Responsible AI
&
AI Governance

Ricardo Baeza-Yates
KTH, Sweden
UPF, Catalonia
UChile, Chile

@PolarBeaRBY

NASEM Workshop
Washington D.C., Feb 2026

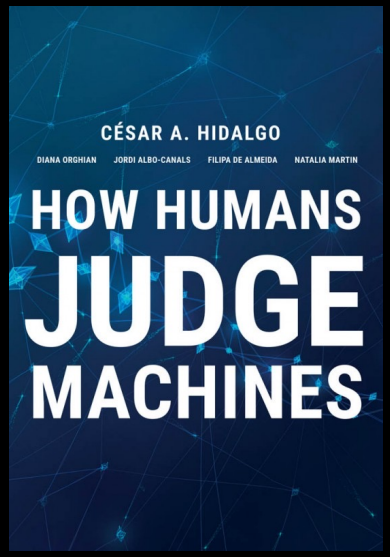
Practice

Why Responsible AI?

Systems do not need to be perfect, but seems that people wants them to be better than us

- Ethical AI?
  - Ethics, justice, trust, etc. are human traits
  - So, we should not associate “ethical” to a machine
- Trustworthy AI?
  - Trust something that does not work all the time?
  - Puts the burden in the user
- AI Safety, **AI Ethics**

[Hidalgo at al., 2021]
Judgingmachines.com

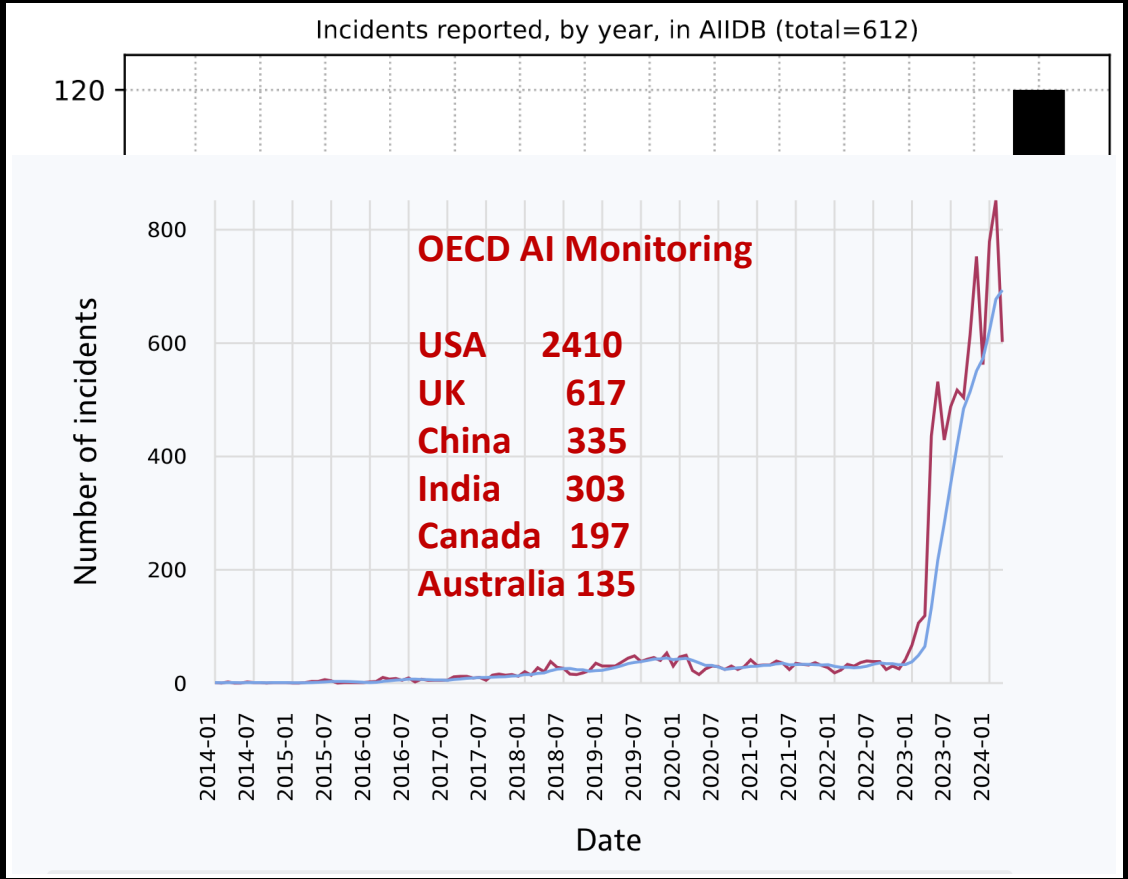


Irresponsible AI

Practice

- Automated discrimination
- Pseudoscience
- Unfair ecommerce
- Environmental impact
- Human incompetence

- Copyright issues
- Disinformation++
- Cognitive loss
- Mental health



incidentdatabase.ai

AI Principles are Instruments

- Belmont Report for biomedical and behavioral research (1979)
- 3 Basic Values
  - Autonomy
  - Beneficial & No harm
  - Justice
- Applications
  - Informed consent
  - Risk & Benefits Assessment
  - Subject selection

Principles Conflict!

[Canca, 2022]

Principles

CORE VALUES / CORE PRINCIPLES	INSTRUMENTAL PRINCIPLES / INSTRUMENTS
Autonomy	human control
	transparency
	agency
	consent
	privacy
	explainability / interpretability
	traceability
Harm-Benefit	competency
	scientific basis
	impact (including environmental)
	well-being
	safety
	security
	accuracy
	reliability
	efficiency
	auditability
Justice	distribution of burden & benefit
	equality / non-discrimination
	protecting the vulnerable
	accessibility
	accountability
	contestability & redress

ACM US TPC Statements

Principles

Algorithm Transparency and Accountability (1/2017)

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing

Responsible Algorithmic Systems (10/2022)

1. Legitimacy and competency
2. Minimizing harm
3. Security and privacy
4. Transparency
5. Interpretability and explanation
6. Maintainability
7. Contestability and auditability
8. Accountability and responsibility
9. Limiting environmental impacts

[Baeza-Yates, Matthews, et al., Oct 2022]

Comparison with Other Influential Proposals

ACM (2017)	OECD (2019)	UNESCO (2021)	White House (2022)	ACM (2022)
Awareness	Human-centered Values & Fairness	Awareness & Literacy	Safe & Effective Systems	Legitimacy & Competency
		Proportionality & Do Not Harm + Fairness & No Discrimination	Algorithmic Discrimination Protection	Minimizing Harm
Data Provenance	Robustness, Security & Safety	Safety & Security + Right to Privacy & Data Protection	Data Privacy	Security & Privacy
	Transparency & Explainability	Transparency & Explainability	Notice & Explanation	Transparency
Explanation				Interpretability & Explainability
Access & Redress + Auditability		Human Oversight & Determination	Human Alternatives, Consideration & Fallback	Contestability & Auditability
Accountability	Accountability	Responsibility & Accountability		Accountability & Responsibility
Validation & Testing				Maintainability
	Inclusive Growth, Sustainable Development & Well-being	Sustainability		Limiting Environmental Impacts
		Multi-Stakeholder and Adaptive Governance & Collaboration		

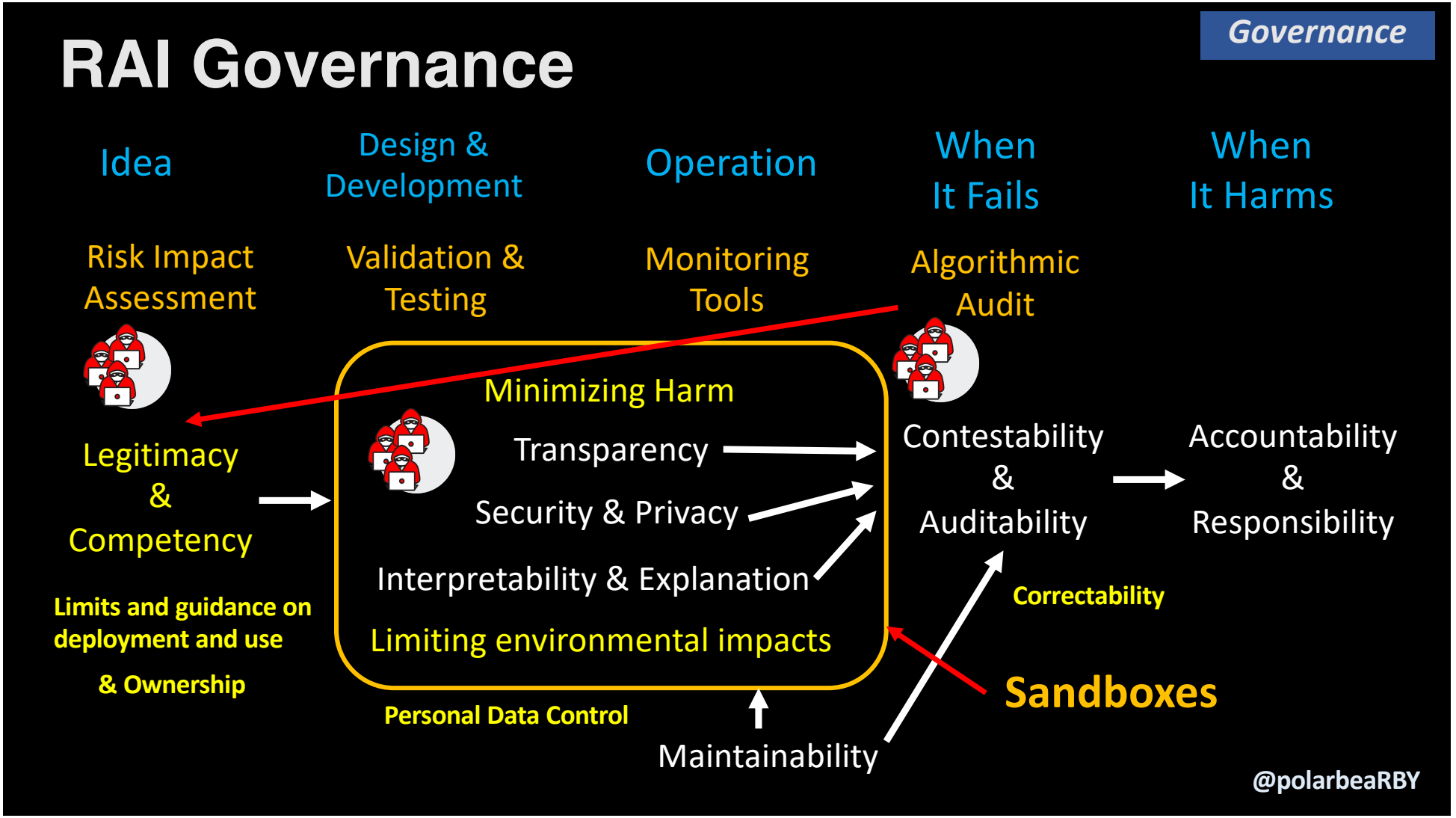
Extension to Generative AI (June 2023)

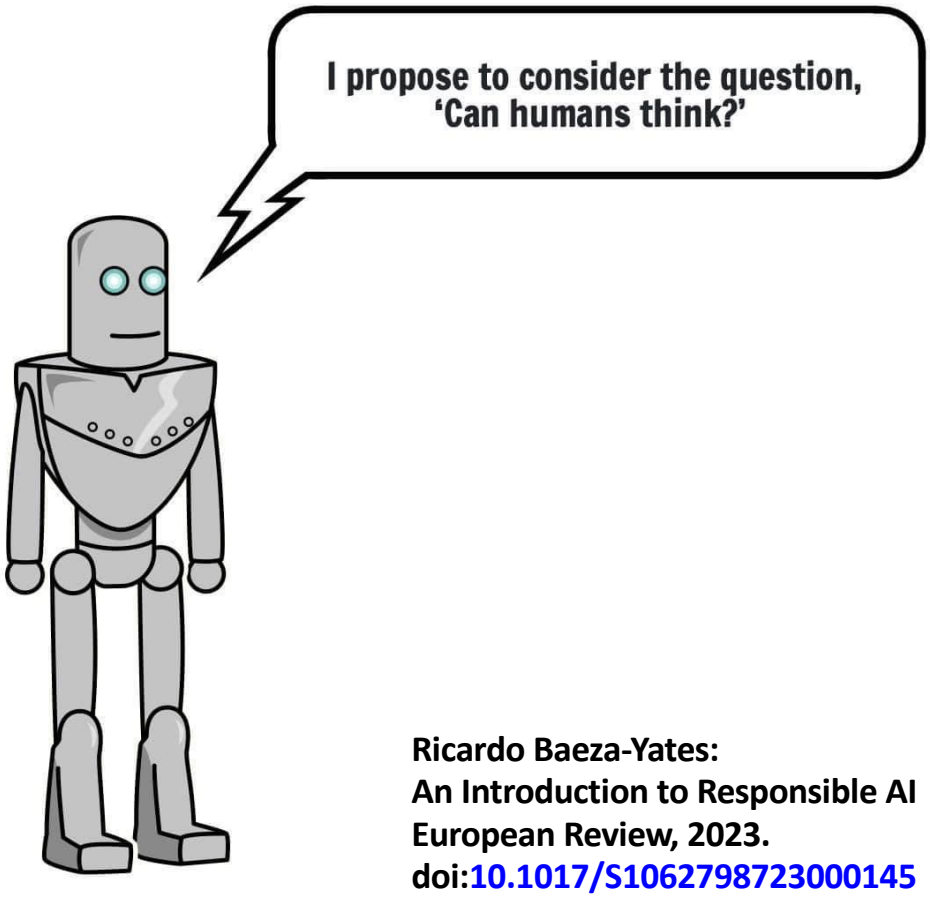
Principles

- **Limits and guidance on deployment and use**
- **Ownership**
- **Personal Data Control**
- **Correctability**

- **Transparency**
- **Auditability and contestability**
- **Limiting environmental impacts**
- **Heightened security and privacy**

Legitimacy and competency
Minimizing harm
Interpretability and explanation
Maintainability
Accountability and responsibility


RAI Governance
Governance
Idea
Design & Development
Operation
When It Fails
When It Harms
Risk Impact Assessment
Validation & Testing
Monitoring Tools
Algorithmic Audit
Minimizing Harm
Transparency
Security & Privacy
Interpretability & Explanation
Limiting environmental impacts
Personal Data Control
Legitimacy & Competency
Limits and guidance on deployment and use
& Ownership
Contestability & Auditability
Accountability & Responsibility
Correctability
Sandboxes
Maintainability
@polarbeaRBY



Ricardo Baeza-Yates:
An Introduction to Responsible AI
European Review, 2023.
doi:10.1017/S1062798723000145

Epilogue