Operationalizing AI Ethics in the Modelling Process

Dr. Stephanie Kelley
June 1st, 2023

IVADO

What is AI ethics?

AI Ethics (in Organizations)

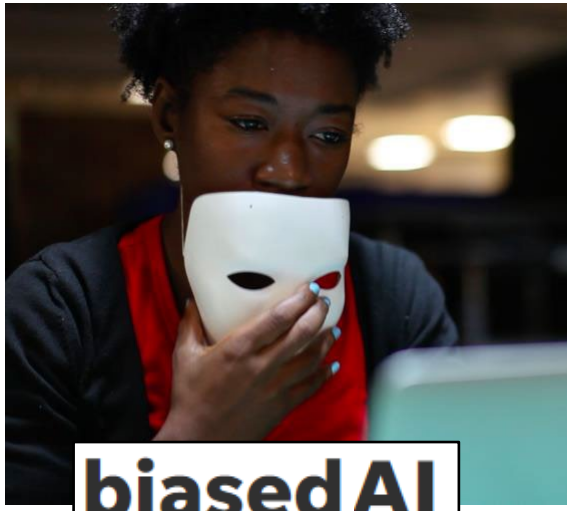
Reports and subsequent investigations of the unethical use of AI in organizations

Apple Card Investigated After Gender Discrimination Complaints

spying on customers' social media

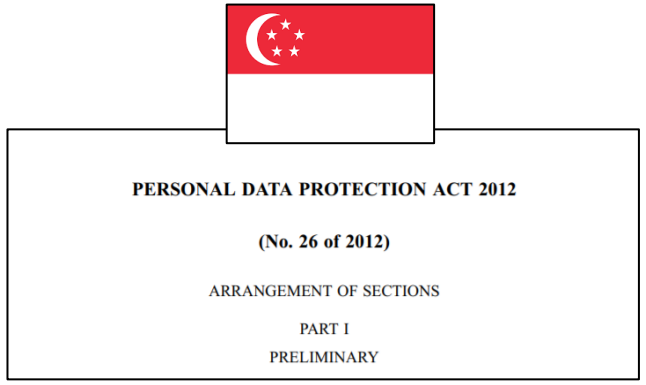
Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

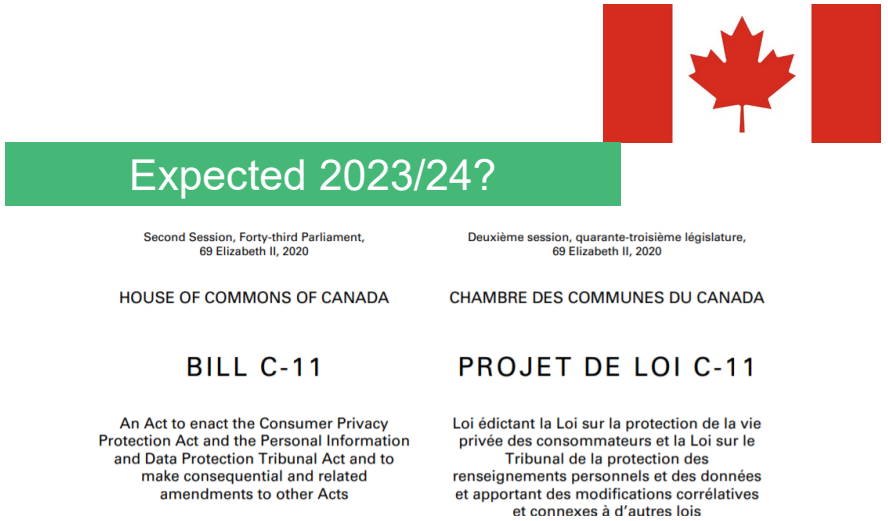
UK class action-style suit filed over DeepMind NHS health data scandal



Laws are slow to catch up to AI development

Expected Q3 2023



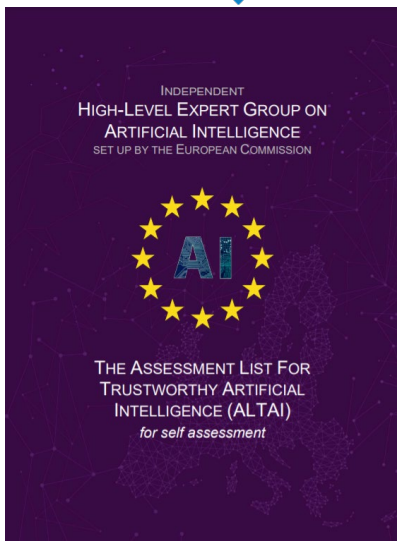


Expected 2023/24?

Proposed October 2022

AI Ethics Principles

- Help fill the gap between existing laws and current uses of AI

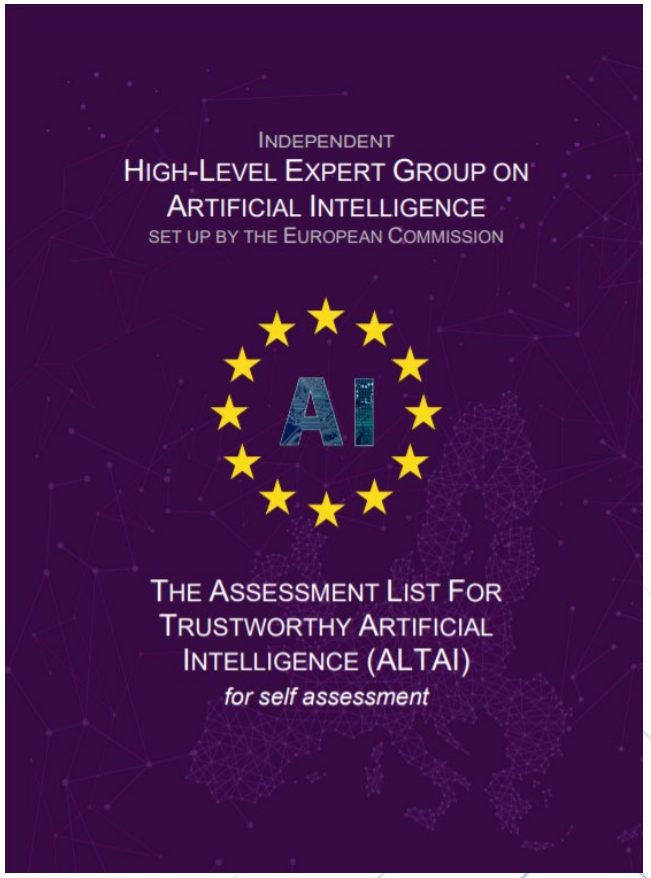


Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector

Microsoft AI principles

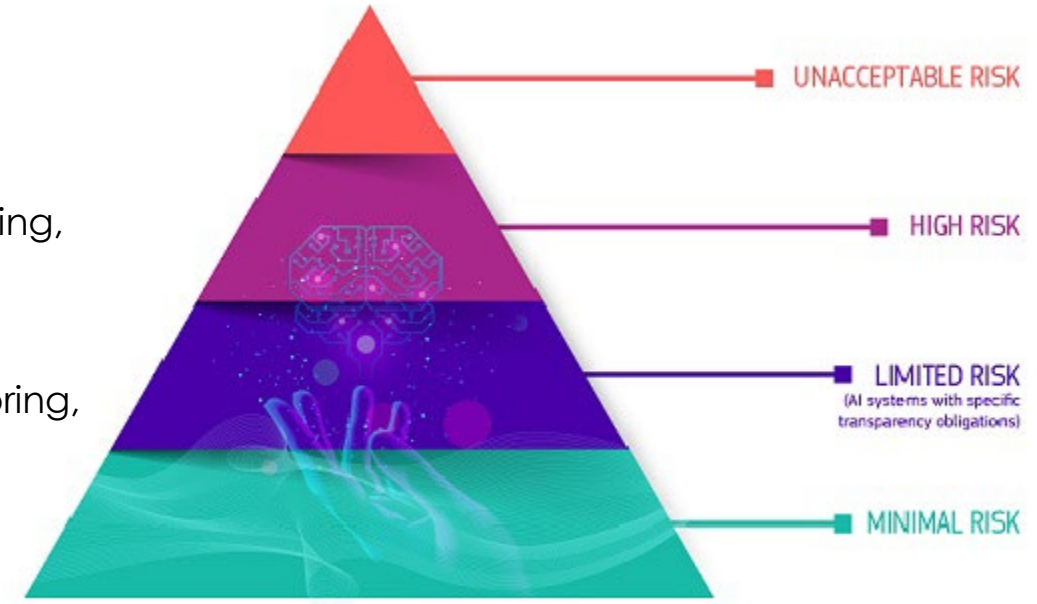
EU Assessment List for Trustworthy AI (ALTAI)

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency (explainability)
- Diversity, non-discrimination and fairness
- Environmental and societal well-being
- Accountability



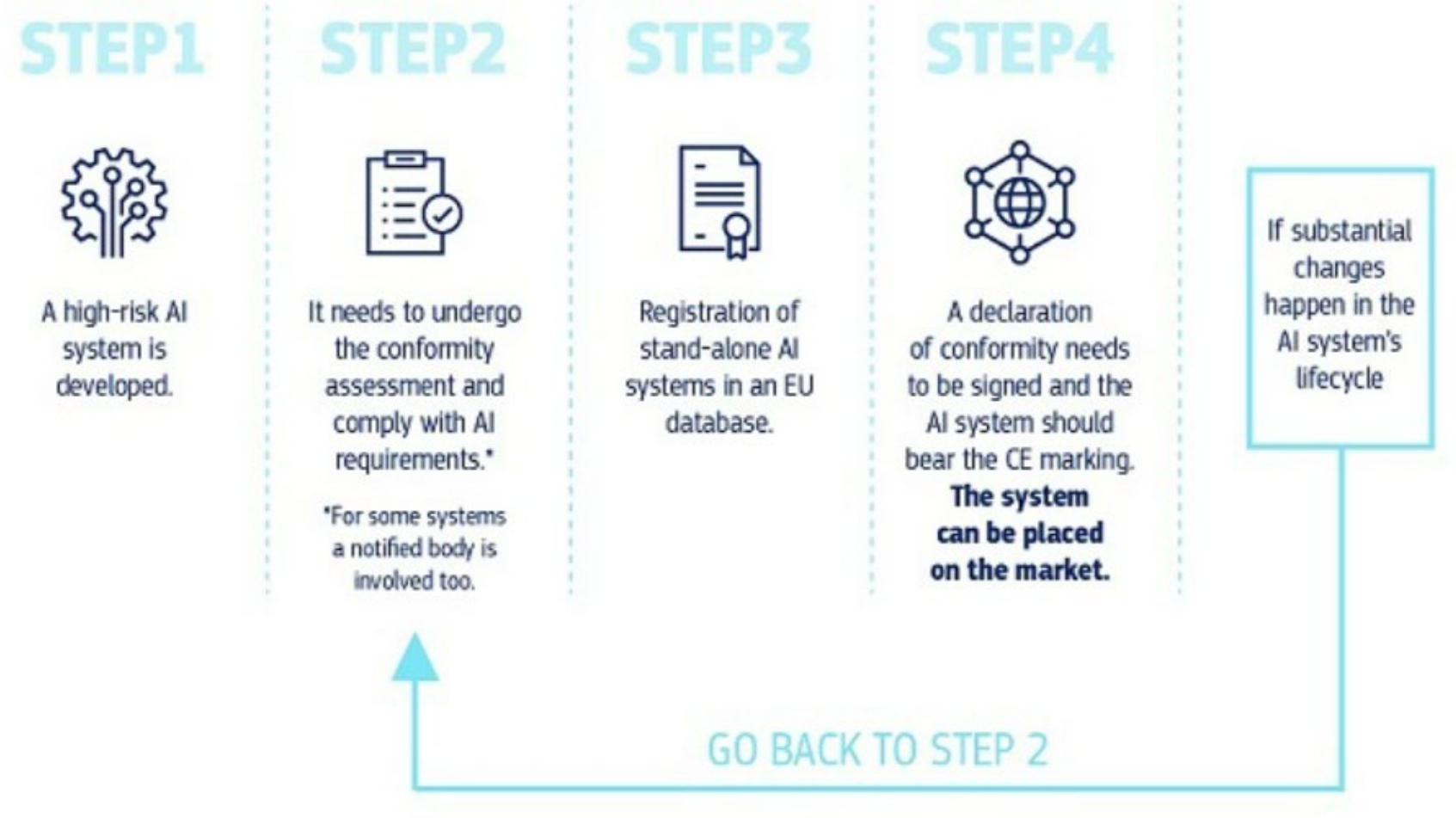
Deep Dive: *Proposed* EU Artificial Intelligence Act

- Values-based: ALTAI Principles
- Integrated with GDPR
- Risk-based approach
  - High-risk AI systems are held to strict requirements with respect to:
    - training data, data governance and explainability, reporting, robustness and accuracy, and human oversight
    - any kind of "safety component", biometric identification/categorization, access to employment/goods/services, law enforcement, social scoring, migration/asylum/border control, justice and democratic process
- Low-risk AI systems are not regulated; voluntary self-regulation
- New regulatory framework: European AI Board
  - Overseeing and co-ordinating enforcement
  - Penalties (per GDPR) for non-compliance



Risk-based Approach (EU regulation example)

*can also be applied to principle implementation

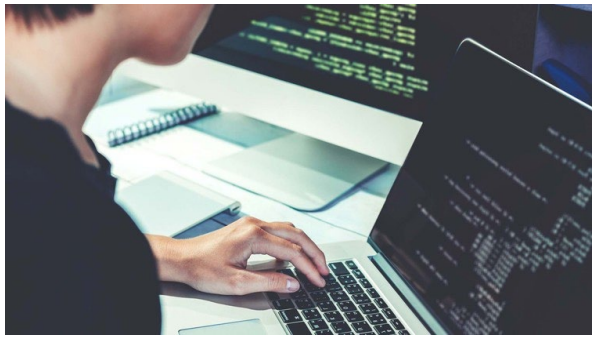


Human agency and oversight

Levels of Autonomy

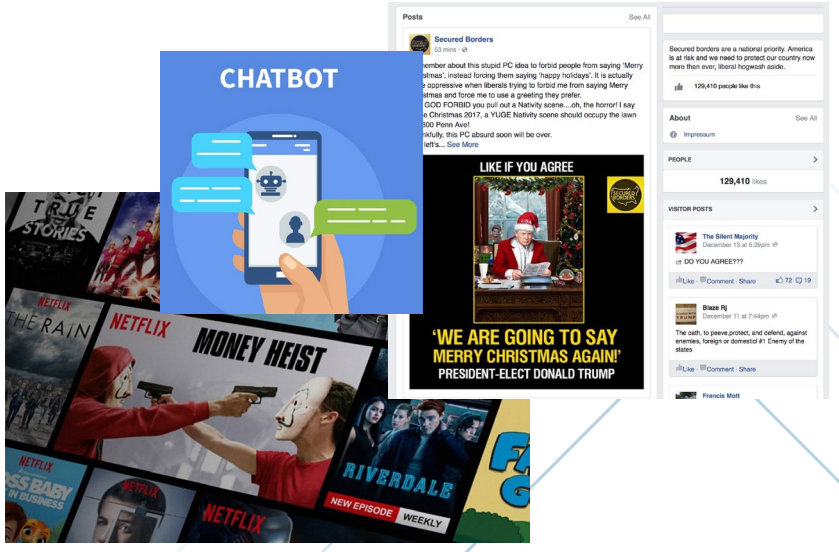
Human-controlled → **Semi-autonomous** → **Fully autonomous**

AI requires real-time human input to perform a task



Some tasks can be performed without input, but not all

AI performs tasks with no real-time human input

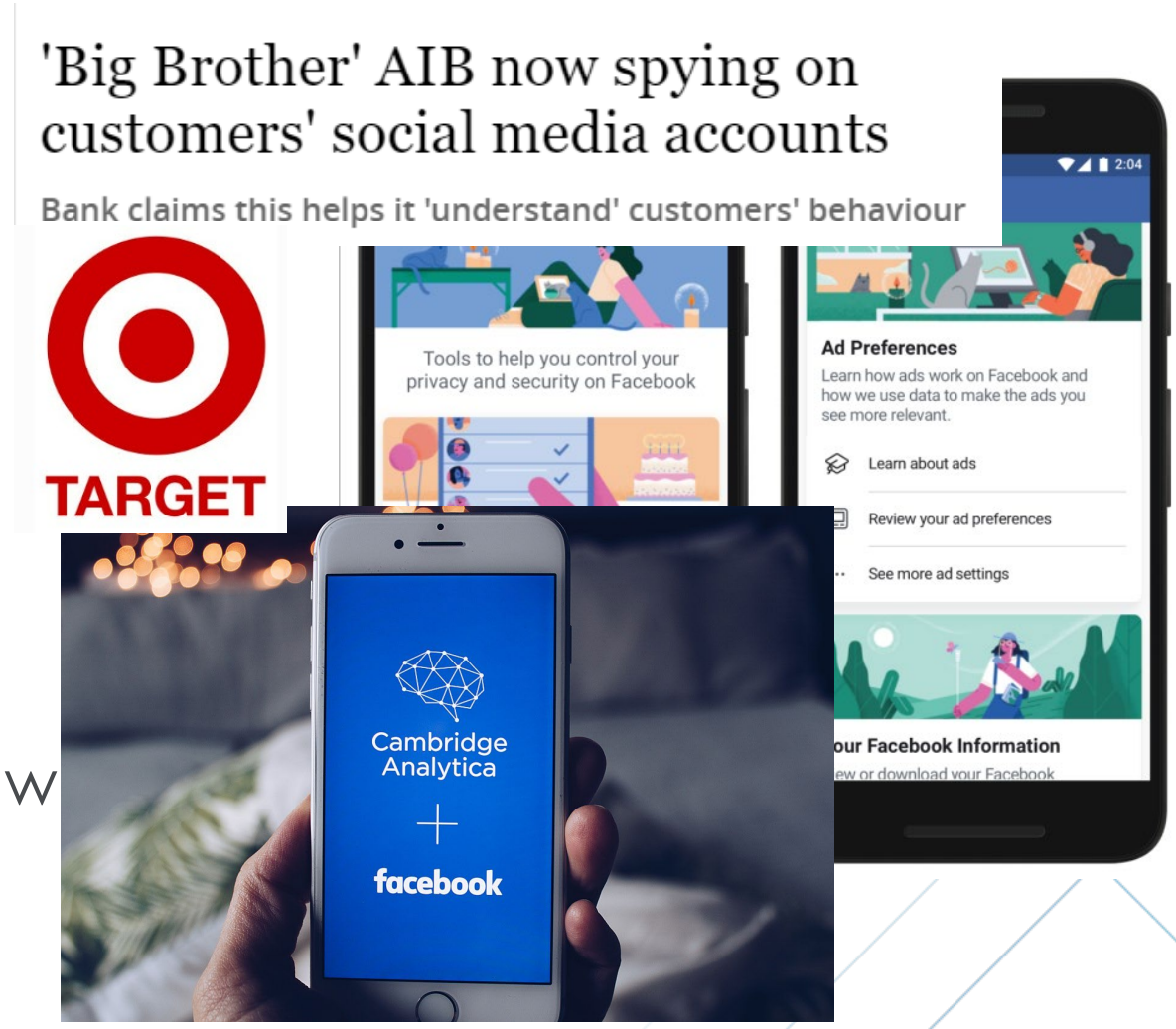


Technical robustness and safety

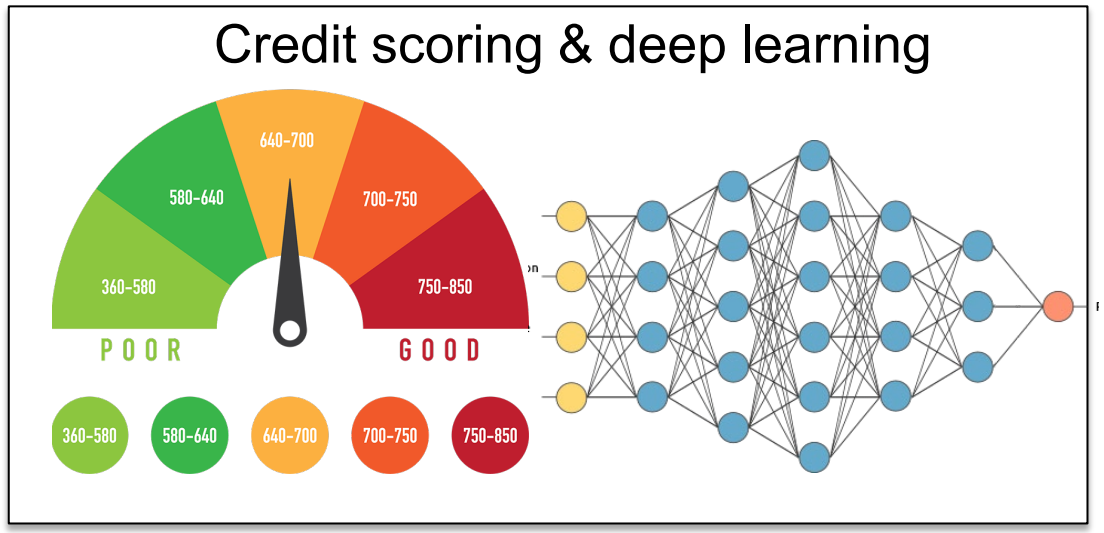
- Resilience to attack
- Security
- General safety
- Accuracy
- Reliability, fall-back plans, and reproducibility

Privacy and data governance

- Consent
- Control over the Use of Data
- Ability to Restrict Processing
- Right to rectification
- Right to erasure
- Privacy by design
- Recommends data protection law



Transparency (explainability)



- Traceability
- Explainability
- Communication

Google AI

"Hi, I'm calling to make a reservation for a client. I'm calling from Google, so the call may be recorded."

Diversity, non-discrimination, and fairness

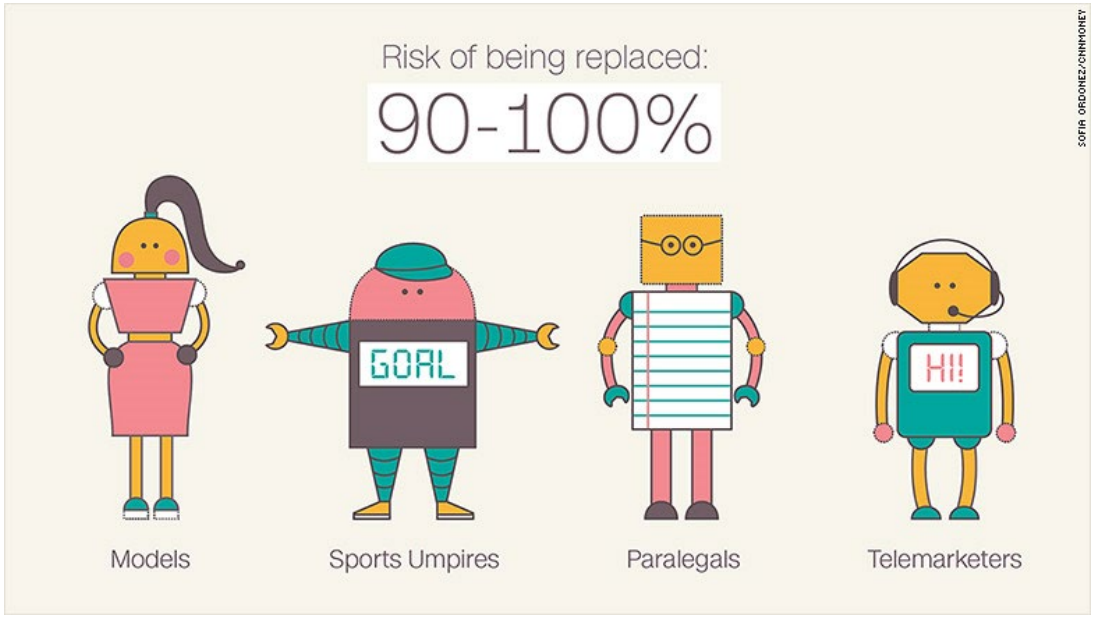
Apple Card algorithm sparks gender bias allegations against Goldman Sachs

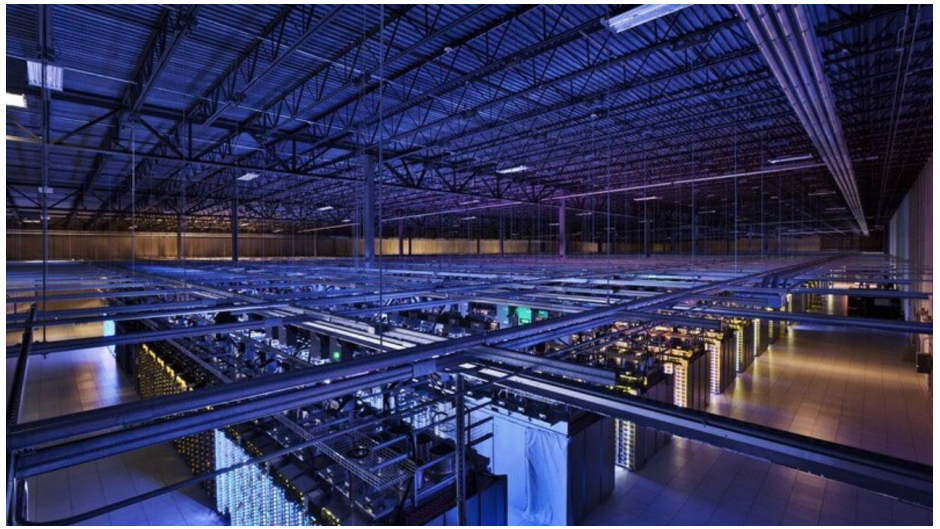
Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the hig

- Reduction of bias/discrimination
- Accessibility and universal design
- Stakeholder participation

How Amazon Accidentally Invented a Sexist Hiring Algorithm

Environmental and societal well-being





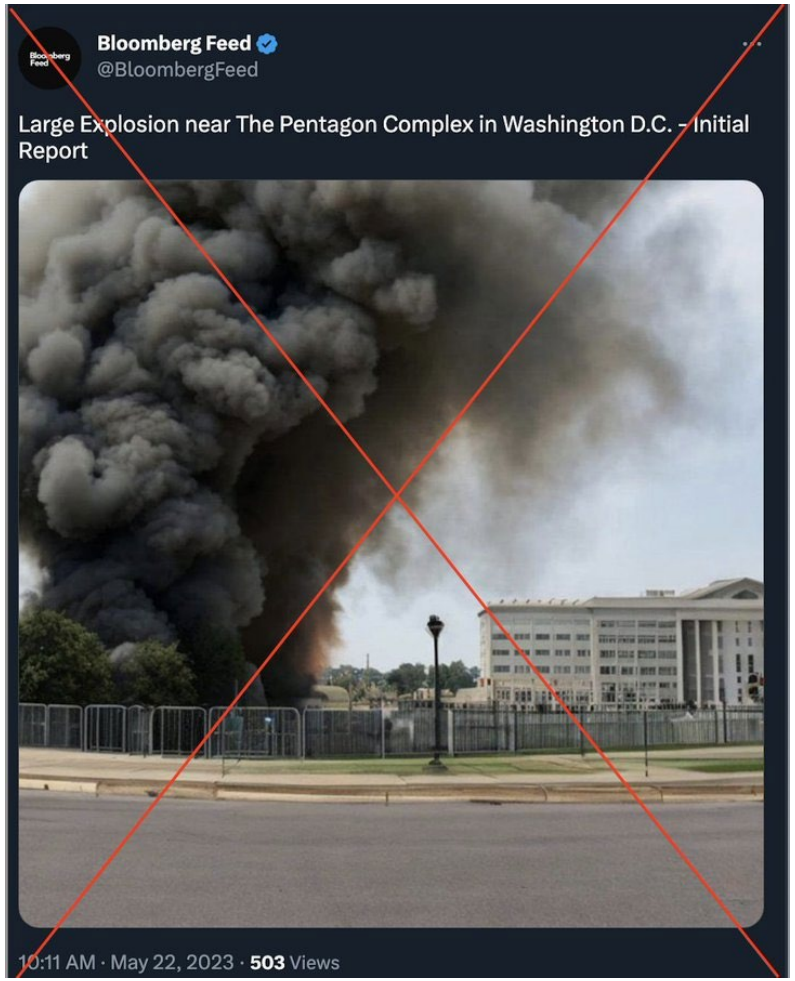
- Environment
- Impact on work & skills
- Impact on society & democracy

MANAGEMENT & CAREERS

Alexa: Don't Let My 2-Year-Old Talk to You That Way

Children are using voice-activated technology at far younger ages than other devices, but some parents worry they're picking up rude habits

Accountability



Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.

How do you operationalize AI ethics principles?

Operationalizing AI Ethics (principles)

Principles

A **document** that states **normative principles** about what employees **should** and **should not do** when it comes to using AI.

Process Governance Framework

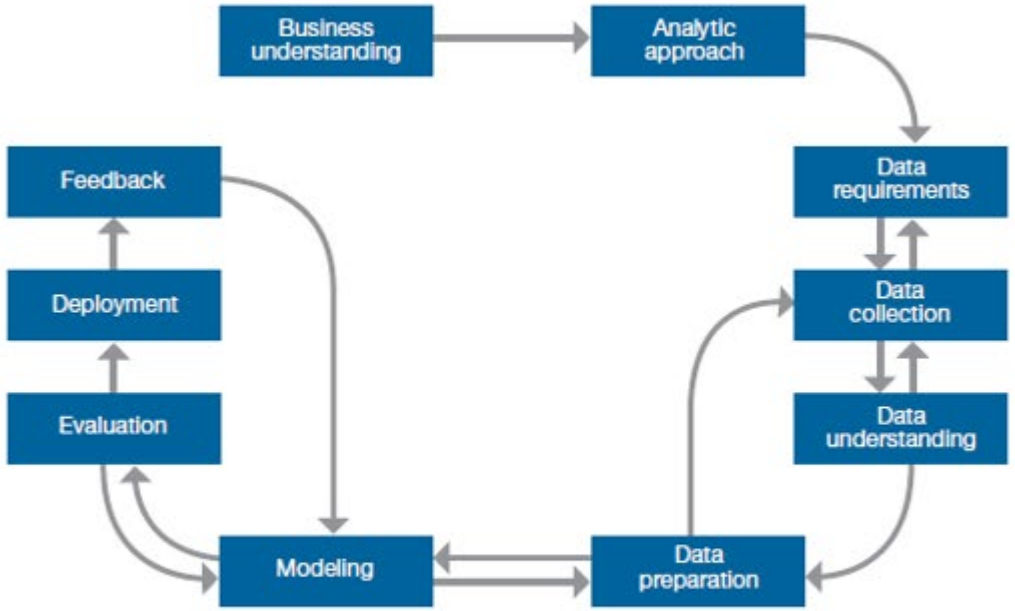
A **standardized set of activities** that occur during the modelling process - used for every AI projects, including developing, testing, launching, monitoring, and auditing AI systems.

Technical Implementation

Specific **technical tools** that help achieve activities set up in the process governance framework and principles.

Examples of existing modelling frameworks

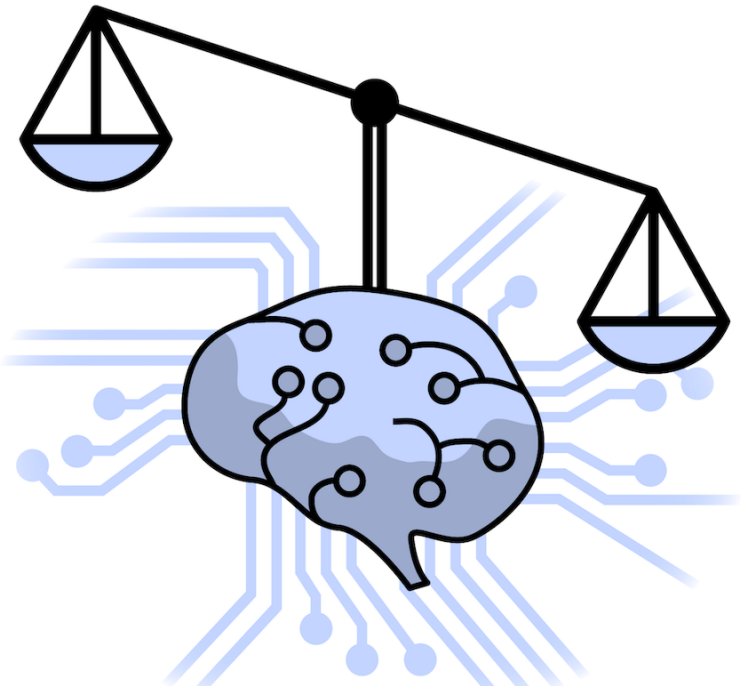
Cross Industry Standard Process for Data Mining



IBM Data Science Methodology

Ethics is not "one size fits all"

- This make operationalizing principles challenging
- AI ethics principles ≠ ethical AI outcomes
- What is "fair"?
- What is "explainable"?
- What is "clear responsibility"?
- ….



Operationalizing AI Ethics (Principles)

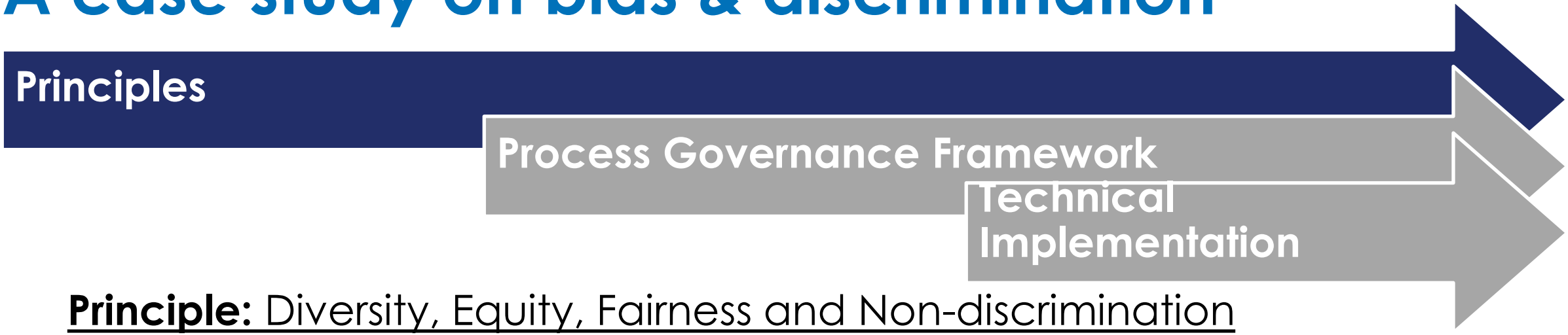
Principles

Process Governance Framework

Technical Implementation

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency (explainability)
- Diversity, non-discrimination and fairness
- Environmental and societal well-being
- Accountability

Operationalizing AI Ethics
A case study on bias & discrimination

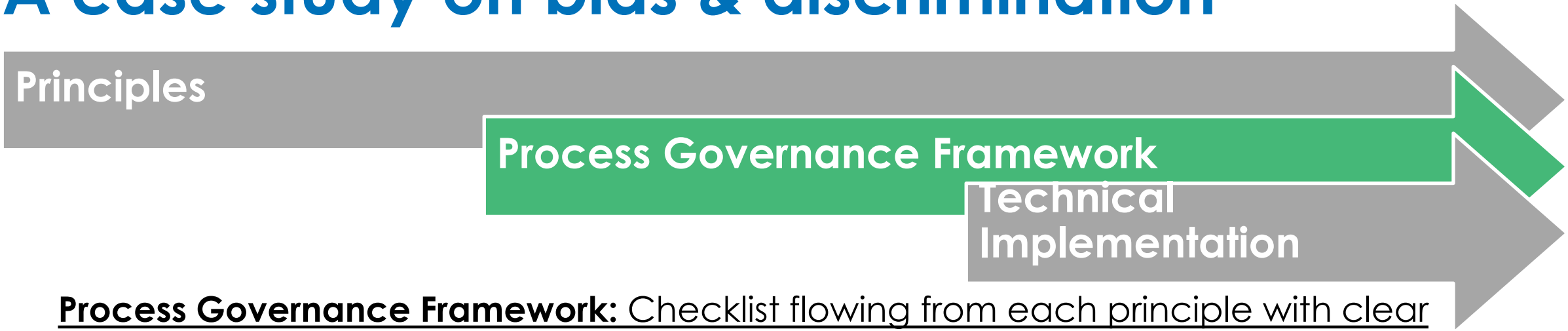


Principle: Diversity, Equity, Fairness and Non-discrimination

Subprinciples

- ***AI systems*** should respect diversity, equity, fairness, and non-discrimination.
- ***AI actors*** should be concerned with prohibited discrimination, as defined by the Canadian Human Rights Act…
- To this end, ***AI actors*** should design ***AI systems*** with diversity, equity, fairness, and non-discrimination in mind.

Operationalizing AI Ethics
A case study on bias & discrimination



Process Governance Framework: Checklist flowing from each principle with clear RACI, integrated into existing data science process.

Example questions from **bias & discrimination section** of the **checklist**

- How will you include a representative group of potential stakeholders in the design and testing processes…?
- How will you provide diverse training data that is representative of potential stakeholders?
- How will you design for adaptive needs, including mobility and different movement modalities?

Operationalizing AI Ethics
A case study on bias & discrimination

Principles

Process Governance Framework

Technical Implementation

Fairness / unjust bias assessment

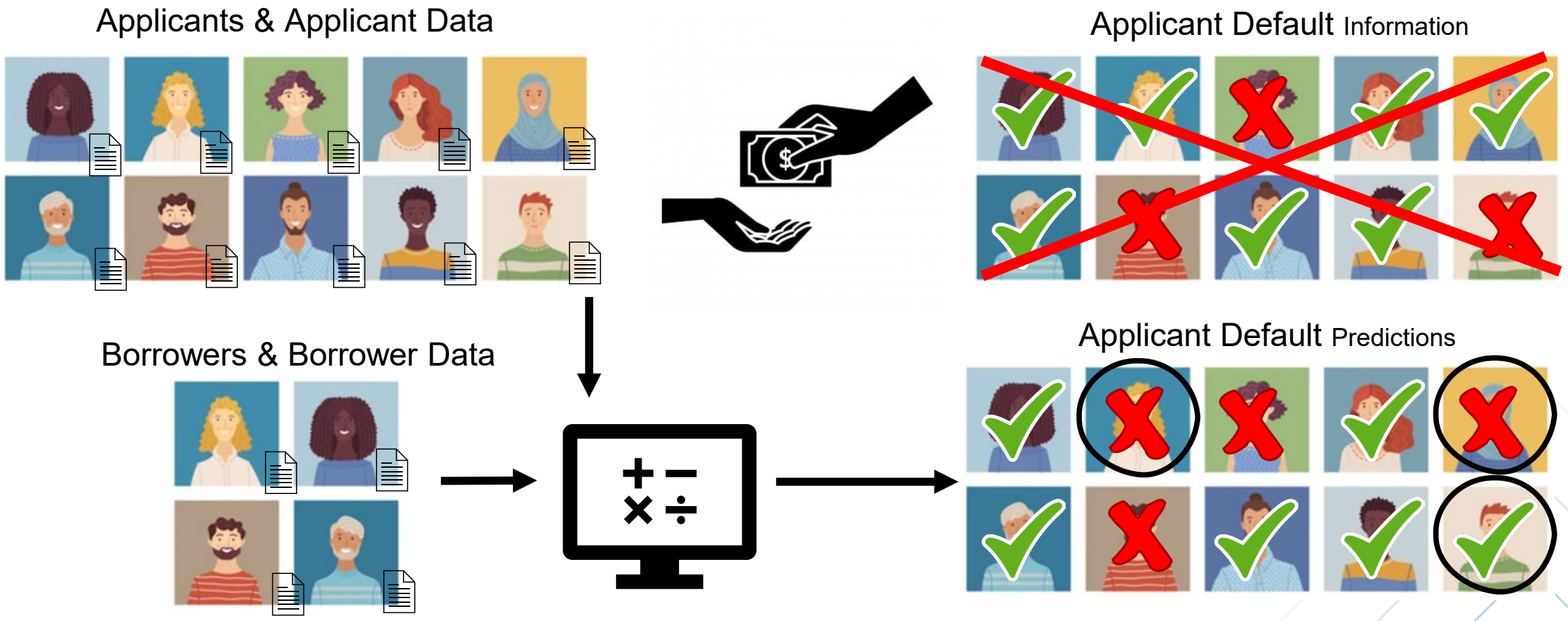
Technical Implementation:

- Example – signed contract with AI ethics SaaS company
- Plug-in for existing model development systems to measure principles
- In-house build *most common today

Credit Lending Case Study: Technical Implementation

Fairness Measurement in Practice

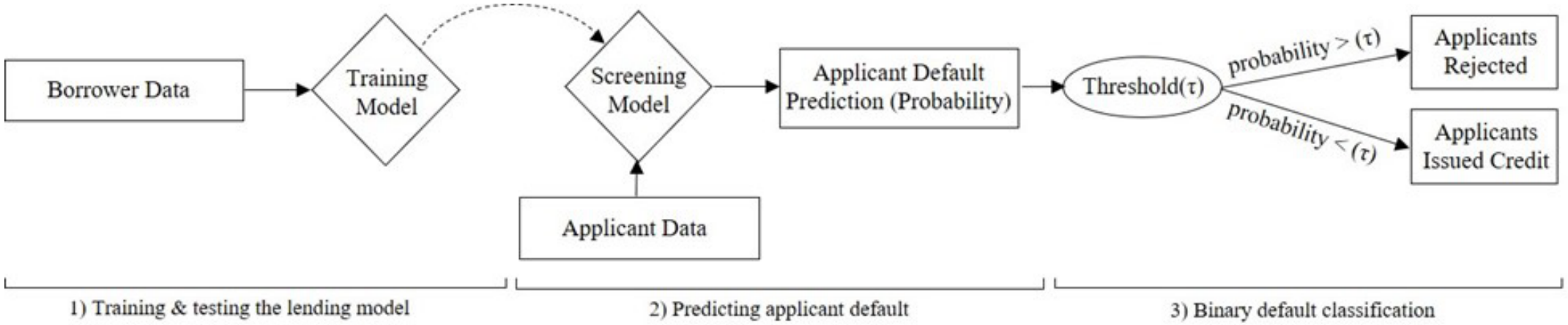
Operationalizing one AI ethics principle in the fintech lending setting



Fairness Measurement in Practice

Using the CRISP-DM framework (as an example modelling process)

1) **Business understanding:** What do we need to predict? How does this integrate into the business process? How high risk is this application (Financial? Reputational? Operational?)



Fairness Measurement in Practice

2) **Data Understanding:** What data do we have? Therefore, what fairness measures can we possibly use? What data might we need?

Discrimination Measures	Reference	Adjusts for unequal base default rates	Calculated without external risk score	Calculated with a known default outcome label
Classification Parity				
Statistical Parity	Berk et al. (2017)		X	X
Equalized Odds	Hardt et al. (2016)		X	X
Treatment Equality	Berk et al. (2017)		X	X
Balance for the Positive Class	Kleinberg et al. (2017)		X	X
Positive Predictive Value	Choudlechova (2017)	X	X	X
Mean Difference	Žliobaitė (2017)		X	
Within Group Mean Difference	Adapted from Žliobaitė (2017)	X	X	
Calibration	Corbett-Davies and Goel (2018)	X		
Anti-classification	Grgic-Hlaca et al. (2016)		X	

Fairness Measurement in Practice

3) **Data Preparation:** cleaning, exploratory data analysis (EDA), iteration, and documentation (data cards)

Use in ML or AI Systems

DATASET USE(S)	NOTABLE FEATURE(S)	USAGE GUIDELINE(S)
*Select **all applicable**:*	*Describe any notable feature distributions or relationships between individual instances made explicit.* *Include links to servers where readers can explore the data on their own.*	*Summarize usage guidelines or policies that consumers should be aware of.* *Use additional notes to capture any other relevant information or considerations.*
Training Testing Validation Development or Production Use Fine Tuning Others (Please Specify)	**Exploration Demo:** [Link to server or demo.] **<Notable Field Name>:** <Describe here. Include links, data examples, metrics, visualizations where relevant.> **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**Usage Guidelines:** <Summarize here. Include links where necessary.> **Approval Steps:** <Summarize here. Include links where necessary.> **Reviewer:** <Provide the name of a reviewer for publications referencing this dataset.> **Additional Notes:** <Add here>
	DISTRIBUTION(S)	**KNOWN CORRELATION(S)**
	Describe the recommended splits and corresponding criteria. *Use additional notes to capture any other relevant information or considerations.*	*Summarize any known correlations with the indicated features in this dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate for each known correlation.)*

Train	62,563	`field_name`,`field_name`
Test	62,563	
Validation	62,563	**Description:** <Summarize here. Include visualizations, metrics, or links where necessary.>
Dev	62,563	
Above: <Provide a caption for the above table or visualization.>		**Impact on dataset use:** <Summarize here. Include visualizations, metrics, or links where necessary.>
Additional Notes: <Add here>		**Risks from correlation:** <Summarize here. Include recommended mitigative steps if available.>
		Additional Notes: <Add here>

SPLIT STATISTICS

Provide the sizes of each split. As appropriate, provide any descriptive statistics for features.

Statistic	Train	Test	Valid	Dev
Count	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456

Above: Caption for table above.

Fairness Measurement in Practice

4) Modeling

- Run many models
- Compare & contrast using evaluation metrics

5) Evaluation

- Discrimination
- Model quality
- Firm profitability

$$WGMD(\tau) = \left[\frac{\sum_{i=1}^{N_W} \hat{Y}_{iW}(\tau)=1}{N_W}\right]_{ModelA} - \left[\frac{\sum_{i=1}^{N_W} \hat{Y}_{iW}(\tau)=1}{N_W}\right]_{ModelB}$$

$$PPV(\tau) = \frac{TP_M(\tau)}{[TP_M(\tau)+FP_M(\tau)]} - \frac{TP_W(\tau)}{[TP_W(\tau)+FP_W(\tau)]}$$

Fairness Measurement in Practice

5) Evaluation

- Discrimination
- Model quality
- Firm profitability

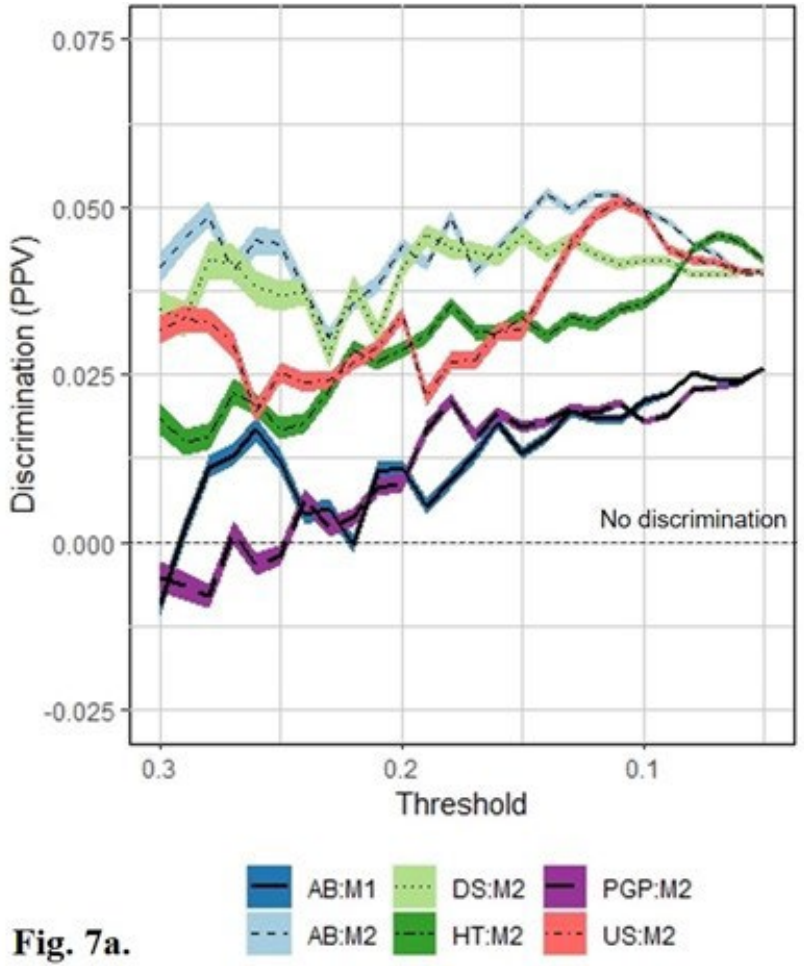


Fig. 7a.

Cross Industry Standard Process for Data Mining

Technical Implementation of Fairness Measurement available on github

https://github.com/stephaniekelley/operationalizingAlethics/blob/0545a20c1088640a2e08b092d45b667f9b092d42/Operationalizing%20AI%20Ethics%20by%20Stephanie%20Kelley.ipynb

Operationalizing AI Ethics (Summary)

Principles

A **document** that states **normative principles** about what employees **should** and **should not do** when it comes to using AI.

Process Governance Framework

A **standardized set of activities** that occur during the modelling process - used for every AI projects, including developing, testing, launching, monitoring, and auditing AI systems.

Technical Implementation

Specific **technical tools** that help achieve activities set up in the process governance framework and principles.

10 Tips* for Operationalizing AI Ethics Principles

- Communication (internal PR)
- Leadership buy-in
- Training across the organization
- Clear responsibility for AI ethics
- Enforcement of principles
- Reporting mechanism
- Reporting mechanism
- Accompanying technical tools
- Updated technical infrastructure
- An interdisciplinary approach

*The “tips” are findings from a study of 24 financial services institutions leading in AI ethics operationalization across 11 countries

IVADO

Thank You!

Dr. Stephanie Kelley
stephanie.kelley@queensu.ca
www.stephaniekelleyresearch.com

CANADA FIRST RESEARCH EXCELLENCE FUND | APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE

Québec Canada