AI ethics:
from principles to practice

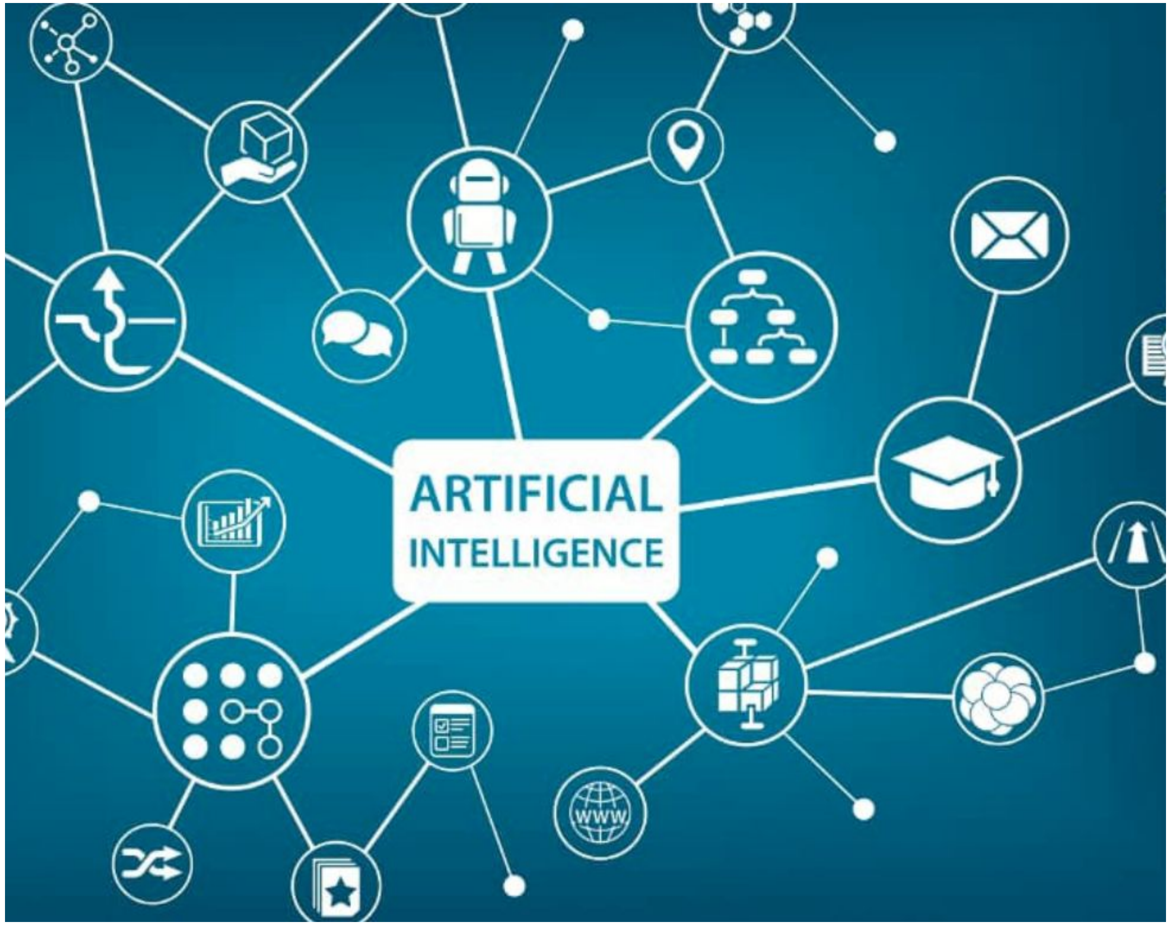
Francesca Rossi

IBM AI Ethics Global Leader

AAAI President

World Summit AI, October 12th, 2022

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

High-stakes decision-making applications

Credit

Employment

Admission

Healthcare

Enterprise Workflows

What can AI be useful for, in a company?

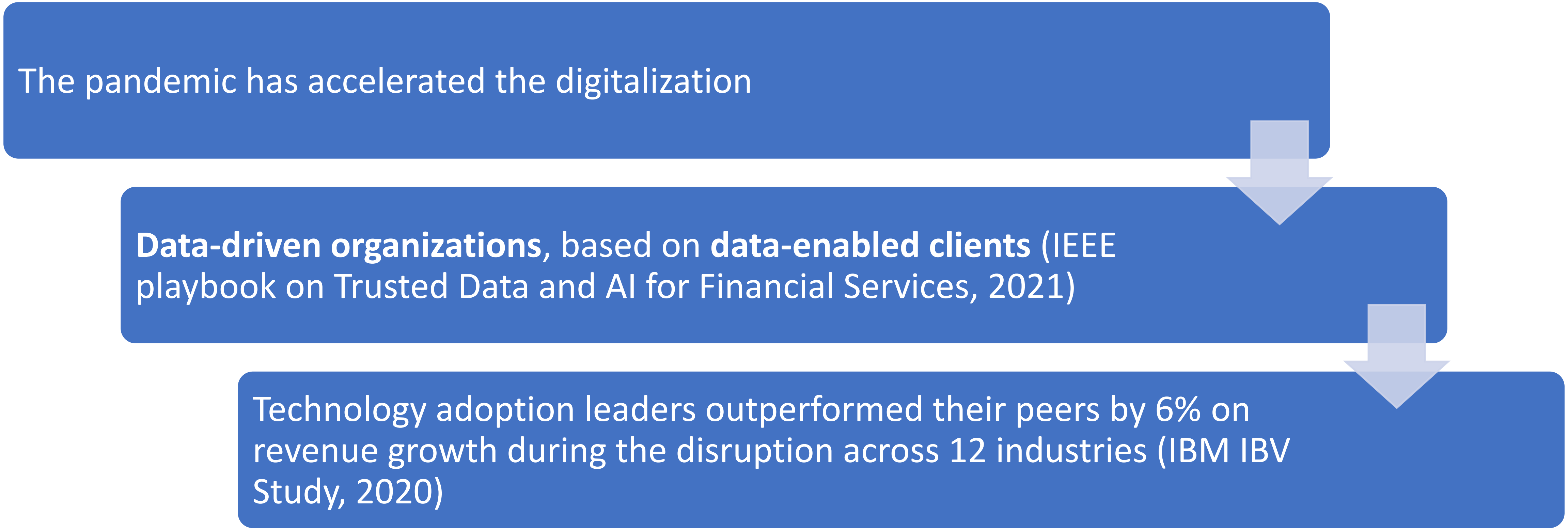
AI can help improve

- 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

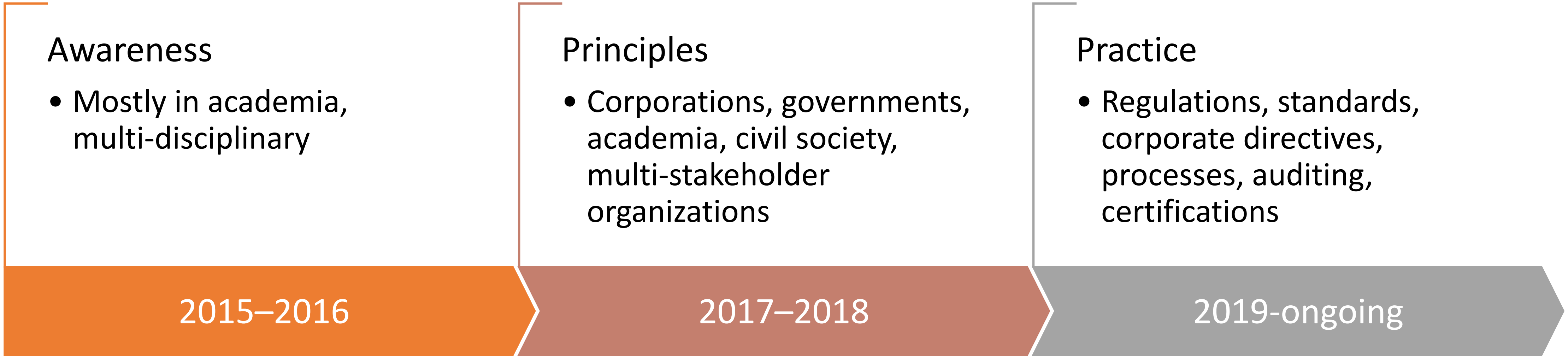
AI Ethics issues -1

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	Autonomous weapons and mass surveillance
	UN Sustainable Development Goals
Environmental impact	Foundation models need huge amounts of energy for training and deployment
Power imbalance	Centralization of data and power

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

AI Ethics in practice

Research

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

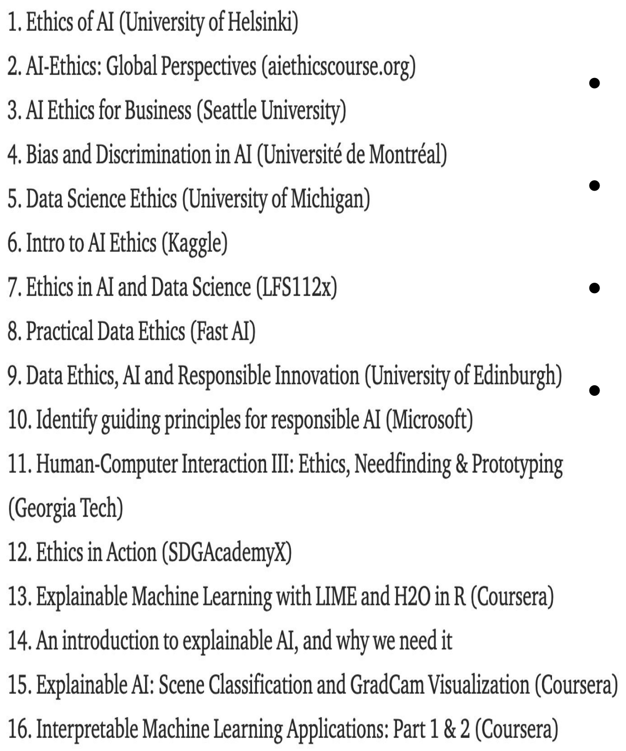
AI companies

- 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

AI Ethics in practice

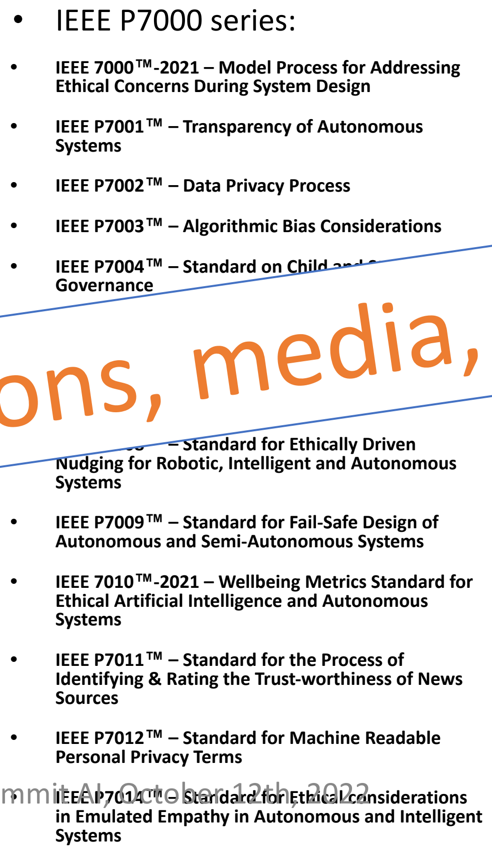
Research

- Fairness
- Explainability
- Interpretability
- Robustness
- Privacy

AI companies

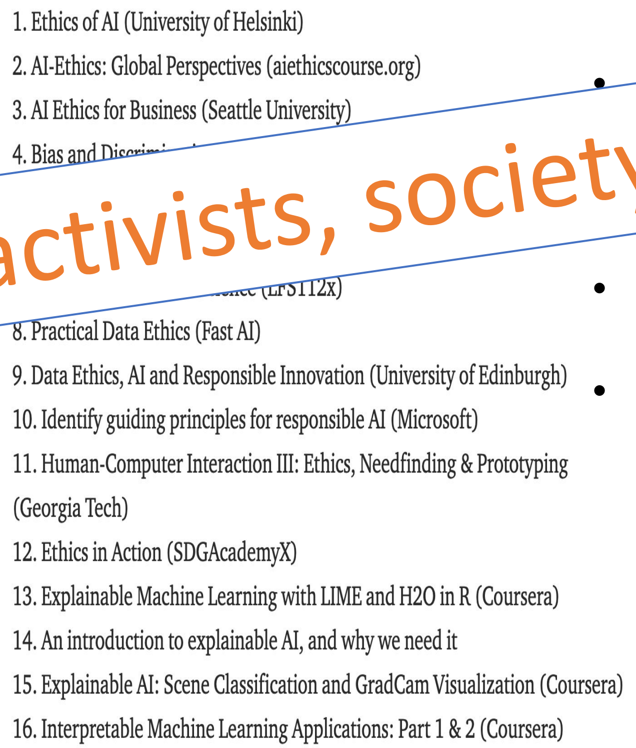
- Governance
- Internal processes
- Tools
- Risk assessment

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 Governance
- – 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 Discrimi

(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

- Ri

risk

- Focus on AI systems
- Obligations for high risk applications, providers and users

Civil society organizations, media, activists, society at large

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

The AI Ethics Holistic ROI

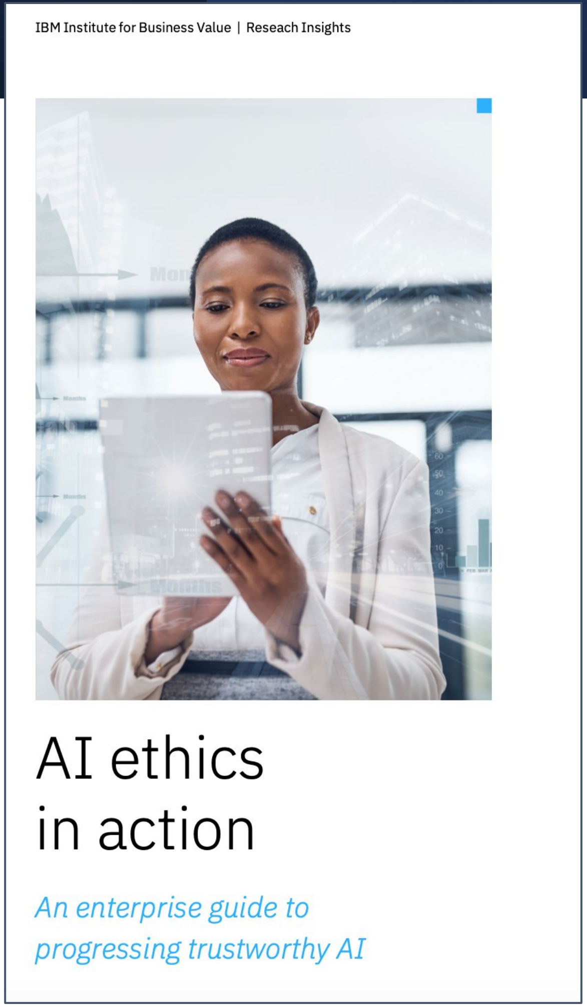
Why should a company building or using AI care about AI ethics?

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

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



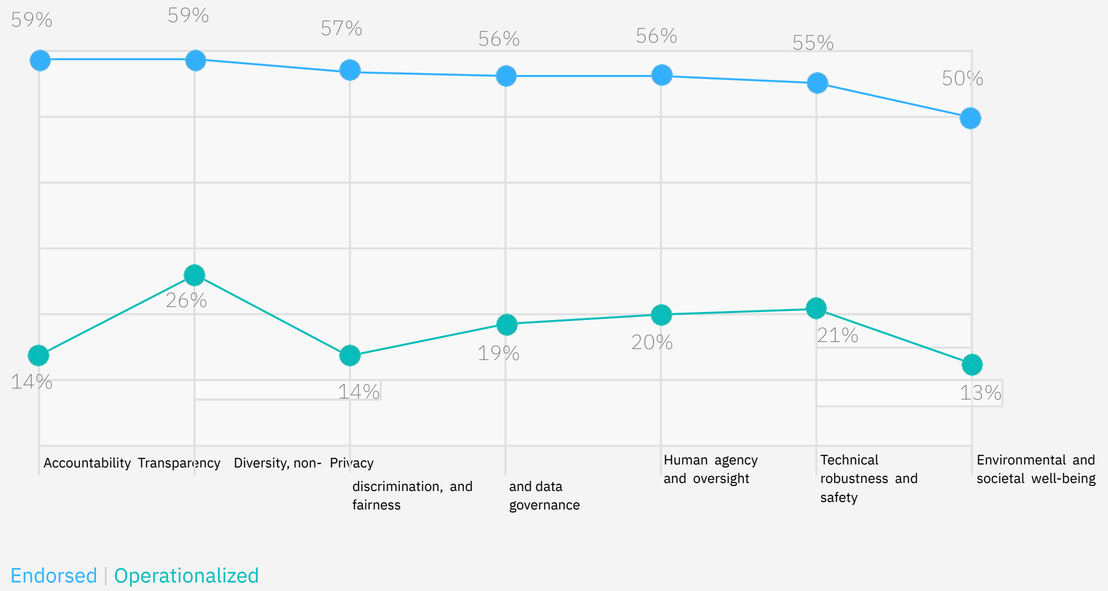
IBM Institute for Business Value | Reseach Insights

AI ethics
in action

An enterprise guide to progressing trustworthy AI

The intention-action gap

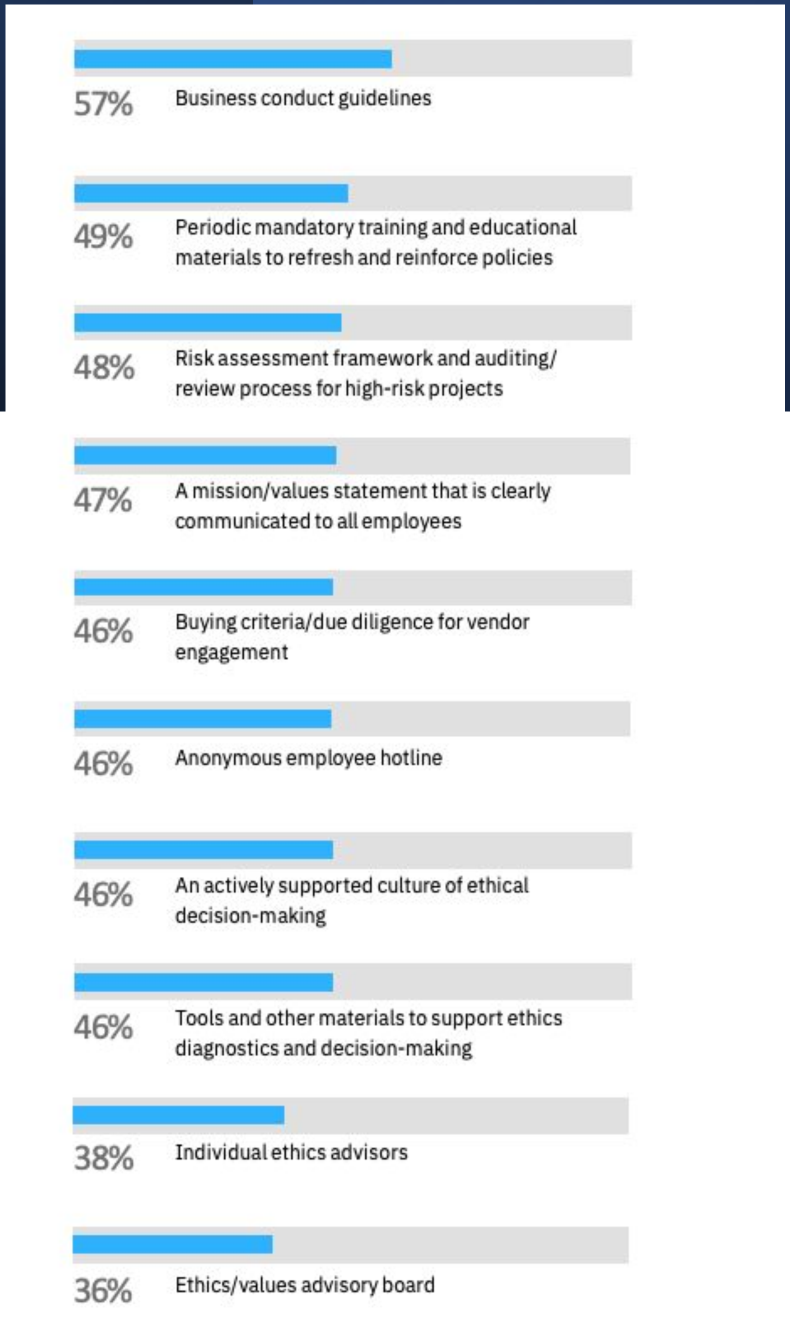
Organizations are endorsing AI ethics principles— but are still catching up on implementing them



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

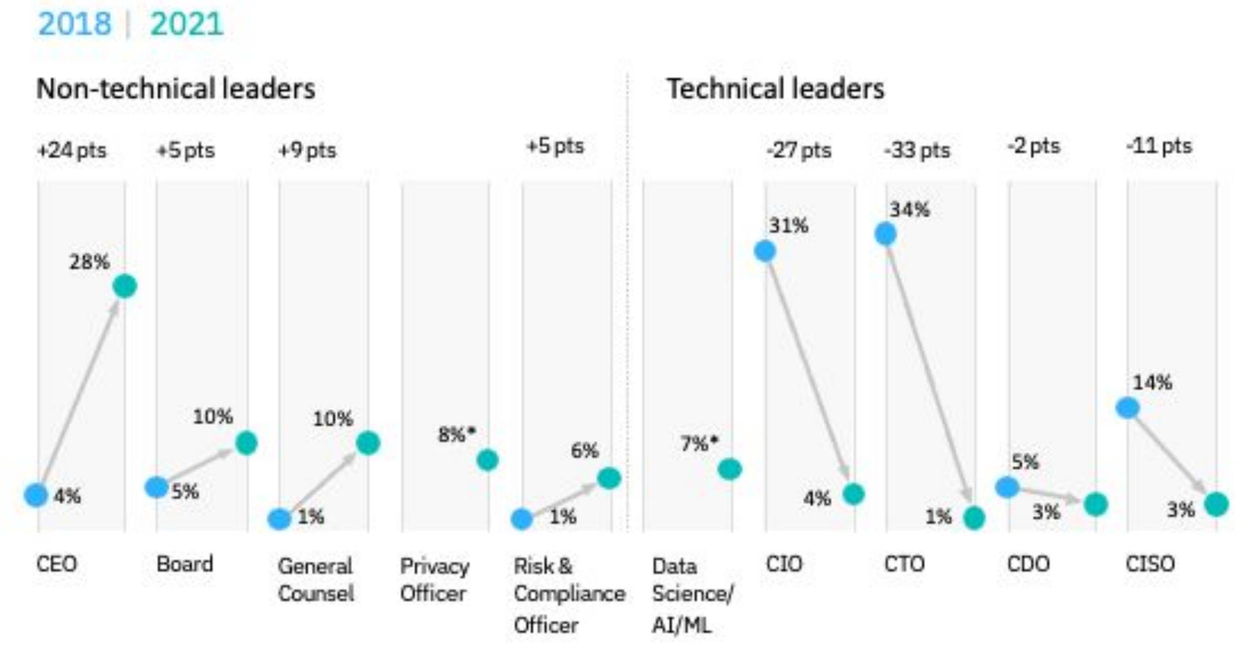
Many organizations are incorporating AI ethics into existing business ethics mechanisms



Not just technical issues

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

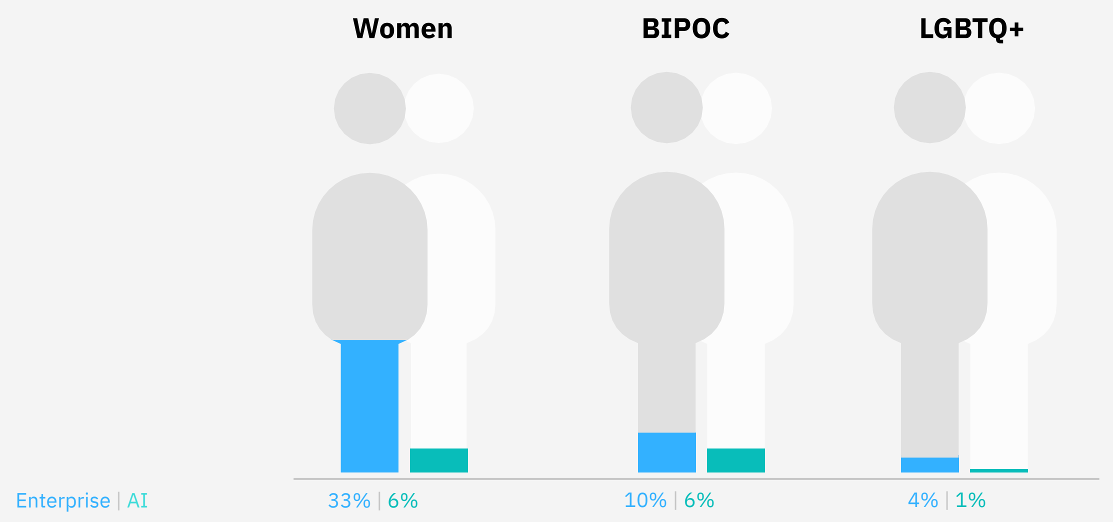
- 2018: IBV study on AI Ethics



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*

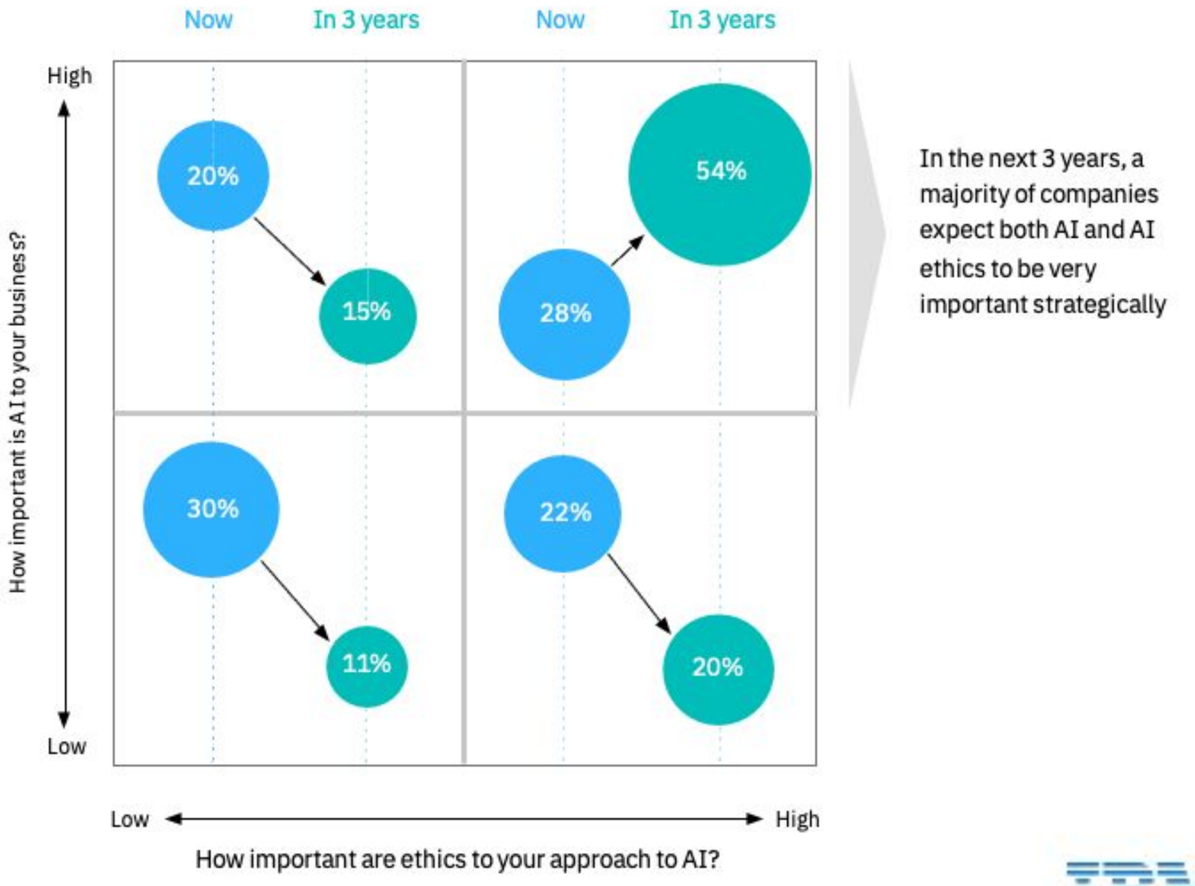
Still a lot of work to do in diversity and inclusion

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



A promising trend

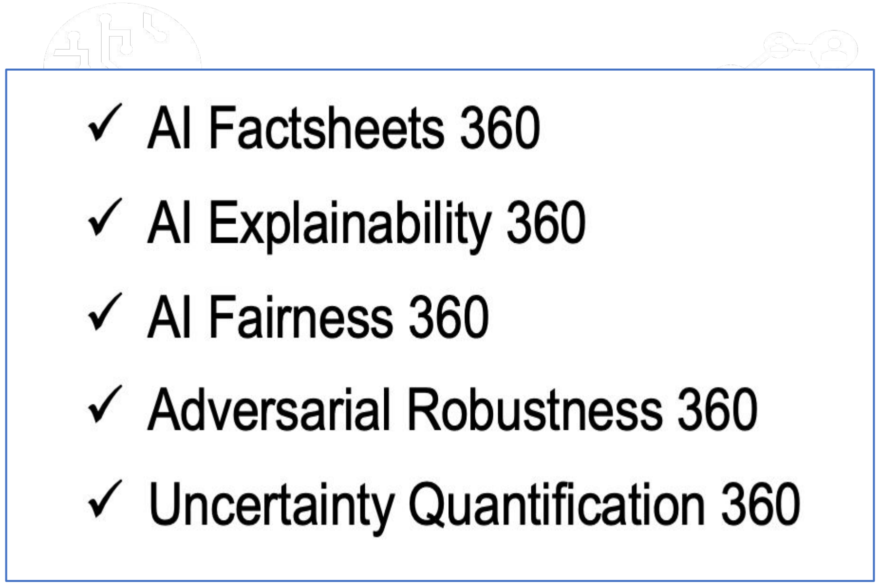
The majority of the organizations expect to increase the importance of AI and AI ethics in the next 3 years



In the next 3 years, a majority of companies expect both AI and AI ethics to be very important strategically

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

Lessons learnt in operationalizing AI ethics principles

Company-wide approach, not just a team

A governance body, with the power to make decisions for the company

Multi-stakeholder partnerships: to learn and to bring experiences/challenges

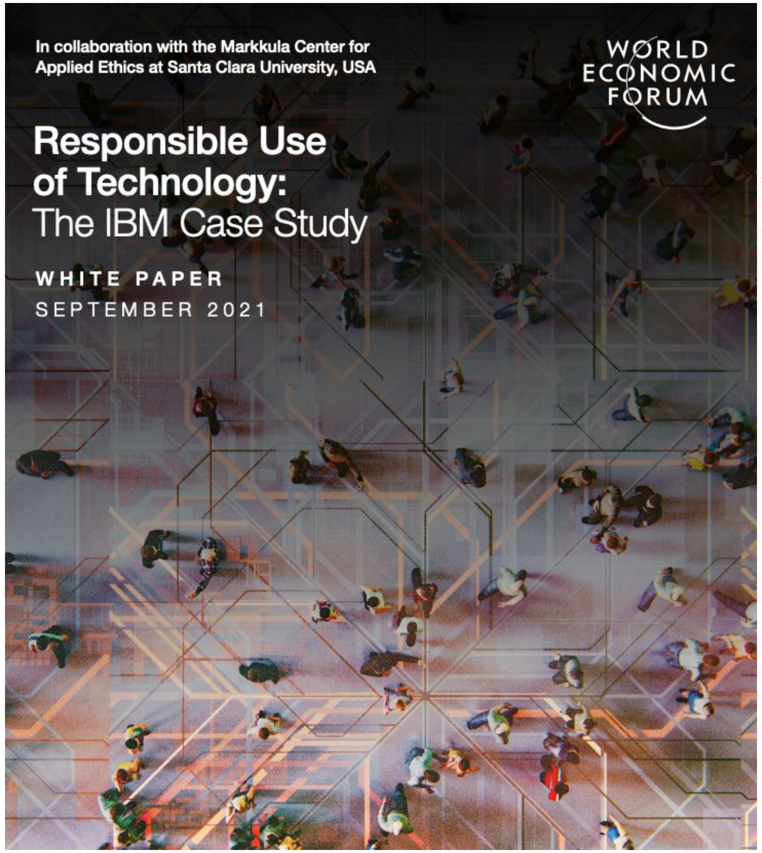
Full operationalization of the principles

Beyond technical tools: also processes, education, risk assessment, and governance

Regulations: beyond compliance

Thanks!

IBM’s approach to AI Ethics



In collaboration with the Markkula Center for Applied Ethics at Santa Clara University, USA

WORLD ECONOMIC FORUM

Responsible Use of Technology:
The IBM Case Study

WHITE PAPER
SEPTEMBER 2021