Ethical Data Science

Ian René Solano-Kamaiko

Cornell Tech &
Center for Responsible AI at NYU

Cornell Bowers C·IS
College of Computing and Information Science

Land Acknowledgement

I would like to acknowledge that I work at Cornell Tech, which occupies part of the unceded homeland of the Lenape people and that Cornell University is located on the traditional homelands of the Cayuga Nation. I want recognize the longstanding significance of these lands for these nations past and present. It is important that we acknowledge the forceful dispossession of both the Lenape and Cayuga people, and to honor them as the original inhabitants of these lands of which we are uninvited settlers.

As we talk about ethics within data science, we will see that throughout this conversation technology is rooted in people and the decisions we make both as a society and as individuals. Therefore, I feel that it is important to acknowledge our past injustices in the United States as well as their systemic nature.

About

Ian René Solano-Kamaiko
él/he/him
@ianrsolano

Special shoutout to:
Dr. Julia Stoyanovich and
Falah Arif Khan

Comics:
dataresponsibly.github.io/comics

Research:

- Ph.D. student at Cornell University co-advised by Dr. Nicola Dell and Dr. Aditya Vashistha.
- Building and evaluating computing technologies that aim to improve the lives of marginalized and underserved populations (specifically in community and in-home healthcare, future of work, and climate resilience)
- M.S. in Computer Science from NYU
- Graduate research fellow at the Center for Responsible AI under the supervision of Dr. Julia Stoyanovich
- Before academia, I worked as a software engineer for various NYC tech startups

AI IS THE FUTURE, AND THE FUTURE IS HERE.

- Every tech article on the Internet (I’ve got real citations if you need it)

Why Data Science? Why Now?

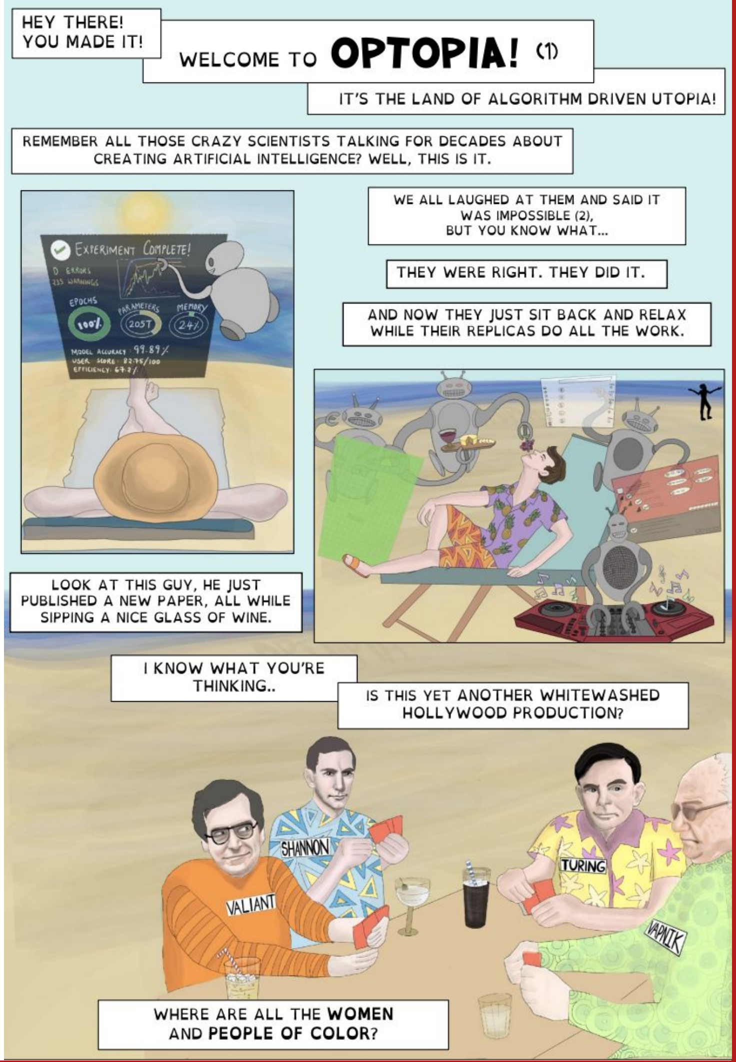
Why DS:

Data science (DS), artificial intelligence (AI), and machine learning (ML) have the potential to impact every facet of our lives from automated vehicles to life saving medicines to targeted advertisements.

Why Now:

(1) Unprecedented data collection capabilities
(2) Increases in computational power and access
(3) A mature field with broader societal acceptance

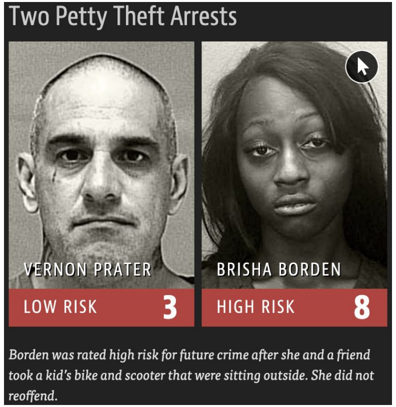
Falaah Arif Khan and Julia Stoyanovich. “Mirror, Mirror”.
Data, Responsibly Comics, Volume 1 (2020)
https://dataresponsibly.github.io/comics/vol1/mirror_en.pdf




WHAT COULD GO WRONG?

Examples of Bias in Algorithms

Criminal Justice



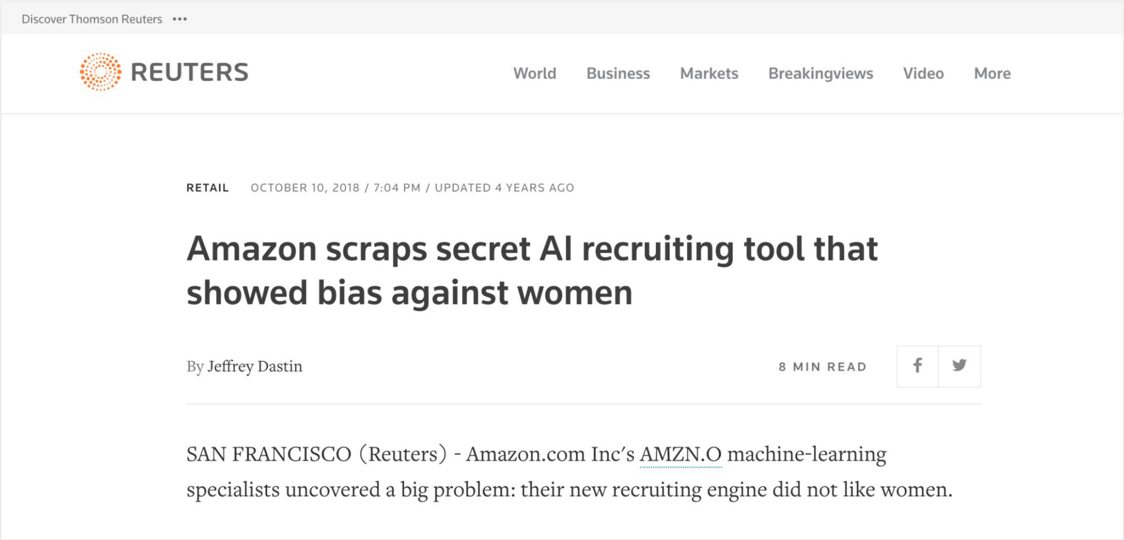
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Automated Hiring



Discover Thomson Reuters •••

REUTERS World Business Markets Breakingviews Video More

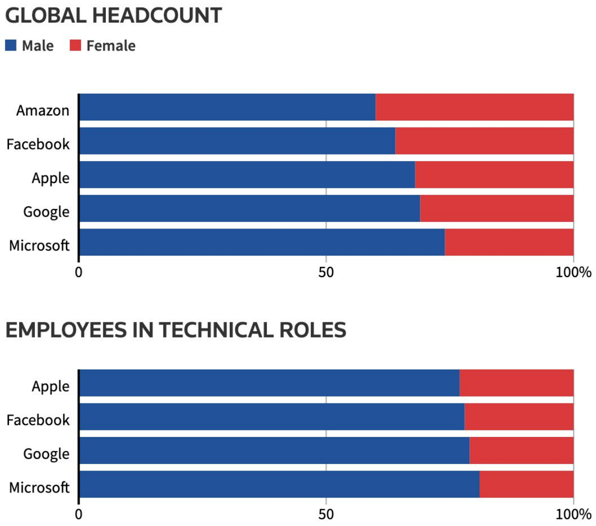
RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 4 YEARS AGO

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/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G



Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

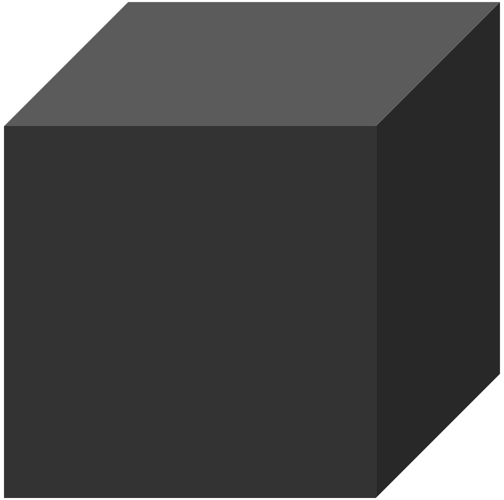
What is Ethical Data Science?

1. **Fairness, Accountability, and Transparency**
2. Data Profiling, Cleaning, and Integration
3. Data Protection and Privacy
4. Legal Frameworks, Codes of Ethics, and Professional Responsibility

Explainable AI (XAI)

XAI is a set of **processes** and **methods** that allows **human users** to **comprehend** and **trust** the results and **output** created by machine learning algorithms. Explainable AI is used to **describe an AI model**, its **expected impact** and **potential biases**. It helps characterize model **accuracy**, **fairness**, **transparency** and **outcomes** in AI-powered decision making.

https://www.ibm.com/watson/explainable-ai (emphasis my own)



Algorithmic Fairness and Bias

Pre-Existing: exists independently and usually prior to the creation of the system, has its roots in society (social institutions, practices, and attitudes)

Technical: introduced or exacerbated by the technical properties of a system (issues in the technical design)

Emergent: arises only in a context of use (a result of changing societal knowledge, population, or cultural values)

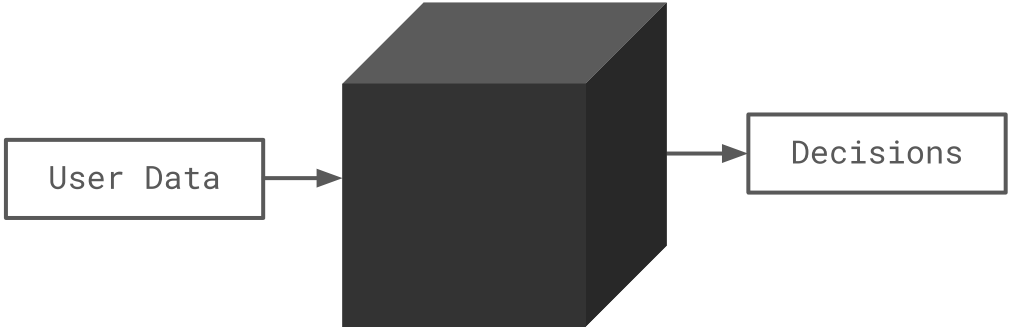
Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on information systems (TOIS), 14(3), 330-347.

Explainability and Transparency

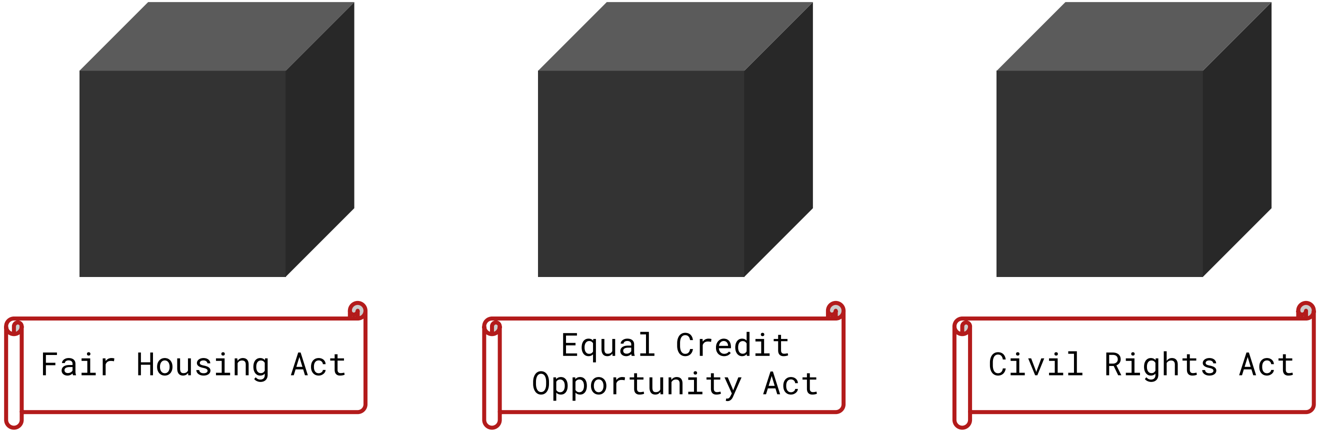
Question	Ways to explain	Example XAI methods
How (global model-wide)	• Describe the general model logic as feature impact*, rules† or decision-trees‡ • If user is only interested in a high-level view, describe what are the top features or rules considered	ProfWeight*†‡ [28], Global feature importance* [71, 105], Global feature inspection plots* (e.g. PDP [49]), Tree surrogates‡ [25]
Why (a given prediction)	• Describe how features of the instance, or what key features, determine the model's prediction of it* • Or describe rules that the instance fits to guarantee the prediction† • Or show similar examples with the same predicted outcome to justify the model's prediction‡	LIME* [89], SHAP* [72], LOCO* [63], Anchors† [90], ProtoDash‡ [47]
Why Not (a different prediction)	• Describe what features of the instance determine the current prediction and/or with what changes the instance would get the alternative prediction* • Or show prototypical examples that have the alternative outcome†	CEM* [27], Counterfactuals* [69], ProtoDash† (on alternative prediction) [47]
How to Be That (a different prediction)	• Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, with minimum effort required* • Or show examples with minimum differences but had the alternative outcome†	CEM* [27], Counterfactuals* [69], Counterfactual instances† [100], DiCE† [78]
How to Still Be This (the current prediction)	• Describe features/feature ranges* or rules† that could guarantee the same prediction • Or show examples that are different from the instance but still had the same outcome	CEM* [27], Anchors† [90]
What if	• Show how the prediction changes corresponding to the inquired change of input	PDP [49], ALE [10], ICE [44]
Performance	• Provide performance information of the model • Provide uncertainty information for each prediction* • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC; Communicate uncertainty of each prediction* [42]; See examples in FactSheets [11] and Model Cards [77]
Data	• Provide comprehensive information about the training data, such as the source, provenance, type, size, coverage of population, potential biases, etc.	See examples in FactSheets [11] and Datasheets [39]
Output	• Describe the scope of output or system functions. • If applicable, suggest how the output should be used for downstream tasks or user workflow	See examples in FactSheets [11] and Model Cards [77]

Table 1. A mapping guidance between categories of user questions in XAI question bank [65] and example XAI methods to answer these questions, with descriptions of their output in "Ways to explain" column. XAI methods are selected based on what are available in current open-source XAI toolkits [1–4]. The last three rows (in *italic*) are broader XAI needs not limited to explaining model processes. This mapping guidance can support identifying appropriate XAI techniques based on user questions.

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790.



Regulating Automated Decision Systems



SHAP and SAGE

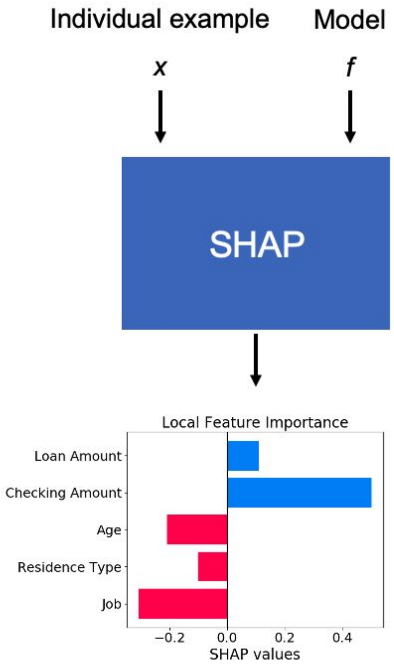
Each method answers a specific type of question:

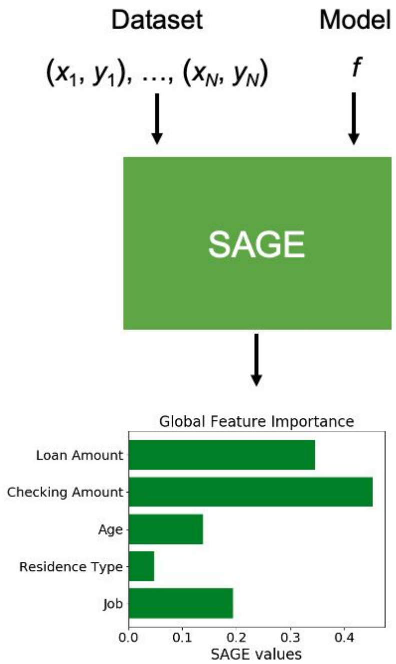
SHAP answers the question how much does each feature contribute to this individual prediction?

SAGE answers the question how much does the model depend on each feature overall?

SHAP is a method for explaining individual predictions (local interpretability), whereas SAGE is a method for explaining the model's behavior across the whole dataset (global interpretability).

https://iancovert.com/blog/understanding-shap-sage/





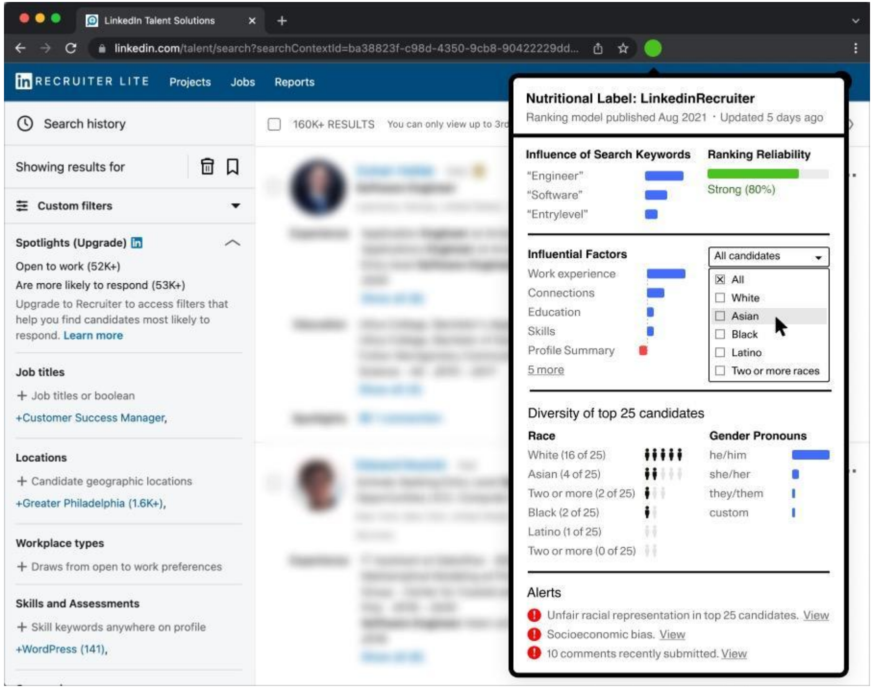
Research Examples of XAI

Nutritional Labels for Recruiting ADS

Comprehensive: short, simple, clear

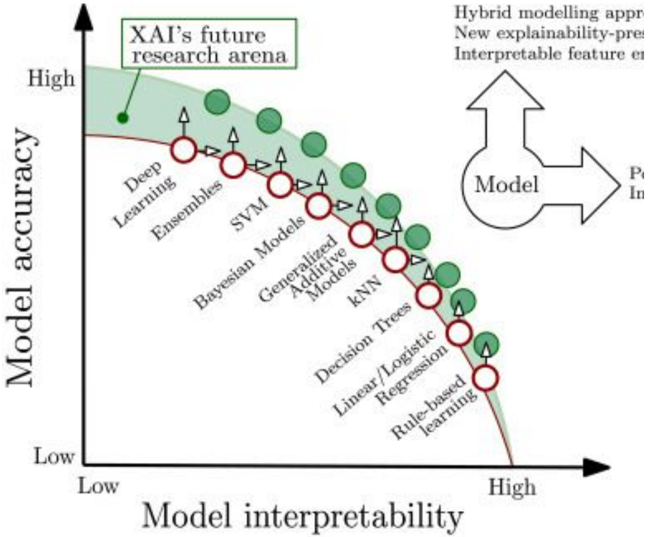
Consultative: provide actionable information

Comparable: implying a standard



Cornell Bowers C·IS

Accuracy-Explainability Trade-off



Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

We empirically quantified of the tradeoff between model accuracy and explainability in two real-world policy contexts (education and housing).

We found that black-box models may be as explainable to a human-in-the-loop as interpretable models and identify two possible reasons: (1) **that there are weaknesses in the intrinsic explainability of interpretable models** and (2) that **more information about a model may confuse users**, leading them to perform worse on objectively measurable tasks.

XAI for Community Healthcare

Motivation: How can XAI tools help support community healthcare workers in the Global South?

Hypothesis: By integrating interactive visual affordances in the risk prediction mobile application, community-healthcare workers are able to better understand what this AI does and how best to operate it.

Takeaways

1. We need regulation.
2. Data science ethics courses should be required in universities.
3. In lieu of cogent regulatory policy, technology professionals have an obligation to hold these technological systems and the people that build them accountable.
4. Many of the problems are socio-technical and cannot be “solved” with technology alone.
5. Some problems shouldn’t involve technology, part of our job is to say “no”.
6. Interdisciplinary collaboration is key. Teams should include collaborators from a variety of disciplines, backgrounds, expertise, and most importantly, where possible include end-users and those most affected by these systems.
7. Incorporate codes of ethics, frameworks, and systems where applicable.

Questions?

irs24@cornell.edu