Ethics

Martha Lewis

Department of Engineering Mathematics
University of Bristol

Outline

- Challenges for AI and Machine Learning
- Steps towards addressing these challenges
- Machines with moral status

Challenges for AI and Machine Learning

- Transparency
- Predictability
- Robust against manipulation
- Responsibility

Amazon Recruitment

RETAIL OCTOBER 11, 2018 / 12:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Strava data publication

Strava's heatmap was a 'clear risk' to security, UK military warned

The Ministry of Defence reissued guidelines after exercise heatmaps revealed bases. In one case, a military sports club with names and photos was revealed

By **MATT BURGESS**

Wednesday 4 April 2018

Credit **Strava**

https://www.wired.co.uk/article/strava-heat-maps-military-app-uk-warning-security

Cambridge Analytica

ByBookCatalog-cambridgeanalyticafacebook,CCBY2.0,
https:
//commons.wikimedia.org/w/index.php?curid=87468247

Autonomous vehicles

CITYLAB

My Fight With a Sidewalk Robot

A life-threatening encounter with AI technology convinced me that the needs of people with disabilities need to be engineered into our autonomous future.

Emily Ackerman

19 November 2019, 17:43 GMT

A Starship Technologies commercial delivery robot navigates a sidewalk. *Wolfgang Ratta*

Copy Link

Facial recognition

Table 1: Overall Error on Pilot Parliaments Benchmark, August 2018 (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Target Corporations									
Face ++	1.6	2.5	0.9	2.6	0.7	4.1	1.3	1.0	0.5
MSFT	0.48	0.90	0.15	0.89	0.15	1.52	0.33	0.34	0.00
IBM	4.41	9.36	0.43	8.16	1.17	16.97	0.63	2.37	0.26
Non-Target Corporations									
Amazon	8.66	18.73	0.57	15.11	3.08	31.37	1.26	7.12	0.00
Kairos	6.60	14.10	0.60	11.10	2.80	22.50	1.30	6.40	0.00

Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019.
`https://dl.acm.org/doi/abs/10.1145/3306618.3314244`

More information

Timnit Gebru: How can we stop artificial intelligence from marginalizing communities
https://www.ted.com/talks/timnit_gebru_how_can_we_stop_artificial_intelligence_from_marginalizing_communities
Joy Buolamwini: How I'm fighting bias in algorithms
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Starting to deal with these challenges

Regulation in government

- UK General Data Protection Regulation
- UK Office for Artificial Intelligence
  `https://www.gov.uk/government/organisations/office-for-artificial-intelligence`
- UK government guide on using artificial intelligence safely
  `https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safet`

Starting to deal with these challenges

The 5 C's

- Consent
- Clarity
- Consistency and Trust
- Control and Transparency
- Consequences

`https://www.ebooks.com/en-uk/book/209925972/ethics-and-data-science/mike-loukides/`

Consent

- At every step of building a data product, it is essential to ask whether appropriate and necessary consent has been provided.
- This means not just consent for the collection of the data, but consent for it to be used and processed

Clarity

Users of a service must have clarity about:

- What data they are providing,
- What is going to be done with the data,
- Any downstream consequences of how their data is used

Consistency and trust

- Trust requires consistency over time.
- Consistency, and therefore trust, can be broken either explicitly or implicitly.
- Failing to safeguard customer data consistently over time breaks trust

Control and Transparency

Once you have given your data to a service, you must be able to understand what is happening to your data

- Can you control how the service uses your data?
- Can you stop the service using your data?
- Can you delete your data?

Control and Transparency

Once you have given your data to a service, you must be able to understand what is happening to your data

- Can you control how the service uses your data?
- Can you stop the service using your data?
- Can you delete your data?

Consequences

- Can data that is being collected cause harm to an individual or a group?
- Can the way that data is being used cause harm to an individual or a group?
- Laws and policies have been put in place to protect specific groups, but these laws do not keep up with the pace of technology.

What should you do?

- Checklists
- Codebooks
- Datasheets
- Model Cards

Checklists https://deon.drivendata.org/

A: Data Collection

- A.1 Informed consent: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
- A.2 Collection bias: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
- A.3 Limit PII exposure: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?
- A.4 Downstream bias mitigation: Have we considered ways to enable testing downstream results for biased outcomes (e.g., collecting data on protected group status like race or gender)?

https://deon.drivendata.org/

Checklists

B. Data Storage

- B.1 Data security: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
- B.2 Right to be forgotten: Do we have a mechanism through which an individual can request their personal information be removed?
- B.3 Data retention plan: Is there a schedule or plan to delete the data after it is no longer needed?

`https://deon.drivendata.org/`

Checklists

C. Analysis

- C.1 Missing perspectives: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
- C.2 Dataset bias: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?
- C.3 Honest representation: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?
- C.4 Privacy in analysis: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
- C.5 Auditability: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

Checklists

D. Modeling

- D.1 Proxy discrimination: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
- D.2 Fairness across groups: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
- D.3 Metric selection: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?
- D.4 Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
- D.5 Communicate bias: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?

`https://deon.drivendata.org/`

Checklists

E. Deployment

- E.1 Redress: Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?
- E.2 Roll back: Is there a way to turn off or roll back the model in production if necessary?
- E.3 Concept drift: Do we test and monitor for concept drift to ensure the model remains fair over time?
- E.4 Unintended use: Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed?

`https://deon.drivendata.org/`

Codebooks

There are seven types of information that a code-book should contain:

- A short description of the study design
- Document all of the sampling information
- Present information on the data file
- The data structure needs to be clearly delineated
- Specific details about the data need to be documented
- The question text and answer categories should be clearly documented along with frequencies of each response option
- If the data have been weighted, a thorough description of the weighting processes should be included

Source: Encyclopedia of survey research methods

Datasheets

Every dataset should be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, etc.

- For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use.
- For dataset consumers, the primary objective is to ensure they have the information they need to make informed decisions about using a dataset
- Transparency on the part of dataset creators is necessary for dataset consumers to be sufficiently well informed that they can select appropriate datasets for their tasks and avoid unintentional misuse.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).

Datasheet Questions

- Motivation
- Composition
- Collection Process
- Preprocessing/cleaning/labelling
- Uses
- Distribution
- Maintenance

Model Cards for Model Reporting

Model Card

- Model Details
- Intended Use
- Factors
- Metrics
- Evaluation Data
- Training Data
- Quantitative Analyses
- Ethical Considerations
- Caveats and Recommendations

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In Proceedings of the conference on fairness, accountability, and transparency, pp. 220-229. 2019. https://arxiv.org/abs/1810.03993

Example model card

https://github.com/openai/gpt-3/blob/master/model-card.md

Summary

- Ethical considerations are far reaching, and need to be embedded into system development at an early stage and throughout
- There are guidelines and both established and new methods to help with doing this
- You can do this too, when you are developing your own projects in both study and work

Machines with moral status

At present, we do not consider AI systems to have moral status.
Potential criteria:

- Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer
- Sapience: a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence, 1, 316-334.

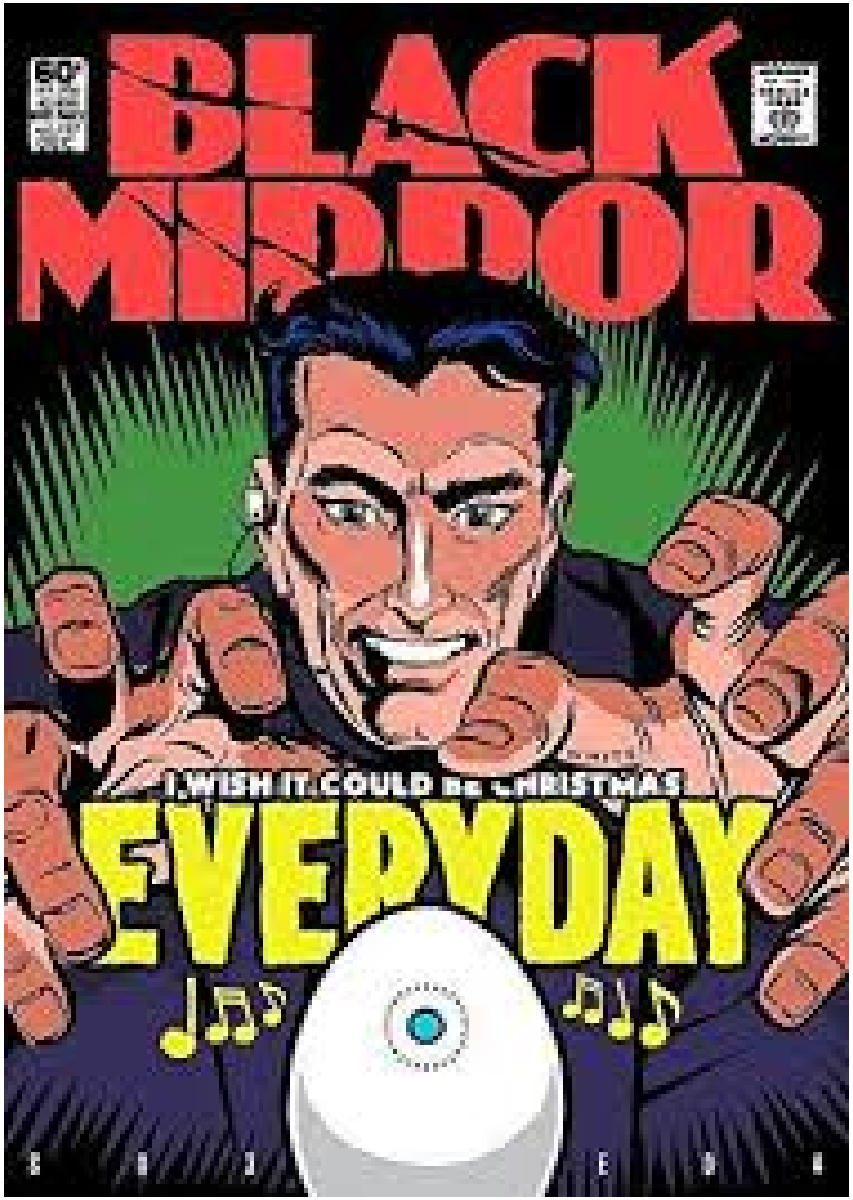
Stanford Encyclopedia of Phil.
- Ethics of AI

Machines with moral status

We can also consider the following criteria:

- Principle of Substrate Non-Discrimination: If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.
- Principle of Ontogeny Non-Discrimination: If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.

Alternative Minds



Summary

- Although concerns with how AI systems are currently being implemented and used are the more pressing, there are also issues to consider for the future
- AI systems may become entities with their own moral status.
- The fundamental differences between humans and AI systems may present us with unusual and difficult questions to consider

Further Reading

- Burr, Christopher, and Nello Cristianini. "Can machines read our minds?." Minds and Machines 29.3 (2019): 461-494.
- Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA.
- `https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety` (long!)
- `https://bristol.ac.uk/golding/what-we-do/data-governance-and-reproducibility/`

- There is a wealth of material available! Let me know if you want help searching for any.