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 may help establish transparent and collaborative auditing as an urgent intervention we must make in the general 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Human decision-making has been replaced by its emulation, interpolated from previously recorded decisions, and during this process is heavily mediated by algorithm owners and designers.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 are the elite of the technocratic hierarchy, and although every rung of that ladder shares some responsibility, they all go largely unaccountable for the quality and social impact of the AIs they deploy. In commercial settings, the only constant, guiding force in the absence of other constraints must be the profit motive.[1] It will take a fight to bring this industry to account.

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 in the USA have no oversight or democratic accountability by which we can fight back yet. AIs operated by companies or governments are closed by default. Companies rely on this information disparity, hiding secrets to protect 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 are blind to their own built-in biases and cannot give an accurate reckoning, and owners hardly ever admit to side-effects they become aware of.[2]

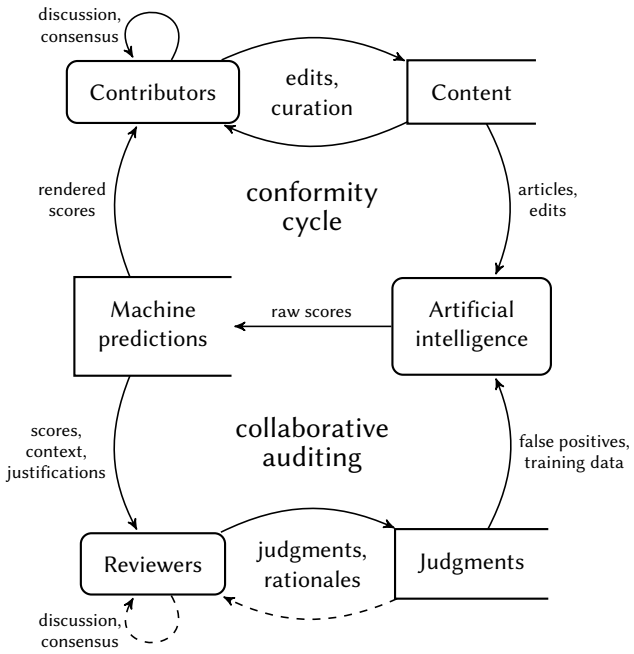

Figure 1: AI feedback with auditing, in Wikipedia. Dashed lines are the data flows unique to JADE.

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 will serve even without privileged access to AI internals. Christian Sandvig (2014) [3] illustrates these methods, but also warns that high-quality information gathering for an independent audit is prohibited by most sites' terms of service, and can even trigger felonies under the United States' Computer Fraud and Abuse Act.

For the moment, we leave it to AI engineers to audit their own systems. The emerging industry standard for audit by users falls well behind the state of the art for rich feedback[4], and will present some sample of users with a dull feedback dialog, posing a question like "Do you agree with this recommendation?". In these systems, the respondent may have the option to leave a comment, but that's the beginning and end of any interaction.[5] 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.

[1] See Dodge v. Ford (1919) for the corporate mandate in a nutshell.

[2] Solon Barocas (2014) [1] 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) [2] optimize the choice of data to include in rich feedback.

[5] For example, Google Ideas's Perspective API https://github.com/conversationai/perspectiveapi/blob/master/api_reference.md#sending-feedback-suggestcommentscore is ahead of the curve only because it allows a client response to include rationales as in Sharma (2015) [4].

WebSci'18 workshop: Understanding the political economy of digital technology, May 2018, Amsterdam, NL
2018.

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 1, JADE is the lower feedback loop involving reviewers, shown here in relation to the upper, "conformity cycle" feedback loop in which AI helps to confirm what editors already believe.

Note in the figure that the bottom, auditing cycle is a mirror image of Wikipedia's existing cooperative workflow, once we add the JADE improvements. We're hoping that established norms of consensus and boldly editing each other's content, are a nautral way to support and accelerate discussion cycles among the reviewers. We believe these principles can be applied to many wiki workflows.

In practice, reviewers using JADE-backed tools 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 to record their personal judgments about wiki content.

JADE provides open access to all data entered by reviewers, each judgment is a wiki page. Crucially, reviewers are welcome to read one another's judgments, discuss their disagreements, and make changes to the judgment of record.

The best outcome would be that JADE performs a disintermediation, taking power away from the algorithm designers and giving it to the reviewers. In this scenario, the data from reviewers would feed back into AI training with minimal mediation by AI technicians. The worst outcome would be that we enable the formation of a new, 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"[6], 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).

[4] Sharma, M., Zhuang, D., and Bilgic, M. Active learning with rationales for text classification. In *North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (2015).

[6] 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.