Collaborative auditing: a challenge to closed AI design

Adam 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). This system expands on current auditing approaches by adding a dimension of human communication and collaborative decision-making between the reviewers.

JADE challenges the hegemony of opaque AIs, providing a platform for holding algorithms accountable and building a community between AI investigators. The outputs can be analyzed, and fed back into AI training data in order to iteratively uncover and mitigate biases. Our hope is to show that JADE improves AI performance, and that transparent auditing should be an urgent and widespread intervention in the public interest.

ACM Reference Format:
Adam Wight. 2018. Collaborative auditing: a challenge to closed AI design. In *Proceedings of Understanding the political economy of digital technology (WebSci'18).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 scale, rather than having humans do the work. Human decision-making has been replaced by an emulation, interpolated from previously recorded decisions, and is mediated by algorithm owners and designers. This last category, the technocracy, is largely unaccountable for the quality and social impact of the AIs they deploy. In a commercial setting, the only constant, guiding force in AI design is the profit motive.

The ownership of powerful AIs which can change people's life chances (mortgages, law enforcement) and exacerbate existing social ills (redlining, racial profiling) is troubling. AIs operated by companies or governments are currently closed by default. Companies rely on an information disparity to protect profit, so have no incentive to make their algorithms or data available for public review.

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, an AI cannot detect its built-in biases. There are research methods available to explore and measure this bias, and some methods can be used independently, without requiring privileged access to the AI internals. Christian Sandvig (2014) illustrates some of these methods, however these are usually outlawed by site terms of service, and can even trigger felonies in the United States under the Computer Fraud and Abuse Act.

WebSci'18, May 2018, Amsterdam, NL

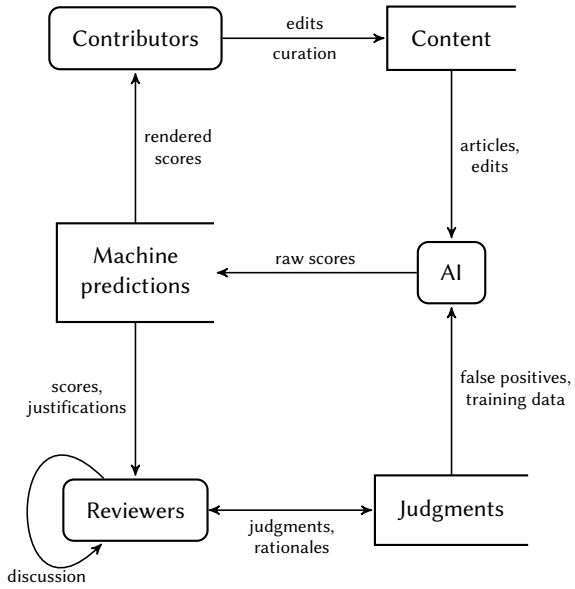

Figure 1: AI feedback with auditing (in Wikipedia)

We're introducing Judgment and Dialogue Engine in order to audit the AIs which are used on Wikipedia and related sites. In figure 1, JADE is the lower feedback loop involving "reviewers". In practice, the reviewers will be looking at wiki content and at machine predictions, and will begin a JADE session when they feel that the prediction is incorrect, or simply to record their judgment about wiki content. The system will provide open access to the data created by reviewers. Crucially, the reviewers are able to read each other's judgments, discuss any disagreements, and make changes to the recorded judgment.

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

If our society can agree that AIs must be regulated[1] by a "right to audit", then JADE may serve as an example for how to accomplish this auditing in a pro-social structure.

REFERENCES

[1] 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).

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