Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook

Yim Register, Emma S. Spiro
Information School, University of Washington
yreg, espiro@uw.edu

Abstract

Facebook users interact with algorithms every day. These algorithms can perpetuate harm via incongruent targeted ads, echo chambers, or "rabbit hole" recommendations. Education around the machine learning (ML) behind Facebook (FB) can help users to point out algorithmic bias and harm, and advocate for themselves effectively when things go wrong. One algorithm that FB users interact with regularly is User-Based Collaborative Filtering (UB-CF) which provides the basis for ad recommendation. We contribute a novel research approach for teaching users about a commonly used algorithm in machine learning in real-world context – an instructive web application using real examples built from the user's *own* FB data on ad interests. The instruction also prompts users to reflect on their interactions with ML systems, specifically Facebook. In a between-subjects design, we tested both Data Science Novices and Experts on the efficacy of the UB-CF instruction. Taking care to highlight the voices of marginalized users, we use the application as a prompt for surfacing potential harms perpetuated by FB ad recommendations, and qualitatively analyze themes of harm and proposed solutions provided by users themselves. The instruction increased comprehension of UB-CF for both groups, and we show that comprehension is associated with mentioning the mechanisms of the algorithm more in advocacy statements, a crucial component of a successful argument. We provide recommendations for increased algorithmic transparency on social media and for including marginalized voices in the conversation of algorithmic harm that are of interest both to social media researchers and ML educators.

Introduction

Imagine a teenager who is struggling with body image. Every day when this teen logs into their Facebook, they see advertisements for beauty and diet products, images of bodies that don't match their own, and posts from friends discussing eating habits. If the teen has no idea that such content is algorithmically curated for them based on their behavior *as well as the behavior of their friends*, how will they ever advocate for a better online environment?

Widespread machine learning (ML) literacy belongs across all levels so that anyone can understand the algorithms they interact with, resist potential harms, and have a say in policy change. Numerous examples illustrate how ML algorithms driving FB advertisement recommendations can either directly or indirectly harm users. On Facebook, the news, posts, and suggested contacts a person sees are all driven by algorithms that make use of the underlying social connections (i.e. network) among users to infer and serve content of possible interest. Despite many positive experiences that result, there is a growing recognition that these systems also contribute to problematic, and pressing, social phenomena. Work on the growth and consequences of echo chambers, in which political and social opinions reflect and reinforce one's own, (Quattrociocchi, Scala, and Sunstein 2016; Bessi 2016) and studies of the perpetuation of gendered and racial biases through content recommendation (Usher, Holcomb, and Littman 2018; O'Callaghan et al. 2015), for example, highlight the ways in which algorithmically driven recommendations based on social network structure can lead to isolated and insular communities that reproduce harmful associations.

A classic algorithm used in ML that often powers recommendations is User-Based Collaborative Filtering (UB-CF), which ranks social contacts according to similarity to serve as a pool for potential recommendations. A key feature of UB-CF is that it relies on the connections among users to search for possible recommendations; in other words, data for ad recommendations can come from your friends and not just your personal characteristics and behaviors. Social media users, on average, do not realize that the content they may see is not only based on their own behavior, but also heavily influenced by what their friends are posting, liking, and clicking on. This process is hidden from the user, whose "likes" may leak into their network or whose social connections may leak into their personal recommendations without ever being notified. This might manifest as getting caught in a cycle of dieting or substance abuse ads, specific political values, conspiracy theories, or ads incongruent to your evolving gender identity.

There is potential for collective advocacy to draw attention to these issues of algorithmic harm via algorithmic resistance (Velkova and Kaun 2021; Karizat et al. 2021). The voice of the user provides us with insight about the ram-

ifications of our AI systems. Therefore, it may be fruitful to explore how ML literacy efforts could work in tandem with self-advocacy skills – a concept primarily talked about in disability studies. Successful advocacy involves identifying the problem and articulating critiques and suggestions in order to adequately meet your needs (Goodley 2005; Wehmeyer, Bersani, and Gagne 2000; Test et al. 2005). A key step in advocating for one's needs is to understand basic properties of the potentially harmful system (Register and Ko 2020). Previous work has demonstrated that using personal data may contribute to better advocacy arguments in ML contexts (Register and Ko 2020; Peck, Ayuso, and El-Etr 2019; Kim et al. 2019).

This work presents a novel research approach to not only teach social media users about algorithmic mechanisms and potential harm using their *own data* to demonstrate common algorithms underlying the platform's behavior, but also as a probe for user reflections on interactions with ML systems. This particular case study explores ad recommendation via UB-CF, a technique employed by Facebook Engineering to provide recommendations to more than a billion people across the globe. While FB modifies standard UB-CF, they likely keep the core idea of using similar users' tastes to identify candidate recommendations to serve to other users (Kabiljo and Ilic 2015). We designed a web app to teach these core ideas to participants as both instruction and as a prompt for surfacing potential harms.

In this intervention, not only do users get to see what kinds of data FB has collected on them, but they are able to trace how that data can be used by algorithms like UB-CF to generate new recommendations. Through investigating one's own data and learning core features of collaborative filtering, we predict that participants may feel empowered to manage the ads they see and be more vocal about what they want from social media companies. They may be motivated to unfollow certain accounts or be more careful about what they click on. They may even take active measures to click on content dissimilar to their interest in order to confuse or diversify what they see, a form of "gaming" the system once they understand the underlying mechanisms. They may even begin to think of their interests as something that can affect their whole network, leveraging the algorithm to promote societal change. We demonstrate how self-advocacy arguments describe these potential solutions post-intervention.

The web app we built teaches FB users (across all levels of Data Science experience) about UB-CF using their own personal Ad Interest data so that we can probe for not only accuracy of comprehension, but also for advocacy arguments about how UB-CF may potentially contribute to harm. While this case study uses the example of UB-CF, our approach can be extended to any common algorithm used by social media platforms. The main idea is to present instruction *in a natural and personal context* of how the user would encounter these algorithms in the real world. Therefore, while the findings in this case may be influenced by FB users' prior notions about and experience with Facebook, they represent a real-world intervention for promoting ML literacy in an ecologically valid context.

Beyond simply learning about a particular algorithm, we care about the voices of users and how they articulate the potential harms caused by such algorithms in the context of FB recommendations. Because FB users are familiar with the platform and their own experiences, they can leverage this history to further understand UB-CF and potential ramifications of such an algorithm. We qualitatively extract themes from the user's advocacy arguments to look at how users express themselves after learning about UB-CF in the Facebook context. Because some of the most egregious harms tend to impact marginalized people specifically ((Shen et al. 2021; Oliva, Antonialli, and Gomes 2021; Alkhatib 2021; Are 2020; Haimson et al. 2021), we particularly consider the voices of marginalized individuals, the kinds of harms they identify and possible solutions they suggest.

We formalize these aims with the following research questions:

RQ1: Does this tool effectively teach how ad recommendation via UB-CF works on Facebook?

RQ2: How do users advocate for themselves regarding potentially harmful ad recommender systems on Facebook after learning about the algorithmic mechanisms?

Foreshadowing the primary contributions of this work, we find that this interactive tutorial is a useful and novel approach to not only teaching social media users about ML algorithms in context, but also as a probe for harms, advocacy, and possible solutions. Through the case study, we provide details on how an ML tutorial can effectively integrate personal user data for instruction, allowing the user to relate more to the domain and use their own expertise to highlight specific concerns they may have about the algorithmic systems and platforms. The application as a probe encourages users to think about algorithmic harm and center *themselves* in this discussion, which can work well alongside researchers inferring user needs.

For this specific case study, we find that ML novices demonstrate high accuracy on a UB-CF comprehension task after participating in the tutorial, and that learning the basics of UB-CF increases the likelihood that the learner's advocacy argument will include mention of the network as a whole, as opposed to just their own behavior. This recognition opens new explanations of harms as well as solutions. We also surface themes of self-advocacy from marginalized users, who often provide personal examples of potential harms of UB-CF on Facebook; from accidental LGBTQ violence to the effects of pervasive dieting ads on individuals with eating disorders. Together these findings offer new directions for work on social media platforms, as well as studies of fairness and bias in these settings, by researchers, designers and policymakers.

Related Work

While research in ML literacy and user empowerment is relatively new, there are a few key spaces that need to be discussed in order to critically engage with the rest of the content in this paper. Current ML literacy efforts range from teaching ML concepts in K-12 (Druga et al.

2019; Hitron et al. 2019; Zimmermann-Niefield et al. 2019) to critically deconstructing data and AI practices in the Science and Technology Studies space (Benjamin 2019; D'Ignazio and Klein 2016; D'Ignazio and Bhargava 2016; D'Ignazio and Bhargava 2015; Prado and Marzal 2013; Schield 2004). One way to facilitate ML literacy is by using personal and relevant data, which has been explored by several research studies in recent years (Kim et al. 2019; Peck, Ayuso, and El-Etr 2019; Bart et al. 2017). However, this work largely centers around formal instruction as opposed to general literacy for users of ML-based systems. Algorithmic systems continue to perpetuate oppression in the form of silencing, censorship, amplifying, and potentially harmful recommendations, particularly due to the structure and mechanisms of social networks (Fabbri et al. 2020; Chakraborty et al. 2017; Samory, Abnousi, and Mitra 2020). Critical works often address these harms and their repercussions (Alkhatib and Bernstein 2019), but there is yet to be much research in the space of empowering novices to *overcome* these harms.

Model interpretability is one aspect of promoting general ML literacy, but explanations of model behavior are not enough to teach novices how ML will function in the future (Smith-Renner et al. 2020; Carton, Mei, and Resnick 2020). Instead, involving the user in a familiar domain and probing for self-advocacy seems to have better results for general literacy (Eslami et al. 2017). The reason why self-advocacy is likely a successful frame for teaching is because when learners are prompted to reflect on their own experiences with a technology, they surface specific harms and must identify how algorithms perpetuated that effect. We know from learning sciences that using relevant and personal examples helps the learner connect to the material; in the case of self-advocacy prompts we are asking learners to not only surface problems of their own but also reason about them, enhancing the learners ability to connect and assemble knowledge. They draw upon the mechanisms of the model to reverse-engineer what is happening to them in their own context. This also allows us to study how marginalized communities are affected differently by the same algorithm.

Especially when studying anyone in a marginalized population, we need to understand that algorithms have differential effects (Boratto, Fenu, and Marras 2019; Edizel et al. 2020; Johnson 2021), from Facebook (Bucher 2012) to health technology (Obermeyer et al. 2019). An app that works perfectly fine for an abled body could be destructive to someone who needs accessibility accommodations (Tanweer et al. 2017). Data collection for those who fit within the gender binary may go unnoticed, where those outside the binary are consistently faced with rejection of their identity in tech(Benhabib 1992; Kingsley et al. 2020). Racial biases in recommender systems, criminal justice, and health algorithms are increasingly part of the AI/ML research space (Abebe et al. 2020; Garcia 2016; Obermeyer et al. 2019; Benjamin 2019). We know that trust from stakeholders matters (Barbosa and Chen 2019; Yin, Wortman Vaughan, and Wallach 2019); practitioners need to learn to design fairly and address the bias of their own systems (Holstein et al. 2019). This paper supports arguments that literacy from the ground up is an important way to facilitate change – stakeholders can speak up for themselves about how ML driven systems, such as social media platforms, are affecting them.

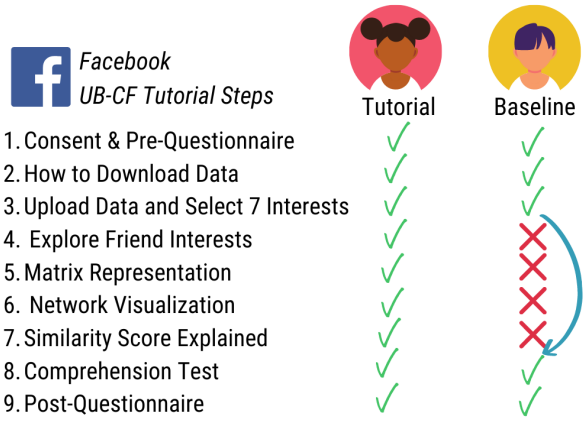


Figure 1: The steps that each group saw in their condition.

Designing a Web Application to Teach UB-CF with Personal Facebook Data

Facebook allows individual users to download a host of personal data, including Facebook's beliefs about the user's Ad Interests. While many algorithms underlie ad recommendation on Facebook, for this case study, we decided to design a tutorial to introduce users to one relatively simple (and commonly used) algorithm – User-based Collaborative Filtering – incorporating the user's own personal data in the instruction. We intend for this case study to be illustrative of the design approach and research probe, providing recommendations for other tutorial designs in the Discussion.

When a user downloads their personal data from Facebook they will find, contained in the files, a list of Ad Interests. The list contains words, products, people, media, and concepts that Facebook believes may be relevant to the user. (Ad Interests are not ranked in any way.) However, it is not explicitly clear how these Interests are curated; they are certainly not selected by the user for the purpose of personalizing advertisement. Our design demonstrated core principles of UB-CF on this data for a person with no prior experience with Data Science. The core principles of UB-CF were identified by synthesizing the simplest form of the algorithm:

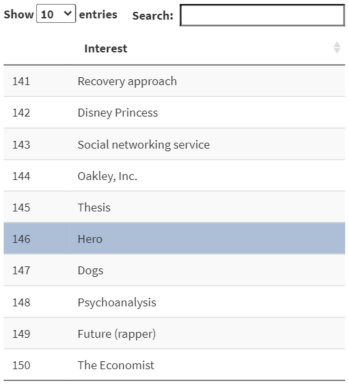
1. Recommendations made to a user can be based on features of other similar users
2. A similarity metric is used to determine which users are the candidates to use for these recommendations
3. The most similar users' interests can be the recommended items

We built the web app using the Shiny package in R to administer the study intervention. The study procedures were reviewed and approved by our university's IRB. Recruited participants visit the domain and encounter on the first page a consent form; this is followed by an introduction with

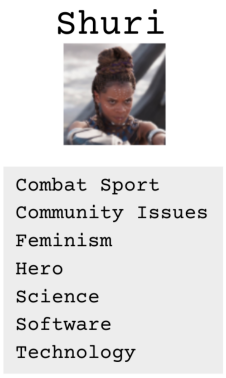
instructions on how to download your Facebook data. We specified that only Ad Interest data was necessary for this tutorial, and we asked users not to explore data prior to the study. Relying on manual data download and upload gives more agency, transparency, and privacy to our participants. We explicitly highlight the fact that no Facebook data would be recorded in the study and that our app only records survey responses. We do not observe the participants' Facebook data. Following the instructions, participants completed a baseline survey while waiting for their Facebook data to download.

In order to further personalize the experience, individuals could upload an image for their own avatar and enter their name. We did not store any personal data. For the purpose of illustrating the participant experience, imagine Shuri from *Black Panther* had a Facebook profile, shown in Figure 2b. In the Marvel Universe, Shuri is a Wakandan princess, scientist, and technologist responsible for much of the technological innovation of Wakanda.

Once a participant uploads their Ad Interests data it appears in a searchable, sortable, and paginated table (see Figure 2a). Interests were presented in random order to allow the participant to see a range of topics. The user was asked to select 7 Interests to use for the remainder of the tutorial. Next, the participant saw their personal avatar and three hypothetical friends, each with listed interests underneath their avatar (all taken from the original data provided by the participant to ensure some overlapping interests). The user could input names for each of the friends and customize their appearance if they desired, see Figure 3.



(a) Facebook interests data is visualized in a clickable, sortable, searchable table. Participants read that items appear in random order.



(b) Selected interests are highlighted and appear in the list of selected interests.

Figure 2: Shuri selects 7 interests actually relevant to her.

Imagine Shuri selected the following interests from her own data: *Combat sport, Community issues, Feminism, Hero, Science, Software,* and *Technology*. Figure 3 shows that Friend1 has two overlapping interests: *Combat sport* and *Hero*. Friend2 has an overlap of five interests: *Combat sport, Hero, Science, Software,* and *Technology*. Friend3 has an overlap of three interests. One of the friends is guaranteed

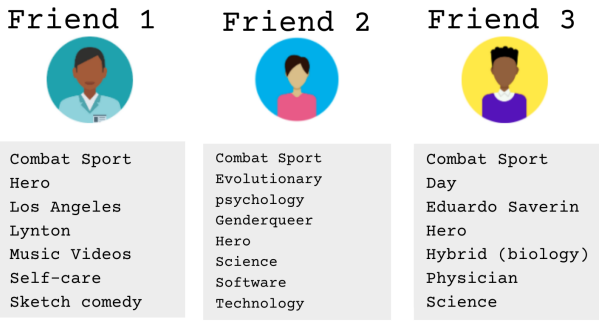


Figure 3: The application interface generates three hypothetical friends, each with some overlapping interests to the study participant. A single friend is guaranteed to have the most overlap. In this example, Friend2 shares the most in common with the participant, Shuri.

to have the most overlap with the study participant (the pre-programmed size of the sample of overlapping interests is unique for each hypothetical friend) and therefore only one friend will be rated as the most similar to the user, avoiding complications with the UB-CF algorithm.

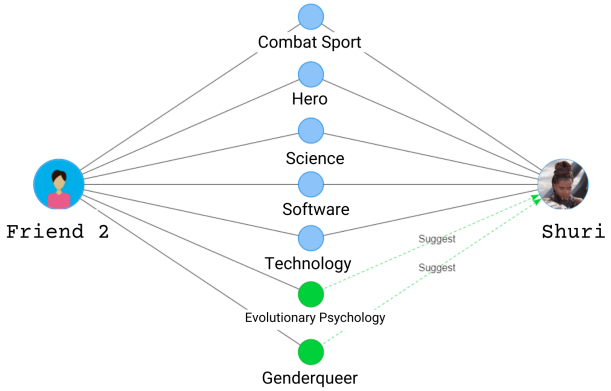


Figure 4: The participant saw a network visualization of shared interests between them and the most similar friend. Interests that may be recommended to Shuri are shown in green (*Evolutionary Psychology and Genderqueer*).

Next, participants saw Ad Interest data represented in matrix form. The columns represented the four people – in this case Shuri and her three friends – and the rows represented all of the possible Ad Interests among the four participants. Cells in the matrix are coded as 1 or 0 if the person in that column is interested in the interest on that row or not, respectively. Providing a matrix representation of the data (as one might use in ML model development) was used as an additional measure of comprehension and engagement with the tutorial. We were curious to see if ML novices who were engaging with their own personal data would be able to draw inferences from this data format. Interpreting the matrix is not trivial, and it is not immediately obvious which of the person pairs are most similar. In order to quantify which friend is the most similar to them, the participant would have

to count the number of rows where both they and their friend had a 1. We asked participants to describe their process in determining which of their friends was the most similar to themselves. Participants were asked to select in a forced-choice response who they thought was the most similar to them. They received feedback on whether or not they were correct. Results from this probe are peripheral to the central arguments of this paper and are therefore not discussed at length, but it is interesting to note that 90% of all participants who saw this question got it correct (N = 35/39) and 87% of *not* Experienced participants got it correct (N=27/31), suggesting that ML novices in the study were able to understand the data presented as a matrix.

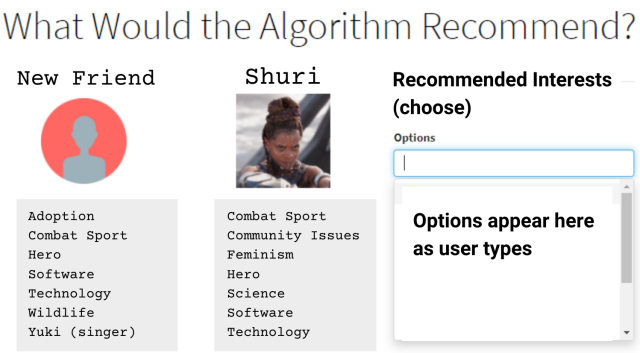


Figure 5: The comprehension test, asking them to report recommended interests based on data from a New Friend.

Next, participants saw a network visualization of their shared interests with their most similar friend, as seen in Figure 4. Figure 4 is a bipartite network, representing users and their interests as two types of nodes. Bipartite networks are a common way to represent users and interests in recommendation based systems, and are often the underlying data structure used in collaborative filtering algorithms. Users are linked or tied by their interests; the visualization thus allows shared interests (those shown in blue) to be easily identified. Interests that are held by their friend, but not themselves, are highlighted in green; these interests of their friend are likely to be recommended to the participant (differentiated with a green dashed arrow). The user can download their own network image as a data keepsake (Lupton 2016). The participants saw a chart detailing the computed similarity metric for each pair of persons, sorted by similarity with the most similar pair at the top. Similarity was computed using cosine similarity (Sarwar et al. 2001), which results in a similarity measure between 0 and 1, with 1 meaning the two vectors are identical. The most similar friend, ranked most highly in this list, is the one whose interests are used as recommendations to the participant. Together, this exercise of exploring one's own interests, a friend's interests, and the overlap of interests provides the basis for a simple UB-CF model.

To test the efficacy of the instruction, and to establish a baseline, participants then took a comprehension test. They saw a New Friend who had 4 overlapping interests with the participants' original 7. Interests for the participant and the New Friend were presented side by side, as in Figure 5. Participants were asked to identify which interests would be recommended to them from the New Friend. They could select from any of the interests on the screen. The correct answer was the one-way "anti join" of the two lists (the Interests that appear on the New Friend's list but not their own). Following the comprehension test, a post-participation survey was the last component of the web application. The entire study took approximately 15 minutes to complete.

Methods

We recruited Facebook users ($N = 77$) of different Data Science Experience levels and tested their comprehension of UB-CF across two conditions: a Baseline Group with no UB-CF instruction (participants saw personal data) and a Tutorial Group where participants received UB-CF instruction on personal data. The Baseline Group included the data viewing page seen in Figure 2 and then proceeded straight to the comprehension task seen in Figure 5 to evaluate if simply *looking* at personal data would give users insight that Ad Interests may be estimated based on their friends. As part of the post-survey, participants responded to the prompt about potentially harmful recommendations on Facebook. These responses are analyzed to identify advocacy themes and to see how participants use new knowledge about UB-CF in their arguments.

Participant Recruitment, Demographics, and Limitations

We employed a quota-based sampling strategy to ensure participation along two axes: participants with a range of data science experience, as well as those in both marginalized and unmarginalized groups (self-identified), as seen in Table 1. We aimed for 15 members of each group, at a minimum. We recruited for this study through authors' social networks and via Facebook groups for various (non-academic) topics, some of them specifically for LGBTQ groups, disability groups, or activism groups. One challenge inherent in working with marginalized people is to establish trust in the researcher-participant relationship. This is difficult without the researcher disclosing their motivations and prior experience. Therefore, the first author relied on word-of-mouth recruitment and disclosure of their own marginalized identities. While this may produce bias towards specific characteristics the participant pool, it allows us access to users who normally would not participate in this kind of research. In fact, several marginalized participants indicated that they would not have done the study for another investigator. We recognize that marginalized people come in with biases, especially against Facebook as captured in our Pre-Survey, but emphasize that these voices are rarely represented in ML literacy research, which often relies on undergraduates or children. Future work is necessary to ensure a larger sample of marginalized individuals, and our results should be viewed with participant bias in mind.

83% of participants were between 21 and 36, with 9% between 37- 45 and 7% over 45. 86% of participants reported that they use Facebook every day, and 11% said they use it about once a week. The remaining 3% said

they use it a couple times a month or less. Gender was not one of our independent variables of interest, though the write-in box for Marginalization status explicitly revealed at least 49% gender minorities in this sample. Other explicitly reported Marginalized identities included race, autism, disability, women working in tech, religion, and sexuality. Some participants simply answered "Yes" or "No".

Participants used their own laptops to complete the study on their own time, ensuring an ecologically valid interaction with the tutorial. Participants were randomly assigned to the Baseline Group or the Tutorial Group as they visited the web application. Every participant received the same pre- and post- surveys; survey responses were stored in a mongodb database. Both Data Science Experience and Marginalization were measured in the survey, as described in Section "Pre- and Post-Surveys" presently. Participants volunteered their time in this study, and were not given any incentive, though several participants did it in order to see their own data (as reported anecdotally to the researchers). Future studies should consider compensation in order to enroll more marginalized individuals.

Data Science Experience	**Marginalized**		**Unmarginalized**		
	Baseline	*Tut*	*Baseline*	*Tut*	**Sum**
Experienced	14	4	5	3	26
Medium	2	7	4	3	16
Novice	10	8	6	11	35
Sum	26	19	15	17	77

Table 1: Participants by Data Science Experience and Marginalization status. Please note that χ^2 tests were not performed on any groups of < 5, but on aggregate groups depending on the research question.

Pre- and Post-Surveys

All participants completed a pre- and post-survey, which asked for basic demographics, willingness to share data, feelings of trust towards Facebook, and a free-response question about how they think Facebook generates recommendations for them. In order to identify participants' level of Data Science Experience, we asked: "Which of the following best describes your experience with Data Science, Computer Science, and/or Machine Learning?" Response options, shown in Table 2, were re-partitioned into Novice, Medium, and Experienced. While difficult to empirically verify, the partition is corroborated by accuracy on the comprehension task in the Baseline Group condition, where we would expect participants coming in with more experience to do better than those without experience.

We asked participants if they identified as marginalized. Participants responded in detail, ranging across sexuality, race, gender identity, neurodivergence, immigration status, and disability. The first author re-coded the responses as a binary variable for anonymity. The survey asked:

☐ I don't know anything about any of those topics.	Novice
☐ I have a vague idea of how some of those things work, but with no formal instruction.	Novice
☐ I have taken classes in any of those subjects.	Medium
☐ I know a fair amount about those topics.	Experienced
☐ My job is in Data Science, Computer Science, and/or Machine Learning (e.g. I have the title of Data Scientist, or do Machine Learning work regularly)	Experienced

Table 2: Survey responses for Data Science Experience, along with partition used in analysis.

> Do you consider yourself a part of a marginalized group? For example, this researcher is nonbinary. Your answer will NOT be shared or linked to your identity. Please describe below. We really appreciate your vulnerability. You may also choose to write "Prefer not to say".

We wanted to know if simply *looking* at your own Facebook data (Baseline Group) affected willingness to share data, and if learning about UB-CF affected that willingness. We asked: "Would you be willing the share the data you just downloaded of what Facebook thinks you're interested in if it were anonymized? Please check all options you would be comfortable sharing anonymized data with." Participants could select multiple choices from: University Researchers, Company Marketing Teams, Other Apps in Your Phone, Political Campaigns , Government Organizations, "I would not be willing to share it", or Other. 13% of participants changed their answers for whether or not they would be willing to share their anonymized Facebook data after looking at their data, and there was no distinct pattern linked to any of the relevant factors (Experience, Condition, Accuracy, Marginalized). Further investigation is needed.

The pre-survey also asked: "Do you trust that Facebook cares about its users and acts with their interests in mind? " following Lankton and McKnight. Only 1 participant said Yes. We also asked: "Do you trust that Facebook's algorithms have the ability to recommend things to you that you actually like?" and 37 participants said Maybe, 16 said No and 24 said Yes. We include these peripheral findings to establish a baseline for our sample and its biases, as well as to offer directions for future work.

In both the pre- and post-survey, we asked for a free-response to the following prompt:

> How do you think Facebook comes up with the list of topics that it thinks you might be interested in? Brainstorm as many ideas as you can. e.g. they gather data from what you click on

We then prompted participants for an advocacy argument:

> Imagine Facebook recommended something harmful to you. Use this space to describe what you think went wrong and what can be done about it.

Evaluating and Comparing Comprehension of UB-CF

We measure comprehension of UB-CF by asking participants which of a list of Interests from a similar *Friend* would be (algorithmically) recommended to them (see Figure 5). The correct answer is any listed Interest of the friend, that the participant does not currently have. We evaluated both the Baseline Group and Tutorial Group for their accuracy on this question. Importantly, the correct answer is a set of 3 Interests. As such, participants could give a partially correct response or a mix of both correct and incorrect responses. To quantify accuracy, we used an F1 Score to capture both Precision and Recall (Sasaki and others 2007). A perfect F1 score is 1. We compare F1 scores between both the Baseline Group and Tutorial Group, and across Experience levels to assess differences in comprehension. We hypothesize for RQ1:

> H1: Those with more Data Science Experience will perform better in the Baseline than Novices, but the Tutorial will improve accuracy for all levels of experience.

An important learning objective was to show to learners how recommendations come from not only their own behavior, but from the behavior of the friends in their network as well. To evaluate this more nuanced signal of UB-CF comprehension we use the textual responses from the survey question on how Facebook generates recommendations. Two coders analyzed the arguments for whether or not they mentioned friends, with an inter-rater reliability (IRR) of .96.

For example, one participant responded: *"The algorithm may have wrongfully computed a suggested interest for me based on interests of those I am **"friends"** with but do not share enough commonalities with."* We further hypothesize for RQ1:

> H2a: More participants will mention the effects of friends on recommendations following the Tutorial.

> H2b: Participants with more Data Science Experience may already know about the effects of friends on recommendations and will mention friend effects in both the Tutorial Group and Baseline Group.

Thematic Analysis of UB-CF Learning Objectives and Potential Harms on Facebook

We are interested in the kinds of potential harms that participants surface after learning about UB-CF on their own data, and how they advocate for themselves when prompted. To extract this from free-response advocacy arguments, we use thematic analysis, following the guidelines of (Nowell et al. 2017). Researchers developed an initial code set by reading a sample of arguments (without knowing the condition, marginalization status, or experience level of the respondent) and identifying recurring words and topics. Next, individual arguments were sorted into affinity groups of the same code, such as "LGBTQ experience" or "politics". At this point, some codes were collapsed in order to produce the final code set shown in Tables 3, 4 and 5. Each theme consisted of a group of arguments with a linking relationship between them that surfaced as a central description for that grouping; we tried to minimize the number of possible groupings to avoid overspecificity. After a period of open coding, two coders labeled the advocacy arguments with their themes, with an inter-rater reliability of .85. In our analysis, we focus on themes with several examples in the data, and discuss these findings.

Results

Participant Comprehension of UB-CF

This study aims first, as stated in RQ1, to evaluate whether participants demonstrate increased comprehension of UB-CF after learning UB-CF with personal data. We evaluate this aim by prompting participants for a fixed choice response with a correct answer, as well as in a free-response question asking participants to describe how they think UB-CF works. Observed F1 scores on the comprehension task, by Data Science Experience and Condition, are shown in Figure 6. As hypothesized (H1), participants in the Baseline Group group with more Data Science experience were more accurate on this task than Novices, with the Medium group in between. This result lends validity to the stratification of participants in terms of Data Science experience at baseline. Following the Tutorial, all levels of Data Science Experience improved to a statistically indistinguishable accuracy ($p = .54$), with a Kruskal Wallis test revealing a significant difference between the Tutorial Group and Baseline Group ($chi^2 = 20.061, df = 1, p < .00001$). To determine if this difference is driven by one level of experience or the other, we also conducted separate Mann Whitney tests that reveal differences between the Tutorial Group and Baseline Group for Medium ($W = 11.5, p = 0.030$) and Novice ($W = 45.5, p < .00001$) levels of experience.

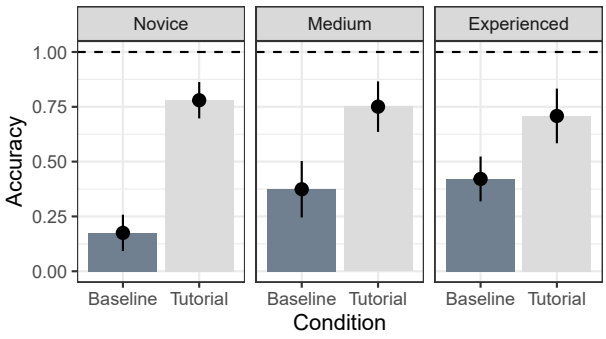


Figure 6: F1 score by Condition and Experience

Testing Hypothesis 2a and 2b, we find that before seeing their own data, 40% of participants included 'friends' in their answers in both the Baseline Group and Tutorial Group, indicating no difference in baseline knowledge before the Tutorial actually occurred. After the completion of the Baseline Group condition, the proportion mentioning friends increases to 60%. In the Tutorial condition, 83% of participants mention friends as part of how Facebook comes up with recommendations. This Tutorial effect is significantly greater than the Baseline ($\chi^2 = 3.9452, df = 1, p = 0.04701$).

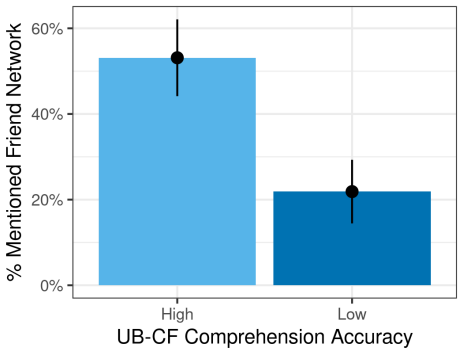


Figure 7: Those with High Accuracy in the UB-CF comprehension task mentioned the effect of friends in their advocacy arguments significantly more than those with Low Accuracy ($\chi^2 = 5.4, p = 0.020$)

Harmful Recommendations and Advocacy

Next, we address the question of how study participants advocate in the face of potentially harmful algorithmic recommendations, as stated in RQ2. We are interested in differences in advocacy arguments between the Tutorial Group and Baseline Group, and across Marginalization status.

Mentioning their Friend Network in their Arguments We stratify participants, grouping those with High accuracy ($F1 >= .6$) on the comprehension task and those with Low accuracy ($F1 < .6$), using the distributions of F1 scores to determine this cutoff. Figure 7 shows that if you understand UB-CF (High Accuracy regardless of whether you learned from the tutorial or if you came in with the knowledge already), you are more likely to use that understanding in your advocacy argument. Recall, mentioning the innerworkings of a problem with actionable points is part of a successful self-advocacy argument (Goodley 2005). We demonstrate here that if you know about the details of UB-CF, you will use those details to advocate.

Participants with High Accuracy mentioned the effect of friends in their advocacy arguments significantly more than those with Low Accuracy ($\chi^2 = 4.35, p = .037$). Figure 7 shows the proportion who mentioned "friends" in their advocacy, such as this iconic argument from the data:

*"A **friend** with other interests in common is into some weird shit and FB assumed that I'm probably into the same weird shit."*

Themes of Advocacy About Risks of UB-CF on Facebook We also surface a non-exhaustive list of themes from participants' advocacy arguments when imagining that Facebook recommended something harmful to them. One category of themes that emerged was specific examples of harm from the influence of peers in the social network, examples seen in Table 3. Next, we see more general commentary about potential harms, generalizable to several behaviors or identities (see Table 4). We also uncovered some themes about solutions to the problem (see Table 5), which is a key part of successful self-advocacy.

Theme	Example Response
Eating Disorders	"Probably recommending diets to someone who has a history of eating disorders. I keep trying to hide those ads and mark them as "sensitive topic". Someone looking up a bunch of diets online is probably interested in diets, but recommending more diets might actually be harmful."
LGBTQ Experience	"Some of my friends/family are still extremely religious. If there's not a way to see if a recommended interest does not work with my current interests, I, a queer person (who fb knows is queer) could get like....recommended conversion therapy because of my conservative family."
Political Ads	"It seems to me that Facebook is factoring the opinions of your Facebook friends into the process. I don't think they should do this at all. People can be friends and not share opinions, especially on Facebook. A liberal person could be friends with a family member who supports Donald Trump. That doesn't mean they support Donald Trump."

Table 3: Themes of specific examples of harm from the influence of network peers.

Each of the themes in Table 3 may provide a basis for further research. Studying the effects of diet and beauty ads on social media users is key to understanding mental health and digital "hygeine" may reveal concerns specific to this context, such as the influence of visual content. The LGBTQ online experience is complex, with the internet often serving as a safe-haven for community but also perpetuating harm (e.g. enforcing strict gender roles or perpetuating violence). Understanding how non-experts perceive political targeting is vital to designing more transparent systems.

Theme	Example Response
"Hate following"	"Hate following happens across industries. Facebook is likely recommending something based on something the user hate followed instead of followed because they actually liked the topic."
Misinformation	"As I'm apparently put in a category of people caring about the environment, I get spammed with all things "natural", so a lot of scam and potentially dangerous "cures". I'd like to see ads making unsubstantial claims gone. Greenwashing should also be forbidden."

Table 4: Themes of algorithmic harm generalizable to many topics, focused on misinterpretation or misrepresentation of network relationships.

Table 4 reveals social media phenomena that can only be understand by representing user voices and what they notice, need, and are concerned about. Current work in this domain does not always capture phenomena that are obvious to actual users. We also surface potential solutions offered in the advocacy arguments, in Table 5, not only to demonstrate potential solutions to the research community but to show that non-experts can be a valuable resource for design solutions.

Theme	Example Response
Personal Behavior Change	"The algorithm took information from my engagement with something similar and used it to make a suggestion it believed would illicit future engagement. Selecting the "see less like this" option or intentionally not engaging in similar content in the future could help correct it. "
Suggestions to FB	"A friend showed interest in (or accidentally clicked or searched for information on) something harmful and it was then recommended to me. Doesn't sound like anything went wrong if it was designed this way. Sounds like it needs to be re-designed to prioritize an ethical process that sees users as human rather than passive money-makers."

Table 5: Themes about possible solutions to the problem

Discussion

"*I would assume that someone else liked or had an interest in that thing, because people can hold harmful values without knowing it.. then those harmful ideas are spread around by Facebook's algorithms and if they go unchecked that can be really problematic. I think more transparency is good, too.*" – P31 (Novice after Tutorial)

In order to increase democratic participation in the design and use of algorithmic-based interactions on social media platforms, we first need to describe what Data Science novices already understand about these systems, and then thoughtfully develop ways for them to learn more about where, how, and why algorithms affect them in their daily lives. It is crucial that we provide pathways for self-advocacy for users to express their needs and potential or experienced harms as they engage with these systems.

This paper contributes: 1) insight into baseline knowledge of a sample of Facebook users about algorithms that underlay recommender systems across all levels of Data Science experience, 2) a successful and personalized intervention design to offer instruction to users about how algorithms work, and 3) empirical demonstration of an association between learning algorithm details and improving self-advocacy arguments when describing potential harms from Facebook's ad recommendation.

At baseline, we find that the majority of our participants do not assume that Facebook uses their friends' data to inform ad recommendations. We see a general distrust in Facebook among participants, but a willingness to explore their own data and provide critique of Facebook's algorithm when prompted. We detail an intervention for teaching users about UB-CF on their own Facebook data, and compare the effects of simply *looking* at your own data to being instructed about UB-CF on that data in terms of comprehension and advocacy. We find that instruction improved comprehension for all levels of Data Science Experience, a boon for both ML pedagogy and widespread ML literacy. We empirically demonstrate an association between accuracy in UB-CF comprehension and using the mechanisms of the algorithm in an advocacy argument about potential harms from the system, providing synchrony between our quantitative and qualitative findings. Successful advocacy often involves negotiation, providing alternative solutions, and understanding how you are being harmed. We show that participants who successfully learned about UB-CF were more likely to use that knowledge in their arguments. Participants advocated for potential solutions, such as unfollowing or hiding ads, exposing themselves to content outside their usual interests to avoid echo chambers, demanding more transparency from Facebook, asking for a feature where they could filter *which* friends are used in recommendation algorithms, and leaving Facebook altogether.

Marginalized users used instruction on UB-CF to help express specific ways that algorithms contribute to harm on Facebook. Surfacing these hypothetical and experienced harms is vital if we – social media researchers and designers – want to meet the needs of marginalized stakeholders and understand differential effects of algorithms in these settings. The insight from Marginalized participants is valuable and detailed, ranging from concerns about religious influence on their queer identity (e.g. *"for instance, I am LGBTQ (closeted) and have friends who are more conservative/fundamentalist Christian. They may have interests that are harmful to my identity (e.g. pray the gay away) that would be pushed to me."*), to pointing out harms perpetuated by the beauty and diet industry (e.g. *"recommending diets to someone who has a history of eating disorders. I keep trying to hide those ads and mark them as "sensitive topic". Someone looking up a bunch of diets online is probably interested in diets, but recommending more diets might actually be harmful."*)

Pathways for Future Work

The results of this study are a starting point for future efforts both on ML literacy, as well as successful advocacy in cases of algorithmic harm. Here we suggest some specific pathways that we believe will yield important research and design; we aim to pursue many of these questions.

Machine Learning Education

Many machine learning lessons rely on unrelatable datasets such as the classic `iris` or `mtcars` datasets, or outdated and implicitly biased, the `Boston housing dataset` from the 1970s. Use of such datasets is perhaps unsurprising – they are clean and nicely demonstrate ML algorithms and models. Our work demonstrates one way to integrate relevant, interesting data into ML instruction. Because learners have expertise about their own experiences, they can apply this knowledge to ask targeted questions about algorithms and data they are learning about in context. ML education and literacy efforts, as well as research on these topics, should leverage relevant contexts and associated data such as social media algorithms, Google searches, face filters, and other ML systems that are being used in the real world. Doing so can increase comprehension and support learners to evaluate the content they see. See the section on Designing Tutorials with Personal Data for further guidance.

Self-Advocacy of Novice ML Users

Work on algorithmic bias and harm often disregards the knowledge and domain expertise of ML Novices, referring to them as "laymen" or "everyday people". Our results suggest that Novices can not only learn about ML topics, but also that they will use what they learn in advocacy arguments. Social media researchers have the ability to probe and highlight these valuable insights. The path to ML literacy involves both *top-down* (designers and engineers programming the systems) and *bottom-up* (users and stakeholders critiquing the systems) approaches. Platforms should provide more opportunities for user feedback specifically regarding the algorithmic portions of the user experience. This could mean offering more user *control* (Burrell et al. 2019) such as sliders, filters, misinformation flags, or access to the "weights" on their newsfeed content, for example. Researchers should take care to include perspectives of users when discussing algorithmic harm, especially because what researchers assume is harmful may differ from what users feel is harmful. Algorithmic bias can be especially damaging for marginalized communities. Researchers have an opportunity to focus on the researcher-participant relationship in order to successfully involve marginalized users in a non-extractive way. For example, see the frameworks presented in *Data Feminism* (D'Ignazio and Klein 2020).

Designing Tutorials with Personal Data as Research Probes

For future design of ML tutorials using personal data, here are a few guidelines to consider based on our experience and research agenda in this space.

Considerations for Using Personal Data The user must consent and know if their data will be stored and how it will be used in any research study. It is also possible, as illustrated here, to make use of personal data while also maintaining participant privacy. For this work, we allowed the user to upload their own data files and did not store their personal data, only their responses to our prompts. Personal data should never be used in a group setting. Any instruction should be mindful of potentially sensitive content that the user may not expect or intend to inspect. This could be explicitly sensitive content such as aspects of identity, or even less obvious content such as recent experiences or events that are recorded in the personal data (e.g. reminds the person of a breakup or death). We recommend that participants filter their own data for what they would like to use for the remainder of the tutorial (as we did in this study); while it is difficult to avoid unanticipated sensitive data altogether, this process can lessen exposure and harm.

Considerations and Ideas for Designing ML Tutorials Our strategy was first to identify platforms that allow the user to download their own data (such as Facebook, Instagram, Google, Twitter, etc.), along with the data formats available. Data is, in fact, often a starting point for research of this nature. One might next consider common ML algorithms used on the focal platform. Social media systems are build with layers of algorithms, and as such there are likely many to consider as topics for instruction, e.g. recommender systems, misinformation or hate speech detection, image recognition, etc. We found it key to consider which algorithms have a significant impact on the user, not only in terms of frequency of interaction but also potential for harm. Algorithms that users may not even realize are powering their experience can be especially fruitful targets (Rader and Gray 2015). A final consideration is the instructional demonstration itself. It needs to realistically incorporate personal data to show the user a basic form of the algorithm; visualizations and other tools can be helpful in communicating about algorithms to participants.

Examples for Future Tutorials In light of the above considerations, and based on our experiences in this project, we believe the following interventions will be promising directions for work:

- demonstrating image classification or object identification on a user's own images, especially with regards to images that get banned or reported
- introducing participants to algorithms powering *DeepFakes* using their own shared videos, and reflect on the dangers of such tools
- show clustering algorithms on a user's personal network on social media platforms

Limitations

Given the sample size and recruitment strategy in this work, results should be replicated on a generalizable sample to confirm the efficacy of this tutorial. This work represents a roadmap for future efforts, including a novel approach to include the user's own data in learning about algorithmic systems and potential harm. A notable limitation in this work was the short responses given for advocacy arguments; making them less fit for in-depth qualitative analysis – necessary to fully understand stakeholder voices. Another limitation is the lack of random sampling of participants; we instead purposely highlight the voices of marginalized groups as a first step in exploring this topic. We did not imitate Facebook's exact algorithm in our instruction, but instead used a rudimentary version. While this may affect the real-world applicability of user advocacy arguments, we believe that core literacies should transfer to other similarity-based recommendation algorithms. Future work might attempt to replicate results across different algorithms, or alternatively, build from real cases of harm before prompting advocacy.

Conclusion

Social media users interact with potentially harmful algorithms every day, but through education, users are able advocate for themselves when such harms occur. This paper represents a step to democratizing ML by promoting understanding and advocacy for Facebook users. Whether its Shuri from Wakanda or a teen scrolling social media, we can all benefit from increased literacy and self-advocacy skills to participate in an ML-driven world.

References

Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Robinson, D. G. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260.

Alkhatib, A., and Bernstein, M. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Alkhatib, A. 2021. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–9.

Are, C. 2020. How instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies* 20(5):741–744.

Barbosa, N. M., and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 ACM Conference on Human Factors in Computing Systems*, 543.

Bart, A. C.; Whitcomb, R.; Kafura, D.; Shaffer, C. A.; and Tilevich, E. 2017. Computing with corgis: Diverse, real-world datasets for introductory computing. *ACM Inroads* 8(2):66–72.

Benhabib, S. 1992. *Situating the self: Gender, community, and postmodernism in contemporary ethics*. Psychology Press.

Benjamin, R. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces*.

Bessi, A. 2016. Personality traits and echo chambers on facebook. *Computers in Human Behavior* 65:319–324.

Boratto, L.; Fenu, G.; and Marras, M. 2019. The effect of algorithmic bias on recommender systems for massive open online courses. In *European Conference on Information Retrieval*, 457–472. Springer.

Bucher, T. 2012. Want to be on the top? algorithmic power and the threat of invisibility on facebook. *New media & society* 14(7):1164–1180.

Burrell, J.; Kahn, Z.; Jonas, A.; and Griffin, D. 2019. When users control the algorithms: Values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–20.

Carton, S.; Mei, Q.; and Resnick, P. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 95–106.

Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

D'Ignazio, C., and Bhargava, R. 2016. Databasic: Design principles, tools and activities for data literacy learners. *The Journal of Community Informatics* 12(3).

D'Ignazio, C., and Klein, L. F. 2020. *Data feminism*. Mit Press.

Druga, S.; Vu, S. T.; Likhith, E.; and Qiu, T. 2019. Inclusive ai literacy for kids around the world. In *Proceedings of FabLearn 2019*, 104–111. ACM.

D'Ignazio, C., and Bhargava, R. 2015. Approaches to building big data literacy. In *Proceedings of the Bloomberg data for good exchange conference*.

D'Ignazio, C., and Klein, L. F. 2016. Feminist data visualization. In *Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE*.

Edizel, B.; Bonchi, F.; Hajian, S.; Panisson, A.; and Tassa, T. 2020. Fairecsys: Mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9(2):197–213.

Eslami, M.; Vaccaro, K.; Karahalios, K.; and Hamilton, K. 2017. "be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Fabbri, F.; Bonchi, F.; Boratto, L.; and Castillo, C. 2020. The effect of homophily on disparate visibility of minorities in people recommender systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 165–175.

Garcia, M. 2016. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal* 33(4):111–117.

Goodley, D. 2005. Empowerment, self-advocacy and resilience. *Journal of Intellectual Disabilities* 9(4):333–343.

Haimson, O. L.; Delmonaco, D.; Nie, P.; and Wegner, A. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–35.

Hitron, T.; Orlev, Y.; Wald, I.; Shamir, A.; Erel, H.; and Zuckerman, O. 2019. Can children understand machine learning concepts?: The effect of uncovering black boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 415. ACM.

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*, 600. ACM.

Johnson, G. M. 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese* 198(10):9941–9961.

Kabiljo, M., and Ilic, A. 2015. Recommending items to more than a billion people. *Retrieved May* 2:2018.

Karizat, N.; Delmonaco, D.; Eslami, M.; and Andalibi, N. 2021. Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–44.

Kim, N. W.; Im, H.; Henry Riche, N.; Wang, A.; Gajos, K.; and Pfister, H. 2019. Dataselfie: Empowering people to design personalized visuals to represent their data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 79. ACM.

Kingsley, S.; Wang, C.; Mikhalenko, A.; Sinha, P.; and Kulkarni, C. 2020. Auditing digital platforms for discrimination in economic opportunity advertising. *arXiv preprint arXiv:2008.09656*.

Lankton, N. K., and McKnight, D. H. 2011. What does it mean to trust facebook? examining technology and interpersonal trust beliefs. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 42(2):32–54.

Lupton, D. 2016. You are your data: Self-tracking practices and concepts of data. In *Lifelogging*. Springer. 61–79.

Nowell, L. S.; Norris, J. M.; White, D. E.; and Moules, N. J. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16(1):1609406917733847.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

Oliva, T. D.; Antonialli, D. M.; and Gomes, A. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture* 25(2):700–732.

O'Callaghan, D.; Greene, D.; Conway, M.; Carthy, J.; and Cunningham, P. 2015. Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review* 33(4):459–478.

Peck, E. M.; Ayuso, S. E.; and El-Etr, O. 2019. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 244. ACM.

Prado, J. C., and Marzal, M. Á. 2013. Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri* 63(2):123–134.

Quattrociocchi, W.; Scala, A.; and Sunstein, C. R. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.

Rader, E., and Gray, R. 2015. Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 173–182.

Register, Y., and Ko, A. J. 2020. Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*, ICER '20, 67–78. New York, NY, USA: Association for Computing Machinery.

Samory, M.; Abnousi, V. K.; and Mitra, T. 2020. Characterizing the social media news sphere through user co-sharing practices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 602–613.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, 285–295.

Sasaki, Y., et al. 2007. The truth of the f-measure. 2007. *URL: https://www. cs. odu. edu/˜ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf [accessed 2021-05-26]*.

Schield, M. 2004. Information literacy, statistical literacy and data literacy. In *Iassist Quarterly (IQ)*. Citeseer.

Shen, H.; DeVos, A.; Eslami, M.; and Holstein, K. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *arXiv preprint arXiv:2105.02980*.

Smith-Renner, A.; Fan, R.; Birchfield, M.; Wu, T.; Boyd-Graber, J.; Weld, D. S.; and Findlater, L. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–13. New York, NY, USA: Association for Computing Machinery.

Tanweer, A.; Bolten, N.; Drouhard, M.; Hamilton, J.; Caspi, A.; Fiore-Gartland, B.; and Tan, K. 2017. Mapping for accessibility: A case study of ethics in data science for social good. *CoRR* abs/1710.06882.

Test, D. W.; Fowler, C. H.; Wood, W. M.; Brewer, D. M.; and Eddy, S. 2005. A conceptual framework of self-advocacy for students with disabilities. *Remedial and Special education* 26(1):43–54.

Usher, N.; Holcomb, J.; and Littman, J. 2018. Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The international journal of press/politics* 23(3):324–344.

Velkova, J., and Kaun, A. 2021. Algorithmic resistance: media practices and the politics of repair. *Information, Communication & Society* 24(4):523–540.

Wehmeyer, M.; Bersani, H.; and Gagne, R. 2000. Riding the third wave: Self-determination and self-advocacy in the 21st century. *Focus on autism and other developmental disabilities* 15(2):106–115.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 279. ACM.

Zimmermann-Niefield, A.; Turner, M.; Murphy, B.; Kane, S. K.; and Shapiro, R. B. 2019. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 121–132.