How Fake News Affect Trust in the Output of a Machine Learning System for News Curation

Hendrik Heuer[1,2](✉) and Andreas Breiter[1,2]

[1] Institute for Information Management Bremen GmbH (ifib), Bremen, Germany
{hheuer,abreiter}@ifib.de

[2] Centre for Media, Communication and Information Research (ZeMKI), University of Bremen, Bremen, Germany
{hheuer,abreiter}@uni-bremen.de

Abstract. People are increasingly consuming news curated by machine learning (ML) systems. Motivated by studies on algorithmic bias, this paper explores which recommendations of an algorithmic news curation system users trust and how this trust is affected by untrustworthy news stories like fake news. In a study with 82 vocational school students with a background in IT, we found that users are able to provide trust ratings that distinguish trustworthy recommendations of quality news stories from untrustworthy recommendations. However, a single untrustworthy news story combined with four trustworthy news stories is rated similarly as five trustworthy news stories. The results could be a first indication that untrustworthy news stories benefit from appearing in a trustworthy context. The results also show the limitations of users' abilities to rate the recommendations of a news curation system. We discuss the implications of this for the user experience of interactive machine learning systems.

Keywords: Human-centered machine learning · Algorithmic experience · Algorithmic bias · Fake news · Social media

1 Introduction

News curation is the complex activity of selecting and prioritizing information based on some criteria of relevance and in regards to limitations of time and space. While traditionally the domain of editorial offices of newspapers and other media outlets, this curation is increasingly performed by machine learning (ML) systems that rank the relevance of content [3,17]. This means that complex, intransparent ML systems influence the news consumption of billions of users. Pew Research Center found that around half of U.S. adults who use Facebook (53%) think they do not understand why certain posts are included in their news feeds [2]. This motivates us to explore how users perceive news recommendations and whether users can distinguish trustworthy from untrustworthy ML recommendations. We also examine whether untrustworthy news stories like fake news

M. van Duijn et al. (Eds.): MISDOOM 2020, LNCS 12259, pp. 18–36, 2020.
https://doi.org/10.1007/978-3-030-61841-4_2

benefit from a trustworthy context, for instance, when an ML system predicts five stories, where four are trustworthy news stories and one is a fake news story. We operationalized the term fake news as "fabricated information that mimics news media content in form but not in organizational process or intent" [28]. Investigating trust and fake news in the context of an algorithmic news curation is important since such algorithms are an integral part of social media platforms like Facebook, which are a key vector of fake news distribution [3]. Investigations of trust in news and people's propensity to believe in rumors has a long history [4,5].

We focus on trust in a news recommender system, which connects to O'Donovan et al. and Massa and Bhattacharjee [35,41]. Unlike them, our focus is not the trust in the individual items, but the trust in the ML system and its recommendations. The design of the study is shaped by how users interact with machine learning systems. Participants rate their trust in the recommendations of a machine learning system, i.e. they rate groups of news stories. Participants were told that they are interacting with an ML system, i.e. that they are not simply rating the content. We focus on trust because falling for fake news is not simply a mistake. Fake news are designed to mislead people by mimicking news media content. Our setting connects to human-in-the-loop and active machine learning, where users are interacting with a live system that they improve with their actions [7,26,52]. In such settings, improving a news curation algorithm by rating individual items would require a lot of time and effort from users. We, therefore, explore explicitly rating ML recommendations as a whole as a way to gather feedback.

An investigation of how ML systems and their recommendations are perceived by users is important for those who apply algorithmic news curation and those who want to enable users to detect algorithmic bias in use. This is relevant for all human-computer interaction designers who want to enable users to interact with machine learning systems. This investigation is also relevant for ML practitioners who want to collect feedback from users on the quality of their systems or practitioners who want to crowdsource the collection of training data for their machine learning models [20,53,58].

In our experiment, participants interacted with a simple algorithmic news curation system that presented them with news recommendations similar to a collaborative filtering system [22,49]. We conducted a between-subjects study with two phases. Our participants were recruited in a vocational school. They all had a technical background and were briefed on the type of errors that ML systems can make at unexpected times. In the first phase, participants rated their trust in different news stories. This generated a pool of news stories with trust ratings from our participants. Participants rated different subsets of news stories, i.e. each of the news stories in our investigation was rated by some users while others did not see it. In the second phase, the algorithmic news curation system combined unseen news stories for each user based on each news stories' median trust rating. This means that the trust rating of a story is based on the intersubjective agreement of the participants that rated it in the first phase.

This allowed us to investigate how the trust in individual stories influences the trust in groups of news stories predicted by an ML system. We vary the number of trustworthy and untrustworthy news stories in the recommendations to study their influence on the trust rating on an 11-point rating scale. Our main goal is to understand the trust ratings of ML output as a function of the trust of individual news items for a machine learning system. In summary, this paper answers the following three research questions:

- Can users provide trust ratings for news recommendations of a machine learning system (RQ1)?
- Do users distinguish trustworthy ML recommendations from untrustworthy ML recommendations (RQ2)?
- Do users distinguish trustworthy ML recommendations from recommendations that include one individual untrustworthy news story (RQ3)?

We found that users are able to give nuanced ratings of machine learning recommendations. In their trust ratings, they distinguish trustworthy from untrustworthy ML recommendations, if all stories in the output are trustworthy or if all are untrustworthy. However, participants are not able to distinguish trustworthy news recommendations from recommendations that include one fake news story. Even though they can distinguish other ML recommendations from trustworthy recommendations.

2 Related Work

The goal of news recommendation and algorithmic news curation systems is to model users' interests and to recommend relevant news stories. An early example of this is GroupLens, a collaborative filtering architecture for news [49]. The prevalence of opaque and invisible algorithms that curate and recommend news motivated a variety of investigations of user awareness of algorithmic curation [17,18,21,46]. A widely used example of such a machine learning system is Facebook's News Feed. Introduced in 2006, Facebook describes the News Feed as a "personalized, ever-changing collection of posts from the friends, family, businesses, public figures and news sources you've connected to on Facebook" [19]. By their own account, the three main signals that they use to estimate the relevance of a post are: who posted it, the type of content, and the interactions with the post. In this investigation, we primarily focus on news and fake news on social media and the impact of the machine learning system on news curation.

Alvarado and Waern coined the term algorithmic experience as an analytic framing for making the interaction with and experience of algorithms explicit [6]. Following their framework, we investigate the algorithmic experience of users of a news curation algorithm. This connects to Shou and Farkas, who investigated algorithmic news curation and the epistemological challenges of Facebook [54]. They address the role of algorithms in pre-selecting what appears as representable information, which connects to our research question whether users can detect fake news stories.

This paper extends on prior work on algorithmic bias. Eslami et al. showed that users can detect algorithmic bias during their regular usage of online hotel rating platforms and that this affects trust in the platform [18]. Our investigation is focused on trust as an important expression of users' beliefs. This connects to Rader et al., who explored how different ways of explaining the outputs of an algorithmic news curation system affects users' beliefs and judgments [45]. While explanations make people more aware of how a system works, they are less effective in helping people evaluate the correctness of a system's output.

The Oxford dictionary defines trust as the firm belief in the reliability, truth, or ability of someone or something [14]. Due to the diverse interest in trust, there are many different definitions and angles of inquiry. They range from trust as an attitude or expectation [48,50], to trust as an intention or willingness to act [36] to trust as a result of behaviour [13]. Trust was explored in a variety of different contexts, including, but not limited to intelligent systems [22,56], automation [29,38,39], organisations [36], oneself [34], and others [50]. Lee and See define trust as an attitude of an agent with a goal in a situation that is characterized by some level of uncertainty and vulnerability [29]. The sociologist Niklas Luhmann defined trust as a way to cope with risk, complexity, and a lack of system understanding [30]. For Luhmann, trust is what allows people to face the complexity of the world. Other trust definitions cite a positive expectation of behavior and reliability [39,50,51].

Our research questions connect to Cramer et al., who investigated trust in the context of spam filters, and Berkovsky et al., who investigated trust in movie recommender systems [9,12]. Cramer et al. found that trust guides reliance when the complexity of an automation makes a complete understanding impractical. Berkovsky et al. argue that system designers should consider grouping the recommended items using salient domain features to increase user trust, which supports earlier findings by Pu and Chen [44]. In the context of online behavioral advertising, Eslami et al. explored how to communicate algorithmic processes by showing users why an ad is shown to them [16]. They found that users prefer interpretable, non-creepy explanations.

Trust ratings are central to our investigation. We use them to measure whether participants distinguish trustworthy from untrustworthy machine learning recommendations and investigate the influence of outliers. A large number of publications used trust ratings as a way to assess trust [32,38,39,43]. In the context of online news, Pennycook and Rand showed that users can rate trust in news sources and that they can distinguish mainstream media outlets from hyperpartisan or fake news sources [43]. Muir et al. modeled trust in a machine based on interpersonal trust and showed that users can meaningfully rate their trust [39]. In the context of a pasteurization plant simulation, Muir and Moray showed that operators' subjective ratings of trust provide a simple, nonintrusive insight into their use of the automation [38]. Regarding the validity of such ratings, Cosley et al. showed that users of recommender system interfaces rate fairly consistently across rating scales and that they can detect systems that manipulate outputs [11].

3 Methods

To explore trust in the context of algorithmic news curation, we conducted an experiment with 82 participants from a vocational school with a focus on IT. In the first phase of the study, participants with a technical background rated individual news stories, one at a time. In the second phase of the study, participants rated ML recommendations, i.e. five news stories that were presented together as the recommendations of an algorithmic news curation system. The study was conducted in situ via a web application that presented the two phases.

We recruited a homogeneous group of participants in a German vocational school. To prevent a language barrier from adding bias, the experiment was conducted in German. In Germany, the performance of students is strongly dependent on socio-economic factors [42]. Students of a vocational school, which starts after compulsory schooling, have a similar background. This allows us to control for age, educational background, and socio-economic background. The mean age of the 82 participants was 21.40 (SD = 3.92). The school had a strong STEM focus: All of the six classes were trained in IT (but they had no formal training in machine learning). The IT focus of the vocational school introduced a gender bias: 73 participants identified as male, 5 as female, 2 chose not to disclose their gender and 2 identified as a non-binary gender. This gender bias is representative of a vocational school with a STEM focus in Germany. In the training year 2016, women only accounted for 7.9% of new IT trainees in Germany [1].

Like Muir et al. and Cramer et al., we adopt Luhmann's definition of trust as a way to cope with risk, complexity, and a lack of system understanding [12,30,39]. Our operationalization focuses on interpersonal and social trust, which can be described as the generalized expectancy that a person can rely on the words or promises of others [50]. When consuming news, a person is making herself or himself reliant on a highly complex system that involves journalists, publishers, and interviewees. When interacting with an algorithmic news curation system, a person is making herself or himself reliant on a highly complex socio-technical system, which cannot be understood entirely and which can malfunction for myriad reasons. Each part of the system poses a risk, either due to mistakes, misunderstandings, or malicious intent. A social media platform that performs algorithmic news curation includes actors like the platform provider, the advertisers, other users, and all the different news sources with different levels of trustworthiness. All add complexity and risk. Understanding and auditing how this socio-technical system works is neither possible nor practical.

Before the experiment, we explained the rating interface, provided Mitchell's definition of ML, and briefly mentioned ML applications like object detection and self-driving cars. According to Mitchell, "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [37]. To illustrate this, we showed participants how an ML algorithm learns to recognize hand-written digits. This was meant to show how and why some digits are inevitably misclassified. Algorithmic news curation was introduced as another machine learning application. The term fake news was illustrated using examples like Pope Francis backing Trump and the German Green party banning meat.

3.1 Rating News Stories (Phase 1)

The task in the first phase was to provide trust ratings for news stories from different sources. In this phase, participants evaluated each piece of content individually. As news stories, we used two days of publicly available Facebook posts of 13 different sources. The study was conducted in May 2017, i.e. before the Cambridge Analytica scandal and before the Russian interference in the 2016 United States elections became publicly known.

We distinguish between seven quality media sources, e.g. public-service broadcasters and newspapers of record, and six biased sources, including tabloid media and fake news blogs. The quality media sources and the tabloid sources were selected based on their reach as measured by Facebook likes. Fake news sources were selected based on mentions in news articles on German fake news [8]. Tabloid newspapers are characterized by a sensationalistic writing style and limited reliability. But, unlike fake news, they are not fabricated or intentionally misleading. For our experiment, a weighted random sample of news stories was selected from all available posts. Each of the 82 participants rated 20 news stories from a weighted random sample consisting of eight quality media news stories, four tabloid news stories, and eight fake news stories. The weighted sample accounted for the focus on fake news and online misinformation. The selected stories cover a broad range of topics, including sports like soccer, social issues like homelessness and refugees, and stories on politicians from Germany, France, and the U.S.

The presentation of the news stories resembled Facebook's official visual design. For each news story, participants saw the headline, lead paragraph, lead image, the name of the source, source logo, source URL, date and time, as well as the number of likes, comments, and shares of the Facebook post. Participants were not able to click on links or read the entire article. The data was not personalized, i.e. all participants saw the same number of likes, shares, and comments that anybody without a Facebook account would have seen if s/he would have visited the Facebook Page of the news source. In the experiment, participant rated news stories on an 11-point rating scale. The question they were asked for each news story was: "Generally speaking, would you say that this news story can be trusted, or that you can't be too careful? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that this news story can be trusted". Range and phrasing of the question are modeled after the first question of the Social Trust Scale (STS) of the European Social Survey (ESS) which is aimed at interpersonal trust and connected to the risk of trusting a person respectively a news story [47]. After the experiment, the ratings of the news stories from Phase 1 were validated with media research experts. Each media researcher ranked the news sources by how trustworthy they considered the source. These rankings were compared to the median trust ratings of the news sources by the users. The experts were recruited from two German labs with a focus on media research on public communication and other cultural and social domains. All members of the two labs were contacted through internal newsletters. In a self-selection sample, nine media researcher (three male, six female) provided their ranking via e-mail (two from lab A, seven from lab B).

3.2 Rating News Recommendations (Phase 2)

In the second phase, participants rated their trust in the output of a news curation system. The task was not to identify individual fake news items. Participants rated the ML recommendations as a group selected by an ML system. In the study, the output of the ML system always consisted of five unseen news stories. We selected the unseen news stories based on their median trust ratings from Phase 1. The median is used as a robust measure of central tendency [24], which captures intersubjective agreement and which limits the influence of individual outliers. We adapted our approach from collaborative filtering systems like GroupLens [22,49]. Collaborative filtering systems identify users with similar rating patterns and use these similar users to predict unseen items. Since our sample size was limited, we couldn't train a state-of-the-art collaborative filtering system. Therefore, we used the median trust rating as a proxy.

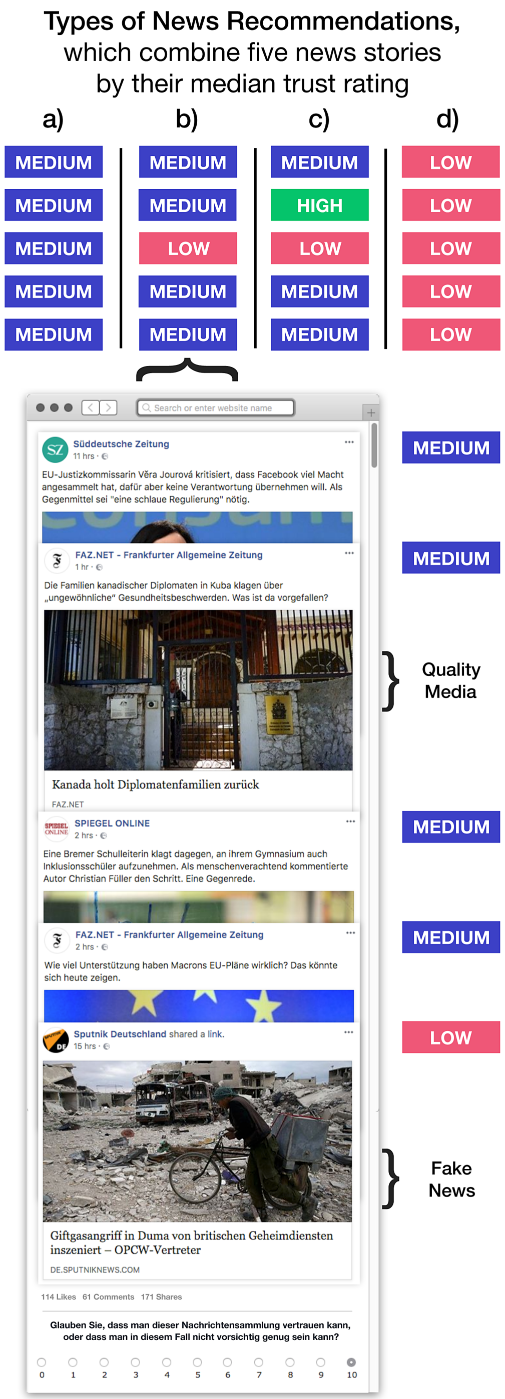
Our goal was to understand how the presence of fake news changes the feedback users give for a machine learning system and whether trust ratings account for the presence of fake news. Our motivation was to explore how fine-grained the user feedback on a system's performance is. This is important for fields like active learning or interactive and mixed-initiative machine learning [7,23,25,55], where user feedback is used to improve the system. While the experiment brief made people believe that they were interacting with a personalized ML system, the recommendations were not actually personalized. We did this to be able to compare the ratings. Unlike in Wizard of Oz experiments, there was no experimenter in the loop. Users freely interacted with an interactive software system that learned from examples.

3.3 Types of News Recommendations

To investigate how the trust ratings of the recommendations change based on the trustworthiness of the individual news stories, we combine five news stories in random order with different levels of trustworthiness. The scale ranges from "can't be too careful (0)" to "can be trusted (10)". We refer to the trustworthiness of a news story as low (if the trust rating is between 0 and 3), medium (4 to 6), and high (7 to 10).

Figure 1 shows the four types of news recommendations that we discuss in this paper as well as the rating interface.

- **a) Medium**—ML output that consists of five news stories with median trust ratings between 4 and 6.
- **b) Medium, 1 Low**—ML output with four news stories with ratings between 4 and 6 and one with a rating between 0 and 3 (shown in Fig. 1).



Types of News Recommendations,
which combine five news stories
by their median trust rating

a)	b)	c)	d)
MEDIUM	MEDIUM	MEDIUM	LOW
MEDIUM	MEDIUM	HIGH	LOW
MEDIUM	LOW	LOW	LOW
MEDIUM	MEDIUM	MEDIUM	LOW
MEDIUM	MEDIUM	MEDIUM	LOW

Süddeutsche Zeitung
11 hrs ·

EU-Justizkommissarin Věra Jourová kritisiert, dass Facebook viel Macht angesammelt hat, dafür aber keine Verantwortung übernehmen will. Als Gegenmittel sei "eine schlaue Regulierung" nötig.

FAZ.NET - Frankfurter Allgemeine Zeitung
1 hr ·

Die Familien kanadischer Diplomaten in Kuba klagen über „ungewöhnliche" Gesundheitsbeschwerden. Was ist da vorgefallen?

Kanada holt Diplomatenfamilien zurück

FAZ.NET

SPIEGEL ONLINE
2 hrs ·

Eine Bremer Schulleiterin klagt dagegen, an ihrem Gymnasium auch Inklusionsschüler aufzunehmen. Als menschenverachtend kommentierte Autor Christian Füller den Schritt. Eine Gegenrede.

FAZ.NET - Frankfurter Allgemeine Zeitung
2 hrs ·

Wie viel Unterstützung haben Macrons EU-Pläne wirklich? Das könnte sich heute zeigen.

Sputnik Deutschland shared a link.
15 hrs ·

Giftgasangriff in Duma von britischen Geheimdiensten inszeniert – OPCW-Vertreter

DE.SPUTNIKNEWS.COM

114 Likes 61 Comments 171 Shares

Glauben Sie, dass man dieser Nachrichtensammlung vertrauen kann, oder dass man in diesem Fall nicht vorsichtig genug sein kann?

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ● 10

Fig. 1. For Phase 2, different types of ML recommendations were generated by combining five news stories from Phase 1 by their median trust rating. Participants rated the trustworthiness of these collections of unseen news stories with a single score on an 11-point rating scale.

- **c) Medium, 1 Low, 1 High**—ML output that consists of three medium news stories, one with a low trust rating and one with a high rating between 7 and 10.
- **d) Low**—ML output where all news stories have a trust rating between 0 and 3.

Our goal was to show as many different combinations of news recommendations to participants as possible. Unfortunately, what we were able to test depended on the news ratings in the first phase. Here, only a small subset of participants gave high ratings. This means that news recommendations like **High, 1 Low**, as well as **Low, 1 High** could not be investigated. Figure 1 shows the different types of ML recommendations that were presented to more than ten participants. In the figure, the five stories are shown in a collapsed view. In the experiment, participants saw each news story in its full size, i.e. the texts, images and the number of shares, likes, and comments were fully visible for each of the five news stories in the news recommendation. The news stories were presented in a web browser where participants were able to scroll. Participant rated the news recommendation on the same 11-point rating scale as the individual news items, where 0 was defined as "you can't be too careful" and 10 as "this collection of news stories can be trusted".

4 Results

In Phase 1, participants were presented with individual news stories, which they rated one at a time. The news stories came from 13 different news sources. Each participant rated 20 news stories (8 quality media, 4 tabloid, and 8 fake news stories). More than half (53.47%) of the trust ratings are rated as low (with a rating between 0 and 3). 28.22% are rated as medium (rated 4, 5 and 6) and 18.32% high (7 and 10).

The first goal of this section is to establish whether our method and the trust ratings are valid. For this, we grouped the news stories by source and ranked them by their median trust rating (Table 1). The most trustworthy news source is a conservative newspaper of record with a median trust rating of 6.0 (N = 256). The least trustworthy news sources is a fake news blog with a median trust rating of 1.0 (N = 129). Participants distinguish quality media (Sources A to F) from tabloid media and fake news blogs (G to M). There is one exception: Rank H is a quality media source - produced by the public-service television - which received a median trust rating of 4.0 and which is ranked between tabloid media and fake news. Unlike the other news sources, this median trust rating is only based on one article and 25 responses. The median ratings of all other news sources are based on four or more news articles and more than 100 ratings per news source (with a maximum of 258 ratings for 10 articles from source G). The fake news outlets are ranked as I (9), K (11), and M (13).

We validated the trust ratings of news items by comparing them to rankings of the news sources by nine media researchers (three male, six female), also

Table 1. Quality media sources (marked in green) are distinguished from tabloid media (yellow) and fake news sources (red) in the Participants' Ranking ($N = 82$, median trust rating) and the Media Researchers' Rankings ($N = 9$).

Rank	Trust Ratings		News Sources Ranked By Media Researchers								
1	6.0	🟩 A	🟩 H	🟩 D	🟩 D	🟩 D	🟩 C	🟩 C	🟩 D	🟩 H	🟩 H
2	6.0	🟩 B	🟩 C	🟩 H	🟩 C	🟩 B	🟩 B	🟩 H	🟩 H	🟩 D	🟩 D
3	6.0	🟩 C	🟩 D	🟩 C	🟩 B	🟩 H	🟩 H	🟩 B	🟩 C	🟩 B	🟩 C
4	5.0	🟩 D	🟩 B	🟩 B	🟩 H	🟩 C	🟩 D	🟩 D	🟩 E	🟩 C	🟩 B
5	5.0	🟩 E	🟩 A	🟩 E	🟩 A	🟩 A	🟩 E	🟩 A	🟩 B	🟥 I	🟩 A
6	4.5	🟩 F	🟩 E	🟩 F	🟩 E	🟩 E	🟩 A	🟩 F	🟩 A	🟩 A	🟩 E
7	4.5	🟨 G	🟩 F	🟩 A	🟩 F	🟩 F	🟩 F	🟩 E	🟩 F	🟩 F	🟩 F
8	4.0	🟩 H	🟨 G	🟨 G	🟨 L	🟨 G	🟨 G	🟨 G	🟨 G	🟥 K	🟥 K
9	4.0	🟥 I	🟨 L	🟨 L	🟨 G	🟨 L	🟨 L	🟨 L	🟨 L	🟨 J	🟥 I
10	3.0	🟨 J	🟨 J	🟥 I	🟨 J	🟥 I	🟥 K	🟨 J	🟥 I	🟩 E	🟥 M
11	3.0	🟥 K	🟥 K	🟥 K	🟥 K	🟥 K	🟥 I	🟥 I	🟥 K	🟨 G	🟨 G
12	2.0	🟨 L	🟥 I	🟨 J	🟥 I	🟨 J	🟨 J	🟥 M	🟨 J	🟥 M	🟨 L
13	1.0	🟥 M	🟥 M	🟥 M	🟥 M	🟥 M	🟥 M	🟥 K	🟥 M	🟨 L	🟨 J

Participants' Ranking **Media Researchers' Rankings**

🟩 Quality Media Sources 🟨 Tabloid Media Sources 🟥 Fake News Sources

shown in Table 1. Unlike the vocational school students, the experts did not rate individual news stories but ranked the names of the news sources by their trustworthiness. With one exception, researchers made the same distinction between quality media and biased media (fake news and tabloid media). Like our participants, the experts did not distinguish tabloid media from fake news blogs. Overall, the comparison of the two rankings shows that the trust ratings of the participants correspond to expert opinion. This validates the results through a sample different in expertise, age, and gender. The experts have a background in media research and two-thirds of the experts were female (which counterbalanced the male bias in the participants).

4.1 Trust Ratings for Algorithmic News Curation (RQ1)

The first research question was whether users can provide trust ratings for recommendations of an algorithmic news curation system. We addressed this question with a between-subjects design where the samples are independent, i.e. different participants saw different news stories and news recommendations [31]. Participants provided their trust ratings for the news stories and the news recommendations on an 11-point rating scale. We analyzed this ordinal data using a non-parametric test, i.e. we made no assumptions about the distance between the different categories. To compare the different conditions and to see whether the trust ratings of the news recommendations differ in statistically significant ways, we applied the Mann-Whitney U test (Wilcoxon Rank test) [31,33]. Like the t-test used for continuous variables, the Mann-Whitney U test provides a

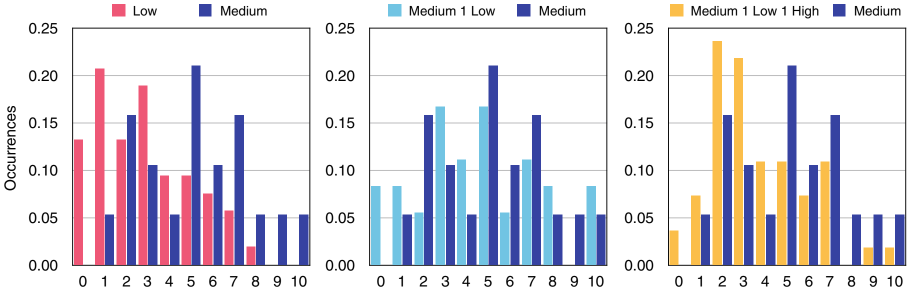


Fig. 2. Histograms comparing the trust ratings of the different recommendations of the ML system.

p-value that indicates whether statistical differences between ordinal variables exist. Participants were told to rate ML recommendations. The framing of the experiment explicitly mentioned that they are not rating an ML system, but one recommendation of an ML system that consisted of five news stories. The results show that participants can differentiate between such ML recommendations. The ranking of the ML recommendations corresponds to the news items that make up the recommendations. Of the four types of news recommendations, a) Medium recommendations, which consist of five news stories with a trust rating between 4 and 6, have a median rating of 5.0. d) Low recommendations with five news stories with a low rating (0 and 3), have a median trust rating of 3.0. The trust ratings of b) Medium, 1 Low recommendations, which combine four trustworthy stories and one untrustworthy, are rated considerably higher (4.5). ML recommendations that consist of three trustworthy news items, one untrustworthy news items (rating between 0 and 3) and one highly trustworthy news story (7 and 10), received a median trust rating of 3.0.

4.2 Trustworthy News Recommendations (RQ2)

Table 2. The Mann-Whitney U test was applied to see whether statistically significant differences between the trust ratings of different news recommendations exist (italic for $p < 0.05$, bold for $p < 0.01$).

Comparison of news recommendations		U	p
Medium	Low	258.50	**.0008**
Medium	M., 1 Low	303.50	.2491
Medium	M., 1 Low, 1 High	358.50	*.0204*
M., 1 Low	Low	619.50	*.0024*
M., 1 Low	M., 1 Low, 1 High	801.50	.0618
M., 1 L., 1 High	Low	1141.50	*.0250*

The second research question was whether users can distinguish trustworthy from untrustworthy machine learning recommendations. To answer this, we compare the trust ratings of a) Medium and d) Low recommendations. The trustworthy a) Medium recommendations have the same median rating (5.0) as the quality media sources D and E. Untrustworthy d) Low recommendations with a median rating of 3.0 have the same rating as the tabloid news source J and the fake news source K. The Mann-Whitney U test shows that participants reliably distinguish between a) Medium and d) Low recommendations (U = 258.5, p = .001). Figure 2 (left) shows the histogram of the a) Medium recommendations, which resembles a normal distribution. 5 is the most frequent trust rating, followed by 8 and 2. The histogram of d) Low is skewed towards negative ratings. Here, 1 and 3 are the most frequent trust rating. Nevertheless, a large number of participants still gave a rating of 6 or higher for d) Low recommendations. A large fraction also gave a) Medium recommendations a rating lower than 5.

4.3 Fake News Stories (RQ3)

The first two research questions showed that technically advanced participants are able to differentiate between trustworthy and untrustworthy ML recommendations in an experiment where they are primed to pay attention to individual fake news stories. The most important research question, however, was whether users distinguish trustworthy ML recommendations from recommendations that include one fake news story in their ratings. For this, we compare the trust ratings of a) Medium recommendations to those of b) 4 Medium, 1 Low recommendations, which have a median trust rating of 4.5 (N = 36). Compared to a) Medium at 5.0 (N = 19), the median is slightly lower. Compared to the news sources, b) 4 Medium, 1 Low at 4.5 is similar to quality media (Source F) and tabloid media (Source G). The Mann-Whitney U test shows that the ratings for b) Medium, 1 Low recommendations are significantly different from d) Low recommendations (U = 619.5, p = .002). However, the difference between a) Medium and b) 4 Medium, 1 Low is not statistically significant (U = 303.5, p = .249). This means that the crucial fake news case, where a recommendation consists of four trustworthy news stories and one fake news story, is not distinguished in a statistically significant way. The histogram in Fig. 2 (center) shows that a) Medium and b) Medium, 1 Low are very similar. Both resemble a normal distribution and both have strong peaks at 5, the neutral position of the 11-point rating scale. a) Medium recommendations have strong peaks at 2 and 7, b) Medium, 1 Low recommendations have peaks at 3 and 7. To see whether participants are able to distinguish the fake news case from other recommendations, we also compare b) 4 Medium, 1 Low recommendations to c) Medium, 1 Low, 1 High recommendations, which consist of three trustworthy news stories (rated between 4 and 6), one highly trustworthy story (7 and 10) and one untrustworthy news item (0 and 3). The c) 3 Medium, 1 Low, 1 High recommendations are rated as 3.0 (N = 55). This is the same as d) Low recommendations (3.0). It is also much lower than the ratings of b) Medium, 1 Low recommendations (4.5). In comparison to the median trust rating of the news sources, this places c) 3 Medium, 1 Low, 1 High between the tabloid source J and the fake news source K. According to the Mann-Whitney U test, participants are able to distinguish c) 3 Medium, 1 Low, 1 High recommendations from a) Medium (U = 358.5, p = .020) and d) Low (U = 1141.5, p = .025) recommendations. c) 3 Medium, 1 Low, 1 High recommendations are not distinguished from the fake news case of c) Medium, 1 Low recommendations (U = 801.50, p = .062). Figure 2 (right) compares the histograms of a) Medium and c) 3 Medium, 1 Low, 1 High recommendations. The largest peaks for c) recommendations are at 2 and 3, with very few high ratings of 7, 8, 9 or 10, but also few ratings of 0 and 1. The difference between the ratings of the two recommendations is clearly recognizable in the histograms.

5 Discussion

The study found that participants with a technical background can provide plausible trust ratings for individual news items as well as for groups of news items presented as the recommendations of an ML system. The ratings of the news recommendations correspond to the news stories that are part of the news recommendations. We further showed that the trust ratings for individual news items correspond to expert opinion. Vocational school students and media researchers both distinguish news stories of quality media sources from biased sources. Neither experts nor participants placed the fake news sources at the end of the rankings. These findings are highly problematic considering the nature of fake news. Following Lazer et al.'s definition of fake news as fabricated information that mimics news media content in form but not in organizational process or intent [28], fake news are more likely to emulate tabloid media in form and content than quality media.

We found that users can provide trust ratings for an algorithmic news curation system when presented with recommendations of a machine learning system. Participants were able to assign trust ratings that differentiated between news recommendations in a statistically significant way, at least when comparing trustworthy from untrustworthy machine learning recommendations. However, the crucial fake news case was not distinguished from trustworthy recommendations. This is noteworthy since the first phase of our study showed that users are able to identify individual fake news stories. When providing trust ratings for groups of news items in the second phase, the presence of fake news did not affect the trust ratings of the output as a whole. This is surprising since prior research on trust in automation reliance implies that user's assessment of a system changes when the system makes mistakes [15]. Dzindolet et al. report that the consequences of this were so severe that after encountering a system that makes mistakes, participants distrusted even reliable aids. In our study, one fake news story did not affect the trust rating in such a drastic way. An untrustworthy fake news story did not lead to a very low trust rating for the news recommendation as a whole. The simplest explanation for this would be that the task is too hard for users. Identifying a lowly trusted news story in the recommendations of an algorithmic news curation system may overstrain users. A contrary indication against this explanation is that trustworthy and untrustworthy recommendations can be distinguished from other news recommendations like the c) Medium, 1 Low, 1 High recommendations.

Our findings could, therefore, be a first indication that untrustworthy news stories benefit from appearing in a trustworthy context. Our findings are especially surprising considering that the users have an IT background and were primed to be suspicious. If users implicitly trust fake news that appear in a trustworthy context, this would have far-reaching consequences. Especially since social media is becoming the primary news sources for a large group of people [40]. The question whether untrustworthy news stories like fake news benefit from a trustworthy context is directly connected to research on algorithmic experience and the user awareness of algorithmic curation.

Our understanding of the user experience of machine learning systems is only emerging [17,18,21,46]. In the context of an online hotel rating platforms, Eslami et al. found that users can detect algorithmic bias during their regular usage of a service and that this bias affects trust in the platform [18]. The question, therefore, is why participants did not react to the fake news stories in our study in a similar way. Further research has to show what role the context of our study - machine learning and algorithmic news curation - may have played. While framing effects are known to affect trust, our expectation was that the framing would have primed users to be overly cautious [32]. This would mean that participants can distinguish them in the experiment, but not in the practice. This was not the case.

In the instructions of the controlled experiment, we define the terms fake news and machine learning. This increased algorithmic awareness and the expectation of algorithmic bias. It could also have influenced the perception and actions of the participants by making them more cautious and distrusting. We show that despite this priming and framing, participants were not able to provide ratings that reflect the presence of fake news stories in the output. If people with a technical background and a task framed like this are unable to do this, how could a layperson? Especially considering that participants were able to distinguish uniformly trustworthy from uniformly untrustworthy output. All this makes the implications of our experiment on the UX of machine learning and how feedback/training data needs to be collected especially surprising and urgent. This adds to a large body of research on algorithmic experience and algorithmic awareness [10,16,57].

6 Limitations

Studying trust in machine learning systems for news curation is challenging. We had to simplify a complex socio-technical system. Our approach connects to a large body of research that applies trust ratings to study complex phenomena [32, 38, 39, 43]. Since no ground truth data on the trustworthiness of different news stories was available, we designed a study that used the median trust ratings of our participants as intersubjective agreement on the perceived trustworthiness of a news story. A real-world algorithmic news curation system is more complex and judges the relevance of postings based on three factors: who posted it, the type of content, and the interactions with the post [19]. Even though we recreated the design of Facebook's News Feed, our setting was artificial. Interactions with the posts were limited, participants did not select the news sources themselves and they did not see the likes, shares, and comments of their real Facebook "friends". We focused on news stories and did not personalize the recommendations of the ML system. Further research could investigate how the different sources affect the trust perception of news stories respectively the trust perception of ML recommendations. However, not personalizing the results and focusing on news was necessary to get comparable results.

We conducted the experiment in a German vocational school with an IT focus. This limits biasing factors like age, educational background, and socio-economic background, but led to a strong male bias. We counteracted this bias by validating the trust ratings of news stories with nine media research experts - a heterogeneous group that is different in age, gender (three male, six female), and background, which confirmed our results. Prior research also implies that the findings from our sample of participants are generalizable despite the strong male bias. A German study (N = 1,011) from 2017 showed that age and gender have little influence on experience with fake news, which is similar for all people under 60, especially between 14-to-24-year olds and 25-to-44-year olds [27]. The participants in this study had a background in IT, which could have influenced the results. Prior work on algorithmically generated image captions showed that technical proficiency and education level do not influence trust ratings [32]. Moreover, even if the technical background of the participants would have helped the task, they were not able to provide nuanced ratings that accounted for untrustworthy news items, which further supports our arguments.

7 Conclusion

Our study investigated how fake news affect trust in the output of a machine learning system for news curation. Our results show that participants distinguish trustworthy from untrustworthy ML recommendations in significantly different trust ratings. Meanwhile, the crucial fake news case, where an individual fake news story appears among trustworthy news stories, is not distinguished from trustworthy ML recommendations. Since ML systems make a variety of errors that can be subtle, it is important to incorporate user feedback on the performance of the system. Our study shows that gathering such feedback is challenging. While participants are able to distinguish exclusively trustworthy from untrustworthy recommendations, they do not account for subtle but crucial differences like fake news. Our recommendations for those who want to apply machine learning is, therefore, to evaluate how well users can give feedback before training active learning and human-in-the-loop machine learning systems. Further work in other real-world scenarios is needed, especially since news recommendation systems are constantly changing.

References

1. Fachinformatiker: IT-Berufsausbildung auf dem Arbeitsmarkt sehr gefragt - Golem.de (2017). https://www.golem.de/news/fachinformatiker-it-berufsausbildung-auf-dem-arbeitsmarkt-sehr-gefragt-1702-126214.html
2. Many Facebook users don't understand its news feed (2019). http://www.pewresearch.org/fact-tank/2018/09/05/many-facebook-users-dont-understand-how-the-sites-news-feed-works/
3. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Tech. rep., National Bureau of Economic Research (2017)
4. Allport, F.H., Lepkin, M.: Wartime rumors of waste and special privilege: why some people believe them. J. Abnorm. Soc. Psychol. **40**(1), 3 (1945)
5. Allport, G.W., Postman, L.: The psychology of rumor (1947)
6. Alvarado, O., Waern, A.: Towards algorithmic experience: initial efforts for social media contexts. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, pp. 286:1–286:12. ACM, New York (2018). https://doi.org/10.1145/3173574.3173860, http://doi.acm.org/10.1145/3173574.3173860
7. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. AI Mag. **35**(4), 105–120 (2014)
8. Bento, Katharina Hölter, S.L.: Fake news in Deutschland: Diese Webseiten machen Stimmung gegen Merkel (2017). http://www.bento.de/today/fake-news-in-deutschland-diese-seiten-machen-stimmung-gegen-merkel-1126168/
9. Berkovsky, S., Taib, R., Conway, D.: How to recommend?: User trust factors in movie recommender systems. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, pp. 287–300. ACM, New York (2017). https://doi.org/10.1145/3025171.3025209, http://doi.acm.org/10.1145/3025171.3025209
10. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, pp. 377:1–377:14. ACM, New York (2018). https://doi.org/10.1145/3173574.3173951, http://doi.acm.org/10.1145/3173574.3173951
11. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing?: How recommender system interfaces affect users' opinions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2003, pp. 585–592. ACM, New York (2003). https://doi.org/10.1145/642611.642713, http://doi.acm.org/10.1145/642611.642713
12. Cramer, H.S., Evers, V., van Someren, M.W., Wielinga, B.J.: Awareness, training and trust in interaction with adaptive spam filters. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009, pp. 909–912. ACM, New York (2009). https://doi.org/10.1145/1518701.1518839, http://doi.acm.org/10.1145/1518701.1518839
13. Deutsch, M.: Trust, trustworthiness, and the F scale. J. Abnorm. Soc. Psychol. **61**(1), 138 (1960)

14. Dictionaries, O.: Trust (2018). https://en.oxforddictionaries.com/definition/trust
15. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. Hum.-Comput. Stud. **58**(6), 697–718 (2003)
16. Eslami, M., Krishna Kumaran, S.R., Sandvig, C., Karahalios, K.: Communicating algorithmic process in online behavioral advertising. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, pp. 432:1–432:13. ACM, New York (2018). https://doi.org/10.1145/3173574.3174006, http://doi.acm.org/10.1145/3173574.3174006
17. Eslami, M., et al.: "I always assumed that i wasn't really that close to [her]": reasoning about invisible algorithms in news feeds. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, pp. 153–162. ACM, New York (2015). https://doi.org/10.1145/2702123.2702556, http://doi.acm.org/10.1145/2702123.2702556
18. Eslami, M., Vaccaro, K., Karahalios, K., Hamilton, K.: "Be careful; things can be worse than they appear": understanding biased algorithms and users' behavior around them in rating platforms. In: ICWSM, pp. 62–71 (2017)
19. Facebook: Facebook news feed (2018). https://newsfeed.fb.com/
20. Gulla, J.A., Zhang, L., Liu, P., Özgöbek, O., Su, X.: The Adressa dataset for news recommendation. In: Proceedings of the International Conference on Web Intelligence, WI 2017, pp. 1042–1048. ACM, New York (2017). https://doi.org/10.1145/3106426.3109436, http://doi.acm.org/10.1145/3106426.3109436
21. Hamilton, K., Karahalios, K., Sandvig, C., Eslami, M.: A path to understanding the effects of algorithm awareness. In: CHI 2014 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2014, pp. 631–642. ACM, New York (2014). https://doi.org/10.1145/2559206.2578883, http://doi.acm.org/10.1145/2559206.2578883
22. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
23. Horvitz, E.J.: Reflections on challenges and promises of mixed-initiative interaction. AI Mag. **28**(2), 3 (2007)
24. Lovric, M. (ed.): Robust Statistics. International Encyclopedia of Statistical Science, pp. 1248–1251. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-04898-2
25. Kim, B.: Interactive and interpretable machine learning models for human machine collaboration. Ph.D. thesis, Massachusetts Institute of Technology (2015)
26. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, pp. 126–137. ACM, New York (2015). https://doi.org/10.1145/2678025.2701399, http://doi.acm.org/10.1145/2678025.2701399
27. Landesanstalt für Medien NRW (LfM): Fake news. Tech. rep., forsa (May 2017). https://bit.ly/2ya2gj0
28. Lazer, D.M., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
29. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors: J. Hum. Factors Ergon. Soc. **46**(1), 50–80 (2004)
30. Luhmann, N.: Trust and Power. Wiley, Hoboken (1979)
31. MacKenzie, I.S.: Human-computer interaction: an empirical research perspective. Morgan Kaufmann, Amsterdam (2013). http://www.sciencedirect.com/science/book/9780124058651

32. MacLeod, H., Bennett, C.L., Morris, M.R., Cutrell, E.: Understanding blind people's experiences with computer-generated captions of social media images. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI 2017, pp. 5988–5999. ACM, New York (2017). https://doi.org/10.1145/3025453.3025814, http://doi.acm.org/10.1145/3025453.3025814
33. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**, 50–60 (1947)
34. Marsh, S.P.: Formalising trust as a computational concept. Ph.D. thesis (1994)
35. Massa, P., Bhattacharjee, B.: Using trust in recommender systems: an experimental analysis. In: Jensen, C., Poslad, S., Dimitrakos, T. (eds.) iTrust 2004. LNCS, vol. 2995, pp. 221–235. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24747-0_17
36. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995). https://doi.org/10.2307/258792. http://www.jstor.org/stable/258792
37. Mitchell, T.M.: Machine Learning, 1st edn. McGraw-Hill Inc., New York (1997)
38. Muir, B.M., Moray, N.: Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**(3), 429–460 (1996). https://doi.org/10.1080/00140139608964474
39. Muir, B.M.: Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics **37**(11), 1905–1922 (1994). https://doi.org/10.1080/00140139408964957, http://dx.doi.org/10.1080/00140139408964957
40. Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D.A., Nielsen, R.K.: Reuters institute digital news report 2017 (2017). https://ssrn.com/abstract=3026082
41. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 167–174. ACM (2005)
42. OECD: PISA 2006 (2007). https://www.oecd-ilibrary.org/content/publication/9789264040014-en
43. Pennycook, G., Rand, D.G.: Crowdsourcing judgments of news source quality (2018)
44. Pu, P., Chen, L.: Trust building with explanation interfaces. In: Proceedings of the 11th International Conference on Intelligent User Interfaces, IUI 2006, pp. 93–100. ACM, New York (2006). https://doi.org/10.1145/1111449.1111475, http://doi.acm.org/10.1145/1111449.1111475
45. Rader, E., Cotter, K., Cho, J.: Explanations as mechanisms for supporting algorithmic transparency. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, pp. 103:1–103:13. ACM, New York (2018). https://doi.org/10.1145/3173574.3173677, http://doi.acm.org/10.1145/3173574.3173677
46. Rader, E., Gray, R.: 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, CHI 2015, pp. 173–182. ACM, New York (2015). https://doi.org/10.1145/2702123.2702174, http://doi.acm.org/10.1145/2702123.2702174
47. Reeskens, T., Hooghe, M.: Cross-cultural measurement equivalence of generalized trust. Evidence from the European Social Survey (2002 and 2004). Soc. Indic. Res. **85**(3), 515–532 (2008). https://doi.org/10.1007/s11205-007-9100-z
48. Rempel, J.K., Holmes, J.G., Zanna, M.P.: Trust in close relationships. J. Pers. Soc. Psychol. **49**(1), 95 (1985)

49. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW 1994, pp. 175–186. ACM, New York (1994). https://doi.org/10.1145/192844.192905, http://doi.acm.org/10.1145/192844.192905
50. Rotter, J.B.: A new scale for the measurement of interpersonal trust. J. Pers. **35**(4), 651–665 (1967). https://doi.org/10.1111/j.1467-6494.1967.tb01454.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6494.1967.tb01454.x/abstract
51. Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C.: Not so different after all: a cross-discipline view of trust. Acad. Manag. Rev. **23**(3), 393–404 (1998). https://doi.org/10.5465/AMR.1998.926617. http://amr.aom.org/content/23/3/393
52. Rubens, N., Elahi, M., Sugiyama, M., Kaplan, D.: Active learning in recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 809–846. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_24
53. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
54. Schou, J., Farkas, J.: Algorithms, interfaces, and the circulation of information: interrogating the epistemological challenges of Facebook. KOME: Int. J. Pure Commun. Inq. **4**(1), 36–49 (2016)
55. Stumpf, S., et al.: Interacting meaningfully with machine learning systems: three experiments. Int. J. Hum.-Comput. Stud. **67**(8), 639–662 (2009). https://doi.org/10.1016/j.ijhcs.2009.03.004. http://www.sciencedirect.com/science/article/pii/S1071581909000457
56. Tullio, J., Dey, A.K., Chalecki, J., Fogarty, J.: How it works: a field study of non-technical users interacting with an intelligent system. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 31–40. ACM (2007)
57. Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warshaw, J.: A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, pp. 656:1–656:14. ACM, New York (2018). https://doi.org/10.1145/3173574.3174230, http://doi.acm.org/10.1145/3173574.3174230
58. Özgöbek, O., Shabib, N., Gulla, J.: Data sets and news recommendation. In: Workshops Proceedings of the 24th ACM Conference on User Modeling, Adaptation, and Personalization, vol. 1181, pp. 5–12 (January 2014)