INVESTIGATION · DIGITAL INVESTIGATION

Investigating Algorithmic Black Boxes: Generating Knowledge on AI Decision-Making

Algorithms now help decide who gets a loan, a job interview, medical care or bail, yet their workings are often hidden. This article explains how researchers, journalists and auditors investigate these "black boxes" from the outside and from the inside, and what such investigations have found.

ModQu research article · 752 words · 8 references (APA 7th edition)

Automated systems increasingly make or shape decisions that matter: which job applicants are shortlisted, which patients are offered extra care, which defendants are considered high risk and which content people see online. Many of these systems are "black boxes". Their code and data are secret, protected as trade secrets, or so complex that even their developers cannot fully explain an individual result. Pasquale (2015) argues that this secrecy shifts power to the organisations that build and use such systems, and makes independent investigation essential.

Investigating an algorithm means generating knowledge about how it behaves, whom it affects and whether its results are fair and accurate, often without access to its internal workings.

Methods of algorithm auditing

Sandvig et al. (2014) set out five broad designs for auditing algorithms, adapted from earlier audit studies of discrimination in housing and employment.

Five ways to audit an algorithm

Each design trades access, cost and legal risk differently

	How it works	Main limitation
Code audit	Experts examine the source code and data	Rarely permitted; code alone may not show behaviour
User survey audit	Ask users what results they received	Relies on recall and self-reporting
Scraping audit	Collect large numbers of outputs automatically	May breach terms of service
Sock puppet audit	Create test profiles that differ in one feature	Fake accounts may be detected or prohibited
Crowdsourced audit	Volunteers share their real results	Volunteers may not be representative

Based on Sandvig et al. (2014) ModQu.com

Figure 1. Five audit designs for investigating algorithms from outside or inside. Investigators often combine several to check one method's weaknesses against another's. Diagram by ModQu.

The sock puppet design is the most direct test of discrimination: two test profiles identical except for one feature, such as gender or postcode, are submitted to the system, and differences in the results are attributed to that feature. Where investigators can obtain the system's outputs together with real outcomes, they can also measure accuracy and error rates for different groups.

What investigations have found

Several investigations have shaped public understanding. Angwin et al. (2016), journalists at ProPublica, obtained risk scores produced by a commercial tool used in criminal courts in Florida and compared them with defendants' later records. They reported that Black defendants who did not reoffend were almost twice as likely as white defendants to be wrongly labelled high risk. The developer disputed the analysis, and the debate showed that different definitions of fairness cannot all be satisfied at once when groups have different underlying rates.

Buolamwini and Gebru (2018) tested commercial facial analysis systems on a balanced set of faces and found that error rates in classifying gender were far higher for darker-skinned women than for lighter-skinned men. Obermeyer et al. (2019) studied an algorithm widely used by United States health systems to identify patients for extra care. Because it predicted future health costs rather than health needs, and less money had historically been spent on Black patients with the same needs, it systematically underestimated how ill Black patients were. With access to the underlying data, the researchers could identify the exact cause and show how changing the prediction target reduced the bias.

Looking inside

When investigators have access to a model but not a simple explanation of it, explanation methods can help. Ribeiro et al. (2016) introduced a technique that explains an individual prediction by testing how the model's output changes when the input is slightly altered, showing which features drove the result. Such methods are approximations and can themselves mislead, so they are best used alongside outcome testing.

Organisations can also investigate their own systems. Raji et al. (2020) propose a framework for internal audits that runs through a system's whole development, with documented checks at each stage, so that problems are found before deployment rather than by outsiders afterwards. Regulation is adding to this: the European Union's Artificial Intelligence Act requires providers of high-risk systems to keep documentation and logs and to allow oversight (Regulation (EU) 2024/1689, 2024).

Practical challenges

Algorithm investigations face particular difficulties. Systems change frequently, so a result found one month may not hold the next, and investigators should record the date and version of what they tested. Many systems are personalised, so two users may see different results for reasons unrelated to the feature being tested; large samples and careful controls are needed to separate a real effect from noise. Access is often contested: platforms may block automated data collection, and researchers have faced legal threats under terms of service and computer misuse laws. Finally, the investigator must decide what counts as a problem. A system can be accurate overall and still make far more errors for one group, and different measures of fairness can point in opposite directions. Stating the chosen measure, and why it suits the decision being made, is part of reporting the finding honestly.

Conclusion

Algorithmic black boxes can be investigated. Outside audits test what systems do, access to data reveals why, and explanation methods give partial views of how. The strongest investigations combine these approaches, state clearly which definition of fairness they apply, and give the system's owners a chance to respond.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union, L series*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014, May). *Auditing algorithms: Research methods for detecting discrimination on internet platforms* [Paper presentation]. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, preconference of the 64th Annual Meeting of the International Communication Association, Seattle, WA, United States.