Incorporating Ethics into the Data Science Curriculum

Royal Statistical Society Conference, Aberdeen 2022

Dr Zak Varty

Imperial College London

Twitter: `@zakvarty` **Email:** `z.varty@imperial.ac.uk`

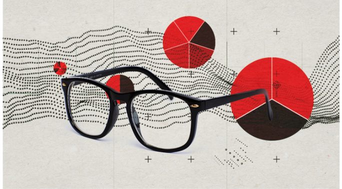
Data Science: Miracle Cure and Sexiest job

- Healthcare
- Ecology & Conservation
- Business & Government
- Environment

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022



HBR Staff/StudioM1/Moritz Otto/Getty Images

Source: Harvard Business Review

Data Science: What could possibly go wrong?

Facial recognition fails on race, government study says

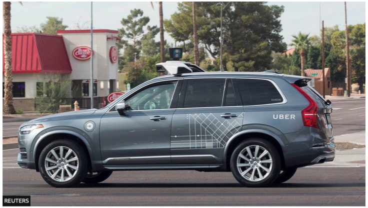
20 December 2019

Facial recognition tools are increasingly being used by police forces

A US government study suggests facial recognition algorithms are far less accurate at identifying African-American and Asian faces compared to Caucasian faces.

Uber's self-driving operator charged over fatal crash

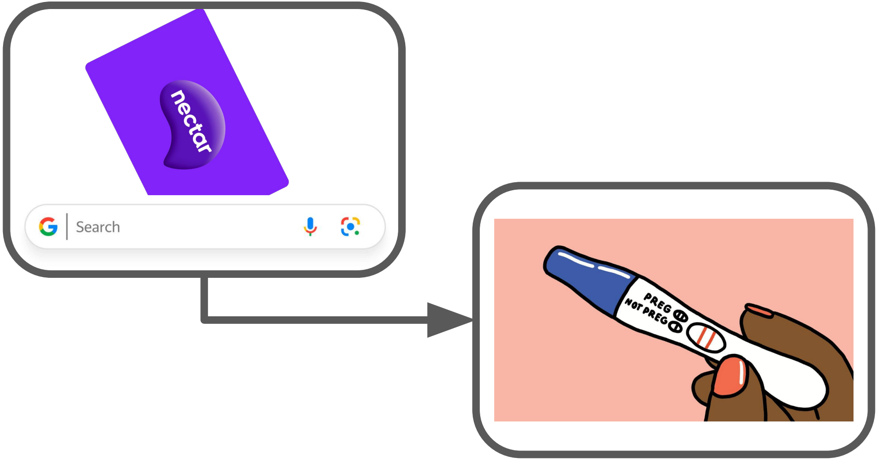
16 September 2020



The self-driving Volvo hit a pedestrian at 39mph, despite the presence of a safety driver

The back-up driver of an Uber self-driving car that killed a pedestrian has been charged with negligent homicide.

Data Science: What could possibly go wrong?



A bit of context

Imperial MSc in Machine Learning and Data Science

Degree

- Postgraduate
- Online
- Part Time

Cohort

- Global
- Numerate
- Professional

Course

- Full Course
- Progression
- Mathematical

5(ish) Principles of Ethical DS

& how you might already teach them

Principle 1: Privacy & Autonomy

Privacy & Autonomy

The right and ability to be unobserved.

Control over collection, storage and use of personal data.

Principle 1: Privacy & Autonomy

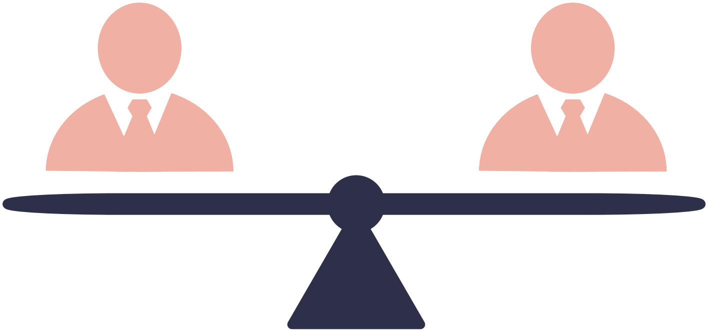
- **Discussion:** Online fingerprints, survey design, GDPR.
- **Technical:** Missing data, non- and randomised-response, hard-to-reach demographics.

Principle 2: Fairness

Fairness

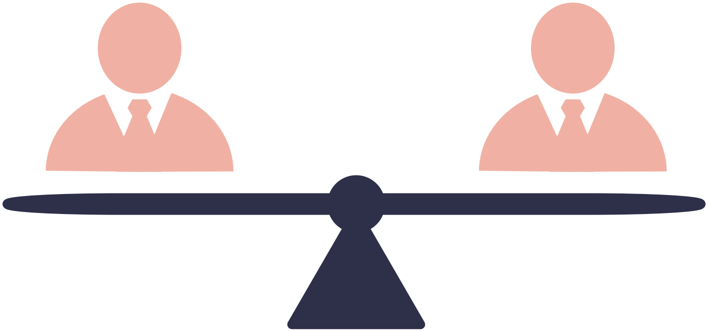
Predictions or decisions should not be influenced by protected characteristics.

Various, conflicting definitions, incl. *error parity* and *equal opportunity*.

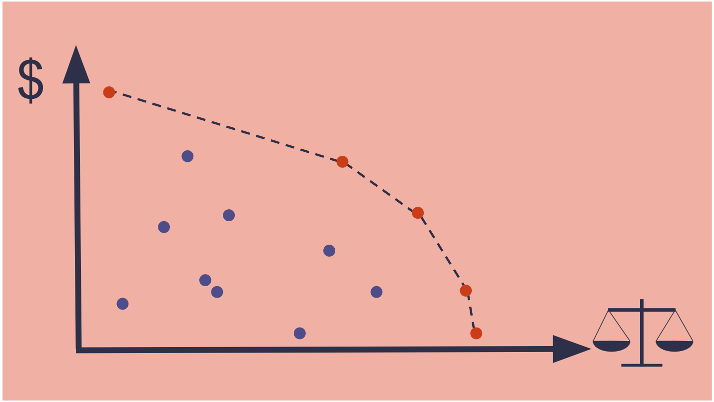


Principle 2: Fairness

- **Case Studies:** Facial Recognition, Income Inequality, Testing Fairness.
- **Technical:** Classification problems, Conditional probability.



Principle 3: Value Alignment

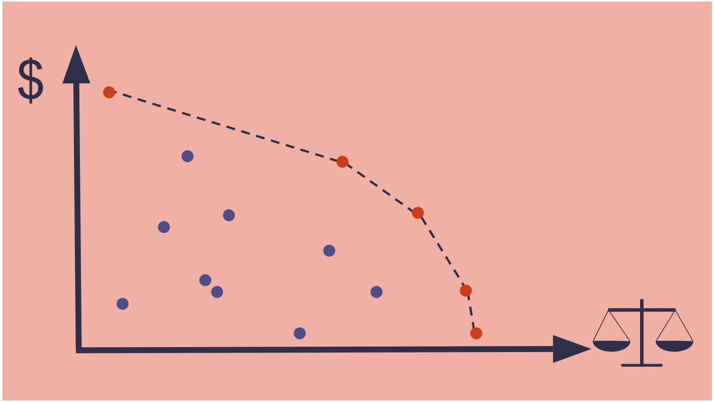


Value Alignment

We have competing objectives.

The trade-off between them should be made in an explicit and considered manner.

Principle 3: Value Alignment



- Loss functions: false + vs false − and the importance of context.
- MOO & Pareto Efficiency.
- Utility functions, decision theory and subjectivity.

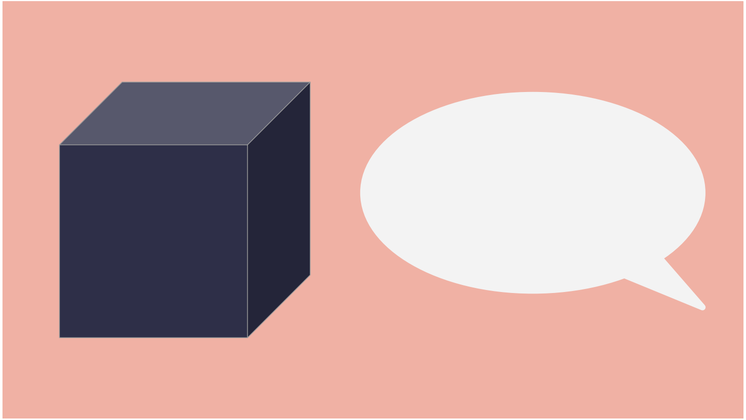
Principle 4: Explainability & Interpretability

Explainability & Interpretability

Opaque vs Transparent models.

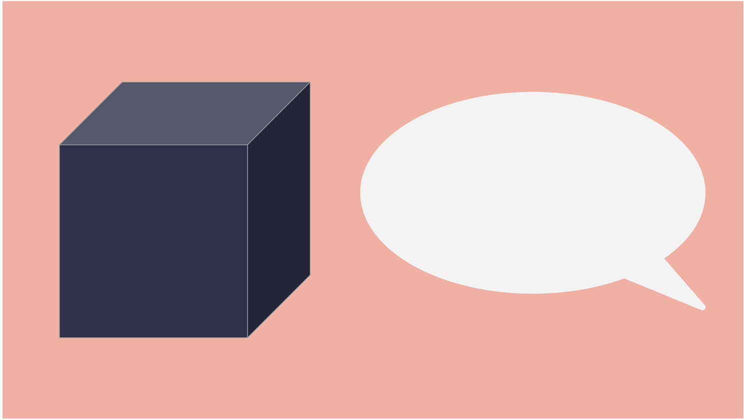
How does the model work? Why did it make this prediction?

Specialist vs end-user.



Principle 4: Explainability & Interpretability

- Interpretation of fitted models.
- Communication in a range of contexts.
- LIME, SHAP and friends.



Principle 5: Safety, Security and Accountability

Safety: Hippocratic Oath for Data Scientists.

Security: Plan ahead for lazy and malicious actors. Data governance, storage, encryption.

Accountability: What happens when it all goes terribly wrong?

Principle 5: Safety, Security and Accountability

- Professional codes of ethics, GradStat and CStat.
- Data Governance, storage and encryption.
- Putting models into production, monitoring and updating.

Wrapping up

Three Take-Aways:

Wrapping up

Three Take-Aways:

1. Models don't hurt people, people hurt people.

Wrapping up

Three Take-Aways:

1. Models don't hurt people, people hurt people.
2. Ethics $\neq$ Essays.

Wrapping up

Three Take-Aways:

1. Models don't hurt people, people hurt people.
2. Ethics $\neq$ Essays.
3. Lean into what you are already doing.

Thank you. Any Questions?

Where to get started? - Books

- Kearns, M., & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.
- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. `https://www.fairmlbook.org`.
- Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. `https://christophm.github.io/interpretable-ml-book/`.

Where to get started? Papers

- Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:77-91.

- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8:141-163.

- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in neural information processing systems, 30.

Where to get started? Other Resources

- RSS Guidelines for Ethical Data Science (2019) [Download]
- Causal Inference from a Machine Learnign Perspective. Brady Neal. [Link to Course and Book]
- Let the algorthim work for you: YouTube, Twitter, TikTok. (Here are a few twitter handles to get you started! Kristian Lum @KLdivergence, Chelsea Parlett-Pelleriti @ChelseaParlett, Joshua Loftus @joftius.)