The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

We are in the midst of a global trend to regulate the use of algorithms, artificial intelligence (AI) and automated decision systems (ADS). As reported by the *One Hundred Year Study on Artificial Intelligence*[1]: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy." Major cities, states and national governments are establishing task forces, passing laws and issuing guidelines about responsible development and use of technology, often starting with its use in government itself, where there is, at least in theory, less friction between organizational goals and societal values.

In the United States, New York City has made a public commitment to opening the black box of the government's use of technology: in 2018, an ADS task force was convened, the first of such in the nation, and charged with providing recommendations to New York City's government agencies for how to become transparent and accountable in their use of ADS. In a 2019 report, the task force recommended using ADS where they are beneficial, reduce potential harm and promote fairness, equity, accountability and transparency[2]. Can these principles become policy in the face of the apparent lack of trust in the government's ability to manage AI in the interest of the public? We argue that overcoming this mistrust hinges on our ability to engage in substantive multi-stakeholder conversations around ADS, bringing with it the imperative of interpretability — allowing humans to understand and, if necessary, contest the computational process and its outcomes.

Remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm[3], and how we can empower humans to exercise agency — to adopt or challenge an algorithmic decision. Consider, for example, scoring and ranking — data-driven algorithms that prioritize entities such as individuals, schools, or products and services. These algorithms may be used to determine credit worthiness, and desirability for college admissions or employment. Scoring and ranking are as ubiquitous and powerful as they are opaque. Despite their importance, members of the public often know little about why one person is ranked higher than another by a résumé screening or a credit scoring tool, how the ranking process is designed and whether its results can be trusted.

As an interdisciplinary team of scientists in computer science and social psychology, we propose a framework that forms connections between interpretability and trust, and develops actionable explanations for a diversity of stakeholders, recognizing their unique perspectives and needs. We focus on three questions (Box 1) about making machines interpretable: (1) what are we explaining, (2) to whom are we explaining and for what purpose, and (3) how do we know that an explanation is effective? By asking — and charting the path towards answering — these questions, we can promote greater trust in algorithms, and improve fairness and efficiency of algorithm-assisted decision making.

Box 1 | Research questions

- **What are we explaining?** Do people trust algorithms more or less than they would trust an individual making the same decisions? What are the perceived trade-offs between data disclosure and the privacy of individuals whose data are being analysed, in the context of interpretability? Which potential sources of bias are most likely to trigger distrust in algorithms? What is the relationship between the perceptions about a dataset's fitness for use and the overall trust in the algorithmic system?
- **To whom are we explaining and why?** How do group identities shape perceptions about algorithms? Do people lose trust in algorithmic decisions when they learn that outcomes produce disparities? Is this only the case when these disparities harm their in-group? Are people more likely to see algorithms as biased if members of their own group were not involved in algorithm construction? What kinds of transparency will promote trust, and when will transparency decrease trust? Do people trust the moral cognition embedded within algorithms? Does this apply to some domains (for example, pragmatic decisions, such as clothes shopping) more than others (for example, moral domains, such as criminal sentencing)? Are certain decisions taboo to delegate to algorithms (for example, religious advice)?
- **Are explanations effective?** Do people understand the label? What kinds of explanations allow individuals to exercise agency: make informed decisions, modify their behaviour in light of the information, or challenge the results of the algorithmic process? Does the nutrition label help create trust? Can the creation of nutrition labels lead programmers to alter the algorithm?

What are we explaining?

Existing legal and regulatory frameworks, such as the US's Fair Credit Reporting Act and the EU's General Data Protection Regulation, differentiate between two kinds of explanations. The first concerns the outcome: what are the results for an individual, a demographic group or the population as a whole? The second concerns the logic behind the decision-making process: what features help an individual or group get a higher score, or, more generally, what are the rules by which the score is computed? Selbst and Barocas[4] argue for an additional kind of an explanation that considers the justification: why are the rules what they are? Much has been written about explaining outcomes[5], so we focus on explaining and justifying the process.

Procedural justice aims to ensure that algorithms are perceived as fair and legitimate. Research demonstrates that, as long as a process is seen as fair, people will accept outcomes that may not benefit them. This finding is supported in numerous domains, including hiring and employment, legal dispute resolution and citizen reactions to police and political leaders[6], and it remains relevant when decisions are made with the assistance of algorithms. A recent lawsuit against Harvard University, filed by Students for Fair Admissions, stems, at least in part, from a lack of transparency and sense of procedural justice among some applicant groups. Similar allegations of injustice were levelled against the New York City Department of Education when only seven black students (out of 895 spots) had been admitted into New York's most selective high school[7]. To increase feelings of procedural justice, interests of different stakeholders should be taken into account when building and evaluating algorithms, prior to observing any outcomes[8].

Data transparency is a dimension of explainability unique to algorithm-assisted — rather than purely human — decision making. In applications involving predictive analytics, data are used to customize generic algorithms for specific situations: algorithms are trained using data. The same algorithm may exhibit radically different behaviour — making different predictions and different kinds of mistakes — when trained on two different datasets. Without access to the training data, it is impossible to know how an algorithm will behave. For example, predictive policing algorithms often reproduce the systemic historical bias towards poor or black neighbourhoods because of their reliance on historical policing data. This can amplify historical patterns of discrimination, rather than provide insight into crime patterns[9]. Transparency of the algorithm alone is insufficient to understand and counteract these particular errors.

The requirement for data transparency is in keeping with the justification dimension of interpretability: if the rules derived by the algorithm are due to the data on which it was trained, then justifying these rules must entail explaining the rationale behind the data selection and collection process. Why was this particular dataset used, or not used? It is also important to make statistical properties of the data available and interpretable, along with the methodology that was used to produce it, substantiating the fitness for use of the data for the task at hand[10].

To whom are we explaining and why?

Different stakeholder groups take on distinct roles in algorithm-assisted decision making, and so have different interpretability requirements. While much important work focuses on interpretability for computing professionals[5] — those who design, develop and test technical solutions — less is known about the interpretability needs of others. These include members of the public who are affected by algorithmic decisions: doctors, judges and college admissions officers who make — and take responsibility for — these decisions; and auditors, policymakers and regulators who assess the systems' legal compliance and alignment with societal norms.

Social identity is key to understanding the values, beliefs and interpretations of the world held by members of a group[11]. People tend to trust in-group members more than out-group members, and if their group is not represented during decision making, they will not trust the system to make judgments that are in their best interest[12]. Numerous identities may play a critical role in how algorithms are evaluated and whether the results they produced should be trusted. One recent case that highlights the contentious role of group identity is the effect of political ideology on search engines and news feeds. Liberal and conservative politicians both demand that technology platforms like Facebook become 'neutral'[13], and have repeatedly criticized Google for embedding bias into its algorithms[14]. In this case, the identity of the programmers can overshadow more central features, such as the accuracy of the news source.

Moral cognition is concerned with how people determine whether an action or outcome is morally right or wrong. Moral cognition is influenced by intuitions, and therefore is often inconsistent with reasoning[15]. A large body of evidence suggests that people evaluate decisions made by humans differently from those made by computers (although this may be changing, see ref. [16]); as such, they may be uncomfortable delegating certain types of decisions to algorithms. Consider the case of driverless vehicles. Even though people approve of autonomous vehicles that might sacrifice passengers to save a larger number of non-passengers, they would prefer not to ride in such vehicles[17]. Thus, utilitarian algorithms designed to minimize net harm may ironically increase harm by making objectively safer technology aversive to consumers. Failing to understand how people evaluate the moral programming of algorithms could thus unwittingly cause harm to large groups of people. The problem is compounded by the fact that moral preferences for driverless vehicles vary dramatically across cultures[18]. Solving these sorts of problems will require an understanding of social dilemmas, since self-interest might come directly in conflict with collective interest[19].

Are explanations effective?

A promising approach for interpretability is to develop labels for data and models analogous to nutritional labels used in the food industry, where simple, standard labels convey information about the ingredients and nutritional value. Nutritional labels are designed to inform specific decisions rather than provide exhaustive information. Proposals for hand-designed labels for data, models or both have been suggested in the literature[20,21]. We advocate instead for generating such labels automatically or semi-automatically as a part of the computational process itself, embodying the paradigm of interpretability by design[10,22].

We expect that data and model labels will inform different design choices by computer scientists and data scientists who implement algorithms and deploy them in complex multi-step decision-making processes. These processes typically use a combination of proprietary and third-party algorithms that may encode hidden assumptions, and rely on datasets that are often repurposed (used outside of the original context for which they were intended). Labels will help determine the 'fitness for use' of a given model or dataset, and assess the methodology that was used to produce it.

Information disclosure does not always have the intended effect. For instance, nutritional and calorie labelling for food are in broad use today. However, the information conveyed in the labels does not always affect calorie consumption[23]. A plausible explanation is that "When comparing a $3 Big Mac at 540 calories with a similarly priced chicken sandwich with 360 calories, the financially strapped consumer […] may well conclude that the Big Mac is a better deal in terms of calories per dollar"[23]. It is therefore important to understand, with the help of experimental studies, what kinds of disclosure are effective, and for what purpose.

Conclusion

The integration of expertise from behavioural science and computer science is essential to making algorithmic systems interpretable by a wide range of stakeholders, allowing people to exercise agency and ultimately building trust. Individuals and groups who distrust algorithms may be less likely to harness the potential benefits of new technology, and, in this sense, interpretability intimately relates to equity. Education is an integral part of making explanations effective. Recent studies found that individuals who are more familiar with AI fear it less, and are more optimistic about its potential societal impacts[24]. We share this cautious optimism, but predicate it on helping different stakeholders move beyond the extremes of unbounded techno-optimism and techno-criticism, and into a nuanced and productive conversation about the role of technology in society. ❐

Julia Stoyanovich[1,2], Jay J. Van Bavel[3,4] and Tessa V. West[3]

[1]Department of Computer Science and Engineering, Tandon School of Engineering, New York University, New York, NY, USA. [2]Center for Data Science, New York University, New York, NY, USA. [3]Department of Psychology, College of Arts and Sciences, New York University, New York, NY, USA. [4]Center for Neural Science, New York University, New York, NY, USA.
e-mail: stoyanovich@nyu.edu; jay.vanbavel@nyu.edu; tessa.west@nyu.edu

Published online: 13 April 2020
https://doi.org/10.1038/s42256-020-0171-8

References

1. Stone, P. et al. *One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel* (Stanford Univ., 2016).
2. *New York City Automated Decision Systems Task Force Report* (NYC.gov, 2019).
3. Rovatsos, M. *Nat. Mach. Intell* **1**, 497–498 (2019).
4. Selbst, A. & Barocas, S. *Fordham L. Rev* **87**, 1085–1139 (2018).
5. Guidotti, R. et al. *ACM Comput. Surv.* **51**, 93 (2019).
6. Bobocel D. R., Gosse, L. *The Oxford Handbook of Justice in the Workplace* (Oxford Univ. Press, 2015).
7. Shapiro, E. *The New York Times* http://www.nytimes.com/2019/03/18/nyregion/black-students-nyc-high-schools.html (2019).
8. Lee, M. K. et al. *Proc. ACM CHI* **3**, 1–35 (2019).
9. Lum, K. & Isaac, W. *Significance* **13**, 14–19 (2016).
10. Stoyanovich, J. & Howe, B. *IEEE Data Eng. Bull* **42**, 13–23 (2019).
11. Van Bavel, J. J. & Pereira, A. *Trends Cogn. Sci.* **22**, 213–224 (2018).
12. Alfano, M. & Huijts, N. *Handbook of Trust and Philosophy* (Routledge, 2019).
13. Feiner, L. *CBNC* https://www.cnbc.com/2019/08/20/republican-report-of-facebook-anti-conservative-bias-suggests-changes.html (2019).
14. Schwartz, O. *The Guardian* https://www.theguardian.com/technology/2018/dec/04/google-facebook-anti-conservative-bias-claims (2018).
15. Haidt, J. *Science* **316**, 998–1002 (2007).
16. Bigman, Y. E., Waytz, A., Alterovitz, R. & Gray, K. *Trends Cogn. Sci.* **23**, 365–368 (2019).
17. Bonnefon, J. F., Shariff, A. & Rahwan, I. *Science* **352**, 1573–1576 (2016).
18. Awad, E. et al. *Nature* **563**, 59–64 (2018).
19. Van Lange, P. A. M., Joireman, J., Parks, C. D. & Van Dijk, E. *Organ. Behav. Hum. Decis. Process* **120**, 125–141 (2013).
20. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. Preprint at https://arxiv.org/abs/1805.03677 (2018).
21. Mitchell, M. et al. in *Proc. ACM FAT** 220–229 (2019).
22. Yang, K. et al. in *Proc. ACM SIGMOD* 1773–1776 (2018).
23. Loewenstein, G. *Am. J. Clin. Nutr* **93**, 679–680 (2011).
24. Zhang, B. & Dafoe, A. *Artificial Intelligence: American Attitudes and Trends* (Center for the Governance of AI, 2019).

Competing interests

The authors declare no competing interests.