Why do we need to be bots? What prevents society from detecting biases in recommendation systems

Tobias D. Krafft[1,2][0000−0002−3527−1092], Marc P. Hauer[1][0000−0002−1598−1812], and Katharina A. Zweig[1][0000−0002−4294−9017]

[1] Algorithm Accountability Lab, Technische Universität Kaiserslautern, Deutschland
[2] krafft@cs.uni-kl.de

Abstract. Concerns about social networks manipulating the (general) public opinion have become a recurring theme in recent years. Whether such an impact actually exists could so far only be tested to a very limited extent. Yet to guarantee the accountability of recommendation and information filtering systems, society needs to be able to determine whether they comply with ethical and legal requirements. This paper focuses on black box analyses as methods that are designed to systematically assess the performance of such systems, but that are, at the same time, not very intrusive. We describe the conditions that must be met to allow black box analyses of recommendation systems based on an application on Facebook's News Feed. While black box analyses have proven to be useful in the past, several barriers can easily get in the way, such as a limited possibility of automated account control, bot detection and bot inhibition. Drawing on the insights from our case study and the state of the art of research on algorithmic accountability, we formulate several policy demands that need to be met in order to allow monitoring of ADM systems for their compliance with social values.

Keywords: ADM systems · Recommendation systems · Black box analysis · Bot detection · Black box audit.

1 Introduction

With new machine learning techniques, more and more decision-making is delegated to machines. Therefore, the potential societal impact of so-called *algorithmic decision making systems* (**ADM systems**) and the information and power asymmetries that they can entail increases accordingly. Among these ADM systems, we count all recommendation systems that filter and rank news and messages on search engines and social media, such as news feeds, or time line curating systems. An entire field of research has emerged that deals with the question of how to safeguard an accountable use of such ADM systems. This field of *algorithmic accountability* [Diakopoulos, 2014] encompass various theoretical, technical, legal, and civil society approaches to contribute to a responsible and transparent handling of algorithmic decision processes.

An important method to monitor some characteristics of an opaque ADM system and to reduce information asymmetries without acquiring insight into the actual decision structures of the ADM system (for example via code audit) is called a *black box analysis* [Diakopoulos, 2014]. When conducting this form of testing, the appropriateness of the ADM system's results is assessed by running various experiments on the system (e.g., varying the input, observations under different conditions) without looking into its code, the implemented decision rules or its statistical model that produces the results.
In recent years, many people started using certain social media platforms as their primary, possibly even sole source of information and news. Websites like Facebook and Google have become power intermediaries between information sources and readers. This has led to a discussion about the rights and responsibilities associated with this position of large tech companies and information intermediaries [Dreyer and Schulz, 2019].
One aspect of this discussion is whether television channels which host a Facebook page can fulfill the principle of neutrality (§11/2 of the *German Interstate Broadcasting Agreement*[3]) on the platform Facebook given the nontransparent behavior of its News Feed algorithm. In collaboration with the *Rhein-Neckar Fernsehen* (RNF)[4] we probed the usefulness of black box analysis by examining whether Facebook displays an unduly polarizing or a balanced selection of news to the subscribers of the respective pages. In this analysis we were faced with two main obstacles, one of which was the very limited access to information and the other the quick banning of our fake accounts (bots). Our results suggest that while bot detection and selective bot inhibition are fundamental to a trustworthy usage of social network platforms, this case study shows that at least for some questions, society might need privileged access in form of fake accounts.

2 Black box analyses

Black box analyses as a form of systematic auditing allow for the evaluation of the overall appropriateness of an ADM system, including indirectly observable effects [Diakopoulos, 2014]. This requires access to interfaces through which the reviewing entity can observe the system as a black box and inspect which outputs are generated based on which inputs. Although this method does not enable a researcher to understand the ADM system completely - since it does not peak inside the black box - it can nevertheless reveal undesired behaviour of an ADM system's results - whether it was intended or not. Hence, this approach is rather superficial and hardly intrusive. It does not inspect the way in which the ADM system has been configured and *how* it produces outcomes, as it does not go beyond what is commonly called *instrumental* or *outcome accountability* [Patil et al., 2014].
Although the details of a black box analyses are highly application-specific, they roughly follow the same five steps (see Figure 1). Depending on the access to

[3] Rundfunkstaatsvertrag
[4] A regional, private television channel in Germany: https://www.rnf.de.

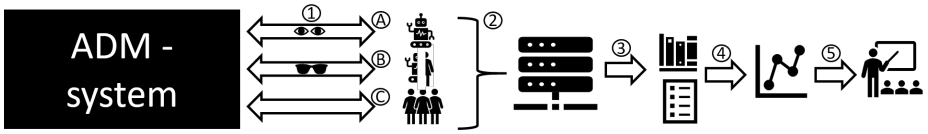

Fig. 1. Conceptualized process of a black box analysis. The numbers represent the different fields in which errors can occur.

the system, requests with previously defined input variables are automatically sent to the system and the results are collected (1 A). This audit form is called **Scraping Audit** [Sandvig et al., 2014]. Researchers issue queries to a platform, observe its reactions and make a statistical evaluation of them. These queries might be very simple and can be issued by either using an application programming interface (*API*) or a browser control system like Selenium[5]. Either way, the automated access does not try to impersonate human behaviour, so that queries can be submitted at a very high frequency and/or can act in a for human atypical manner, as long as the interface allows this [Sandvig et al., 2014]. If this form of auditing does not work, e.g. because the specific usage behaviour is part of the required input variables (like human typing or clicking), the automated query must pretend to be a real user before or during the data collection. This can be done, for example, by simulating an organic user behaviour (1 B). This form of audit, in which a computer imitates human behaviour, is called **Sock Puppet** or **Bot Audit** [Sandvig et al., 2014]. It is similar to a Scraping Audit, but aims at impersonating realistic user behavior. By simulating human interaction, including personal characteristics, the behaviour of the platform towards its actual users can be captured. Since actual human behaviour is part of the input to the ADM system, there is normally no API over which these inputs can be submitted. A third approach is the so called **Crowdsourced Audit** [Sandvig et al., 2014] which makes use of actual users of a platform, by either letting them enter queries, or by interposing an algorithm that pretends to be the actual user that is logged in[6]. Finding enough users that agree to participate is a major obstacle of this kind of audit— as is the possible self-selection of these users into the auditing procedure[7]. No matter which form of auditing is chosen, it is always followed by a central data collection (2) as well as a processing of the collected results which are fed into data cleaning (3). Data analysis methods (4) can then only be applied to structured and verified data sets. The last step is the presentation of the results of the data analysis (5). All these steps involve their

[5] https://selenium.dev/

[6] Even when no login is needed, this approach yields a great advantage in certain cases, for example, when geospatial data like the IP address might be relevant [Krafft et al., 2019].

[7] Self-selection refers to the self-enrollment in these kinds of studies. It almost always biases the sample such that it is not representative of all users of a system.

own challenges. This paper focuses on the intricacies and challenges encountered in the first stage.

3 Case study Facebook

In a private conversation, a member of the private TV station RNF in Germany told us that the followers of their Facebook account complained about the selection of news issued by the RNF in their time line. Some of them expressed the feeling that they only got to see news of the "Blood & Crime" type, while the RNF has a wide range of regional and global news, from the weather forecast over municipal to police and societal news.
The first question was whether and, if so, to which extent followers of RNF's page see only a part of the news content. The second question was how such a selection developed in time: Did followers at the beginning see a fair sample of all news, maybe weighted by the frequency of the corresponding category? Was it then influenced and more selective by the way in which the followers interacted with the content? A follower who is interested in all categories but—inadvertently—clicks more on the "Blood & Crime" news might induce a positive feedback cycle with the recommendation system that increasingly prioritized those news and suppresses news from other categories.

3.1 Page owner perspective

We obtained full access to the Facebook account of the RNF, but still could not answer any of the questions meniontedabove with the information provided directly to the account. The type of information provided by Facebook is highly aggregated and does not allow to track the delivery of single news items to followers. It was thus clear, that we needed to do a black-box analysis.

3.2 Appropriate forms of audit

For the black box analysis of Facebook, we examined the three audit approaches previously introduced, and evaluated their feasibility. In general, we wanted to see whether the filtering and ranking of the news items by the personalized recommendation system would change the fraction of news in each category in a user's News Feed, e.g., towards a heavy fraction of "Blood & Crime" news. At the time of the study, there was no API available, which we could have used to address that question. Extended API access can be granted by Facebook, for example to support the solution of research tasks, but a corresponding request remained unanswered. We also ruled out a Crowdsourced Audit in which we would ask users to open up their accounts to us. We would then have been able to scrape the RNF news from their News Feed, however, the privacy problem would have been massive without any possibility of filtering only those news items from the otherwise very personal stream of messages. For the same reason, crowdsourcing of only political ads in Facebook is impossible today which heavily

impedes the analysis of how political elections might be influenced by those ads. In general, any Crowdsourced Audit is highly problematic on Facebook as long as there are no fine-grained filtering approaches that enable a selective access to a user's News Feed.
Based on these considerations, we concluded that: To respect the privacy of real users and without access to information via the page owner's account or a suitable API, we needed to implement a Sock Puppet Audit. To make initial validations in a pre-study, we generated 30 fake accounts by hand based on email addresses from various providers. Each account has been manually set up to follow only the Facebook group of the RNF. Every day, our software logged in with each of the accounts, scrolled through the respective News Feed and saved the displayed posts in a database. The software was developed in a way that the behavior was as realistic as possible to avoid bot detection [Yang et al., 2014]. After the first day it already became obvious that even accounts that have been created in a seemingly identical fashion are treated differently in terms of selection of posts for the respective News Feed. For the next days none of the News Feeds displayed the same posts in the exact same order. After three days the selection of posts didn't match for any two News Feeds at all, independent of their order. From the fourth day on, bot detection could not be avoided and thus, several of our Sock Puppets got banned every day because we could not provide a telephone number for account verification—after 10 days none of them remained. As a result, further analyses were neither qualitatively nor quantitatively feasible.
While it might have been possible to create even more realistic bots by, e.g. faking telephone numbers or by hiring real people to navigate our fake accounts, the effort necessary to ask this simple but important question on Facebook's News Feed recommendation system is exceedingly high. To assess the appropriateness of personalized recommendation systems and to ensure algorithmic accountability even through non-intrusive procedures, such as black box analysis, society needs a reliable, efficient, and not too costly access. In the following, we will quickly sketch the general scope of this demand.

3.3 Broader scope

While the RNF case study provides a sketch of the problems of black box analyses in one important question, namely the question of news diversity, this is by far not the only application where society needs to analyze personalized recommendation systems. Other applications are:

1. Webshops with dynamic prices like Amazon or Trivago have the option of offering personalised prices on the basis of recommendation systems. This involves the risk of personalization based on protected characteristics and thus of discrimination.
2. Do headhunters on career platforms like LinkedIn, Xing, Monster, Stepstone or others get a personalized selection of possible candidates? Might this lead to a biased selection towards a certain gender or ethnicity over time? This

would be problematic because national law in many countries regulates a fair access to job opportunities.
3. Analysis of the personalized roll-out of political ads on Facebook, Instagram or Twitter. A biased roll-out might hinder democratic processes, as indicated by the Cambridge Analytica scandal [Schneble et al., 2018].

The last section sketches possible solutions on the political actions that need to be taken in order to give society the ability to reveal illegal, illegitimate or unethical biases in recommendation systems.

4 Demands for a legal framework for black box analyses

For monitoring black box systems, privileged, legally guaranteed and continuous access is needed. In order to make this possible, politics must intervene and create a legal framework for black box analyses. This section points to the requirements of such necessary accesses. Many problems with opaque systems can be countered with provisions that establish transparency and allow for the scrutiny of ADM systems. Such provisions should be demanded if a sufficiently great danger to democratic values is possible. The following demands address the obstacles that currently hinder an inspecting instance in trying to reveal illegal or immoral behaviour of recommendation systems.

I. Set up of a suitable machine Interface (API)

There are two perspectives on the monitoring of recommendation systems in which granting suitable API access is useful.

The first requirement concerns the users of the recommendation systems. In our case study for example, a preferable solution to the problem would be a more comprehensive access to relevant information for page operators on Facebook. We found that the existing API does not give insights into what posts are displayed to whom in their respective News Feed. As demanded by van Drunen, Helberger and Bastian, it must become clearer how user behaviour affects selection [van Drunen et al., 2019]. Still, it is important to comply with data protection and privacy regulations such as the General Data Protection Regulation (GDPR). Some aggregation of user data may therefore be necessary.

Another option is privileged access for accredited researchers/auditors acting on behalf of the state or a regulating instance. Some questions such as which political party orders which kind of advertisement for which target group can only be answered by accessing the system-wide or aggregated information of the recommendation system. Facebook's disclosed information platform for political ads, which actually should answer those questions, is currently under criticism for not revealing all relevant information[8].

II. Allow conditional use of bots

A platform that makes use of a recommendation system must allow the automated control of accounts by accredited scientists. This may include the use of

[8] https://www.propublica.org/article/facebook-blocks-ad-transparency-tools

bots, at least as long as it is assured, that user manipulation by such bots (for example by enforcing trending topics) is prevented. Bots that are specifically authorized by the platform operator raise yet another problem, since it is important that they are treated equally to a human user. Independently, to allow representative monitoring of such a platform, there needs to be a way to automatically generate a large number of bots for scientific purposes.

III. Provide selective access for normal users
An insightful monitoring method for opaque systems needs, as already presented, the active participation of users of the system via Crowdsourced Audit. One obstacle to such participation, however, is the issue of data protection. As stated in the case study, it is not possible for Facebook users to share only parts of their News Feed for the purpose of analysis. There is only full access to the account or none at all. In addition, the API access to the News Feed has been discontinued, which means that access is only possible by reading it from browser sessions. A selective access would therefore be necessary to create a low-entry threshold for such an audit. An important aspect in this regard is the possibility of anonymization or pseudo-anonymization, which could be achieved by allowing adequately configuring access. Scientific analyses would then be significantly simplified.

IV. Legal certainty for automated audits
The attempt to examine a recommendation system for researching activities without any criminal intent must not be criminalised by the terms of use or other legal regulations. Platform terms of service are often written to prohibit the automated downloading of any information from a Website, even if that information is public. For instance, exploiting security vulnerabilities to raise public awareness may result in legal consequences by the US Computer Fraud and Abuse Act (CFAA)[9]. The same legal basis would currently apply if a scientist performs a black box analysis. These are two very different kinds of actions which should be treated differently, and scientists should be allowed to carry out research within a secure legal framework when examining such systems for unaccountable behaviour. Otherwise, there is no possibility to level existing information and power asymmetries.

Of course, the above-mentioned demands raise questions of objectivity, because the platform operators are aware of the required and provided access. This would allow the platform to issue unequal treatment vis-à-vis the reviewing agency, similar to what happened with the Dieselgate affair, where cars recognized that they were in a test stand and then operated differently than under normal conditions [Bovens, 2016]. Another important aspect to consider is the risk of an abusive use of ADM systems by the state. It may ultimately be the state that is enabled to tap into and understand all black boxes that intervene into the public sphere. Great care must thus be taken not to create a set of instruments that would allow total surveillance of citizens. Rather, the state should enable other

[9] https://www.wired.com/2013/03/att-hacker-gets-3-years/

stakeholders to independently ensure the accountability of ADM systems. Only this way it is possible to achieve a balance between the interests of the platforms and the interests of society as well as to avoid a concentration of possibly unaccountable power.

Acknowledgement

We wish to thank Ralph Kühnl for presenting us with the issue of the perceived unequal roll-out of content from Facebook pages and his trust to give us access to the Facebook account of the Rhein Neckar Fernsehen.

References

[Bovens, 2016] Bovens, L. (2016). The ethics of dieselgate. *Midwest Studies In Philosophy*, 40(1):262–283.

[Diakopoulos, 2014] Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism*.

[Dreyer and Schulz, 2019] Dreyer, S. and Schulz, W. (2019). Künstliche Intelligenz, Intermediäre und Öffentlichkeit. Technical report, Alexander von Humboldt Institut für Internet und Gesellschaft & Leibniz-Institut für Medienforschung.

[Krafft et al., 2019] Krafft, T. D., Gamer, M., and Zweig, K. A. (2019). What did you see? A study to measure personalization in google's search engine. *EPJ Data Science*, 8(1):38.

[Patil et al., 2014] Patil, S. V., Vieider, F., and Tetlock, P. E. (2014). Process versus outcome accountability. *The Oxford handbook of public accountability*, pages 69–89.

[Sandvig et al., 2014] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22.

[Schneble et al., 2018] Schneble, C. O., Elger, B. S., and Shaw, D. (2018). The cambridge analytica affair and internet-mediated research. *EMBO reports*, 19(8).

[van Drunen et al., 2019] van Drunen, M., Helberger, N., and Bastian, M. (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law*.

[Yang et al., 2014] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. (2014). Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):1–29.