The Explainability Imperative

Implications of AI for Automated Decision-Making

Julian Thomas

ARC Centre of Excellence for Automated Decision-Making and Society

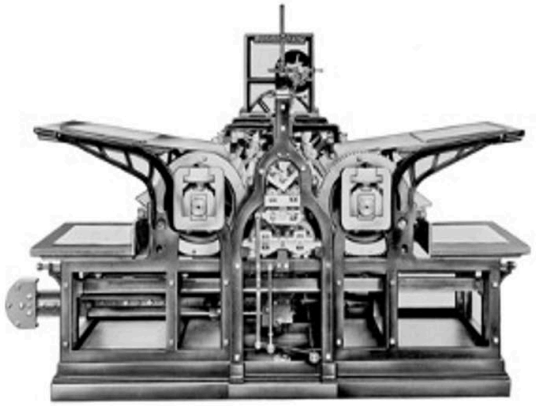
RMIT University

admscentre.org.au

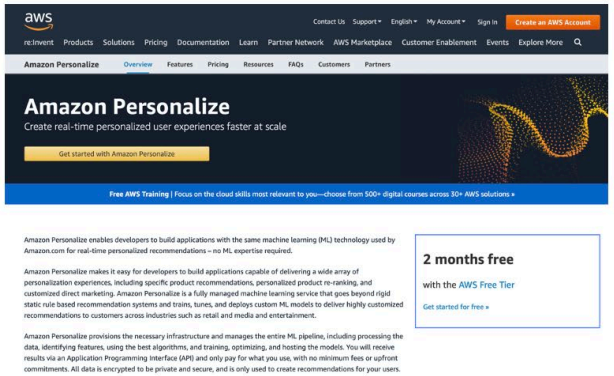
The explainability imperative: driven by a new wave of automation

From machines making things to machines making decisions

in industry...

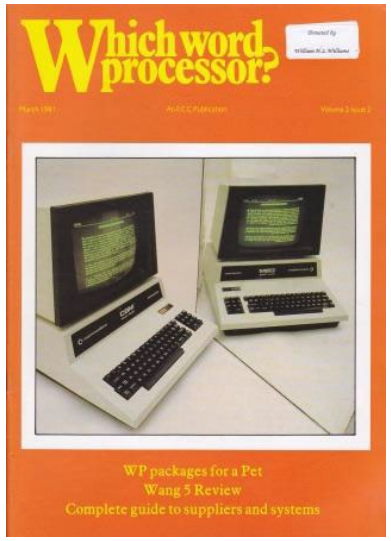


First wave automation: Machines making things
The Times: Koenig Steam press, 1814

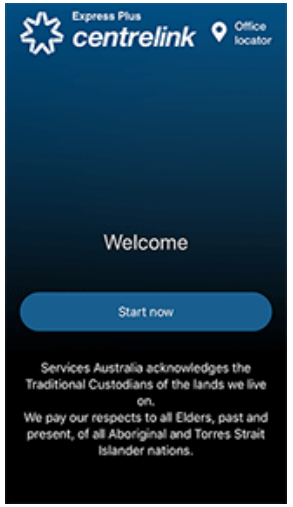


Second wave automation: Machines making decisions
Amazon Web Services, *Amazon Personalize,* 2018-

and in government



First wave automation:
word processors, 1981



Second wave automation:
Smartphone apps for service delivery, 2022

A multi-dimensional challenge: the *intelligibility* of decisions made by machines

Explanation: "An account of the system, its workings, the *implicit and explicit* knowledge it uses to arrive at conclusions in general and the specific decision at hand, that is *sensitive* to the end-user's *understanding, context, and current needs*." (Chari et al.)

Automation offers many benefits for governments and citizens. It also carries significant risks.

How do we explain the outcomes of AI-driven decision-making systems?

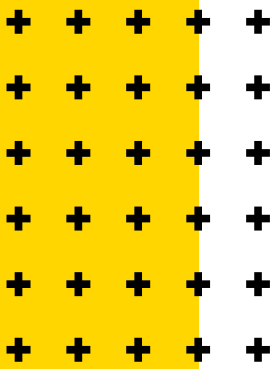
- A new(ish) problem; now arising in many high risk domains (defence, news, health, transport, social services…)
- A threshold requirement for high-risk systems?
- Numerous different forms of explainability and adjacent terms
- Different and potentially competing contexts and imperatives: legal, ethical, technical
- Different research agendas in computer science, law, social science, ethics, cognitive/behavioural studies
- Explanation is a social activity: A more-than-technical problem, involving institutions, business models, data sources and circulation, human design and use (Miller)

Good things but not the same things

Key adjacent terms and policy objectives

(Fjeld et al: *Principled Artificial Intelligence*)

Transparency	• enabling system oversight
Interpretability	• Visibility of how a model is producing particular results: 'Important and slippery' (Lipton)
Accountability	• explainability a necessary condition?
Justification	• the merits of a decision
Responsible disclosure	• a constrained form of communication
Trust	• honesty, reliability, competence generating trustworthiness (O'Neill) • communication, not transparency



Not only *what*, *how* and *why*

Nine kinds of explanation

(Chari et al 2020)

- Case-based — analogies with previous similar situations
- Contextual — explanations derived from the broader circumstances
- Contrastive — why this outcome rather than another?
- Counterfactual — would the decision change with different information?
- Everyday — explanations framed by lived experience, real world situations
- Scientific — derived from scientific theories, concepts or observations
- Simulation-based — explanations derived from 'what if' scenarios
- Statistical — explanations derived from statistical evidence
- Trace-based — a line of reasoning, identifying key steps

Note that there

Obligations to explain?

Artificial Intelligence Ethics Framework (2019) – Australian Government

Transparency and explainability: There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

Automated Decision-Making Better Practice Guide (2019) - Commonwealth Ombudsman

- Ensuring compliance with administrative law requirements.
- Ensuring the transparency and accountability of the system

EU General Data Protection Regulation Recital 71 (2016)

Automated processing “should be subject to suitable safeguards, which should include … the right to obtain an explanation of the decision reached after such assessment and to challenge the decision."

OECD/G20 AI principles (2019)

AI Actors should commit to transparency and responsible disclosure

- 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 based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

Human Rights and Technology Final Report (2021) - Australian Human Rights Commission

Recommended measures to improve transparency:

- notification of the use of AI
- stronger right to reasons for decisions and independent review

Explanations for the people who need them most?

A highly stratified Australian internet

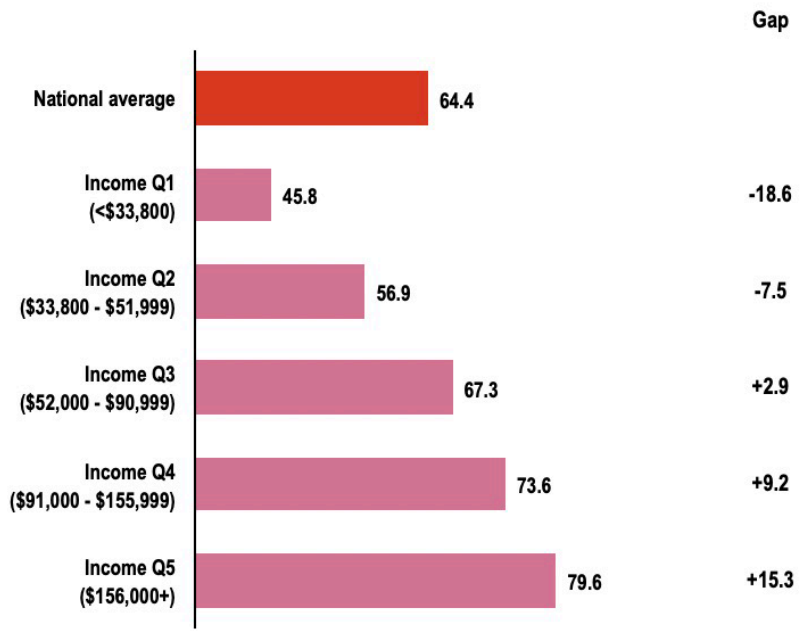
Measures of Digital Ability: Australian Digital Inclusion Index, 2021

Digital Ability comparison — Operational basic

Subgroup	Digital ability Score	Digital ability Gap	Operational basic Score	Operational basic Gap
75+	27.2	-37.2	35.7	-37.3
Did not complete secondary school	36.3	-28.1	43.1	-30.0
65-74	41.8	-22.6	52.7	-20.4
Income Q1 (<$33,800)	45.8	-18.6	51.4	-21.7
Not in labour force	50.6	-13.8	58.8	-14.2
Receives income support	52.3	-12.1	59.1	-14.0
People with disability	52.3	-12.0	59.0	-14.1
Single person	52.4	-12.0	59.7	-13.3

1 / 5

Digital Ability comparison — Digital Ability



Source: **www.digitalinclusionindex.org.au**

Tactics, Tools, Trade-offs and Questions

Emerging problem solving:

Explainability vs predictive accuracy

Interpretability vs performance

Analytic tools that provide insights into particular results

Iterative experimentation

Cross-disciplinary connections

Building capability and skills

Larger questions:

What do we want to know? What sorts of explanations are we seeking, and when?

What kinds of explanations are we obliged to provide, and for whom? What do we want users and citizens to know?