Collaborative auditing: a challenge to secrecy in artificial intelligence

Adam Roses Wight
awight@wikimedia.org
Wikimedia Foundation

ABSTRACT

We introduce a system for detecting and mitigating bias in artificial intelligence, called Judgment and Dialogue Engine (JADE). Expanding on current auditing approaches, JADE adds a dimension of human communication and collaborative decision-making between the auditors.

Collaborative auditing challenges the hegemony of opaque AIs, offering a platform for holding algorithms accountable and building community among AI investigators. Audit results will be analyzed and fed back as AI training data in order to iteratively uncover and mitigate biases. Our hope is that JADE will improve AI fairness and performance, and to show that transparent and collaborative auditing should be insisted upon as an intervention in the public interest.

1 INTRODUCTION

Artificial intelligence has become indispensable to the largest digital businesses, from search engines and email hosts, to shopping, mortgage, insurance, and law enforcement services. AI is used to provide quality control, curation, and analytics at massive scales, rather than having humans do the work. Human decision-making has been replaced by its emulation, interpolated from previously recorded decisions, and during this process is heavily mediated by algorithm owners and designers. These last two categories, the technocratic elite, are largely unaccountable for the quality and social impact of the AIs they deploy. In commercial settings, the only constant, guiding force in AI design must be the profit motive, in absence of any other constraints.[1]

Corporate ownership of powerful AIs which affect people's life chances (mortgages, law enforcement) and exacerbate existing social ills (redlining, racial profiling) is beyond troubling, because we have no oversight or democratic accountability by which we can fight back. AIs operated by companies or governments are closed by default. Companies rely on this information disparity, hiding secrets to protect their profit margins, and have shown no interest in public review of algorithms or data.

All AIs learn from humans, and will replicate our prejudices or group polarization. Just as it's difficult for humans to be self-critical and diagnose our own prejudices, AIs cannot report their own built-in biases, and owners hardly ever give an account of side-effects, even if they're known.[2] Playing with fire, in the dark and uncertain of the consequences, AI practitioners really are Ali Rahimi's modern alchemists.[3]

Exploring and measuring biases is imperative, and researchers have developed methods which work even without privileged access to the AI internals. Christian Sandvig (2014) [**?**] illustrates these methods, but also warns that high-quality techniques for independent auditing are prohibited by most sites' terms of service, and can even trigger felonies under the United States' Computer Fraud and Abuse Act.

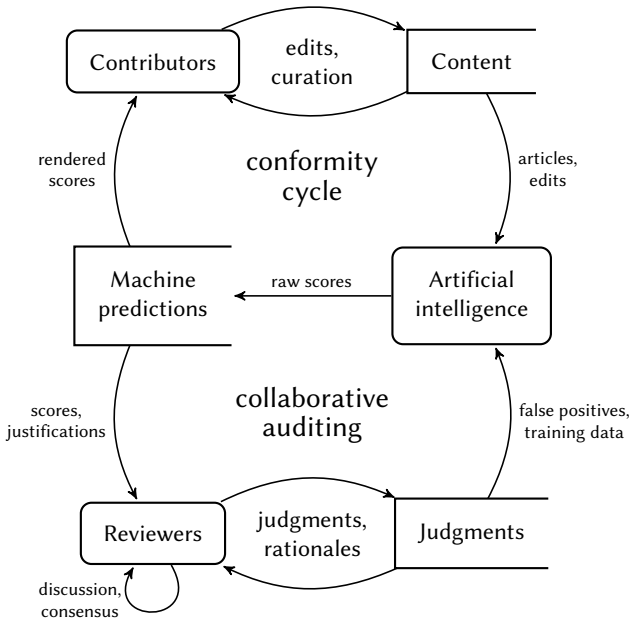

Figure 1: AI feedback with auditing (JADE in Wikipedia)

For the moment, AI engineers must audit their own systems. The emerging industry standard is well behind the state of the art[4], and will usually present a sample of users with a feedback dialog, posing a question like "Do you agree with this recommendation?". In these systems, the lucky respondent may be able to leave a comment, but that's the beginning and end of any interaction. Reviewers are alienated from one another, cannot read or discuss other users' responses, and have no agency. They are reduced to mere processes gathering and corroborating data in silent redundancy.

2 JUDGMENT AND DIALOGUE ENGINE

We're introducing Judgment and Dialogue Engine in order to audit the AIs used on Wikipedia and its sister sites, and to humanize this process through transparency and consensus, already strong traditions among these communities. In figure **??**, JADE is the lower feedback loop involving reviewers, shown here in relation to the upper, "conformity cycle" feedback loop in which AI will tend to confirm what editors already believe.

In practice, reviewers using JADE will be reading through wiki content and looking at machine predictions, and will begin a JADE session either to flag an incorrect prediction, or simply

[1] See Dodge v. Ford (1919) for corporate responsibility in a nutshell.
[2] Solon Barocas (2014) [**?**] gives an overview of the types of bias and in section 2.5 demonstrates a positive feedback loop between AI predictions and iterations on the sample frame.
[3] http://www.argmin.net/2017/12/05/kitchen-sinks/
[4] Stephanie Rosenthal and Anind Dey (2010) [**?**] optimize the choice of data to include in rich feedback.

WebSci'18, May 2018, Amsterdam, NL
2018.

to record their judgments about wiki content. JADE will provide open access to the data created by its reviewers. Crucially, reviewers are able to read one another's judgments, discuss any disagreements, and make changes to the recorded judgment.

The best outcome would be a disintermediation that takes power away from the algorithm designers and gives it to the reviewers. In this scenario, the data from reviewers would feed back into AI training without mediation by AI technicians. The worst outcome would be that we enable the formation of a small group which undergoes polarization, and pushes the AI toward even deeper biases.

If our society can agree that AIs must be regulated by a "right to audit"[5], then JADE may serve as an example for how to accomplish this auditing in a pro-social environment.

REFERENCES

[1] Barocas, S. Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining* (2014).

[2] Rosenthal, S. L., and Dey, A. K. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2010), IUI '10, ACM, pp. 259–268.

[3] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).

[5] A precedent for this sort of regulation is the U.S. Federal Aviation Administration's Aircraft Certification Service, which reviews and certifies all software and hardware to be used in aircraft.