Explanations as Mechanisms for Supporting Algorithmic Transparency

Emilee Rader
Media and Information Dept.
Michigan State University
emilee@msu.edu

Kelley Cotter
Media and Information Dept.
Michigan State University
cotterk6@msu.edu

Janghee Cho
Media and Information Dept.
Michigan State University
chojangh@msu.edu

ABSTRACT

Transparency can empower users to make informed choices about how they use an algorithmic decision-making system and judge its potential consequences. However, transparency is often conceptualized by the outcomes it is intended to bring about, not the specifics of mechanisms to achieve those outcomes. We conducted an online experiment focusing on how different ways of explaining Facebook's News Feed algorithm might affect participants' beliefs and and judgments about the News Feed. We found that all explanations caused participants to become more aware of how the system works, and helped them to determine whether the system is biased and if they can control what they see. The explanations were less effective for helping participants evaluate the correctness of the system's output, and form opinions about how sensible and consistent its behavior is. We present implications for the design of transparency mechanisms in algorithmic decision-making systems based on these results.

Author Keywords

algorithmic decision-making; transparency; explanations

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

An algorithm is "a finite, discrete series of instructions that receive an input and produce an output" [26]. Algorithms now contribute to decisions affecting millions of people, related to employment, housing, healthcare, education, and criminal justice, among many others [2, 44]. Negative consequences can result from decisions that depend upon algorithms, such as economic or social disadvantaging of already marginalized populations [2, 8, 30, 44]. In a social media context, algorithmic curation—automated selection and ranking of content—acts as a gatekeeper, defining what is relevant, knowable, and authoritative [27, 6, 23]. Personalization via algorithmic curation in social media can lead to a lack of information diversity or echo chambers of ideas in which users are closed off from opposing points of view [6].

As important responsibilities and processes are increasingly delegated to algorithmic decision-making systems [13, 56], more attention is being paid to algorithmic transparency [12, 43]. Researchers and policy-makers argue that transparency is valuable, and advocate greater transparency as a remedy for identifying and preventing various potential negative effects of these systems [12]. Transparency involves encountering non-obvious information that is difficult for an individual to learn or experience directly, about how and why a system works the way it does and what this means for the system's outputs. Transparency mechanisms provide opportunities for users to gain familiarity with aspects of a system that are usually hidden [19], and can change people's beliefs about a system and their interactions with it. Greater transparency allows people to question and critique a system in order to develop appropriate reliance, rather than blind faith [2, 56].

Several different types of mechanisms have been identified that contribute to greater transparency in algorithmic decision-making. Users may become aware of an algorithm through repeated experiences with a system [46]. In some cases, users encounter unexpected or confusing information that violates expectations [11, 45] and hints at algorithmic bias [17]. In others, users are motivated to become more knowledgeable about the algorithmic outputs so that they can create workarounds in an effort to avoid negative outcomes [32]. However, such "organic" awareness is not systematic, or spread evenly through the user population.

Another type of mechanism for transparency is algorithm audits, which investigate both how an algorithmic decision-making system works, and what its impacts are [38]. Sandvig et al. [48] describe several different levels at which audits of algorithms could operate, each providing a different type of visibility and accountability. However, audits must generally be undertaken without the cooperation of the system providers, who often include prohibitions of audit techniques in their terms of service. Some have argued that platforms are intentionally opaque regarding details about their operation as a form of self-protection from competitors or others who attempt to "game" the system [7].

A third type of mechanism to promote greater transparency is providing explanations, a common approach in recommender systems [51] that may help solve problems caused by lack

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04 ...$15.00.
http://dx.doi.org/10.1145/3173574.3173677

of transparency in algorithmic decision-making systems [32]. This paper presents the results of an experiment investigating how different types of explanations discussed in the literature and used "in the wild" might affect people's beliefs and judgments about Facebook's News Feed, an algorithmic decision-making system currently in widespread use. We introduce new explanation types, designed based on information in official blog posts about the News Feed, and define a new measurement framework for transparency outcomes, focused on the different functions that transparency is thought to serve (awareness, correctness, interpretability, and accountability). This paper contributes novel results to the research literature about transparency and algorithmic decision-making systems by showing through a controlled experiment that even brief explanations can affect user beliefs about hidden aspects of a system they have been using for many years.

RELATED WORK

Explanations and Transparency

Algorithms that make autonomous decisions and provide recommendations are a "mission critical" aspect of many online content and e-commerce platforms, including Facebook, Google, Netflix, and Amazon [28, 29, 39]. Algorithmic decision-making systems like Facebook, and recommender systems like Netflix, fundamentally use the same kinds of technologies and perform very similar functions: both involve matching users with items. Recommender system output is presented to users as a set of options to choose from, which provides evidence of the existence of the matching process. However, in an algorithmic decision-making system, users are not explicitly informed that the information they see is a subset of what is available. Unlike recommender systems, algorithmic decision-making systems typically do not provide visibility into what the technologies are doing [39].

Many recommender systems also provide explanations, or short persuasive texts, along with the recommendations [3]. Explanations present information about how the system produced the recommendation and the reasons behind it [3, 24, 52], and are in service to the system's purpose and goals [20]. They help the user understand and act upon the recommendation [47]. Previous research has focused on different aspects of explanations, including data sources [42, 5], cognitive fit [22], modality (e.g. text vs. visual) [40, 21], "completeness" and "soundness" of information provided [31], and content type [24, 37]. This last aspect, content, represents the most fundamental consideration in explaining a system. Explanations improve system usability and overall performance, and promote more positive user perceptions and acceptance of the system [10, 24, 21, 37]. They also help users to know what the limitations of a system are, and when they can rely on it [37].

Types of Explanations

Friedrich and Zanker [20] classified explanations into two types, "white box" and "black box". *How* explanations are "white box" descriptions of a system's inputs and outputs and the steps it takes to arrive at a particular outcome. They provide information about how a system produces a recommendation, particularly focusing on the system's reasoning and data source [50, 51]. These explanations disclose important details about the functioning of the system to the user, and fill a knowledge gap between a user's experiences with and intuitions about a system and the system's actual internal processes [54]. Explanations that help users understand how a system works have demonstrated a positive relationship with user satisfaction with the explanation [21]. *How* explanations can also increase beliefs in the competence and benevolence of a system [54] and the perceived usefulness of a system [55].

Why explanations treat systems as "black boxes", providing justifications for a system and its outcomes and explaining the motivations behind the system, but not disclosing how the system works. These explanations fill an intention gap between a user's needs and interests and the system's goals [54], but do not provide any visibility into how the system works. In other words, *Why* explanations allow users to determine whether their goals match those of the system. When users believe they understand why a system makes a recommendation, they feel more comfortable and satisfied with recommendations [49], and are more willing to accept a recommendation [10].

We introduce two additional types of explanations in this study: *What* and *Objective* explanations. Many Facebook users do not realize that the content in their News Feeds is curated by an algorithm [16], so they cannot react to stories in the News Feed like they would a recommendation. Thus, *What* explanations reveal the existence of algorithmic decision-making, without providing additional information about the system. We use a *What* explanation in our experiment to measure the effect of becoming aware of the algorithm separately from the types of information provided by other explanations.

Though technology companies typically use an iterative process of testing and development to produce and maintain systems, explanations in recommender systems usually do not explicitly include information about this. However, a content analysis of Facebook's 'News Feed FYI' blog[1]—the primary venue through which Facebook explains its News Feed algorithm to users—revealed a strong emphasis on the impartial, data-driven design and testing of the algorithm as evidence for the argument that the algorithm serves the interests of users [9]. This kind of information about how the system is developed might be helpful for supporting judgments about it. As such, our study introduces *Objective* explanations, which describe the process by which a system comes into being and is continually improved. "Objective" is used here in the sense of the adjective meaning "unbiased", not the noun meaning "goal". Neither *What* nor *Objective* information is typically included in traditional *How* or *Why* explanations.

Functions of Transparency Mechanisms

Algorithmic transparency often refers to the act of making a system knowable or visible [18, 34, 1, 13]. This conceptualizes transparency as a mechanism or process that brings about changes in user behavior or system governance. However, transparency is also sometimes treated as a state that is the outcome of a process. For example, the Association for Computing Machinery recently included explanations in its list of 'Principles for Algorithmic Transparency and Accountability'

[1] https://newsroom.fb.com/news/category/news-feed-fyi/

as a mechanism for making systems more transparent [2]. It can be difficult to disambiguate in the literature whether "transparency" refers to the mechanism or the outcome, the cause or the effect. In this paper we treat transparency as the mechanism, and define the effects brought about by transparency in terms of different types of functions that transparency mechanisms are thought to perform. This allows us to begin to identify what types of information and arguments explanations might provide that could bring about changes in specific knowledge and beliefs about systems that involve algorithmic decision-making.

Transparency performs the basic function of providing visibility that there is an algorithm that is making decisions, thereby creating *awareness* that interactions with the system are mediated by an algorithm [1]. Telling people what the system is doing makes the aspects of its behavior that may not be visible or detectable able to be perceived and known [38]. An explanation alerting users to the actions of an algorithm can be especially powerful for informing users who may be unaware.

Transparency mechanisms also function to help users to learn about how the system works, so they can evaluate the *correctness* of the outputs they experience and identify outputs that are incorrect. Correctness judgments are a function of transparency in that a mechanism that can support an understanding of how the inputs produce the outputs [31] is necessary before an individual can evaluate for themselves "whether a system is working as intended and what changes are required" [1]. An explanation about how the system works should help users understand what outputs the system is supposed to produce, and recognize when it makes mistakes or errors.

In addition to judgments about correctness, transparency can also function as support for judgments about how sensible the outputs are, and convey that the system's behavior is not arbitrary or random [56]. Understanding that there are reasons why a system behaves the way it does, and evaluating whether the system is acting consistently with those reasons, makes the system's behavior *interpretable* and helps users feel comfortable acting on the outputs [35]. An explanation supporting interpretability would help users better understand the system's behavior based on seeing the "truth and motives" or reasons behind the system's actions [14], and to identify when the system is not acting in support of those motives.

Much of the literature on transparency also emphasizes the goal of governing a system through *accountability* [1]. Transparency mechanisms can convey a sense of iterative control, or individual users feeling like they are in some way responsible for the outputs of the algorithm. In order for a system to be directly accountable to users, an explanation would need to provide information that helps them believe and understand that they can directly affect the outputs of the system. Ideally, transparency mechanisms also enable users to identify biases that may result in negative consequences [13], and empower users to question and critique the system, providing grounds for demanding remediation [1]. However, many of the outcomes of transparency as an accountability intervention are beyond the scope of an individual user's ability to influence the system or the corporations operating the platforms.

The goal of this experiment was to identify the effects of four different types of explanations (What, How, Why and Objective) on user beliefs about an algorithmic decision-making system, measured in terms of the functions that transparency mechanisms perform (awareness, correctness, interpretability, accountability). Each function of transparency reflects a qualitatively different understanding of the system, and some functions may be more beneficial for mitigating potential negative effects of algorithmic decision-making than others. This experiment is an important first step toward identifying how different types of information about a system might bring about changes in specific transparency-related user beliefs.

METHOD

Participants

Data collection took place online from August 10–24, 2017. Participants were recruited from a panel provided by Qualtrics. Eligible participants lived in the United States, were 18 or older, had been using Facebook for at least two months, had more than 50 Facebook friends, and reported visiting Facebook at least once per month (88.55% visited daily). We used quotas for age (35% over age 55) and gender (52% women) to ensure greater diversity in our sample on these two dimensions. We excluded participants who identified themselves as social media experts (managing a Page on Facebook; job responsibilities including posting content on social media, communicating with clients or customers via social media, or working on an organization's social media strategy) or computing experts (job responsibilities including computer programming, quality assurance and testing, IT security, or network administration). We believed that these areas of expertise would be related to greater knowledge of the Facebook News Feed ranking algorithm than would likely be found among the general population, and we wanted to focus on non-experts.

6842 potential participants started the survey by viewing the consent form. 285 declined consent and 5056 were determined to be ineligible. To ensure data quality, we excluded participants who answered one of four attention-check questions incorrectly. 820 participants were excluded before completing the survey for reasons such as failing an attention or manipulation check, taking too long to complete the survey, or submitting poor quality answers to an open-ended question included in the survey for data quality purposes (see the supplementary file for more information). After data cleaning, there were 681 participants in the final dataset for analysis. Participants ranged in age from 18 to 88, with a mean of 43 (*SD*=16). Fifty-two percent of participants were women. A large majority of participants reported "white" as one of the ethnicity categories that described them (84.29%). Seventy-five percent said they had been using Facebook for more than 5 years, and 62% had posted at least one story in the past week. The average number of Facebook friends per participant in our sample was 339 (*SD*=456, *Max*=4958, *Median*=201). Further information about participant demographics is available in the supplementary file.

Procedure

Participants were randomly assigned to one of four explanation conditions or a control condition, making this a between-

Condition	*Word Count*	*Grade Level*	*N*
What	198	10.6	141
How	194	10.7	141
Why	202	10.7	134
Objective	207	10.6	139
Control	189	10.8	126

Table 1: Characteristics of each experiment condition. Grade level is the Flesch-Kincaid Reading Level, based on length and complexity of sentences and number of syllables per word.

subjects experiment. See Table 1 for the number of participants in each condition. We allowed participants no more than 60 minutes to complete the experiment, starting after they had completed the consent and screening questions. With this time limit, we ensured that all participants answered the questions within a similar timeframe after exposure to the experiment manipulation. The study took an average of 21.57 minutes for participants to complete, and the minimum completion time was 6.08 minutes.

Potential participants received a study invitation via an email message, and clicked on a link that directed them to the Qualtrics platform which hosted the experiment. Participants first saw the online consent form which described the study, including its expected duration to complete and the time limit for completion. Participants who consented to the study were directed to a series of screening questions to determine their eligibility to participate. Participants who were eligible then answered questions about themselves and their Facebook use.

Next, participants read a short text about Facebook that was different in each experiment condition, and immediately afterward answered three manipulation check questions. Participants were given two chances to answer three factual questions about the text they had read. One hundred eighty participants did not answer all three questions correctly after two attempts, and were excluded from the experiment. And, 88 participants failed the manipulation check the first time, but answered it correctly the second time and were allowed to proceed with the rest of the experiment. After the manipulation check, participants were asked about how new and surprising the information in the text they read was to them.

The next four pages of the survey consisted of questions designed to measure the functions of transparency discussed in the related work section: awareness, correctness, interpretability, and accountability, in that order. The question order on each page was randomized. The last page of the survey consisted of three final demographics questions about income, ethnicity, and region of the US where the participant lived. At the end of the experiment, participants who had completed the survey received points from the online panel service worth approximately US$1-$2 that could be combined with the incentives from other surveys and redeemed for items like gift cards, frequent flyer miles, credit for online games, etc.

Explanation Conditions

We designed four explanations of the Facebook News Feed that each consisted of two short paragraphs of around 200 words. They were based on a content analysis of blog posts from Facebook's 'News Feed FYI' Blog through December 15, 2016 [9]. This provided greater external validity to the experiment; we did not speculate or guess about how the News Feed ranking algorithm works, so that we were not deceiving our participants. Therefore, the four explanation conditions in our experiment use information that is based on and resembles what Facebook the company is already willing to disclose about its platform to end users. However, our explanations were not personalized to the individual preferences and characteristics of our participants, as they often are in recommender systems [20]. We also designed a short paragraph to use as a control condition, containing general information about Facebook adapted from text on Wikipedia[2] and modified for length and to use more neutral language.

We did three rounds of piloting the explanations and the control condition text and revised them after each round. We did this to ensure that the explanations did not vary across conditions in terms of their tone, clarity, and credibility. Each explanation and the control condition also had three corresponding manipulation check questions, and in the pilots we also tested and revised these questions to ensure that they did not vary in difficulty across the conditions.

One challenge in designing the explanations was how to discuss where the agency lies for what stories users see in their News Feeds. It is difficult to differentiate Facebook the company and its employees from the News Feed feature or the ranking algorithm when writing short 200-word texts for a general audience about who or what is responsible for which stories a user sees when they visit the platform. Also, users are part of a feedback loop, as producers of both content and data that serve as inputs to the News Feed ranking algorithm, and also consumers of the output of the algorithm.

All four explanation conditions contained the information that there is an algorithm that guesses which stories people will want to see most, and decides the order the stories are presented in. The information unique to each condition is briefly described below, and the characteristics of each explanation as well as the number of participants per condition is presented in Table 1. The full text of the explanations is available in the supplementary file.

- *What*: Reveals that stories are not shown in chronological order; the News Feed is personalized by an algorithm that chooses which stories will be at the top, and people are more likely to see stories that are higher up.
- *How*: Informs participants that the ranking algorithm uses data collected about users and their behaviors to calculate a score for each story; the score is used to put stories in order, and the stories higher up are the ones the algorithm guesses users will like the most.
- *Why*: Describes information overload (too many stories to see them all) as the reason the ranking algorithm is necessary, and that Facebook's goal when deciding how stories should be ranked is to prioritize the interesting and relevant high quality stories that users want to see most.

[2] https://en.wikipedia.org/wiki/Facebook

- *Objective*: Presents information about how sometimes the algorithm doesn't rank stories appropriately, so Facebook evaluates the News Feed using behavioral data and feedback from users, and then updates the algorithm based on what they learn.
- *Control*: Includes facts about Facebook the company, its history, and what the News Feed is. It does not mention the algorithm, or the ranking of stories.

Measuring Transparency Mechanism Effects

We developed questions to measure participant knowledge and beliefs related to four functions that transparency mechanisms are believed to perform. Because this is a survey-based experiment, all of the questions are self-report, and participants' responses reflect their knowledge, beliefs and behavioral intentions, but not their actual behaviors. For each transparency function, we asked one to three standalone questions that used a 7-point Strongly disagree to Strongly agree Likert scale, and also a block of 10–15 related statements that participants were asked to rate either using a 7-point agreement Likert scale, or a 0–100 semantic differential scale. We performed an exploratory factor analysis on each block of related statements to group the items into factors that we then combined into composite variables for each transparency function. The full text of all of the questions as well as descriptive statistics for each question are available as supplementary file. Descriptive statistics and Cronbach's α for the composite variables are available in Table 2.

- *Awareness* questions: measure participants' basic awareness of the News Feed algorithm, and their understanding of where the agency lies behind what they see when they visit their News Feeds.
- *Correctness* questions: measure how well participants think the outputs they experience—the stories they see in their News Feeds—agree with what they expect the system to produce, and are not a mistake or incorrect.
- *Interpretability* questions: measure how sensible, and not arbitrary or random, participants think the performance of their News Feed is, given what they know about the goals and reasons behind what the News Feed does.
- *Accountability* questions: measure the extent to which participants think the system is fair and they can control the outputs the system produces.

RESULTS

To determine the causal effect of the four explanation types on outcome variables measuring the functions that transparency mechanisms perform, we conducted an OLS regression for each outcome variable. All models used the experiment condition and variables controlling for participant demographics and Facebook use as predictors. Note that we measured other controls, such as participants' number of Facebook friends (see the supplementary file for more details). We did not include these variables in the models because they did not have a meaningful relationship with the outcome variables, defined as a non-zero and statistically significant coefficient. All control variables were centered at their means for the regression analyses. Descriptive statistics for all of the non-categorical variables used in the regressions are available in Table 2.

Variable	*Type*	*Mean*	*SD*	*Range*	α
Knowledge After	Aw	3.43	1.70	1–7	–
System Agency	Aw	4.14	1.00	1–7	0.70
User Agency	Aw	4.28	1.30	1–7	0.74
Missed Stories	Co	4.43	1.71	1–7	–
Wanted Stories	Co	55.28	15.05	0–100	0.68
Unwanted Stories	Co	54.66	19.14	0–100	0.72
Understand Why	In	5.02	1.37	1–7	–
Interpersonal Goals	In	67.29	16.70	0–100	0.79
Informational Goals	In	56.14	17.22	0–100	0.75
Fairness	Ac	3.86	1.25	1–7	0.64
Content Actions	Ac	72.01	17.14	0–100	0.75
UI Controls	Ac	67.55	18.72	0–100	0.66
Knowledge Before	PK	4.38	1.61	1–7	–
New Info	PK	4.53	1.87	1–7	–
Surprising Info	PK	3.57	1.76	1–7	–

Table 2: Descriptive statistics for the variables used in the analyses. α = Cronbach's. Aw = awareness, Co = correctness, In = interpretability, Ac = accountability, PK = prior knowledge.

Many Users Believe They See Every Available Story

Before they were exposed to the explanations, we asked participants to agree or disagree with the statement, "Facebook shows me every story created by my Facebook friends and the Pages I've 'liked'." This question was based on Rader and Gray [46] and was used as a baseline control. Responses on the *Knowledge Before* variable were on a 7-point Likert agreement scale from Strongly disagree (1) to Strongly agree (7). Fifty-six percent answered with some level of agreement to that statement, while 18% were neutral, and only 26% disagreed. In other words, over half of our participants, who were non-experts both in computing and social media, were generally unaware that their News Feed does not show them every available story. This is important as further evidence that algorithmic-decision making is often invisible to users.

In order to better understand what participant demographics and Facebook use characteristics were associated with less initial awareness, we used a regression model to estimate how control variables were associated with *Knowledge Before*. The *Internet Literacy* control variable is based on the Web Use Skills survey reported in Hargittai and Hsieh [25]. It is the average of self-reported familiarity with a list of internet-related terms on a scale of "No understanding" (1) to "Full understanding" (5). The questions that comprise the *Trust Propensity* composite variable are based on Li et al. [33], modified to refer to "social media" instead of "information systems", and averaged together. We created the *Routine FB Behavior* composite variable by averaging responses to six questions about routine interactions with Facebook, based on Ellison et al. [15], Marino et al. [36], and Oldmeadow et al. [41]. Finally, four questions about participants' satisfaction with Facebook were based on Bhattacherjee [4] and Venkatesh et al. [53], and averaged to create a composite variable (*FB Satisfaction*).

The model results are presented in Table 3. The control variables with the largest coefficients in this model are *Trust Propensity* and *FB Satisfaction*, and these are the only sta-

Control Var.	*Knowledge Before*	*(SE)*
Age	-0.006	(0.00)
Gender (Woman)	-0.129	(0.12)
Internet Literacy	-0.136 •	(0.08)
Trust Propensity	0.217 **	(0.08)
FB Satisfaction	0.246 ***	(0.06)
Routine FB Behavior	0.019	(0.06)
Posted Last Week (Yes)	-0.027	(0.13)
Intercept	4.470 ***	(0.12)
	$R^2 = 0.097$	

'•' p<0.1; '*' p<0.1; '**' p<0.05; '***' p<0.01

Table 3: Regression coefficients (and standard errors) for the *Knowledge Before* model; these controls were used in all models in this paper. FB = Facebook.

tistically significant predictors. This means that people who are more trusting of social media sites and who are more satisfied with Facebook, on average, reported slightly higher levels of agreement that Facebook shows them every story created by their Friends and Pages they 'like'. Internet literacy also had an impact, but in the opposite direction, and the coefficient was not statistically significant; people with higher *Internet Literacy* reported slightly lower responses for *Knowledge Before*, meaning greater disagreement that they see everything their friends post. In other words, trusting social media and being satisfied with Facebook is associated with less knowledge about how the News Feed works, and greater internet literacy is associated with greater knowledge about what it is doing and how it works. The results of this model are correlational, so we cannot draw any conclusions about directionality of this effect. It could be that people who know that they don't see every story are less satisfied, or that less satisfied people are more likely to notice that they're not seeing everything their friends post.

All Explanations Provided New and Surprising Info

After the manipulation, we asked two questions of all participants about whether the information in the explanation was new to them (*New Info*), and had surprised them (*Surprising Info*). These questions were designed to help determine whether the explanations, on average, were simply telling participants things they already knew. If so, it is unlikely that they would be effective transparency mechanisms. Responses to both of these questions were on a 7-point Likert agreement scale. We used two regression models to analyze the effect that the experiment manipulation had on whether participants felt the information in the explanation they read was something they were already aware of, and if it seemed unexpected to them. The results of these regressions are presented in Table 4, and also shown in the heatmap in Figure 1 which depicts the effect sizes of the experiment conditions and controls for all of the outcome variables. Note that all of the models shown in the heatmap have one additional control for participants' *Knowledge Before*.

Compared with the control condition, all of the explanation conditions increased agreement that the information provided was new and surprising. All coefficients for the four explanation conditions increased *New Info* by over 1 point, which

	New Info		*Surprising Info*	
What	1.11 ***	(0.21)	1.06 ***	(0.20)
How	1.13 ***	(0.21)	0.80 ***	(0.20)
Why	1.33 ***	(0.21)	1.14 ***	(0.20)
Objective	1.55 ***	(0.21)	1.00 ***	(0.20)
Intercept	3.48 ***	(0.19)	2.68 ***	(0.18)
	$R^2 = 0.18$		$R^2 = 0.17$	

'•' p<0.1; '*' p<0.1; '**' p<0.05; '***' p<0.01

Table 4: Experiment condition regression coefficients (and standard errors) for the new and surprising information models. The Intercept represents the control condition in the experiment.

is a fairly large effect on a 7-point scale. Responses to the *Surprising Info* variable also indicated agreement, although less so than on the *New Info* variable. The *Objective* condition produced the biggest increase in agreement that the information was new. We suspect that this is because most non-expert Facebook users do not know that Facebook does user testing and frequently updates the algorithm, despite the strong emphasis placed on this information in Facebook's public statements about the News Feed. The *Why* condition was the most surprising to our participants. We think this is because it is the only explanation that focuses on the information overload problem, and it emphasizes that Facebook's solution is to use an algorithm to prioritize stories differently for each person. Stronger agreement that Facebook shows every post (*Knowledge Before*) and lower *Internet Literacy* both led to more agreement that the information was new and surprising in these models. These results indicate that the explanations did provide new information that participants felt in some cases was somewhat surprising, and the explanations are indeed providing new understanding to participants.

All Explanations Increased Awareness

We asked two types of questions designed to measure changes in participants' basic awareness of the News Feed ranking algorithm after reading the explanations. The first, *Knowledge After* (M=3.43, SD=1.7), is the same question that we asked before the manipulation (*Knowledge Before*, M=4.38, SD=1.61). On average, participants agreed less that they see all available stories in their News Feeds after being exposed to any of the explanations. The second type of question consisted of a series of statements describing possible reasons why participants may not see every available story when they visit their News Feeds. These statements were grouped based on an exploratory factor analysis, and averaged to create composite variables. We used two of the factors as outcome variables in regression models, *System Agency* and *User Agency*; the first related to the system as the entity responsible for choosing which stories the user sees, and the second related to the user determining which stories they see based on actions such as how far they scroll or how much time they spend on Facebook.

The *What* and *Why* conditions both had a medium-to-large, statistically significant effect on the *Knowledge After* outcome variable. In these conditions, participants' agreement that they see every story posted by friends and Pages decreased, as compared to the control condition (see Table 5 for the condition coefficients). The effect was strongest in the *What*

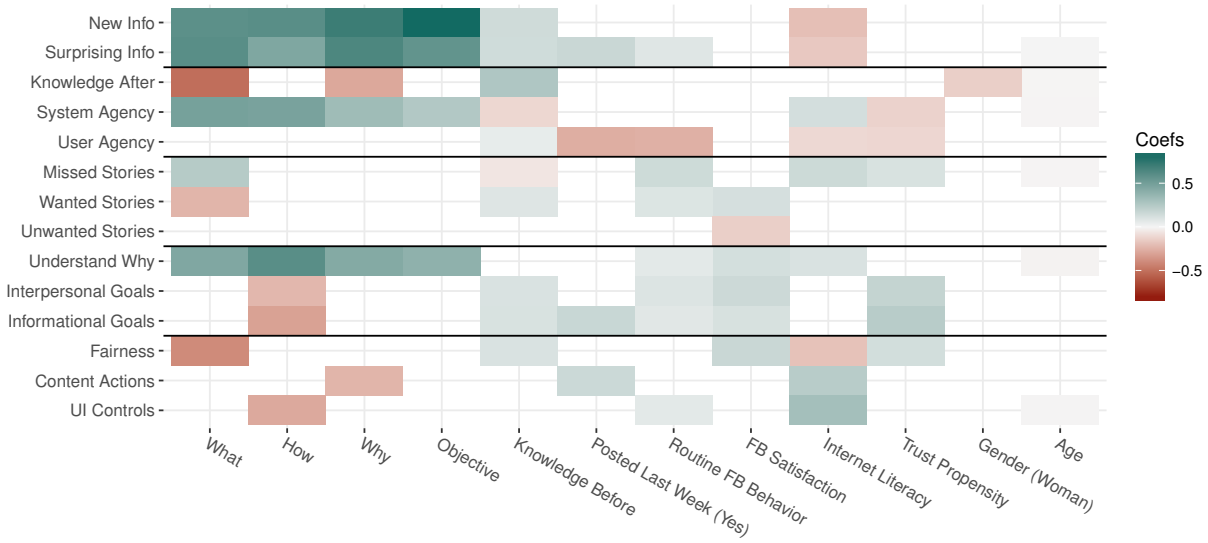

Figure 1: Heatmap showing all coefficients, including experiment conditions and controls, for the predictors in each model having p-values less than 0.05. Outcome variables were standardized so that they are directly comparable. Awareness variables = Knowledge After, System Agency, User Agency. Correctness variables = Missed Stories, Wanted Stories, Unwanted Stories. Interpretability variables = Understand Why, Interpersonal and Informational Goals. Accountability variables = Fairness, Content and User Actions.

condition, which explicitly states that Facebook users may miss stories that are placed lower down in the News Feed. The *Why* explanation contains similar information, but presents it as a solution to the information overload problem that users experience, which may be why the effect was not as strong.

All explanations affected participants' beliefs about the role of *System Agency* in influencing the composition of their News Feeds, indicating that participant awareness increased. The coefficients for all of the explanation conditions were positive and statistically significant, compared to the control condition. The effect sizes were largest in the *What* and *How* conditions, which were both nearly 0.5 points on a 1–7 Likert agreement scale. The intercept of the *System Agency* model, representing the mean for the control condition after controlling for the other variables in the model, is 3.88 (4 is neutral); this means that the effect sizes were large enough in the *What* and *How* conditions to change participants' responses from slight disagreement to slight agreement regarding *System Agency* on average.

In contrast, the *User Agency* outcome variable showed no differences from the control condition. Participants in all conditions expressed slight agreement that their own behaviors can cause them to not see all of the available stories. This means that while the explanations were able to change participants' responses on the questions that were related to the agency of the News Feed ranking algorithm, they did not cause participants to shift any responsibility for the stories they see from themselves onto the system. This is likely because while all explanations mentioned that an algorithm decides the order of stories in the News Feed, none of the explanations made a connection between the ranking algorithm and user actions such as scrolling behaviors or frequency of visiting Facebook.

A explanation successful at promoting awareness would inform people about the role of the algorithm, and all four explanation types in our experiment did this in a measurable way. However, only the *What* and *Why* conditions, which focus more on outputs of the algorithm than inputs, changed participants' specific beliefs about whether they see every story. This means that there are different implications for users depending on whether the transparency mechanism is focused on the design and testing of the system or what the inputs are, versus what the algorithm does and the reasons it is necessary.

What Explanations Affected Beliefs about Correctness

We measured correctness-related beliefs using questions that asked participants to think about whether the News Feed's outputs are correct given what they believe they should be seeing. One question asked participants if they intended to go to their friends' Facebook Timelines to look for stories they had missed (*Missed Stories*). An additional measure consisted of a block of related statements asking participants to evaluate the frequency with which they see different types of stories in their News Feeds, on a scale from "Not Often Enough" (0), to "Too Often" (100). We grouped the statements using exploratory factor analysis, and the factors that we used in the analyses represent *Wanted Stories* from people the participant wants to keep in touch with or that they find interesting or informative, and *Unwanted Stories* from people the participants don't want to hear from, stories they don't want to see, or stories posted by people they don't know.

Only the *What* condition had a statistically significant but fairly small effect on two of the three correctness-related beliefs and behavioral intentions; there were no effects from the other explanations. The *What* explanation caused a small increase on the *Missed Stories* outcome variable, indicating that participants agreed more strongly that they would go look

	Knowledge After	*System Agency*	*User Agency*	*Missed Stories*	*Wanted Stories*	*Unwanted Stories*
What	-0.87 *** (0.18)	0.49 *** (0.12)	-0.09 (0.14)	0.41 * (0.20)	-3.61 * (1.75)	3.27 (2.26)
How	-0.30 • (0.18)	0.48 *** (0.12)	0.14 (0.14)	0.16 (0.20)	-3.20 • (1.76)	0.33 (2.27)
Why	-0.49 ** (0.18)	0.33 ** (0.12)	0.12 (0.14)	0.27 (0.20)	-2.65 (1.79)	0.83 (2.31)
Objective	0.21 (0.18)	0.26 ** (0.12)	0.06 (0.14)	0.24 (0.20)	-2.64 (1.76)	1.07 (2.27)
Intercept	3.83 *** (0.16)	3.88 *** (0.10)	4.50 *** (0.12)	4.01 *** (0.17)	56.68 *** (1.55)	54.93 *** (2.00)
	$R^2 = 0.26$	$R^2 = 0.12$	$R^2 = 0.23$	$R^2 = 0.13$	$R^2 = 0.12$	$R^2 = 0.09$

'•' p<0.1; '*' p<0.1; '**' p<0.05; '***' p<0.01

Table 5: Explanation condition regression coefficients (and standard errors) for the *Awareness* (Knowledge After, System Agency, User Agency) and *Correctness* (Missed Stories, Wanted Stories, Unwanted Stories) models. The Intercept represents the control condition.

for stories on friends' Timelines, that they may not have already seen. In addition, participants in the *What* condition reported stronger beliefs that they do not see *Wanted Stories* often enough, compared with the control condition. However, there was no difference between the control condition and any of the explanation conditions regarding *Unwanted Stories*. These regression results are presented in Table 5.

A successful explanation would support correctness judgments by providing information to users that helps them understand better what the system is supposed to be doing, so they can recognize when it makes mistakes or errors. The only explanation that caused any changes to our measures of correctness judgments was the *What* explanation, which seemed to create some skepticism about whether the system was showing participants the stories they wanted to see. This was somewhat surprising because intuitively, information presented in the *How* explanation about the signals used as input to the algorithm, or information about the data-driven design process in the *Objective* explanation, seems like it should be more useful for actually evaluating correctness rather than simply creating uncertainty.

How Explanations Affected Beliefs about Interpretability

Interpretability-related beliefs are different from correctness beliefs in that they are more focused on the higher-level goals of the News Feed, and how well the News Feed supports those goals. To measure interpretability beliefs, we asked participants for their agreement with a statement about understanding the reasons why they see the stories they do in their News Feeds (*Understand Why*). A second question asked them to evaluate how consistent or random the News Feed is at helping them meet a series of common goals for using the News Feed, on a scale of "Completely Random" (0) to "Completely Consistent" (100). These items were grouped into factors using an exploratory factor analysis, and we used two of those factors in the regression models. One factor represents goals related to entertainment, experiencing a variety of content, and keeping in touch with people (*Interpersonal Goals*) and the other represents goals related to staying informed about news, events, and job opportunities (*Informational Goals*).

All of the explanations had a positive and statistically significant effect on the *Understand Why* outcome variable. The intercept is 4.50 (4 is neutral), and all explanations caused understanding to increase by half a point or more. The largest effect was in the *How* condition which raised agreement by 0.83 points, which is a fairly large effect. However, the *How* explanation was the only condition that differed from the control on both *Interpersonal* and *Informational Goals*. The relationship between the *How* explanation and these outcome variables was negative, indicating that this explanation caused participants to believe that the News Feed's behavior is more random when compared with the control condition. In other words, the *How* explanation decreased the perception that the News Feed consistently helps people meet *Interpersonal* and *Informational Goals*. See Table 6 for these regression results.

The relationship between the *How* explanation and the two interpretability *Goals* outcome variables is somewhat surprising, because the *How* explanation does not discuss goals at all, whereas the *Why* explanation does. A successful explanation for algorithmic transparency would help people make interpretability judgments about how sensible and consistent the system's behaviors are given the motivations behind what it is doing and how it works. The *How* explanation, with its focus on automatic ranking based on data, may have caused participants to doubt the algorithm's ability to be consistent.

Objective Explanations Did Not Affect Accountability

Our final set of measures focused on accountability. These questions asked participants to consider various ways that the system might be accountable to them, as individual users. We asked questions designed to measure whether participants thought the News Feed is fair, and whether they felt like they could influence or control what they see. The first question is a composite of three items measuring participants' beliefs about how fair and unbiased the News Feed is, which we averaged together (*Fairness*). We also asked participants to report how likely they believed a series of different actions would be to affect what they see in their News Feeds. After an exploratory factor analysis, we created two composite variables from their responses, one representing content-related actions such as 'liking' or commenting on a story, or following a person or a Page (*Content Actions*) and the other related to the controls that are provided for Facebook users to prioritize who to 'see first' or sort their News Feeds (*UI Controls*).

Three of the explanations, *What*, *How*, and *Why*, had an effect on participants' responses, each on a different outcome variable. Only the *What* condition differed from the control on *Fairness*. The coefficient for the *What* condition was medium-sized, and statistically significant (see Table 6). The average for the control condition after taking the control variables into account is 4.04, which is almost exactly neutral. The *What* condition decreased this by 0.5 points, indicating that

	Understand Why	*Interpersonal Goals*	*Informational Goals*	*Fairness*	*Content Actions*	*UI Controls*
What	0.62 *** (0.16)	-1.97 (1.79)	-1.96 (1.84)	-0.50 *** (0.14)	-0.70 (2.04)	3.24 (2.11)
How	0.83 *** (0.16)	-3.85 * (1.79)	-5.40 ** (1.85)	-0.18 (0.14)	-0.27 (2.04)	-5.31 * (2.12)
Why	0.60 *** (0.16)	-1.38 (1.82)	-2.24 (1.88)	-0.14 (0.14)	-4.10 * (2.08)	-2.97 (2.15)
Objective	0.54 *** (0.16)	-2.55 (1.80)	-3.60 • (1.85)	-0.24 • (0.14)	-1.58 (2.05)	-1.99 (2.12)
Intercept	4.50 *** (0.14)	67.24 *** (1.58)	57.49 *** (1.63)	4.04 *** (0.12)	70.66 *** (1.80)	66.12 *** (1.87)
	$R^2 = 0.16$	$R^2 = 0.25$	$R^2 = 0.25$	$R^2 = 0.20$	$R^2 = 0.08$	$R^2 = 0.17$

'•' p<0.1; '*' p<0.1; '**' p<0.05; '***' p<0.01

Table 6: Explanation condition regression coefficients (and standard errors) for the *Interpretability* (Understand Why, Interpersonal Goals, Informational Goals) and *Accountability* (Fairness, Content Actions, UI Controls) models. The Intercept represents the control condition.

participants who read this explanation subsequently believed that the News Feed is less fair than participants in the control condition.

The *How* and *Why* conditions both affected participants' perceptions that their actions can affect which stories they see in their News Feeds. Participants who saw the *Why* explanation felt less like their behaviors on Facebook affect the stories they see than those in the control condition, and participants who saw the *How* explanation felt less like the UI controls have an effect on what they see. This may be because the *How* explanation emphasizes the automatic data collection and the score the algorithm creates for every story, but does not discuss how input from the user interface controls may be accounted for by the algorithm, so participants may have been unsure how their use of those controls would be taken into account. The *Why* explanation emphasizes that quality signals are important for prioritizing stories, but it is not specific about how those signals are determined, and does not specify whether they are related to actions users have control over such as 'liking' or commenting on stories. These three explanations brought about changes in accountability-related beliefs, in that participants felt that the News Feed is less fair (*What*), and that they have less control through the UI (*How*) and through their content-related behaviors (*Why*).

The *Objective* explanation, in contrast to the others, was no different from the control condition on all of the accountability-related measures. This is somewhat surprising, because the *Objective* explanation presented information about the data-driven methods that Facebook uses to hold itself accountable to its users, by conducting user testing and revising the algorithm. It is possible that any accountability provided via user testing is too far removed from individual users' experiences with Facebook for their accountability-related beliefs to be affected. In fact, the *Objective* condition was the least impactful of all of the conditions on any outcome variable in the experiment. This is clear from looking at the heatmap in Figure 1 and the summary of the results in Table 7. This indicates that learning about Facebook's user testing may not have been meaningful for our participants, and that the *System Agency* effects of the *Objective* condition may be more related to its similarities to the other conditions than its unique content. However, we cannot make causal claims about this because the content of the explanations differed from each other in multiple ways.

	What	*How*	*Why*	*Objective*
Awareness	X	X	X	X
Correctness	X			
Interpretability		X		
Accountability	X	X	X	

Table 7: Summary of the results. Cells with an "X" indicate functions of transparency that were affected by an explanation.

DISCUSSION

In this study, we focused on the effects of four types of explanations of the Facebook News Feed on user beliefs related to the functions that transparency performs. Our goal was to conceptualize transparency as a mechanism for bringing about reflection and change, and measure the potential effects of providing new information to users about the system. We found that all of the explanations we created contained information that participants believed was new and surprising, made them more aware of the effects of the algorithm, and caused them to feel more like they understood why they see the stories they do in their News Feeds. However, our results show that some impacts of transparency are more amenable to brief, easy-to-read explanations than others. The *What*, *How*, and *Why* explanations all supported both awareness and accountability, but the interpretability and correctness functions of transparency were harder to actualize (see Table 7 for a summary).

The intuition behind calls for greater algorithmic transparency is that providing more information about a system will allow users to be "better able to judge whether a [it] is working as intended and what changes are required" [1]. Our experiment provides evidence that short explanations based on information that a corporation is already willing to provide about its system may not be helpful for achieving change-related transparency objectives. Awareness is possible just by being exposed to new information, and it is a encouraging that all of our brief explanations caused awareness beliefs to change. But the awareness effects were only possible because there are so many people who use Facebook who still are unaware of the influence of the algorithm as a gatekeeper for the information they have access to, and for their interactions with others through the system. Awareness on its own is necessary but not sufficient to bring about the ultimate goal of transparency mechanisms: enabling users to take action to change the system in some way.

In a recommender system, explanations are presented as a way to help users make a choice or take an action. In that context, explanations are a support mechanism for a specific task that the user is performing. However, in an algorithmic decision-making system like Facebook, the ultimate goal of providing a transparency mechanism is less clear and immediate because users are not presented with an explicit choice to make, or an action to take. Instead, users must do additional work to connect the new information with their own past experiences in the system. But, once they have done that work, the next steps to enact changes are unclear.

The algorithmic decision-making that takes place in systems like Facebook's News Feed is invisible to users, so correctness and interpretability judgments are necessarily more difficult to support through explanations than increased awareness. Correctness judgments require having an idea of what the "correct" output would look like and being able to identify when the system has made a mistake. Interpretability judgments rely on the ability of users to form a greater understanding of the goals behind the system's behavior, based on the information provided in the explanation. Our explanations improved awareness, but left participants with beliefs that their News Feeds behave more randomly, show them less of what they expect to see, and that they have less control, than participants who read the neutral control text. These do not seem to be the kinds of beliefs that would empower users to change their own behaviors, or to seek change through other means. However, because we measured only short-term, self-reported effects we cannot say how the changes in beliefs we identified might affect behavior in the short or long term. It is possible that if users feel deprived of self-determination, they may seek actions that would allow them to regain a sense of control.

Accountability is often discussed as the ultimate goal of transparency; it is thought to be a means of shifting the balance of power [1, 14] via increased scrutiny [13, 56]. Our explanations were successful in bringing about increased scrutiny; still, in a system where the algorithm has a greater degree of agency than the user, transparency is "disconnected from power" [1]. Individuals have little recourse in their current relationship with the system for exerting control over it—other than to stop using it, which is something we did not ask participants about. The transparency mechanism itself is sometimes believed to do work that produces understanding; but, explanations in an algorithmic decision-making system are only a first step. Because it is difficult for explanations in algorithmic decision-making systems to provide clear actions for users who want to enact changes, they "place a tremendous burden on individuals" [1] to interpret the new information and figure out for themselves what it means for them and how important and relevant it is to how they use the system. Our correctness and interpretability measures asked participants to consider what characteristics of a hypothetical News Feed without missed stories, and where users' goals are consistently met might look like. There is certainly currently a need for more support for correctness judgments in social media, and both Facebook[3] and Google[4] have recently implemented mechanisms to support this in the form of fact-checking tools for "fake news". But beyond mere exposure to a transparency mechanism, users must do extra work to investigate and take action—work that they may be unwilling to do.

Finally, we based the content of the explanations on information from Facebook's 'News Feed FYI' blog, which we condensed but tried to represent faithfully so as not to deceive our participants. It is notable that the explanations tended to change beliefs in a direction that was less favorable to Facebook, such as causing participants to decrease agreement that the News Feed is fair and unbiased. This was true even though the explanations overall were focused on the beneficial effects of the algorithm: that it ranks stories so as to show users the stories they will want to see (*How*), prioritizes stories that are high quality and important to users (*Why*), and is evaluated to ensure that it continues to improve (*Objective*)[5]. However, all of the explanations were somewhat surprising to participants. This could represent a feeling of violated expectations, which tends to decrease satisfaction and could result in users perceiving that their goals are not being met and that the system is unfair [11, 17]. This seems like a dilemma for explanations as a mechanism for algorithmic transparency; if the aim is to provide information that users are not aware of, then it seems inherently difficult to ensure that the new information does not violate user expectations. Determining what aspects of the explanations were surprising, and how to mitigate effects of expectation violations, is left for future work.

Limitations

For the sake of external validity, we designed the explanations to only contain the ideas and facts expressed in Facebook's News Feed FYI blog, but we might have found larger and different types of effects if we had presented information that cast the News Feed in a different light. Also, while our method allows us to identify causal effects, this is an exploratory study, and the explanations we designed differ from each other in multiple ways. We can attribute differences to which explanation participants read, but we cannot draw causal conclusions about which parts of the texts caused which effects. In addition, the effect sizes are generally small as are the R^2 values for the models. This is to be expected in a study of this nature, where our goal is identifying patterns and not prediction, but it could mean that the practical significance of the differences we observed is limited. Finally, our sample is not representative so generalizability is limited to the characteristics of our sampling frame.

ACKNOWLEDGMENTS

We thank Chankyung Pak, Nick Gilreath, and the BITLab @ MSU research group for helpful discussions and feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1217212.

[3] https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

[4] https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/

[5] We piloted the explanations to make sure they did not differ in tone, and the pilot participants rated the texts as neutral to slightly positive.

REFERENCES

1. Mike Ananny and Kate Crawford. 2017. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 33, 4 (2017), 1–17. DOI: http://dx.doi.org/10.1177/1461444816676645
2. Association for Computing Machinery U.S. Public Policy Council. 2017. Statement on algorithmic transparency and accountability. (2017). https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
3. Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to recommend? User trust factors in movie recommender systems. In *International Conference on Intelligent User Interfaces*. 287–300. DOI: http://dx.doi.org/10.1145/3025171.3025209
4. Anol Bhattacherjee. 2001. Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly* 25, 3 (2001), 351–370. DOI: http://dx.doi.org/10.2307/3250921
5. Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*, Vol. 5. 153.
6. Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: Democracy and design. *Ethics and Information Technology* 17, 4 (2015), 249–265. DOI: http://dx.doi.org/10.1007/s10676-015-9380-y
7. Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 1–12. DOI: http://dx.doi.org/10.1177/2053951715622512
8. Federal Trade Commission. 2016. Big Data: A tool for inclusion or exclusion? Understanding the issues. (2016). https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf
9. Kelley Cotter, Janghee Cho, and Emilee Rader. 2017. Explaining the news feed algorithm: An analysis of the News Feed FYI blog. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1553–1560. DOI: http://dx.doi.org/10.1145/3027063.3053114
10. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455–496. DOI: http://dx.doi.org/10.1007/s11257-008-9051-3
11. Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3163–3174. DOI: http://dx.doi.org/10.1145/3025453.3025659
12. Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62. DOI:http://dx.doi.org/10.1145/2844110
13. Nicholas Diakopoulos. 2017. Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens. In *Transparent Data Mining for Big and Small Data*, Tania Cerquitelli, Daniele Quercia, and Frank Pasquale (Eds.). Vol. 32. Springer International Publishing, Cham, Switzerland, 25–43. DOI: http://dx.doi.org/10.1007/978-3-319-54024-5_2
14. Nicholas Diakopoulos and Michael Koliska. 2016. Algorithmic transparency in the news media. *Digital Journalism* 5, 7 (2016), 809–828. DOI: http://dx.doi.org/10.1080/21670811.2016.1208053
15. Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168. DOI: http://dx.doi.org/10.1111/j.1083-6101.2007.00367.x
16. Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "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*. 153–162. DOI: http://dx.doi.org/10.1145/2702123.2702556
17. Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *The International AAAI Conference on Web and Social Media*. 62–71.
18. Kenneth R. Fleischmann and William A. Wallace. 2005. A covenant with transparency: Opening the black box of models. *Commun. ACM* 48, 5 (2005), 93–97. DOI: http://dx.doi.org/10.1145/1060710.1060715
19. Mikkel Flyverbom. 2016. Transparency: Mediation and the Management of Visibilities. *International Journal of Communication* 10 (2016), 1–13.
20. Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* 32, 3 (2011), 90–98. DOI: http://dx.doi.org/10.1609/aimag.v32i3.2365
21. Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382. DOI: http://dx.doi.org/10.1016/j.ijhcs.2013.12.007

22. Justin Scott Giboney, Susan A. Brown, Paul Benjamin Lowry, and Jay F. Nunamaker. 2015. User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems* 72 (2015), 1–10. DOI: http://dx.doi.org/10.1016/j.dss.2015.02.005

23. Tarleton Gillespie. 2014. The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). The MIT Press, Cambridge, Mass., 167–194. DOI:http://dx.doi.org/10.7551/mitpress/9780262525374.003.0009

24. Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. 23, 4 (1999), 497–530. DOI: http://dx.doi.org/10.2307/249487

25. Eszter Hargittai and Yuli Patrick Hsieh. 2011. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1 (2011), 95–107. DOI: http://dx.doi.org/10.1177/0894439310397146

26. Bernie Hogan. 2015. From invisible algorithms to interactive affordances: Data after the ideology of machine learning. In *Roles, Trust, and Reputation in Social Media Knowledge Markets: Theory and Methods*, Elisa Bertino and Sorin Adam Matei (Eds.). Springer International Publishing, 103–117. DOI: http://dx.doi.org/10.1007/978-3-319-05467-4_7

27. Lucas D. Introna. 2016. Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science, Technology, & Human Values* 41, 1 (2016), 17–49. DOI:http://dx.doi.org/10.1177/0162243915587360

28. Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systems: Beyond matrix completion. *Commun. ACM* 59, 11 (2016), 94–102. DOI:http://dx.doi.org/10.1145/2891406

29. Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123. DOI: http://dx.doi.org/10.1007/s11257-011-9112-x

30. Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review* 165, 3 (2016).

31. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 3–10. DOI: http://dx.doi.org/10.1109/VLHCC.2013.6645235

32. Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1603–1612. DOI: http://dx.doi.org/10.1145/2702123.2702548

33. Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71. DOI: http://dx.doi.org/10.1016/j.jsis.2008.01.001

34. Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. 195–204. DOI: http://dx.doi.org/10.1145/1620545.1620576

35. Zachary C Lipton. 2016. The mythos of model interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. 96–100.

36. Claudia Marino, Livio Finos, Alessio Vieno, Michela Lenzi, and Marcantonio M. Spada. 2017. Objective Facebook behaviour: Differences between problematic and non-problematic users. *Computers in Human Behavior* 73 (2017), 541–546. DOI: http://dx.doi.org/10.1016/j.chb.2017.04.015

37. Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 3 (2016), 401–415. DOI: http://dx.doi.org/10.1177/0018720815621206

38. Brent Mittelstadt. 2016. Automation, algorithms, and politics: Auditing for transparency in content personalization systems. *International Journal of Communication* 10 (2016), 4991–5002.

39. Sayooran Nagulendra and Julita Vassileva. 2016. Providing awareness, explanation and control of personalized filtering in a social networking site. *Information Systems Frontiers* 18, 1 (2016), 145–158.

40. Kenya Freeman Oduor and Eric N. Wiebe. 2008. The effects of automated decision algorithm modality and transparency on reported trust and task performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 4 (2008), 302–306. DOI: http://dx.doi.org/10.1177/154193120805200422

41. Julian A. Oldmeadow, Sally Quinn, and Rachel Kowert. 2013. Attachment style, social skills, and Facebook use amongst adults. *Computers in Human Behavior* 29, 3 (2013), 1142–1149. DOI: http://dx.doi.org/10.1016/j.chb.2012.10.006

42. Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583. DOI: http://dx.doi.org/10.1007/s10618-011-0215-0

43. Frank Pasquale. 2015. *The black box society: the secret algorithms that control money and information*. Harvard University Press, Cambridge, MA.

44. John D. Podesta, Penny Pritzker, Ernest J. Moniz, John P. Holdren, and Jeffrey D. Zients. 2014. Big data: Seizing opportunities, preserving values. (2014). https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

45. Emilee Rader. 2017. Examining user surprise as a symptom of algorithmic filtering. *International Journal of Human Computer Studies* 98 (2017), 72–88. DOI: http://dx.doi.org/10.1016/j.ijhcs.2016.10.005

46. Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook News Feed. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 173–182. DOI: http://dx.doi.org/10.1145/2702123.2702174

47. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the Predictions of Any Classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144. DOI: http://dx.doi.org/10.1145/2939672.2939778

48. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014).

49. Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *Proceedings of the 2002 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 830–831. DOI: http://dx.doi.org/10.1145/506443.506619

50. Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*. 801–810. DOI: http://dx.doi.org/10.1109/ICDEW.2007.4401070

51. Nava Tintarev and Judith Masthoff. 2011. Designing and Evaluating Explanations for Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer-Verlag, New York, New York, 479–510. DOI: http://dx.doi.org/10.1007/978-0-387-85820-3_15

52. Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439. DOI: http://dx.doi.org/10.1007/s11257-011-9117-5

53. Viswanath Venkatesh, James Y. L. Thong, Frank K. Y. Chan, Paul Jen-Hwa Hu, and Susan A. Brown. 2011. Extending the two-stage information systems continuance model: incorporating UTAUT predictors and the role of context. *Information Systems Journal* 21, 6 (2011), 527–555. DOI: http://dx.doi.org/10.1111/j.1365-2575.2011.00373.x

54. Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246. DOI: http://dx.doi.org/10.2753/MIS0742-1222230410

55. Markus Zanker and Daniel Ninaus. 2010. Knowledgeable explanations for recommender systems. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. 657–660. DOI: http://dx.doi.org/10.1109/WI-IAT.2010.131

56. Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132. DOI:http://dx.doi.org/10.1177/0162243915605575