Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization

Gemma Galdon Clavell
Eticas R&C
Barcelona
gemma@eticasconsulting.com

Mariano Martín Zamorano
Eticas R&C
Barcelona
martin@eticasconsulting.com

Carlos Castillo
Department of Information and Communications Technologies
Pompeu Fabra University
Barcelona
carlos.castillo@upf.edu

Oliver Smith
ALPHA Telefonica
Barcelona
oliver.smith@telefonica.com

Aleksandar Matic
ALPHA Telefonica
Barcelona
aleksandar.matic@telefonica.com

ABSTRACT

In this paper, we present the Algorithmic Audit (AA) of REM!X, a personalized well-being recommendation app developed by Telefónica Innovación Alpha. The main goal of the AA was to identify and mitigate algorithmic biases in the recommendation system that could lead to the discrimination of protected groups. The audit was conducted through a qualitative methodology that included five focus groups with developers and a digital ethnography relying on users comments reported in the Google Play Store. To minimize the collection of personal information, as required by best practice and the GDPR [1], the REM!X app did not collect gender, age, race, religion, or other protected attributes from its users. This limited the algorithmic assessment and the ability to control for different algorithmic biases. Indirect evidence was thus used as a partial mitigation for the lack of data on protected attributes, and allowed the AA to identify four domains where bias and discrimination were still possible, even without direct personal identifiers. Our analysis provides important insights into how general data ethics principles such as data minimization, fairness, non-discrimination and transparency can be operationalized via algorithmic auditing, their potential and limitations, and how the collaboration between developers and algorithmic auditors can lead to better technologies

CCS CONCEPTS

• Human-centered computing~Human computer interaction (HCI) • Human-centered computing~Empirical studies in HCI.

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02...$15.00.
DOI: https://doi.org/10.1145/3375627.3375852

KEYWORDS

GDPR; algorithms; AI, recommender systems; data ethics; bias

ACM Reference format:

Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith and Aleksandar Matic. (2020). Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. In *Proceedings of 2020 ACM AI, Ethics, and Society Conference (AIES'20), February 7-8, 2020, New York.* ACM, NY, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375852

1 Introduction

Increased availability of human behavioral data combined with advanced machine learning techniques opens the door to new applications and services that can help in everyday activities. Yet, algorithmic decision-making has raised a great deal of criticism with respect to its potential discrimination and opacity. In this regard, the General Data Protection Regulation (GDPR), Article 22, specifically defines the regulatory framework for automated individual-level decision-making to address some of these concerns. This regulation, in conjunction with the publication of a number of incidents showing negative social impacts caused by machine learning (ML) [2,3], are prompting companies and public organizations to audit their algorithmic services, technologies and procedures. Public organizations are also increasingly asking for stronger safeguards concerning the right to non-discrimination and, at the same time, setting the accuracy bar high for automatic decision-making systems. Private companies are increasingly following the same trend of integrating new protocols and methodologies aimed at addressing algorithmic fairness, accountability, and transparency "by design". The objective is to avoid opacity in the design and use of algorithmic systems [4,5].

However, establishing new procedures and safeguards in algorithm development to address these issues is still in its infancy. Algorithm developers are typically not trained on the relevant methods, such as the identification of protected attributes or contextual social factors that may lead to discrimination [6]. This is also part of an ongoing ontological and sociological debate

reflected in the multiple definitions of fairness used by experts in the field, which are not always contained in its existing statistical characterizations [7, 8, 9]. Effective methodological tools to evaluate training data in terms of the demographic characteristics of the sample groups need to be integrated, and this lack of standardization also affects the relative capacity to "reconstruct" the causes of bias once a system is already in operation.

This paper explores one of the most relevant sources of algorithmic bias, namely the lack of information about group membership of the data processed by the systems. In this study, the Algorithmic Audit (AA) was conducted by Eticas Consulting and Alpha on a case-study of the REM!X app. REM!X was a recommender system with an algorithm based on popularity, which used artificial intelligence (AI) and aimed at helping its users establish healthier everyday routines through customized recommendation.[1]

2 Algorithmic bias and "color blindness"

In order to frame algorithmic bias, it is essential to distinguish between different forms of discrimination. Generic discrimination refers to the unfair treatment of a person (A), with respect to another person (B), due to specific properties that person A has and person B does not have [10]. Group discrimination happens when such a property belongs to a socially salient group, and is founded, either explicitly or implicitly, on animosity against this group, or the belief that people in this group are inferior, or the belief that they should not intermingle with others or have the same rights and opportunities. Statistical discrimination is group discrimination based on a fact that is statistically relevant[2]. A classic example of statistical discrimination is an interviewer recommending not hiring a highly-qualified woman because the interviewer believes women have a higher probability of taking parental leave. Non-statistical discrimination occurs when the interviewer recommends not hiring a highly-qualified woman because she says explicitly that she intends to take parental leave [10]. If we disregard animosity and consider that any feature used by a learning algorithm is considered by the algorithm as statistically relevant, we can say that algorithms can discriminate [11].

In line with the above, we define algorithmic discrimination or algorithmic bias as disadvantageous differential treatment of (or impact on) an already disadvantaged group[3]. Social groups with these protected attributes can be either legally protected (e.g., people with disabilities) or not, for instance in the case of the participation of women or minorities who might be underrepresented in certain professions or positions. The criteria by which what constitutes bias is defined also need to be framed from a social and ethical standpoint. This is important because some attributes may be legal and considered legitimate for differential treatment, but still seen as discriminatory in some social contexts due to cultural, historical or ethical reasons [12,13]. It is also important in understanding the relevance of a statistically significant feature.

Although quantitative techniques for measuring disparate impact/treatment in predictive systems are constantly improving, the qualitative and procedural aspects of these analyses remain poorly systematized. This is more evident in systems that have a more complex interaction with users, such as apps powered by a recommender system, the design of which requires a holistic understanding of fairness as a philosophical and ethical notion [6].

During the algorithmic design stage, developers should try to minimize risks of discrimination either by eliminating categories involving protected groups, when they are not needed for achieving the purposes of the system, or by removing possible discriminatory links between recommendations and protected groups. In this regard, one way of reducing such risks includes eliminating data corresponding to protected attributes. This approach lies at the core of the data minimization principle defined by the GDPR, Article 5(1) (c). This principle is defined in the following manner: "1. Personal data shall be: [...] (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization')". Therefore, the principle of data minimization aims to establish a clear link between the personal data collected by data controllers and the purposes for which data are collected. The collection of personal data must be minimized, albeit within the context of what data is required by the controller to accomplish its data processing goals.

As an example, information about gender may be removed from the data collection and modeling; however, this data minimization poses two main problems, which we encountered in our work on the REM!X system. First, many systems cannot deliver precise and useful outputs without knowing the user's gender (or their performance improves with this information). Therefore, it might justifiable to collect some categories of data concerning protected groups in order to produce better outputs. Second, and more importantly, not collecting data on gender or the other protected attributes can make it challenging or impossible to identify discrimination against those protected groups once the system has been implemented and the machine-learning algorithm has been deployed. Specifically, it has been already demonstrated that the lack of data about a social identifier – an attribute enclosing the information of disadvantaged groups (race, gender, religion, etc.)-, can lead to bias [14]. This form of bias is defined as "color blindness"[4] and it occurs when variables used to identify sensitive information or protected attributes are removed from the training data. This is typically performed with

[1] It should be noted that the app was a prototype and has been closed down after testing.

[2] We remark that the above definitions are different from standard definitions of statistical bias, which involve distortions of a statistic resulting from biased samples or estimators whose calculation is not correct in relation to the right or expected value of a parameter (Turney, 1996), and hence statistical bias cannot (always) be an adequate criteria of algorithmic fairness.

[3] These disadvantaged groups can be defined in relation to the attributes mentioned in Article 21 (Non-discrimination), of the EU Charter of Fundamental Rights.

[4] It should be noted that we consider the term "color blindness" to be problematic due to its potential confusion with the term for the inability of a person's vision to distinguish between certain colors. For this reason we refer to "color blindness" as it pertains to data by using speech marks.

the objective to render the resulting system bias free [15]. However, in practice this does not reduce the risk of bias due to the presence of proxies for sensitive attributes, while at the same time making it difficult to audit such systems. The issue of proxies has been amply reported in the literature. During the learning process, systems based on ML are often able to infer these protected categories from the data by using proxies embedded in the other variables, and thus indirectly learning about sensitive attributes and incorporating them in the decision-making process. The extent to which data minimization may also be detrimental to transparency and accountability efforts such as the ability to conduct rigorous algorithmic audits, has not yet been raised in the literature as far as we know, and is one of the findings that has led Eticas to change how it provides Algorithmic Auditing services.

Limiting the collection of data to information only directly relevant for the specified purpose is highly encouraged and defined by the GDPR, and referred to as data minimization. However, applying data minimization before considering other basic legal principles and accountability requirements may also lead to "color blindness" and result in major limitations to achieving other legal principles, such as the actual accuracy (GDPR, Art 5, 1, d) of the system, and tackling ethical concerns. From a methodological standpoint, developing a system to be "color-blind" by removing protected attributes can lead to an increase in the opacity of algorithmic processing by limiting the mechanisms for identifying bias.

In this paper, we shed light onto potential algorithm bias in the context of minimizing the collection of personal information. We show that data minimization should not be applied without consideration of, but rather in concert with consideration of other key GDPR principles – accuracy as well as transparency and accountability, lawfulness and fairness. In this regard, we claim that collecting specific personal information can be adequate from the data minimization standpoint when required to deliver an accurate, fair and transparent algorithm, and when the collection is carried out solely for the purposes of auditing and accountability.

3 The REM!X app – Building a Wellbeing Recommender System

REM!X was developed as an AI-based recommender app for the Spanish market that offered customized advice on healthy habits through small exercises and wellness challenges. The main objective of the project was as a prototype to test how to improve users' well-being, facilitating their personal development goals, helping them to overcome anxiety and stress, and ultimately, making them happier. Users were asked to report how they felt at different moments (e.g. sad, bored, anxious, etc.) and how they wanted to feel (e.g. happy, calm, content, etc.). Based on that, REM!X offered simple tricks, mini tutorials, and challenges that were designed to make them feel better. The REM!X database included more than a thousand activities created based on the well-being literature and in consultation with experts in psychology, behavioral economics, and psychiatry.

The process of designing the REM!X application was iterative, with continuous user feedback. User feedback and the observed drop-out rates were two of the main criteria for shaping the app design, which also had a major impact on the collected data categories. Importantly, gender and age were collected in the first versions of the prototype, mainly with the intention to adjust the language (e.g. in Spanish, verbs and nouns are declined differently based on the gender), but were later removed in the subsequent versions based on user feedback and the data minimization principle. Similarly, sensor data was initially collected with the aim of improving the recommendations based on user context, but the use of sensor data was removed due to user privacy concerns, and a careful consideration of usability/privacy trade-offs.

Thus, the REM!X recommendation algorithm relied on the user's impressions, views, bookmarks, likes and performed activities, and additional data logs included an anonymized user ID, and activity tags (such as social, cognitive, sport, travel, food, etc.).

The initial recommendation system was developed based on popularity, which was turned into a collaborative filtering algorithm. This algorithm was the only one in use at the time of performing the AA reported in this paper. The output represented a ranked list of activities that a user (a) is likely to undertake, and (b) is likely to value positively or have a positive impact on a their wellbeing. Systems designed in this collaborative way establish a set of item preferences per user [16] so that the algorithms are able to match one user to others by identifying those who have historically had a similar taste or followed similar patterns. This process is at the root of REM!X's algorithmic design for the collaborative filtering recommender system analyzed in this paper. The scalability of these systems, namely their capacity to process large data sets, and the quality of their recommendations constitute two of their major challenges [17].

Due to this popularity-based method, the recommender system outputed the most popular activities first. Importantly, it also introduced randomness to generate serendipitous recommendations and to avoid the "filter-bubble" of putting users in feedback loops in which popular activities became even more popular. Random recommendations were also used in an A/B experimental fashion to evaluate the performance of the newly developed recommender systems.

While recommendation algorithms can have many advantages, the literature has also shown how such algorithms may introduce biases that favor the most popular options [18]. Often, the capacity to recommend options beyond a certain "band of popularity" will determine if a recommender system can introduce users to new options, given that a limited number of choices are likely to be highly popular among many users. In the case of REM!X, users were presented with a large range of recommendations, on the basis of how a series of categories correlated, such as current and expected status, practice or mood. In order to mitigate the bias risks of popularity-based models, a

measure of randomness and serendipity was also introduced in the process and results.

Besides promoting a "filter-bubble", algorithmic recommendations may also be inaccurate and unfair.. Fairness in recommender systems is an increasingly popular line of research[5] in which most authors stress the need to increase awareness and transparency on how different recommendations affect different groups in order to audit how systems make decision and whether bias is being reproduced and amplified. Inaccuracy and unfairness can arise as a result of the association of specific social groups to certain tastes, practices and attributes, which can lead to discriminatory or biased recommendations [19]. Machine learning can establish such associations based on the previously mentioned reproduction of historical selection trends by protected groups or collectives. These processes may also be based on discriminatory assumptions and stereotypes. If specific social groups are more inclined towards specific recommendations or are more likely to have certain preferences (for instance, women looking for certain jobs), this may determine the kind of recommendations that the system suggests to them, thus replicating and amplifying bias.

4 AA methodology

Evaluating whether algorithmic decision-making is based on unfair grounds and/or can lead to discriminatory outcomes requires the use of pre-processing, in-processing, and post-processing methods [20]. Specifically, the AA methodology applied to the REM!X app consisted of four main steps: (1) defining an assignment of elements in the data to groups, (2) defining a protected group; (3) determining a set of metrics aimed at measuring bias; (4) measuring and comparing across groups.

The qualitative analysis conducted to support this audit relied on four data collection tools:

1. The analysis of the recommendations provided by the REM!X app,
2. Five focus groups, which included Eticas and different Alpha team members, where an analysis of the app design and development process was performed. Together with researchers, engineers and product team members at Alpha, we identified relevant variables, described the algorithms and we unpacked potential proxies in the data that could lead to discriminatory outcomes for individuals belonging to protected groups. Alpha's initial design process did not analyze the categories of users or analyze social impact related to human-machine interaction since it was considered that this would be adequately addressed at a later stage through an AA, and this was the specific expertise brought in by the Eticas team.
3. Desk research consisting of a thorough review of Alpha documentation on the desired algorithm development process. Since the initial design process did not include the need to analyze protected attributes on that basis that such data was not collected, the AA team led the effort to identify gaps and make suggestions to modify the process to ensure that all identified bias and fairness concerns were addressed.
4. A digital ethnography using 206 messages left as feedback by users in Google Play Store, which were categorized and analyzed to understand the app's performance and its social implications.[6]

The Eticas team also collected information about data gathered and processed by REM!X during its first stage of development, when gender and possibly other protected attributes were integrated, as described above. After removing these data categories, none of the collected training datasets included preselected protected attributes.

Overall, the available sources of information that were useful to analyze the impact of algorithmic processing in terms of bias or differential impact included the app's data retention information and the messages sent by users using Google App. These two sources were used to obtain indirect evidence about differential treatment, particularly concerning age, cultural background and gender. Even though these sources of information are not systematic nor representative, they were used with full awareness of their limitations to support the analysis of algorithmic bias. Relying on these sources, the AA analyzed the societal factors that could be impacting and being impacted by biases in REM!X.

5 Analysis of bias in REM!X

Two elements should be considered in the analysis of potentially biased recommendations in REM!X. On the one hand, the app was intended to be used internationally, albeit after further development to localize content and tailor recommendations to each market. This breadth of scope meant that recommendations needed to be able to be universal while at the same time sensitive and responsive to different cultural, historical, legal and social contexts and value systems. On the other hand, protected categories were not collected nor processed by the app. This, combined with the lack of indirect evidence of discrimination, led us to examine the recommendations primarily from a qualitative standpoint. While this is a useful approach, we have learned in the process that collecting personal data to test and audit bias is important, and that data minimization may go against transparency if implemented without taking into account the elements we bring up in this analysis.

[5] See, e.g., the FATREC Workshop at RecSys'18 https://piret.gitlab.io/fatrec2018/).

[6] This method, which has been combined with the other described data collection tools, has been used as a complementary instrument for the analysis of the case. Following [21] online information has been properly contrasted, triangulated with other sources of information and placed within broader societal knowledge.

Regardless of these limitations, the analysis was conducted by defining four domains that may be problematic in terms of algorithmic discrimination, as explained above.

5.1 Socioeconomic barriers and implications

Recommendations provided by REM!X may contain certain socio-economic implications in terms of access and/or inequality. On the one hand, access relates to the capacity of users to effectively carry out the recommended activities (like riding a horse or improvising a trip) as such recommendations might be beyond a user's means. This has the potential to lead to property-based discrimination, which could also be deepened and expanded by machine learning, since some racial, gender or religious groups could have preferences that are correlated with socio-economic factors. On the other hand, distributing recommendations based on popularity could lead to reproducing the socioeconomic status of specific groups. Interestingly, customizing recommendations based on socio-economic status can lead to a sort of economic segregation, whereas its absence can have financial and psychologically negative consequences.

5.2 Cultural barriers and implications

REM!X introduced users to several activities with cultural implications. In particular, the app could suggest activities, habits or goods that were linked to race, religion, or place of origin. For instance, "Enjoy Christmas" was one of the recommended activities, yet the app did not offer the same for other religious celebrations, which clearly made it less sensitive to the expectations of other faiths. Even though this form of discrimination is not a result of algorithmic processing but a design feature, it should be considered, since once inserted into the design of the algorithm, the app and its algorithm would have reproduced and thus amplified religious discrimination.

In addition, the use of certain dialects as opposed to others could alienate certain users or worsen their user experience. This is exemplified by a comment by one user who rated the app in Google Apps and highlighted the issue of adapting the app to the Spanish spoken in Latin America:

> "I love this application, I am from Argentina and I was able to download and use it. The only thing that I would change is to modify some words and adapt it more to Latin America"

In some contexts, language can reveal class, race and gender, among other personal information, and so lead to discrimination. Therefore, language should not be considered as a neutral choice. In an algorithmic context, an ML system could pick up on relations between income and language, for instance, and make discriminatory decisions on this basis. While this was not observed in REM!X, it is one of the aspects that was addressed to raise awareness within the developer team.

5.3 Gendered recommendations and gender inclusiveness

A specific set of activities recommended by REM!X could be considered as gender driven if they follow dominant social practices and stereotypes in specific social contexts. One example includes a REM!X recommendation of "nail polishing" with an image of female hands, which is a gender-normative recommendation. However, we could not analyze whether this recommendation had been delivered disproportionately more frequently to women than to men due to the lack of gender information in REM!X logs. Interestingly, REM!X also used gender neutral language[7] to address the inclusiveness criteria yet it provoked negative feedback on the Google Play Store.

The ethnographic analysis of the users' comments on Google Play Store indicated that women were overrepresented among users. In total (at the time of this analysis), 61 comments were made by users identified as female, versus 23 comments made by users identified as male, out of 84 comments in which the gender of the user leaving the comment was identifiable (from 206 comments). This might suggest that the app had been more popular among women, which could cause recommendations to become attuned to supposedly female preferences (such as polishing their nails, doing yoga, making smoothies and eating salads). This, in turn, may have made the application less attractive to male users. Potentially, this feedback loop could turn REM!X into an app used mainly, in practice, by women. This would eliminate issues of gender discrimination within the app, as differential treatment received by men would cease to be a major concern, but it would certainly come at the cost of the level of inclusiveness of the app.

5.4 Accuracy of the recommendations

The amount and types of data processing categories can also affect the quality of the recommendations made in terms of accuracy and personalization. This is open to abuse by service providers, who may collect personal information that is not directly relevant for the main purpose of their app. This is the reason why the GDPR introduced data minimization as one of its key principles. In practice, however, it is difficult to find an appropriate trade-off in data collection due to the "cold-start" problem; initially it is not clear which data categories will be the most valuable for personalizing recommendations in the future, and so it is difficult to prioritize them for the purposes of data minimization. In the case of REM!X, data minimization resulted in the reduction of data points and of the categories of data potentially useful for profiling users and providing recommendations. As discussed above, not

[7] Most nouns in Spanish are either feminine (ending with an-a) or masculine (ending in -o). The generic use of masculine when we refer to both sexes has been criticized by feminist academics as a way of excluding women. Inclusive or gender neutral language (using an -@or an -e instead of the -o has been proposed and is being used to foster inclusion.

collecting personal information can minimize risks of discrimination; but it can also limit the potential to personalize recommendations, and make auditing more difficult. With respect to personalization, comments that we found on Google Play Store indicated the need for enhancing the accuracy of the system in order to recommend more personalized activities. In particular, the lack of location data appeared to have a negative consequence on the system's accuracy.

Still, in the case of REM!X, the trade-off between data protection rights and reputational risks derived from unfair bias, on the one hand, and accuracy on the other hand, seems to be balanced; the app is able to reach most of its aims without having to integrate extra categories of data. Moreover, in judging this balance, we have been mindful of Alpha's aim to be grounded in ethics.

6. Discussion and conclusion

Recommender systems can be optimized based on an analysis of benefits and risks. The REM!X application aimed to recommend daily activities to establish healthier habits and ultimately to improve users' health and subjective well-being. The AA performed on the REM!X app suggested that the system was able to achieve a solid performance in recommending relevant activities (i.e. click-through and rates of bookmarked activities that the users planned to do) while reducing the amount of data collected, particularly sensitive data. No data on gender, race, age, religion or other protected attributes were collected or processed by the algorithm. These variables were used neither to train the system during its design, nor to assess its performance and social impact once in operation. As a result, this minimized the potential discriminatory implications of the algorithmic processing outcomes, but it also increased opacity with respect to identifying differential impact, particularly considering that the system was developed as a collaborative algorithm.

Hence, the AA focused on the analysis of recommendations and the development of hypothesis on bias and not on measuring disparate impact/treatment and feedback loops (as would be the case when conducting a traditional AA) since the information about protected groups could not be statistically correlated to other variables or proxies. This is why most of the fieldwork analysis was geared towards finding indirect evidence or sources of bias in order to determine the appropriate strategies to mitigate it.

Based on the above indirect sources of information, we developed a series of hypotheses about the risk of bias in order to consider the need for conducting a trial. As far as property discrimination is concerned, risks were low since only a few recommendations would be relatively expensive for users and the system was not designed to allocate or limit material resources to specific groups. With respect to religious or cultural discrimination, the algorithm was not able to personalize per user based on linguistic or religious grounds. Religious implications within the observed recommendations were very limited and the issue concerning the dialectal variants of the languages used did not appear to represent a major barrier for users. Finally, the potential for gender bias was only detected in the way in which certain activities or challenges were recommended to women on the basis of historical bias. Across the first three hypotheses, we did not observe major ethical challenges in REM!X, and so we consider that the risk of discrimination in the app is low, especially in the case of gender discrimination.

With respect to our hypothesis on recommendation accuracy, the lack of specific personal data categories may have had a negative impact on the app's performance and in its ability to provide more personalized recommendations. The REM!X design traded-off the benefits of improving recommendations and being able to measure accuracy for different groups to reduce the privacy-related risks associated with collecting sensitive data (such as location data).

As our findings indicated a low level of risk, carrying out a specifically designed trial to test bias in human-machine interaction, which is best practice when auditing a high-risk algorithm, was not deemed necessary. Such a trial would have required gathering sensitive information from a representative sample of users in order to measure disparate treatment and impact.

Even with such limitations and circumstances (lack of personal data to test bias and low ethics risks overall), the AA resulted in a set of important proposals that have been useful to improve Alpha's products, and may be helpful for other related initiatives that seek to develop algorithmic services and products in ways that are legally compliant but also responsible, accountable, and ethical.

Firstly, that when implementing best practice and regulations on data minimization, this should be done in concert with other ethical data principles. The REM!X case illustrates the risk of over-indexing a focus on data minimization without sufficient and timely consideration of the principles of fairness and transparency, as defined by the GDPR. A more balanced assessment might result in more data being collected to ensure that algorithms can be assessed *a posteriori* for transparency and for the identification of potential biases. This data can and should of course be collected just for testing purposes, and the risk of some personal attributes capturing and reproducing bias should always be taken into account and mitigated since, as we have seen, personal attributes can also be inferred from pseudonymous data.

Secondly, the methodological lesson drawn from the AA is the necessity of applying a series of methods and fieldwork activities from the very beginning (design and pre-processing phases). This is in line with the literature on assessing algorithmic fairness by industry practitioners [6]. When other ways to test

algorithms for fairness and bias are not feasible (especially in the long run), it is important to "document the model" [22] in its training phase. Collecting interaction logs can help algorithm auditors infer and tackle potential bias prior to the system's deployment. In general, the practice of continuously documenting models is highly encouraged as it provides an instrument for auditing the algorithm's functions and for avoiding biases.

Lastly, very early in the development process, app developers need to decide on how to evaluate fairness with respect to sensitive data categories. The most appropriate approach will depend on how the service provider wishes to balance the demands of evaluating fairness with other principles, such as data minimization. Depending on the balance chosen, credible responses may range from less robust, indirect methods (such as the digital ethnography applied to REM!X); through relying on advanced documenting methods; to more rigorous methods such as the collection of personal information from at least a sample of users to be able to perform a quantitative analysis comparing, e.g., recommendation accuracy for protected and non-protected groups. This latter method is what Eticas currently implements when conducting algorithmic audits.

Analysis of property-based discrimination, cultural bias, gender bias, and accuracy issues in personalization, as undertaken with REM!X, form a core part of an AA. Taken together, the three proposals described above point to the need to work with product teams to raise awareness not only on what happens to personal data inside an algorithm, but also broader issues related to social, historical or economic dynamics. Often, it is the choices and assumptions made in the early stages of the design of an algorithm that lay the ground for bias and discrimination. Likewise, an early understanding of these broader issues and their translation into responsible data choices can avoid many of the problems that many algorithms are currently facing (in terms of lack of transparency, fairness and accountability).

Further research into how organizations implement the three proposals set out here will help to ensure not only that the relevant legal frameworks are upheld in the technical specifications of algorithmic systems, but also that citizens and their data are effectively protected. Moreover, a practical approach to data responsibility and ethics (like the one put forward by algorithmic auditing) is a robust step towards opening the "black box" of algorithms and machine learning processes and ensuring that society as a whole has ways to inquire about how algorithms work, make decisions on their willingness to share data, and seek redress mechanisms when needed.

ACKNOWLEDGMENTS

Alpha hired Eticas Consulting to conduct the AA, and there is an ongoing professional collaboration. C. Castillo thanks La Caixa project (LCF/PR/PR16/11110009) for partial support.

REFERENCES

[1] European Union. 2016. The General Data Protection Regulation, 2016/679. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[2] Eubanks, Virginia. 2018. Automating Inequality. New York: St. Martin's Press.
[3] O'Neil, C. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Broadway Books.
[4] Pasquale, Frank. 2015. The Black Box Society. Cambridge: Harvard University Press.
[5] Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society. doi: 10.1177/2053951715622512.
[6] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 600). ACM.
[7] Woodruff, A.; Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018). ACM, 656.
[8] Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, p.1085.
[9] Narayanan, A. 2018. Tutorial: 21 definitions of fairness and their politics. Conference on Fairness, Accountability, and Transparency, NYC Feb 23.
[10] Lippert-Rasmussen, K. 2013. Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford: Oxford University Press.
[11] Castillo, C. 2018. Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavor. In Proceedings of HUMAINT Workshop.
[12] Binns, R. 2018. Algorithmic Accountability and Public Reason, Philos. Technol., 31 (4), 543-556.
[13] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). ACM.
[14] Shimao, H.; Komiyama, J.; Khern-am-nuai, W; Kannan, Karthik Natarajan. 2019. Strategic Best-Response Fairness in Fair Machine Learning Algorithms. Available at SSRN: https://ssrn.com/abstract=3389631
[15] Barocas, S. and Selbst, A. 2016. Big Data's Disparate Impact. California Law Review.
[16] Resnick, P., and Varian, H. R. 1997. Recommender Systems. Special issue of Communications of the ACM. 40(3).
[17] Sarwar, B.; Karypis, G., Joseph Konstan, and John Riedlfsarwar. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In Proc. of the 10th International WWW Conference.
[18] Abdollahpouri, H.; Burke, R., and Mobasher, B. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. Paper presented in AAAI Florida Artificial Intelligence Research Society (FLAIRS '19), May18–22.
[19] Tsintzou, V., Pitoura, E., & Tsaparas, P. 2018. Bias Disparity in Recommendation Systems. arXiv preprint arXiv:1811.01461.
[20] Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2125-2126). ACM.
[21] Baer, T. 2019. Understand, Manage, and Prevent Algorithmic Bias. Berkeley, CA: Apress.
[22] Krieg, L. J; Berning, M and Hardon, A. 2017. Anthropology with algorithms? An exploration of online drug knowledge using digital methods, Medicine Anthropology Theory 4 (3): 21–52.