Toward User-Driven Algorithm Auditing: Investigating users’ strategies for uncovering harmful algorithmic behavior

Alicia DeVos (adevos@cmu.edu) with Aditi Dhabalia, Hong Shen, Ken Holstein, & Motahhare Eslami
Human-Computer Interaction Institute, Carnegie Mellon University

INTRODUCTION

Efforts to detect biased, discriminatory, or otherwise harmful behaviors in algorithms — known as algorithm audits — are typically led by experts such as industry practitioners, researchers, activists, or government agencies. However, in recent years we’ve seen an interesting phenomenon appear: there have been a number of cases in which everyday users detect, investigated, & raised awareness about potential issues in algorithmic systems where experts failed to do so. 

For example, one heavily publicized case involved Twitter users discovering bias in Twitter’s image cropping algorithm. Images of this & some of the other cases that have received news coverage are shown here.

This behavior has been described as everyday algorithm auditing, a term we’ve coined to define as the ways everyday users detect, understand, and/or interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems.

Though this past work helped us to characterize the phenomenon & dynamics that these everyday algorithm audits have historically followed & to begin to understand some audit failures in having impact where others failed, we still don’t know much about how people actually go about searching for & making sense of these behaviors, let alone how they sometimes are able to be so effective.

We are interested in understanding how we might design tools & platforms to help these kinds of user-driven auditing behaviors be more effective & increase their potential for mitigating algorithmic behavior.

RESULTS

Our process model adapts existing models of information search & sensemaking such as that of Profili & Card to help understand how people find & make sense of biased & harmful algorithmic behaviors. Our model captures the process by which people find & make sense of biased & harmful algorithmic behavior. Our model adapts existing models of information search & sensemaking to help us understand this phenomenon.

In our open coding, 3 connected high-level stages of people’s bias search & sensemaking process emerged: search inspiration, sensemaking, & remediation.

Search inspiration encompasses the ways that participants came up with ideas for where to look for potential biases in the systems. In one search inspiration theme, patterns often helped signal to participants the presence of potential biases, inspiring them to look closer: e.g., one participant noted that when searching for images of black men, the website would immediately display images of black men (P22) many times in the autocomplete suggestions, which prompted them to look deeper.

Sensemaking, the second stage, involves how participants understood & evaluated algorithmic behaviors that might be harmful. In one sensemaking theme, participants evaluated bias based on their perceptions of reality: e.g., one participant’s notion of a search result was that “if you look at the surface there are indeed more African Americans being arrested or reported violent … but this is institutional bias, right?” (P14).

Finally, remediation covers participant actions & desires to mitigate the ramifications of harmful algorithmic bias. In one remediation theme, participants described wanting to foster awareness of bias: e.g., some participants thought that users should be “taught how to use the technology … what are the do’s and don’ts, what are the ethics” (P18).

DISCUSSION

Informed by the insights gleaned from user-driven algorithm auditing thus far, one valuable avenue for future work centers around designing & developing tools & platforms that support user-driven algorithm auditing. Our design implications indicate that people’s lived experiences have a large influence on the biases they are able to surface. So, we could assign people particular auditing roles based on their specific exposure & experience to boost future auditing efforts.

Second, a large question lies in determining what it might actually look like to fix the issues revealed by user-driven algorithm auditing. In many cases, there is no clear alignment as to what bias is or as to how to fix the issues. Especially when fixing biases requires a large, collaborative effort, it may not be possible, promoting awareness of algorithmic biases & other harmful behaviors can be especially valuable. User awareness helps users adapt their behaviors in & around algorithmic systems, in turn fostering user-driven auditing & supporting remediation of harmful issues by other means.

METHODS

We need to better understand these behaviors — how users currently surface & evaluate harmful issues, as well as how they might be able to more effectively design user-driven algorithm audits, amplifying what users already do well & assisting where they might need help.

So we asked the question: How do people find, make sense of, & evaluate potentially harmful algorithmic behavior? To investigate this question, we ran a 3-part study with a diverse set of 12 participants.

We first conducted think-aloud interviews that provided an opportunity to closely observe participants’ thought processes & ask clarifying questions in a relatively controlled setting.

In these, participants actively searched for & made sense of potentially harmful behaviors in the Google Images search engine. They were tasked with evaluating images as they were returned, as well as potentially contain bias as well as looking for new cases where image search might be biased.

Next, participants took part in a 14-day diary study to complement the interviews’ live observations by enabling observation of participants’ search & sensemaking strategies in a more naturalistic setting & more longitudinally, which allowed for the possibility of chance encounters with algorithmic systems.

In this, participants documented & reported potential issues they found through active searching or during the course of their day-to-day interactions with algorithmic systems.

Finally, since user-driven audits frequently consist of many users, we invited participants who completed the think-aloud interview & diary study to participate in group workshop sessions.

In these, participants worked together to evaluate & discuss some of the cases that were uncovered during the preceding diary study.

To analyze the data collected, we conducted bottom-up coding, a series of interpretation sessions, & a thematic analysis of our codes, from which we synthesized a process model to understand how people surface these harmful algorithmic behaviors.

In our open coding, 3 connected high-level stages of people’s bias search & sensemaking process emerged: search inspiration, sensemaking, & remediation.

Search inspiration encompasses the ways that participants came up with ideas for where to look for potential biases in the systems. In one search inspiration theme, patterns often helped signal to participants the presence of potential biases, inspiring them to look closer: e.g., one participant noted that when searching for images of black men, the website would immediately display images of black men (P22) many times in the autocomplete suggestions, which prompted them to look deeper.

Sensemaking, the second stage, involves how participants understood & evaluated algorithmic behaviors that might be harmful. In one sensemaking theme, participants evaluated bias based on their perceptions of reality: e.g., one participant’s notion of a search result was that “if you look at the surface there are indeed more African Americans being arrested or reported violent … but this is institutional bias, right?” (P14).

Finally, remediation covers participant actions & desires to mitigate the ramifications of harmful algorithmic bias. In one remediation theme, participants described wanting to foster awareness of bias: e.g., some participants thought that users should be “taught how to use the technology … what are the do’s and don’ts, what are the ethics” (P18).

These connections capture how people move through the process of bias search & sensemaking. Connected to each stage, we uncovered 2 common themes of the process across participants: knowledge & beliefs on the left and platform affordances on the right. Smaller arrows represent interactions between the process stages in the middle & the process influences.

Knowledge & beliefs encompass participants’ expectations & values, their folk theories of algorithms, their exposures & experiences, & their conceptions of bias. This poster will describe just one of these.

Participants’ prior knowledge & beliefs about biased & harmful algorithmic behavior to make sense of problematic behaviors they encountered, & they gained new knowledge & beliefs along the way. This is represented by the arrow leading from sensemaking into knowledge & beliefs. E.g., some participants’ perceptions of reality were shaped by algorithmic outputs, as they used search results as a basis for making inferences about reality: one participant said about a search, “I see a lot of the black men, not a lot of men” and used this to conclude, “I guess it’s predominantly a women’s field” (P10). These are some of the ways that knowledge & beliefs influence — and are influenced by — the process of bias search & sensemaking.

Shown on the right, participants capabilities to detect and make sense of such behaviors were enabled and guided by platform affordances. Platform affordances influenced all 3 stages of the process. E.g., sometimes Google’s autocomplete suggestions inspired participants to search terms they might not have otherwise. “As I was typing the word ‘women’, I could see the choices that other users were tapping up … so that quickly reoriented me to type my search differently” (P19). This is one of the ways that platform affordances influence the process of bias search & sensemaking.

Additionally, we need better design tools & platforms to help us understand how to augment & try to fix algorithmic systems when vs to kill & completely remove power from algorithmic systems.