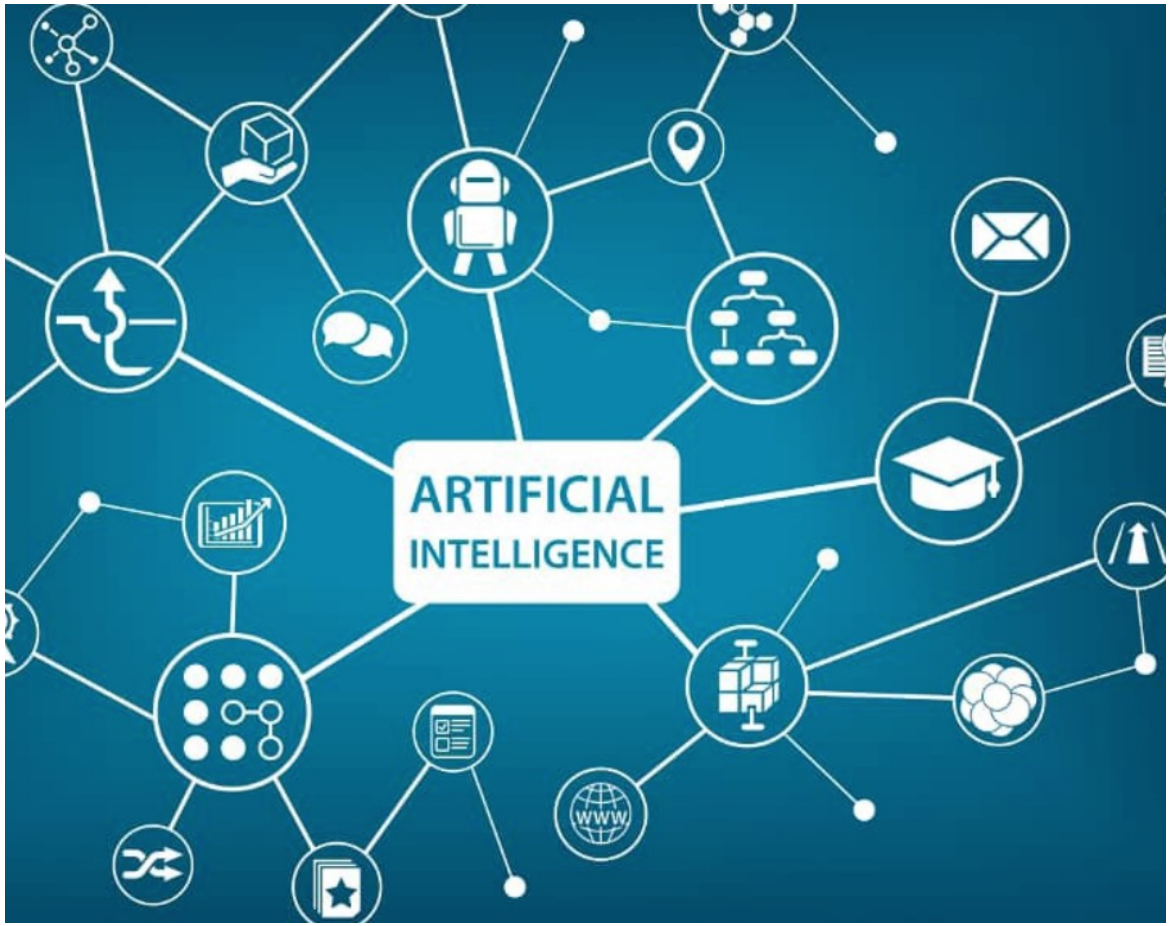
AI ethics

Francesca Rossi

IBM AI Ethics Global Leader

AAAI President

Pervasive AI applications

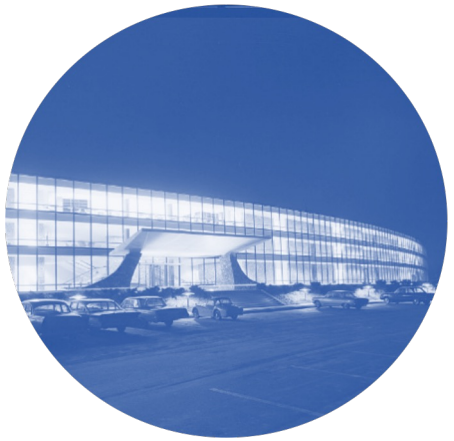


- Digital assistants: travel and home
- Driving/travel support: auto-pilot, ride sharing
- Customer care: chatbots
- Online recommendations: friends, purchases, movies
- Media and news: add placement, news curation
- Healthcare: medical image analysis, treatment plan recommendation
- Financial services: credit risk scoring, loan approval, fraud detection
- Job market: resume prioritization
- Judicial system: recidivism prediction

IBM

High-stakes decision-making applications

Credit



Employment

Admission

Healthcare

Enterprise Workflows

IBM

What can AI be useful for, in a company?

AI can help improve

- All business functions and processes
- Client relationship, engagement, and experience
- Credit loss reduction
- Growth
- Better business decisions
- Risk management

In most areas of operations

- Payments
- Personalized services/policies
- Digital Assets
- Client and investment risk management
- Internal and external audit
- Data governance and privacy
- Insurance
- Customer relationship
- Fraud prevention and detection

Especially now

The pandemic has accelerated the digitalization

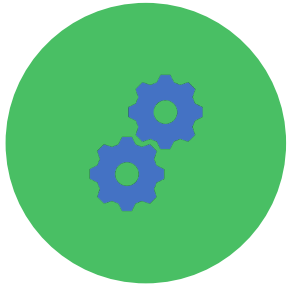
Data-driven organizations, based on **data-enabled clients** (IEEE playbook on Trusted Data and AI for Financial Services, 2021)

Technology adoption leaders outperformed their peers by 6% on revenue growth during the disruption across 12 industries (IBM IBV Study, 2020)

AI Ethics

Multidisciplinary field of study

Main goal: how to optimize AI's beneficial impact while reducing risks and adverse outcomes



Tech solutions: How to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios

Socio-tech approach: To identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society

IBM

AI Ethics issues -1

Data privacy and governance	AI needs data
Fairness	AI can make or recommend decisions, and these should not be discriminatory
Inclusion	Use of AI should not increase the social gaps
Explainability	AI is often opaque
Transparency	More informed use of AI
Accountability	AI is based on statistics and has always a small percentage of error
Social impact	Fast transformation of jobs and society

AI Ethics issues -2

Human and moral agency	AI can profile people and manipulate their preferences
Social good uses	UN Sustainable Development Goals
Environmental impact	Foundation models need huge amounts of energy for training and deployment
Power imbalance	Centralization of data and power

Semantic web specific issues

Trust	Who defined the notion of truth?
Data Privacy	How to avoid data and privacy leakage?
Human engagement and oversight	Should machine-readable metadata replace human interpretation?
Decentralization	No centralized regulatory entity: how to regulate crime, harmful content, youthful abuse, etc?
Non-compositionality	Ethical components do not assure an ethical composition

AI Ethics 3.0

Awareness

- Mostly in academia, multi-disciplinary

2015–2016

Principles

- Corporations, governments, academia, civil society, multi-stakeholder organizations

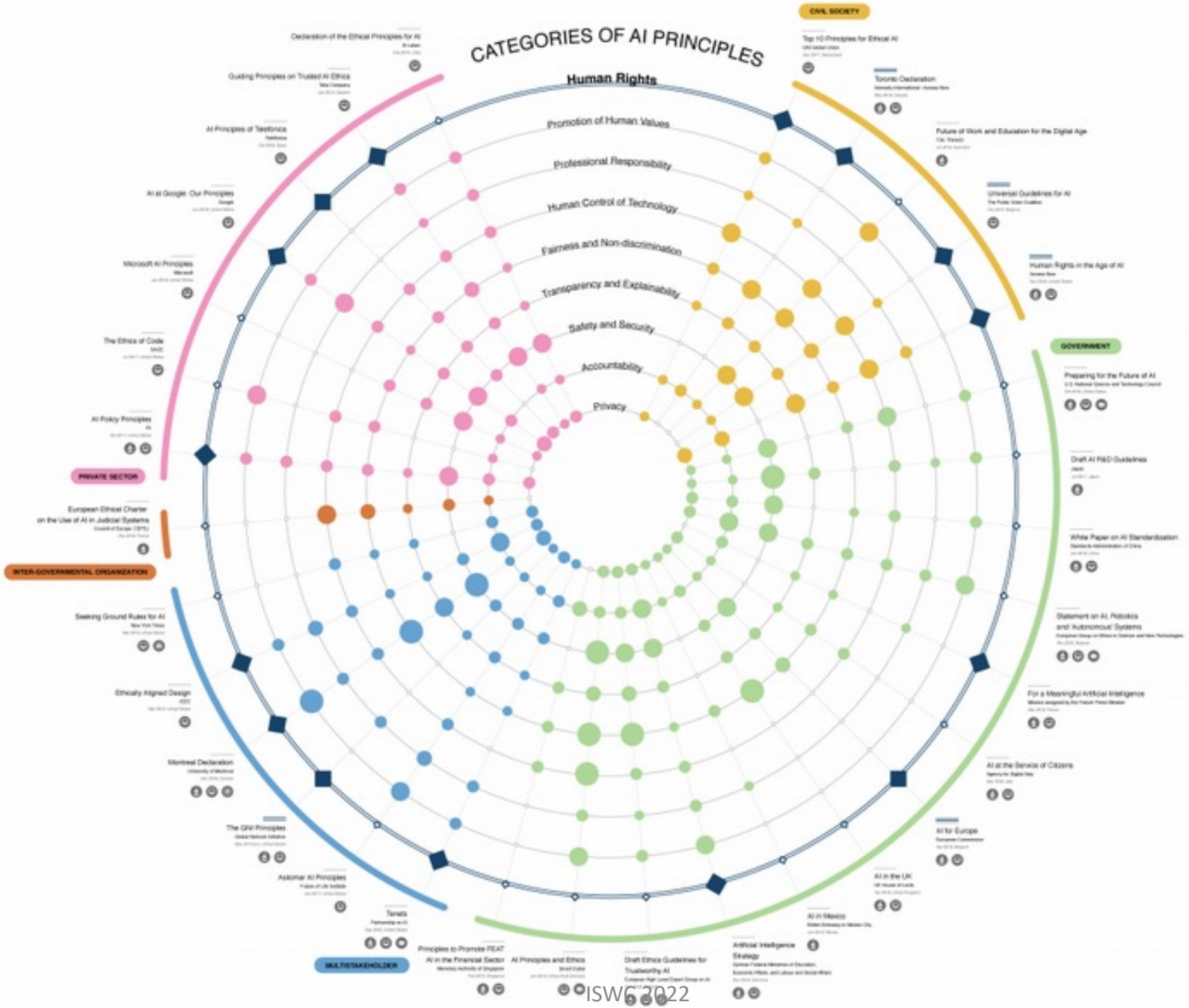
2017–2018

Practice

- Regulations, standards, corporate directives, processes, auditing, certifications

2019-ongoing

IBM



AI Ethics principles

Actors:

- Private sector
- Inter-governmental
- Multistakeholder
- Governments
- Civil society

Main themes:

- Human rights
- Human values
- Responsibility
- Human control
- Fairness
- Transparency and explainability
- Safety and Security
- Accountability
- Privacy

Principled AI Project, Berkman Klein's Cyberlaw Clinic, 2019

AI Ethics in practice

Research

- Fairness
- Explainability
- Interpretability
- Robustness
- Privacy
- Value alignment

AI companies

- Governance
- Internal processes
- Tools
- Risk assessment
- Training

Standard bodies

- IEEE P7000 series:
- **IEEE 7000™-2021 – Model Process for Addressing Ethical Concerns During System Design**
- **IEEE P7001™ – Transparency of Autonomous Systems**
- **IEEE P7002™ – Data Privacy Process**
- **IEEE P7003™ – Algorithmic Bias Considerations**
- **IEEE P7004™ – Standard on Child and Student Data Governance**
- **IEEE P7005™ – Standard on Employer Data Governance**
- **IEEE P7007™ – Ontological Standard for Ethically driven Robotics and Automation Systems**
- **IEEE P7008™ – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems**
- **IEEE P7009™ – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems**
- **IEEE 7010™-2021 – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems**
- **IEEE P7011™ – Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources**
- **IEEE P7012™ – Standard for Machine Readable Personal Privacy Terms**
- **IEEE P7014™ – Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems**

Educational institutions

1. Ethics of AI (University of Helsinki)
2. AI-Ethics: Global Perspectives (aiethicscourse.org)
3. AI Ethics for Business (Seattle University)
4. Bias and Discrimination in AI (Université de Montréal)
5. Data Science Ethics (University of Michigan)
6. Intro to AI Ethics (Kaggle)
7. Ethics in AI and Data Science (LFS112x)
8. Practical Data Ethics (Fast AI)
9. Data Ethics, AI and Responsible Innovation (University of Edinburgh)
10. Identify guiding principles for responsible AI (Microsoft)
11. Human-Computer Interaction III: Ethics, Needfinding & Prototyping (Georgia Tech)
12. Ethics in Action (SDGAcademyX)
13. Explainable Machine Learning with LIME and H2O in R (Coursera)
14. An introduction to explainable AI, and why we need it
15. Explainable AI: Scene Classification and GradCam Visualization (Coursera)
16. Interpretable Machine Learning Applications: Part 1 & 2 (Coursera)

Nerd for Tech, 2021

Governments

Example: EU AI Act

- Risk-based approach
- Four levels of risk
- Focus on AI systems
- Obligations for high risk applications, providers and users

IBM

AI Ethics in practice

Research

- Fairness
- Explainability
- Interpretability
- Value alignment

AI companies

- Governance
- Internal processes
- Tools

Standard bodies

- IEEE P7000 series:
- **IEEE 7000™-2021 – Model Process for Addressing Ethical Concerns During System Design**
- **IEEE P7001™ – Transparency of Autonomous Systems**
- **IEEE P7002™ – Data Privacy Process**
- **IEEE P7003™ – Algorithmic Bias Considerations**
- **IEEE P7004™ – Standard on Child and Student Data Governance**

for Robotic, Intelligent and Autonomous Systems

- **IEEE P7009™ – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems**
- **IEEE 7010™-2021 – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems**
- **IEEE P7011™ – Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources**
- **IEEE P7012™ – Standard for Machine Readable Personal Privacy Terms**
- **IEEE P7014™ – Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems**

Educational institutions

1. Ethics of AI (University of Helsinki)
2. AI-Ethics: Global Perspectives (aiethicscourse.org)
3. AI Ethics for Business (Seattle University)
4. Bias and Discrimination in AI (Université de Montréal)
5. Data Science Ethics (University of Michigan)

9. Data Ethics, AI and Responsible Innovation (University of Edinburgh)
10. Identify guiding principles for responsible AI (Microsoft)
11. Human-Computer Interaction III: Ethics, Needfinding & Prototyping (Georgia Tech)
12. Ethics in Action (SDGAcademyX)
13. Explainable Machine Learning with LIME and H2O in R (Coursera)
14. An introduction to explainable AI, and why we need it
15. Explainable AI: Scene Classification and GradCam Visualization (Coursera)
16. Interpretable Machine Learning Applications: Part 1 & 2 (Coursera)

Nerd for Tech, 2021

Governments

Example: EU AI Act

- Risk-based approach
- Four levels of risk
- Obligations for high risk applications, providers and users

Civil society organizations, media, activists, society at large

IBM

Research: a personal journey on value alignment

Embedding ethical principles in collective decision making systems, IBM+MIT+Harvard+other univ., 2016-2017

- How to make collective decisions in a way that is aligned to some ethical principles

Ethically bounded AI, IBM 2018-2019

- Reinforcement learning + ethical policy, orchestration

Engineering morality, IBM+MIT, 2019-2021

- Modelling and reasoning with human switching between deontology and consequentialism

Embedding and learning ethical properties in collective decision systems, IBM+RPI, 2020-2022

- Tradeoffs between privacy, social welfare, and fairness

Thinking fast and slow in AI, 2020-

- Fast and slow solvers, metacognition
- Human-like decision modalities
- Support for human decision making

IBM

The AI Ethics Drivers

Why should a company building or using AI care about AI ethics?

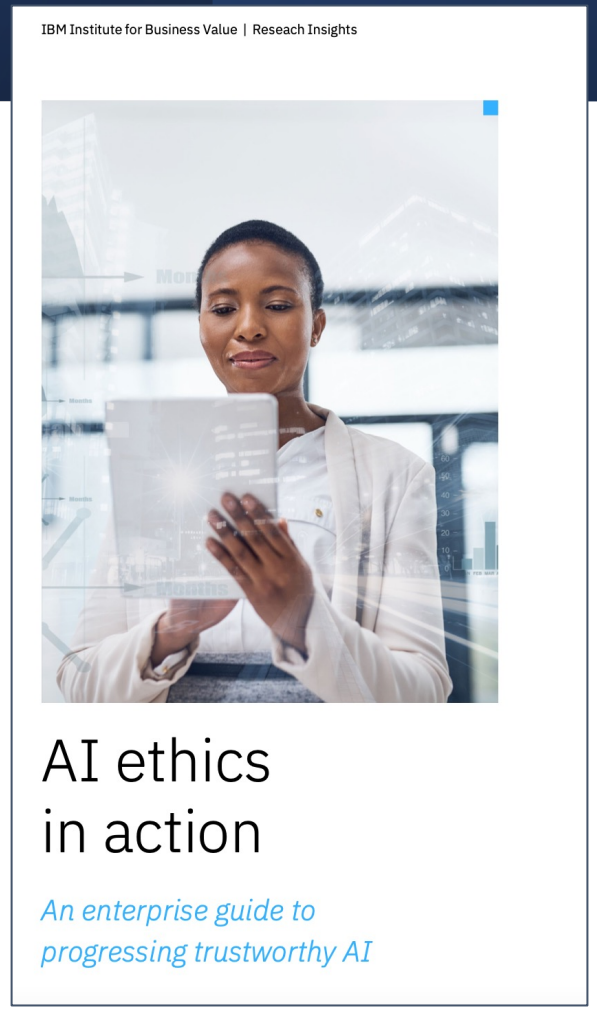
Company values	Company reputation and trust	Existing or expected regulations	Social justice and equity
Client requests	Media coverage	Differentiators	Business opportunities

IBM

What are companies concretely doing to address AI Ethics issues?

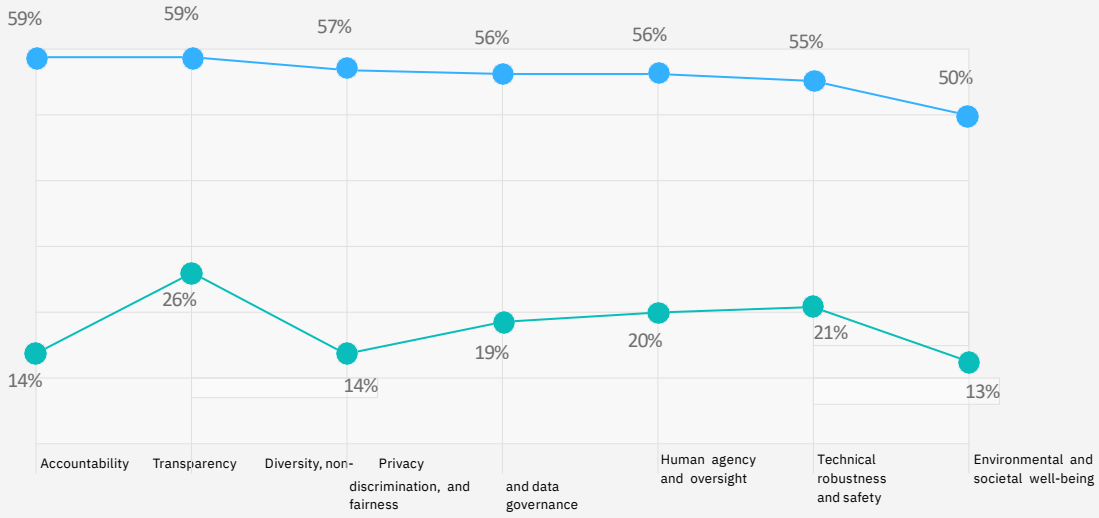
- An IBM Institute for Business Value study, 2022

- 1,200 executives and AI developers
- 22 countries



The intention-action gap

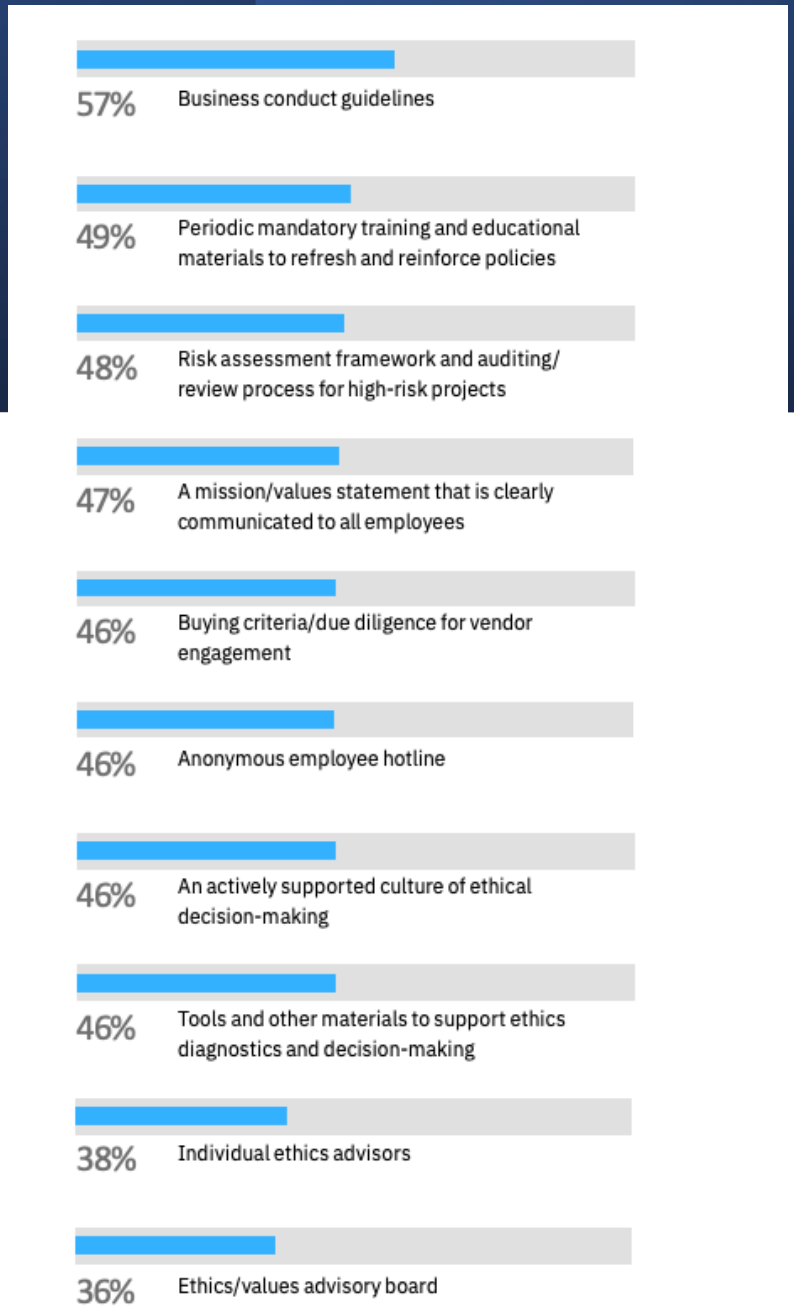
Organizations are endorsing AI ethics principles— but are still catching up on implementing them



Note: AI ethics principles as defined by the European Commission High-Level Expert Group on AI in "Ethics guidelines for trustworthy AI." April 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

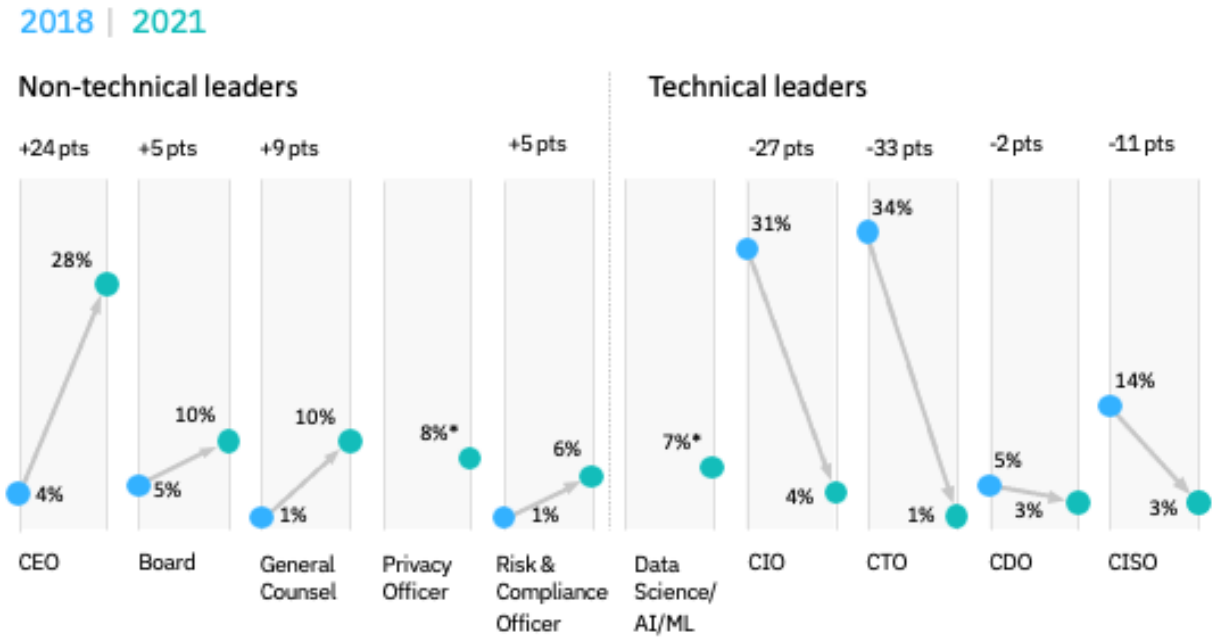
First steps

Many organizations are incorporating AI ethics into existing business ethics mechanisms



Not just technical issues

Good news: from 2018 to 2021, those primarily accountable for AI ethics have shifted from technical to non-technical leaders

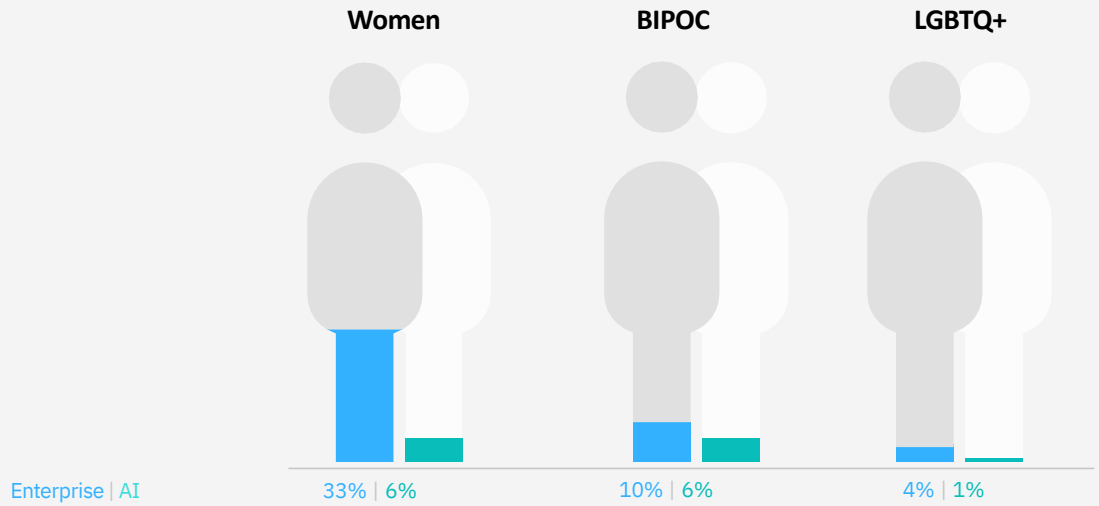


Q: Which function is primarily accountable for AI ethics?
Source for 2018 survey data: Goehring, Brian, Francesca Rossi, and Dave Zaharchuk. "Advancing AI ethics beyond compliance: From principles to practice." IBM Institute for Business Value. April 2020.
**Position was not included in 2018 data*

- 2018: IBV study on AI Ethics

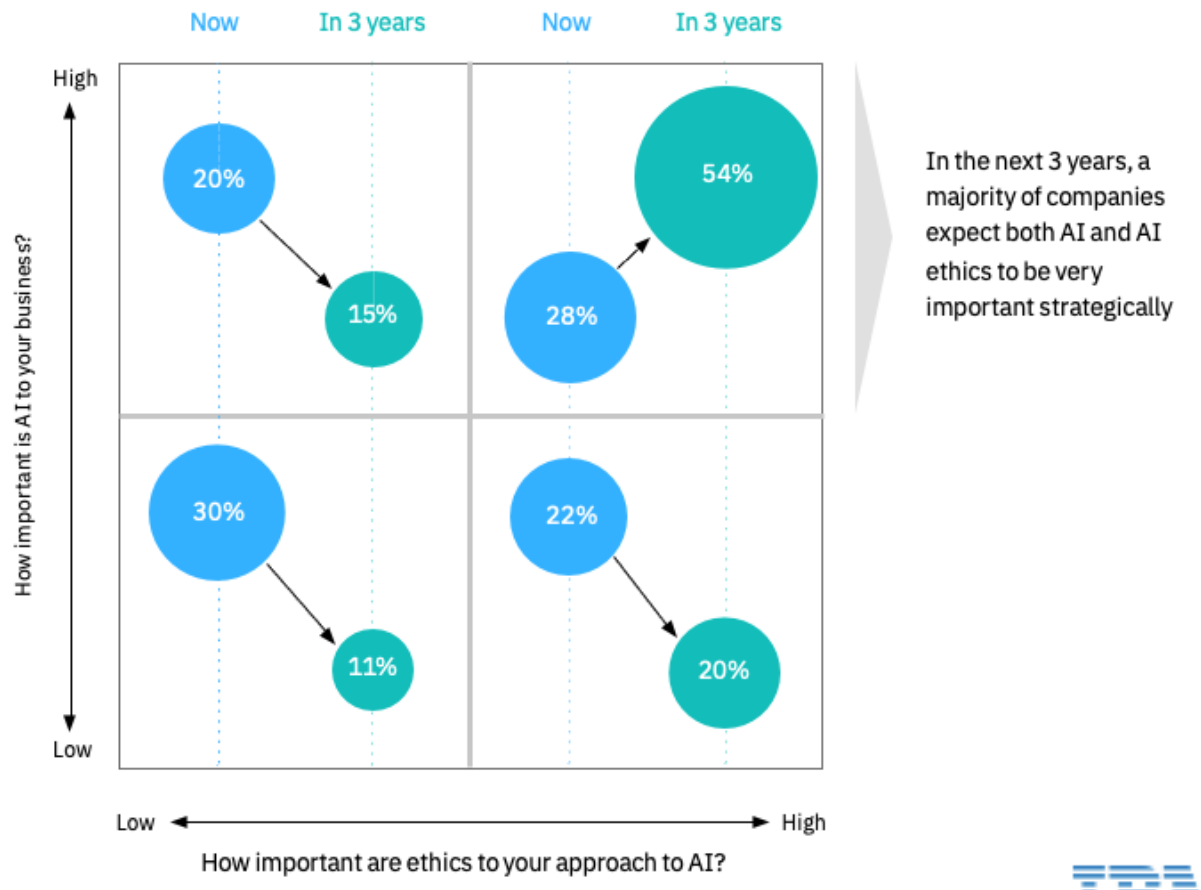
Still a lot of work to do in diversity and inclusion

Organizations' AI teams are significantly less diverse than their enterprise workforces



A promising trend

The majority of the organizations expect to increase the importance of AI and AI ethics in the next 3 years



AI Ethics at IBM: not just tools

Principles: augmentation, data, transparency	Trustworthy AI: fairness, transparency, robustness, explainability, privacy	Governance: the AI Ethics board
Use case risk assessment process	Education modules	Ethics by Design playbook
Adoption strategies	AI lifecycle governance	Team diversity
Multi-stakeholder consultations	Partnerships: academia, companies, civil society orgs, policy makers	Other emerging technologies: neurotech, quantum computing

- ✓ AI Factsheets 360
- ✓ AI Explainability 360
- ✓ AI Fairness 360
- ✓ Adversarial Robustness 360
- ✓ Uncertainty Quantification 360

IBM

AI Ethics at IBM: Principles and Pillars

Principles for Trust and Transparency

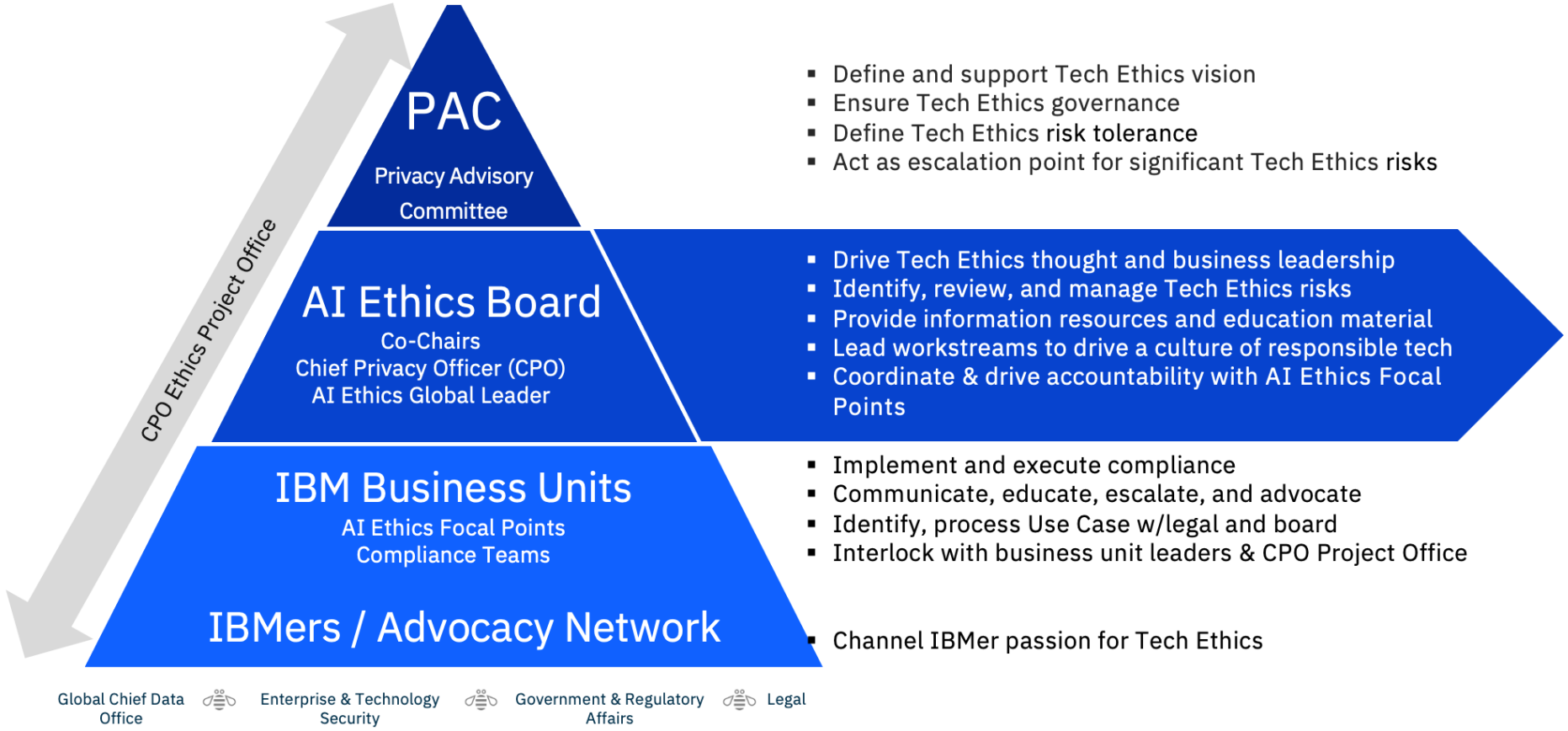
The purpose of AI is to augment human intelligence	Data and insights belong to their creator	New technology, including AI systems, must be transparent and explainable.

Trustworthy AI Pillars

Explainability	Fairness	Robustness	Transparency	Privacy
AI system's ability to provide a human-interpretable explanation for its predictions and insights.	Equitable treatment of individuals or groups of individuals by an AI system. Fairness for an AI system depends on the context in which it is used.	AI system's ability to handle exceptional conditions, such as abnormalities in input, effectively.	AI system's ability to include and share information on how it has been designed and developed.	AI system's ability to prioritize and safeguard consumers' privacy and data rights.

IBM

Governance structure



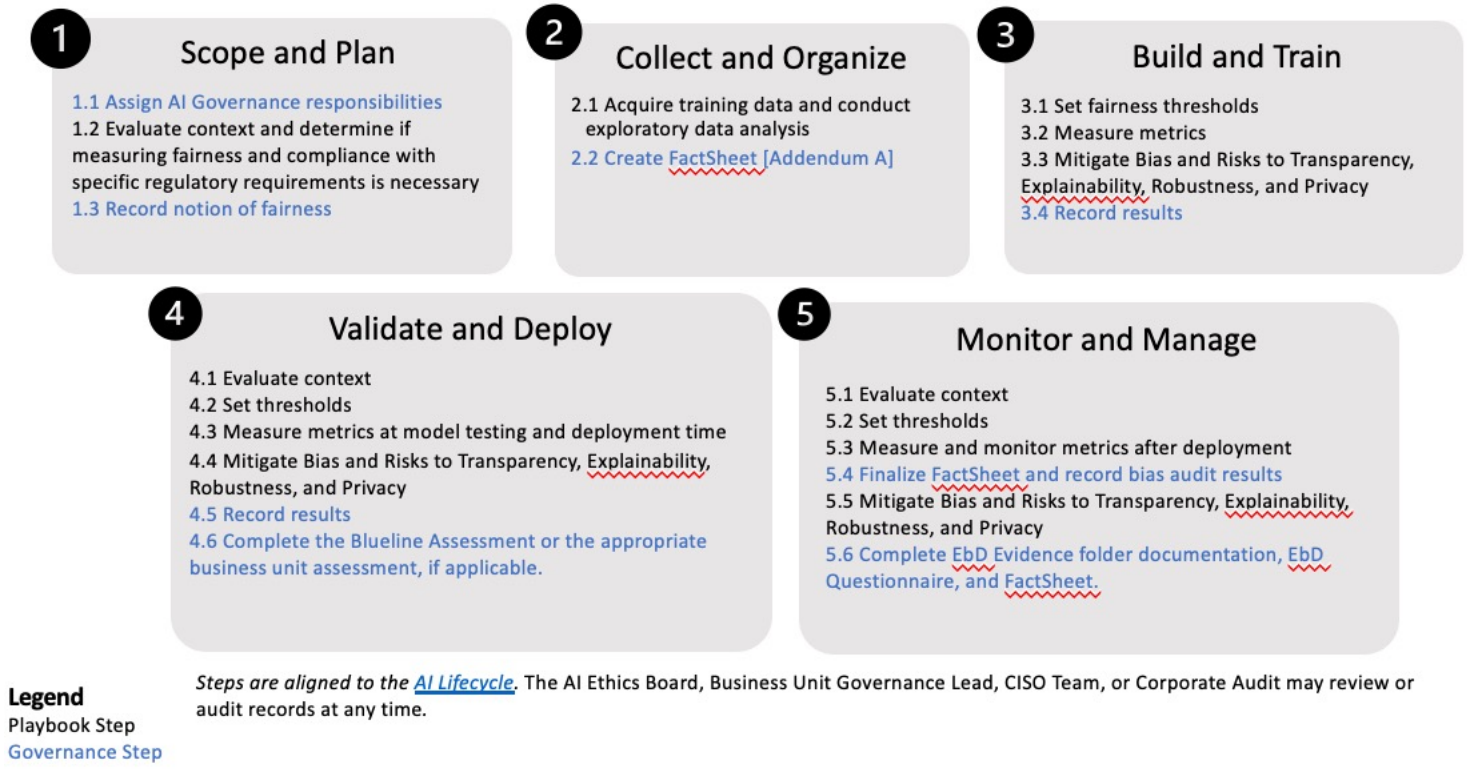
IBM AI Research

Neuro-symbolic AI	Machine learning combined with knowledge reasoning
Secure and Trusted AI	Fairness, explainability, robustness, transparency
AI engineering	Tools to simplify and automate key tasks in the AI pipeline
AI hardware	Energy-efficient hardware, quantum computing

IBM

Ethics by Design Playbook

- Specific guidelines, thresholds, goals, etc on how to build trustworthy AI
- Along the five pillars
- Integrated with security and privacy by design



Use Case Risk Assessment and Review Process

Education

Trustworthy AI and AI Ethics Foundations Badge

Earn this badge to learn how to think critically about AI ethics in your everyday work and how to help clients implement trustworthy AI.

START LEARNING

Ethics by Design Learning Plan

Ethics by Design (EbD) is a structured framework to fully integrate tech ethics in the technology development pipeline, including AI systems. Here, find resources to help you and your team adopt EbD.

START LEARNING

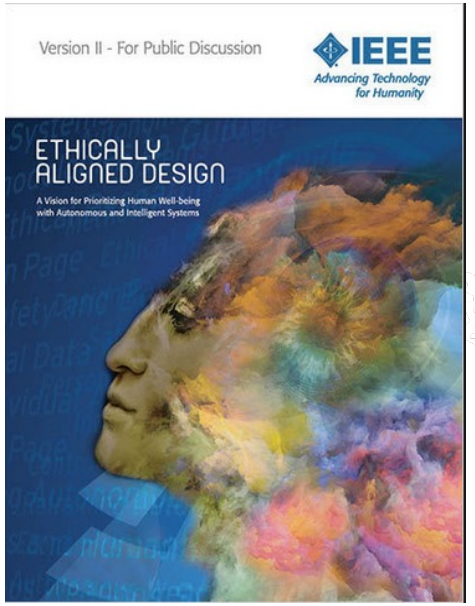
IBM

Partnerships
Multi-disciplinary and multi-stakeholder

IBM | MIT

AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY

AI for Good
Global Summit
An ITU experience



GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Rome Call for AI Ethics

IBM | Tech Ethics LAB | UNIVERSITY OF NOTRE DAME

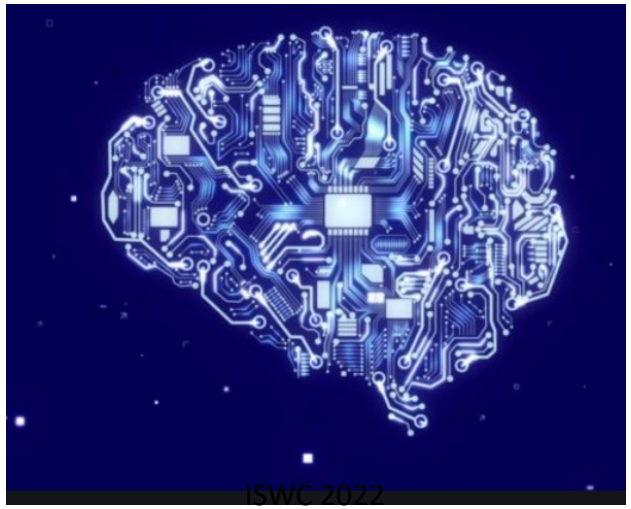
WORLD ECONOMIC FORUM

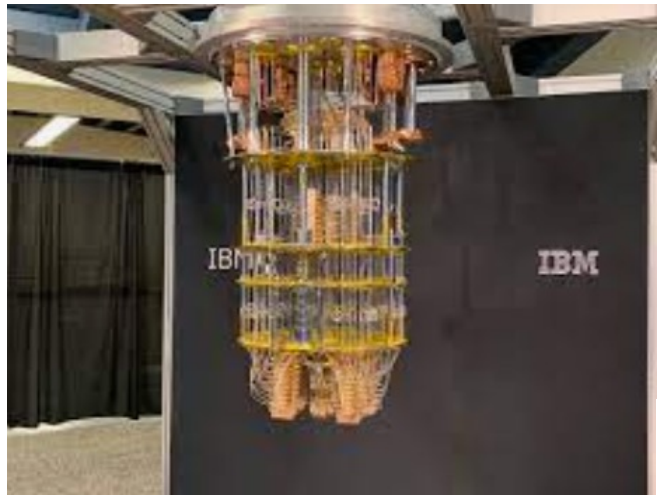
Not just AI

Addressing neuroethics issues in practice: Lessons learnt by tech companies in AI ethics, Neuron, 2022.

Artificial Intelligence and Neurotechnology: Learning from AI Ethics to Proactively Address an Expanded Ethics Landscape, to be published in CACM, Oct. 2022.

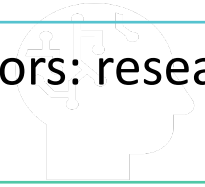
- Neurotechnologies
  - Huge potential for wellbeing
  - Reading/writing neurodata
  - Additional issues around mental privacy, human agency and identity
- Quantum computing
  - How to responsibly use such a huge computing power?





Lessons learnt in operationalizing AI ethics principles

Great progress: from awareness to principles to practice in few years



Complementary roles for different societal actors: researchers, companies, governments, civil society orgs

Multi-stakeholder, multi-culture, multi-disciplinary, and proactive approach

Operationalizing AI ethics in a company: company-wide approach, governance body, partnerships, beyond tech tools

Anticipate and inform regulations: beyond compliance

Expand to include other technologies and decentralized frameworks

Thanks!

IBM’s approach to AI Ethics