Ethics in Data Science

Case 1: AI Boosts Discovery

Can you spot any issues?

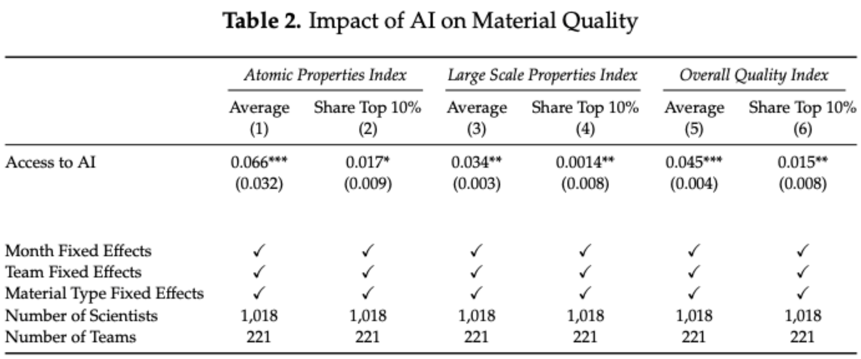

Table 2. Impact of AI on Material Quality

	Atomic Properties Index		*Large Scale Properties Index*		*Overall Quality Index*	
	Average (1)	Share Top 10% (2)	Average (3)	Share Top 10% (4)	Average (5)	Share Top 10% (6)
Access to AI	0.066^{***} (0.032)	0.017^{*} (0.009)	0.034^{**} (0.003)	0.0014^{**} (0.008)	0.045^{***} (0.004)	0.015^{**} (0.008)
Month Fixed Effects	✓	✓	✓	✓	✓	✓
Team Fixed Effects	✓	✓	✓	✓	✓	✓
Material Type Fixed Effects	✓	✓	✓	✓	✓	✓
Number of Scientists	1,018	1,018	1,018	1,018	1,018	1,018
Number of Teams	221	221	221	221	221	221

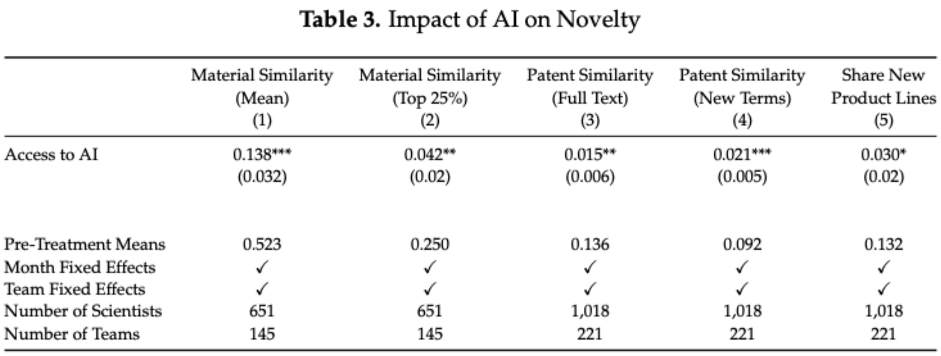

Table 3. Impact of AI on Novelty

	Material Similarity (Mean) (1)	Material Similarity (Top 25%) (2)	Patent Similarity (Full Text) (3)	Patent Similarity (New Terms) (4)	Share New Product Lines (5)
Access to AI	0.138^{***} (0.032)	0.042^{**} (0.02)	0.015^{**} (0.006)	0.021^{***} (0.005)	0.030^{*} (0.02)
Pre-Treatment Means	0.523	0.250	0.136	0.092	0.132
Month Fixed Effects	✓	✓	✓	✓	✓
Team Fixed Effects	✓	✓	✓	✓	✓
Number of Scientists	651	651	1,018	1,018	1,018
Number of Teams	145	145	221	221	221

Read more about this story in original article, nature report, first post on possible fraud, blog post by Andrew Gelman, MIT retraction note

Reproducibility in Scientific Research

- There are many more cases of data fogery in scientific publications.
- "An analysis of past studies indicates that the cumulative (total) prevalence of irreproducible preclinical research exceeds 50%" Freedman et al. (2015)

Solution:

- Making code and data available Reproducible Research in Computational Science by Roger Peng

- ...

Case 2: Facebook’s Secret Emotion Experiment

In 2014, Facebook conducted an experiment on nearly **700,000 users**—without telling them. It manipulated the emotional tone of their news feeds to see if users’ own posts became more positive or negative. A paper was published using this dataset.

The result? Emotional contagion was real. But users were **never informed** they were part of an experiment.

References:

- Original paper in PNAS
- New York Times article

Experiments with Human Subjects

The 1991 Common Rule requires Institutional Review Board (IRB) review and approval for certain types of human subjects research.

References:

- IRB FAQ by FDA
- IRB at UC Davis

Case 3: The Netflix Prize and the Mystery of Re-Identification

- The Neflix Prize: predict movie ratings based on anonymized dataset containing millions of user movie ratings.
  No names, usernames, emails in the dataset.
- Researchers managed to re-identify some of the users by comparing their Netflix ratings to *public reviews on IMDb*.
  Example: someone rated The Godfather five stars on both platforms, rated Finding Nemo three stars the same day, and only used certain genres. These patterns are surprisingly unique that can be used as a digital fingerprint.

Even if names are removed, patterns can still reveal identities!

Differential Privacy

- Differential Privacy: mathematically rigorous framework for releasing statistical information about datasets while protecting the privacy of individual data subjects.

- Example: Apple uses differential privacy in iOS to protect user privacy while collecting useful data. Instead of sending raw keystrokes or emoji usage, devices send **noisy versions** of the data. This way, Apple can learn general trends without knowing exactly what any one person typed.

- Privacy protections are challenging in the era of Large Language Models
  Q: Will your chat history with ChatGPT reveals who you are?

Case 4: When AI Struggles to See Faces

In 2018, Joy Buolamwini and Timnit Gebru discovered that commercial facial recognition systems had much higher error rates for Black women (30%) than white men (1%).

The reason? Most training data came from white male faces.

This is called **representation bias**—where some groups are underrepresented in the data the model learns from.

More info:

- Gender Shades Project
- NPR facial recognition bias article

Case 5: The Algorithm That Decided Freedom

- The [COMPAS algorithm](https://en.wikipedia.org/wiki/COMPAS_(software) predicts a defendant’s likelihood of committing a future crime. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions.

- A 2016 ProPublica investigation found racial bias:

  - Black defendants were more likely to be **falsely** labeled “high risk.”
  - White defendants were more likely to be **falsely** labeled “low risk.”
- Even though race wasn't used as an input, variables like zip code and prior arrest history created **indirect bias**.

Algorithmic Fairness

- Algorithmic fairness refers to principles and techniques that ensure machine learning models treat individuals or groups equitably, especially across categories like race, gender, or age.

- Common definitions include demographic parity (equal outcomes across groups), equalized odds** (equal error rates), and individual fairness (similar individuals receive similar predictions).
- Fairness constraints are often added during model training, post-processing, or through data rebalancing.
- There are trade-offs between different fairness definitions and between fairness and model accuracy, requiring careful design choices.

Reference:

- Algorithmic Fairness by Kleinberg et al.
- Wikipedia on Fairness
- Google’s PAIR Guidebook