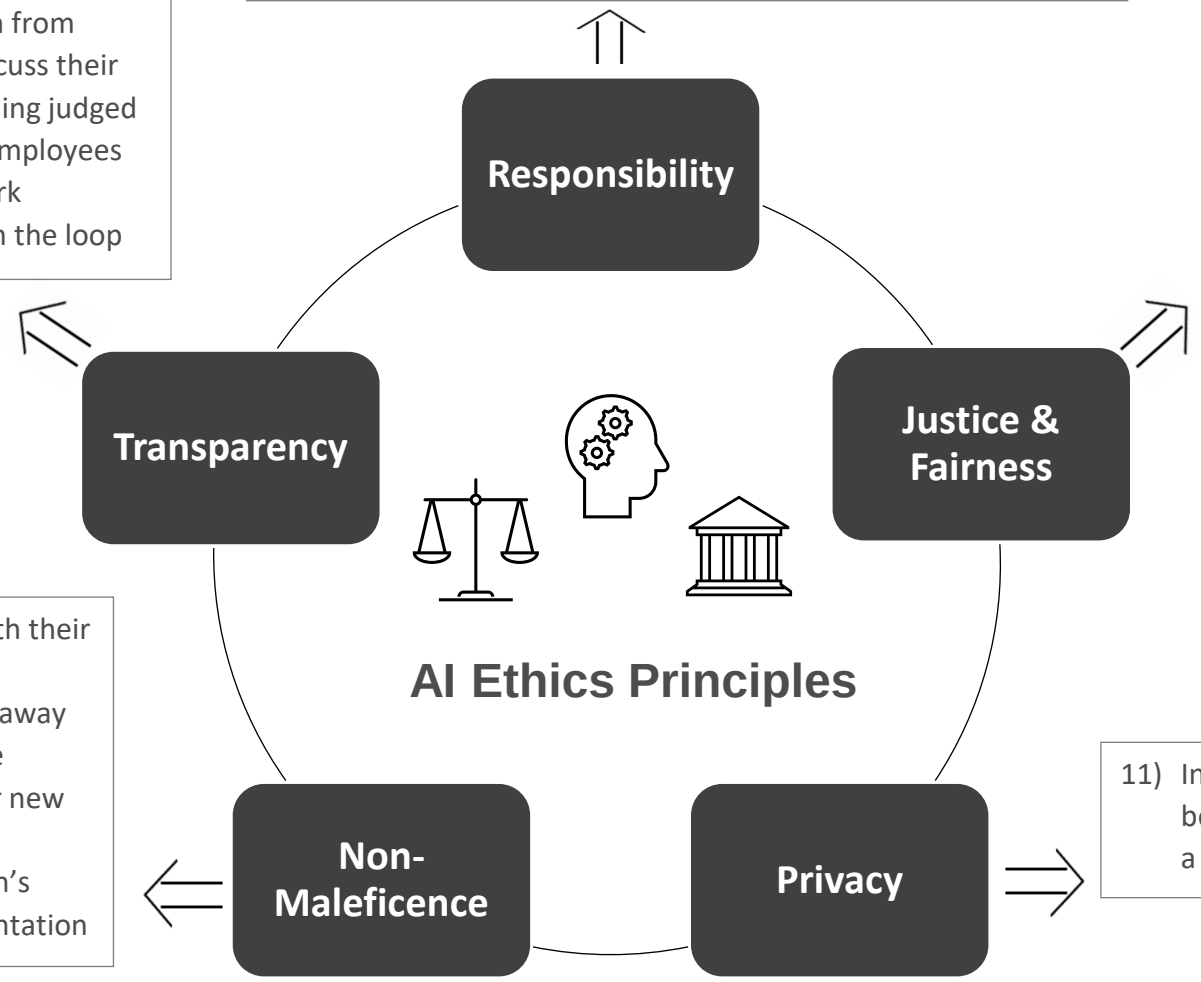
Schulte-Derne & Gnewuch 2024

Recommendations for Translating Ethical AI Principles into Practice

1) Define and assign roles for selecting, implementing, supervising, and optimizing AI-automated processes
2) Avoid delegating the overall responsibility for (negative) consequences to those developing or implementing AI

5) Enable open and honest communication from trusted peers to allow employees to discuss their concerns and ask questions without feeling judged
6) Make the invisible visible by providing employees the opportunity to observe the AI at work
7) Keep employee representation bodies in the loop

3) Strive for a fair distribution of AI's benefits and costs among all employees
4) Understand and address personal tensions in a diverse workforce with different socio-demographic backgrounds and levels of technical expertise



8) Help employees find strategies to cope with their fears
9) Avoid overwhelming employees by taking away all their routine tasks before they have the opportunity to develop confidence in their new responsibilities
10) Examine who else besides the organization's employees is impacted by the AI implementation

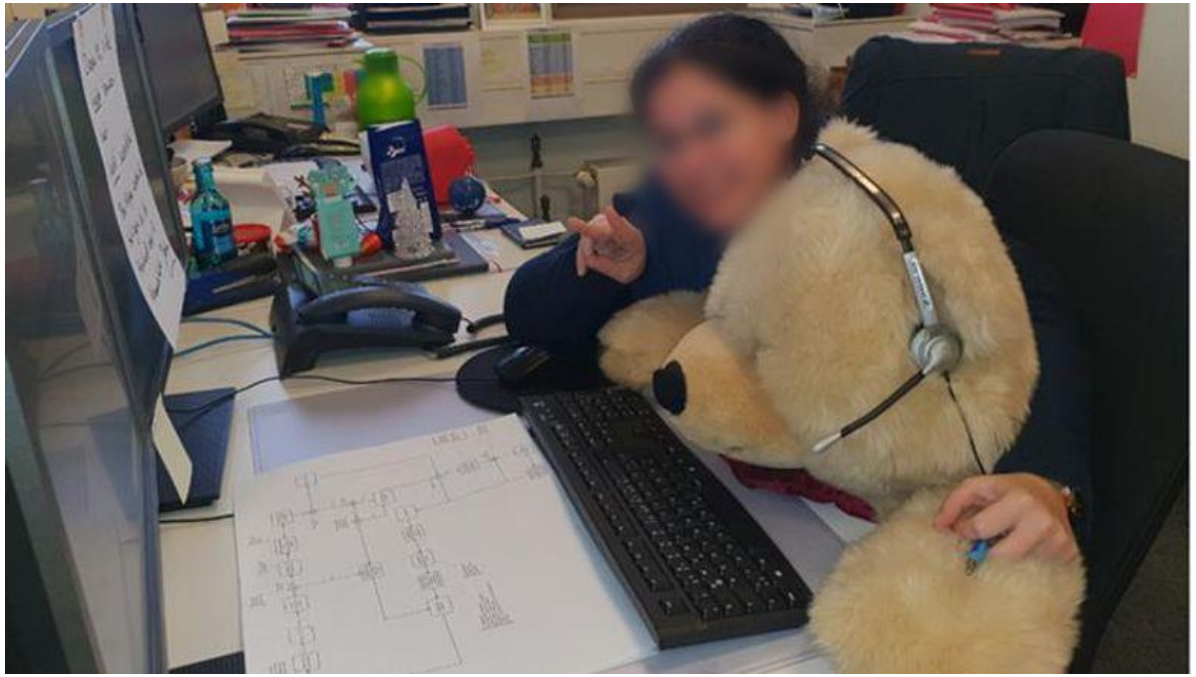
11) Involve data protection officers before investing time and effort in a new AI initiative

Schulte-Derne & Gnewuch 2024

Example: Transparency

Strategy #6: *Make the invisible visible by providing employees the opportunity to observe the AI at work*

- The RPA team set up the RPA software on a normal desktop PC placed in one of the offices and invited employees and works council members to observe its functioning at any time
- This not only reduced employees' uncertainty and works council members' concerns, but also turned fear into curiosity



Teddy Bear "Robbie" Sitting in Front of the RPA Workstation

Schulte-Derne & Gnewuch 2024

Legislative vs. Principled Approach to AI Ethics

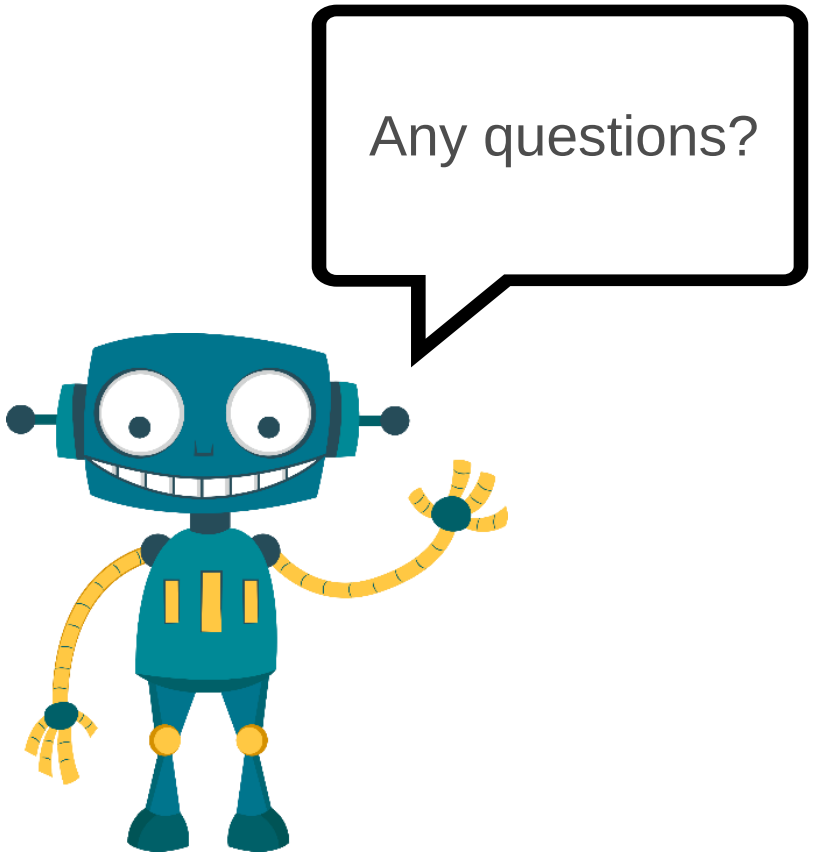
In your opinion, which approach to AI ethics is better? Should governments formally regulate AI? Should we trust companies to follow their ethics principles?

→Discuss these questions with a partner for **~5 minutes** and be ready to share your opinions

Key Takeaways From This Lecture

- The development, deployment, and use of AI raise many ethical concerns, such as job displacement, bias, lack of control, and privacy issues
- The field of AI ethics aims to address these concerns by exploring and guiding what is morally right or wrong
- There are two main approaches to AI ethics
- 1) The legislative approach focusing on top-down regulation
  - Example: EU AI Act
- 2) The principled approach focusing on bottom-up “soft laws”
  - Example: Ethical AI principles such as transparency, justice and fairness, and responsibility
- Both approaches have their benefits and challenges:
  - Regulation may impose inflexible requirements and hinder innovation
  - Principles are non-binding and difficult to put into practice

Thank you for your attention!



References

Asimov, I. (1942). Runaround. In: I. Asimov (Ed.), I, Robot Collection (pp. 40–54). New York, NY: Doubleday.

Bradford, A. (2023). Digital empires: The global battle to regulate technology. Oxford University Press.

EU's AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence (AI HLEG). https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. The Berkman Klein Center for Internet and Society. https://cyber.harvard.edu/publication/2020/principled-ai

Hassan, N. R., Mingers, J., & Stahl, B. (2018). Philosophy and information systems: where are we and where should we go?. European Journal of Information Systems, 27(3), 263-277.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature machine intelligence, 1(11), 501-507.

Munn, L. (2023). The uselessness of AI ethics. AI and Ethics, 3(3), 869-877.

Nussbaum, M. C. (2009). The therapy of desire theory and practice in Hellenistic ethics. Princeton, NJ: Princeton University Press.

Schulte-Derne, D., & Gnewuch, U. (2024). Translating AI ethics principles into practice to support robotic process automation implementation. MIS Quarterly Executive, 23(2), 187-203.