Placebo Effect of Control Settings in Feeds Are Not Always Strong

Silas Hsu
silash2@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Vinay Koshy
vkoshy2@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Kristen Vaccaro
kv@ucsd.edu
University of California San Diego
San Diego, California, USA

Christian Sandvig
csandvig@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Karrie Karahalios
kkarahal@illinois.edu
University of Illinois at
Urbana-Champaign
Urbana, USA

Abstract

Recent work has catalogued a variety of "dark" design patterns, including deception, that undermine user intent. We focus on deceptive "placebo" control settings for social media that do not work. While prior work reported that placebo controls increase feed satisfaction, we add to this body of knowledge by addressing possible placebo mechanisms, and potential side effects and confounds from the original study. Knowledge of these placebo mechanisms can help predict potential harms to users and prioritize the most problematic cases for regulators to pursue. In an online experiment, participants (N=762) browsed a Twitter feed with no control setting, a working control setting, or a placebo control setting. We found a placebo effect much smaller in magnitude than originally reported. This finding adds another objection to use of placebo controls in social media settings, while our methodology offers insights into finding confounds in placebo experiments in HCI.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **User studies**; *Empirical studies in HCI*.

Keywords

control settings, placebo, social media, Twitter, deception, dark pattern

ACM Reference Format:
Silas Hsu, Vinay Koshy, Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2025. Placebo Effect of Control Settings in Feeds Are Not Always Strong. In *CHI Conference on Human Factors in Computing Systems (CHI'25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3714197

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714197

1 Introduction

Placebos, long studied in medicine, have emerged as a compelling area of inquiry in HCI. For instance, around 2014, the dating site OkCupid showed users high compatibility scores indicating a good match with a potential partner, regardless of the true scores [61]. Users informed of high compatibility were almost twice as likely to initiate and sustain communication compared to those informed of low compatibility. It appeared that a placebo effect of perceived compatibility, rather than computable facets of compatibility, played a key role in bringing users together. Kosch et al. define the placebo effect in HCI as changes in users' behaviors and subjective evaluations arising from their expectations of system functionality, as opposed to actual functionality [37].

OkCupid's experiment with deceptive "placebo" scores and lack of informed consent raises ethical concerns. Intentional and unethically deceptive interfaces are frequent enough that scholarship has named deception as one key characteristic of dark patterns [14, 24, 25, 42]. Literature in placebo interfaces has shown deception can be quite effective [15, 35, 37, 57, 59]. However, there is limited knowledge around the mechanisms of placebo effects and the consequences in HCI. A more nuanced understanding of the design elements and environmental factors underlying a system's ability to fool users brings several benefits. First, it can guide designers in designing placebo interfaces in the rare cases where they are ethically justified. Second, it can help experimenters control for placebo effects to avoid biased user evaluations. Third, we can gain clarity in which design elements are the most manipulative and harmful, which regulators in the US and EU use as key signals for taking action [41, 42, 49].

To that end, we examine mechanisms of placebo effects in HCI, specifically placebo control settings for social media content. Social media is widely used, with companies like Facebook often publicizing new ways of controlling their systems [19, 27]. Yet, many social media settings have a vague, delayed, or unverifiable impact, e.g. "See fewer posts like this" or "Turn off personalized ads." We build on prior work by Vaccaro et al. [59], who showed that a "popularity" slider increased user satisfaction towards a Twitter feed, even when it randomized the feed. This study addresses limitations and confounds of the original study by modeling alternative explanations for a placebo effect in a feed popularity slider, including

participants' expectations of a better feed (*Expectations Mechanism*), classically-conditioned boosts in satisfaction from clicking (*Clicking Mechanism*), users valuing the ability to take control even if they do not utilize the control (*Sovereignty Mechanism*), and randomization from the placebo setting surfacing more novel and surprising content (*Randomness Mechanism*).

To test these mechanisms we designed an experiment where each of our mechanisms made a distinct prediction of the outcome. A probability sample of the US population (N=762) browsed Twitter feeds with or without placebo control settings, in combination with chronological or randomized feed sorting. Bayesian modeling revealed the placebo "popularity" slider increased feed satisfaction only slightly. Follow-up experiments using only Twitter users (N=123) revealed the same small effect. Increases in feed satisfaction occurred only in participants who used our sliders, corresponding to the Expectations Mechanism's prediction. The other mechanisms' predictions, meanwhile, did not come true. With a better understanding of the placebo effect's origin and size, we conclude with future directions for the research and design of controls for social media feeds.

2 Related Work

Our work with placebo control settings intersects multiple areas, including control settings in social media, dark patterns, and placebo effects. In this section we situate our work in each of these areas, which will motivate our investigation of placebo controls.

2.1 Controls in Social Media

Many works in HCI tout the foundational goal of providing people control [3, 20, 55]. Control settings – widgets like sliders, buttons, and switches for configuration – often implement this advice. On social media, well-implemented settings can allow users to better personalize their feeds when algorithmic curation may not reflect their values [17, 52, 53]. Many platforms provide settings like "More like this" (Google[1]), "Not interested in this post" (Twitter), and "Hot" and "Best" (Reddit). However, platforms seldom document the exact functionality of such controls, nor provide ways to easily validate that the controls worked. To our knowledge, only Vaccaro et al. has explored how users react when social media settings do not meet their expectations, finding that even settings that do not work increase feed satisfaction [59]. We add to this body of knowledge by addressing *what aspects* of vague and unverifiable settings improve user experiences.

2.2 Deceptive and Dark Patterns

Brignull first coined the expression *dark pattern* in his 2010 website that taxonomized problematic design patterns [8]. The website now uses the name *deceptive design*, and curates "tricks used in websites and apps that make you do things that you didn't mean to" [9]. Examples include fake messages about low stock to pressure consumers into purchases (Fake Scarcity), or disguising important information through Visual Interference. Dark patterns encompass more than deception [24, 25, 42, 45, 62]. For instance, the Forced Action pattern requires users to take some undesirable action to accomplish their goal, such needing to subscribe to a newsletter to access a site: this might subvert users' intentions, but it does not rely on deception. More generally, definitions of dark patterns in academic and legal literature cover a wide variety of facets, including deceptive/misleading interfaces; undermining of user intent, preferences, and autonomy; designer intentions; benefits to the service; and harms to the user [42].

We examine control settings that mislead or outright lie, a pattern the Norwegian Consumer Council calls the Illusion of Control [21]. For example, Facebook deployed settings that could restrict data sharing to "Friends Only," but still shared data with the third-party apps that friends used. A lawsuit and $5 billion dollar settlement followed [22]. In addition, platforms have deployed settings that users have trouble finding and understanding, especially privacy and ad settings [21, 28, 30]. For instance, Hsu et al. found that some Facebook users believed opting out of personalized ads also stopped online tracking [30]. Our study, in contrast, documents not misconceptions stemming from potentially deceptive settings, but how deceptive settings directly shape subjective feed experiences.

2.3 Placebos in HCI

Outside the dark patterns literature, another set of "placebo" literature discusses deceptive design. Placebos in medicine describe sham or inert treatments like sugar pills which nonetheless heal patients. HCI practitioners use *placebo* to indicate system descriptions or interfaces which imply functionality, while the system does not perform that function or behaves randomly. Given that expectations are a key driver of the placebo effect in medicine [50], Kosch et al. define the *placebo effect* in HCI as more positive user evaluations "due to heightened expectations in the system's capabilities" [37].

HCI practitioners have evaluated many placebos. An "emotion meter" for writing worked randomly, yet people evaluated it as accurate [57]. Players' subjective experiences of an online game improved with the display of a lower network latency, regardless of the actual latency [29]. Power-ups that did nothing increased players' perceived immersion [16]. Informing participants that an AI adapted the difficulty of a game or anagram puzzles increased feelings of immersion [15] and perceived performance [35, 37], even when no such AI existed. These boosts happened even with the portrayal of an unreliable AI, possibly because prevailing positive preconceptions of AI [35].

2.4 Ethics and Applications of Placebos

In medicine, placebos go against Kantian moral principles that reject deception, invite fears that patients will lose trust in doctors, and arguably infringe on patient autonomy [4]. Guidelines of the American Medical Association prohibit placebo treatments without the patient's informed consent [5]. Open-label placebos, where patients are informed they will receive a placebo, sidestep ethical concerns while still having medical efficacy [50].

On the other hand, a patient may *want* to be deceived for the sake of achieving relief from a disease. 76% of participants in a 2013 survey thought that a placebo treatment was acceptable if the doctor was certain it would help the patient, and around 21% thought that placebos were never acceptable [31].

In HCI, many of the same arguments for and against placebos apply. Adar et al. discuss the notion of the user wanting a system

[1] https://myadcenter.google.com/home

to deceive them, terming it "benevolent deception" [1]. Placebo interfaces already exist in fake crosswalk buttons and elevator close door buttons, justified by their relatively inconsequential nature and ability to provide people a sense of control [51]. For fake power-ups [16], one could argue that players prioritize fun, and they would endorse deceiving placebo power-ups. Other practical concerns like reduction in players' trust may arise, however.

Our study examines placebo social media controls stated to adjust the information that a user consumes, but do something else. Such placebo controls are rarely ethical. Users have diverse goals on social media, such as entertainment, catching up with close ties, reading news, finding recommendations, learning, etc [17]. A one-size-fits-all deceptive setting may not help fulfill everybody's goals, decreasing the likelihood of benevolent deception. Furthermore, many people use social media as a news source [40]. Controls that mislead information-seekers about a source's characteristics, such as its popularity, may undermine their ability to critically evaluate the source. However, benevolent deception or an overriding societal interest may justify placebo settings in rare cases. For instance, imagine a user with an eating disorder who frequently searches for dieting and exercise videos [26]. The platform might opt to ignore the user's requests to "see more like this." Or, imagine a platform has detected a user has spent an excessive amount of time browsing their feed, especially with negative content ("doomscrolling"). The platform might temporarily have the refresh option display the deceptive message that there is no new relevant content. The argument for both these placebos is that the user would endorse these plans to protect their own health.

2.5 Mechanisms of Placebo Effects

The HCI placebo literature has sought primarily to quantify how user expectations bias system evaluations so that experimenters can control for expectations; and secondarily to identify how designers might exploit placebo effects to improve user experiences. Our investigation of placebo controls can benefit both these applications. In addition, given the danger of unethical deceptive controls, our research can complement dark pattern research, which seeks to identify deceptive designs and quantify their harms. While the US and EU already prohibit many forms of deceptive design [41, 42, 49], regulators may not have the resources or mandate to pursue every instance of deception. The US's FTC evaluates whether a design has caused "substantial injury" to consumers [41]. A better understanding of the harms of a particular design can help prioritize enforcement and argue the case for substantial injury.

Much of the prior placebo work in HCI studies exposure to system descriptions, making it straightforward to conclude that user expectations or beliefs were responsible for the placebo effects. But certain placebo interventions cause side effects, making this conclusion not as straightforward. For instance, Denisova and Cook found that fake power-ups increased player immersion [16], but their study did not address whether players felt more immersed because they expected a boost in power (what we would term an expectation effect), or if they simply enjoyed collecting the flashy objects that represented the power-ups (an effect of the system's objective behavior that does not rely on prior expectations). This concern about side effects, in addition to the importance of control settings in social media, motivates our revisiting of Vaccaro et al [59]. Below, we introduce several mechanisms that could explain their finding that controls intended to adjust content popularity in a news feed increased user satisfaction, even when the controls actually randomized the content of the feed.

2.5.1 Expectations Mechanism. The Expectations Mechanism most closely aligns with the consensus understanding of placebo effects in the medical field [18, 50]. This mechanism predicts two specific outcomes of user expectations that a popularity slider will improve the quality of a feed. First, users with higher expectations will be more likely to use the setting. Second, the expectations cause users to feel that they have meaningfully controlled the feed to have better content, regardless of whether the setting worked as advertised.

How higher expectations cause users to feel the feed has better content could happen via several avenues. Users might expect the popularity slider to improve the feed because they assume popular content is better, and preferentially filter the feed to contain more popular content. Recommender systems researchers have documented that displaying ratings of content causes users to rate the content in the same direction [2, 13]; in our context, the expectation that popular content is better might also positively bias user perceptions of the feed.

Simultaneously, higher expectations could cause stronger confirmation bias [47] – a tendency to find, pay more attention to, and recall evidence that the feed contains desired content. This could explain how participants in Vaccaro et al. easily found ways to explain how even content that violated their expectations still fulfilled the setting's intent.

If expectations of a better feed play a strong role, then it signals that strong trust in the platform and its settings will increase the effectiveness of the settings. It would additionally invite more scrutiny towards platforms' messaging around their settings, e.g. "We care about your privacy" or "Your ads, your choice"[2]. This kind of messaging can invoke a greater scope of control than what the platforms actually provide [21].

2.5.2 Clicking Mechanism. Alternately, the act of clicking a setting and seeing a change might satisfy users, no matter the change, and regardless of whether they believe they have meaningfully controlled anything. This mechanism borrows the idea of classical or Pavlovian conditioning, where an organism has learned to associate a stimulus with an outcome. Much clinical research has demonstrated such effects [18, 50] – for instance, pairing immunosuppressive drugs with an inert drink confers some immunosuppressive effects to the inert drink, even with patients informed of the drink's inert nature [33]. In our case, the Clicking Mechanism posits that users have learned to subconsciously associate clicking with satisfaction though their past experiences with on-screen interactive elements.

Petrie and Rief argue that classical conditioning is a form of expectation, as it involves people learning to *expect* a result from a stimulus, even if the learning happens subconsciously [50]. Nonetheless, we have separated the Clicking and Expectations Mechanisms

[2]From Google's ad settings, https://myadcenter.google.com/home

due to the nature of expectation generation. The Clicking Mechanism not only implies that we should be wary of settings that encourage users to click on them, but also suggests that we should search for other stimuli that users have subconsciously associated with feed satisfaction.

2.5.3 Sovereignty Mechanism. In contrast, it may be that users value having the ability to take control (sovereignty), so seeing a control setting on-screen produces feelings of control and satisfaction, even if the user does not use the setting. At least one extension of the Technology Acceptance Model (TAM) suggests that the flexibility of information systems, which control settings would enhance, improve satisfaction [60]. Users might adopt the attitude that a setting might prove useful in the future, even if it is not needed now. Hsu et al.'s user evaluation of Facebook settings also inspired this hypothesis: Facebook users taken on a tour of control settings declined to change most settings but they still appreciated learning about the settings [30]. If mere awareness of a control setting causes a strong effect, then we should be more wary when platforms add and advertise their controls, as even controls that consumers don't use could manipulate them.

2.5.4 Randomness Mechanism. Lastly, we propose that a randomly-ordered Twitter feed might satisfy users more than the default chronological ordering. Randomization may surface more diverse or serendipitous content, and prior work suggests that diversity and serendipity of content improves user satisfaction [36, 38]. Put another way, we propose that a control setting that randomizes the feed is not an inert placebo; it improves the objective quality of the feed. If so, then feed designers should put more emphasis on improving content diversity/serendipity.

2.6 Research Questions

To recap, prior work found that a "placebo" slider that randomizes a feed increases user satisfaction towards the feed. But it did not address why, and left open the possibility of non-placebo-effects for increased user satisfaction. Disentangling the mechanisms by which a placebo setting satisfies users will guide experimenters, designers, and regulators working with feed control settings. To disentangle the available explanations, we must first observe placebo controls improving users' subjective experiences (e.g. feelings of control, satisfaction), and then test our four mechanisms. Thus, we ask:

- **RQ1: How does a placebo control setting in a social media feed affect users' subjective experiences of the feed?**
- **RQ2: What explains the placebo effect?**

3 Methods

We designed an experiment that could both demonstrate a placebo effect of settings (RQ1) and form a context in which our four proposed mechanisms each made a distinct, falsifiable prediction (RQ2). Showing that a prediction did *not* come true provides strong evidence for ruling out a mechanism. At a high level, participants browsed a Twitter feed alongside different control settings (or no setting), and then they rated their feelings of control over the feed and feed satisfaction.

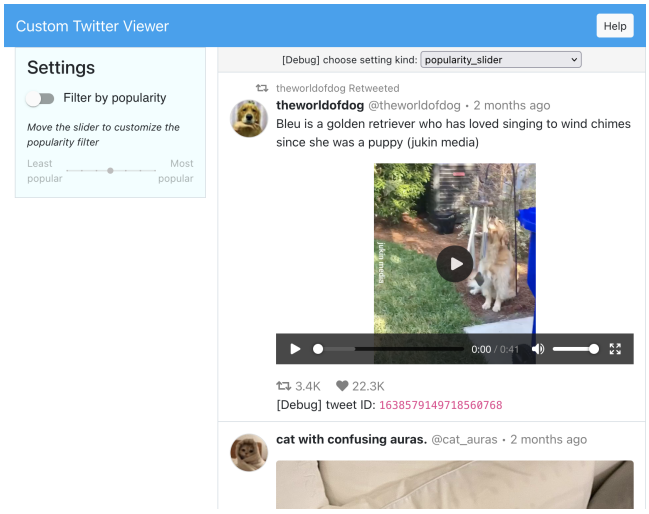

Figure 1: The interface that participants saw during the study. The Popularity Slider appears near the upper left corner.

3.1 Participant Recruitment and Survey Context

To improve generalizability we purchased a probability sample of the adult U.S. population to run our experiment on. 1464 people completed an online survey between March 17, 2023 and May 1, 2023. We employed NORC AmeriSpeak, a panel consisting of US households randomly sampled from a national list of addresses derived from the USPS Delivery Sequence File, with augmentations by NORC to better cover rural areas [48]. NORC contacts sampled households by US mail, telephone, and field interviewers (face to face) if necessary.

To increase the consistency of the feed interface, we restricted our recruitment from this panel to only include laptop/desktop computer users. Participants received a survey containing tasks and questions pertaining to several different research projects, including the current study, which occurred near the midpoint of the survey. Participants took a median of 29.5 minutes to complete the entire survey, with our study's tasks expected to take about 5 minutes. The survey completion rate was 14.5% and panelists were offered the cash equivalent of $2 for completing this survey.[3]

3.2 The Feed

Participants browsed a Twitter feed on a custom-made web app for 60 seconds, a screenshot of which appears in Figure 1. Tweets displayed the same fundamental information as found on Twitter (author, threads, videos/photos, number of retweets and likes, etc.); however, the app did not support interactions such as liking or replying. Like on Twitter and many of today's other feed systems, scrolling down loaded more tweets in an "infinite" fashion.

[3]We recognize the problematic nature of compensation well below U.S. federal minimum wage [56]. Unfortunately, we had no direct control over NORC's compensation, nor was it disclosed to us until after the completion of the survey.

Participants were first asked if they had an active Twitter account and were willing to use it for the study. If so, the feed viewer web app used the Twitter API[4] to pull up to 200 of participants' most recent home timeline tweets to populate the feed. Otherwise, participants selected up to six topics for their feed. The topics were designed to have broad appeal: Entertainment/Celebrities, Technology, News, Funny/Interesting, Sports, and Cute/Beautiful Photos. Selecting a topic populated the feed with tweets sampled from accounts that we designated as relevant to that topic. Appendix A contains more details on the account curation and tweet sampling process.

3.3 Experimental Treatments

Participants were randomly assigned to one of five experimental treatments in between-subjects fashion, summarized in Table 1. Three control setting types were possible: no setting, Popularity Slider, or Swap Button.

#	Control Setting	Default Feed Sort	Associated RQs & Mechanisms	Prob.
1	No setting	Chronological	All, as it functions as a baseline	10%
2	No setting	Random	RQ2: Randomness	10%
3	Popularity Slider	Chronological	RQ1, RQ2: Expectations, Sovereignty	30%
4	Placebo Slider	Chronological	RQ1, RQ2: Expectations, Sovereignty	30%
5	Swap Button	Chronological	RQ2: Clicking	20%

Table 1: Overview of experimental treatments. The Prob. column contains probability of assignment. Note experimental treatments with settings have increased probability of assignment to balance the number of participants using and not using settings.

3.3.1 The Popularity Slider (and Placebo Popularity Slider). Vacarro et al. designed a Popularity Slider that we partially replicated, shown in Figure 1. Vaccaro et al's slider always actively filtered the feed – to avoid a potential confound, we added a switch labeled "Filter by popularity" that started in an inactive state. While inactive, tweets appeared chronologically. This meant that participants with no setting and participants that did not use the slider would both see the feed in a chronological order, enabling estimation of the effect of setting existence (Sovereignty Mechanism) under the constant condition of chronological sort.

Turning the switch on or moving the slider activated the popularity filter. Instead of operating on tweets, the filtering algorithms for the slider operated on *threads*, chains of tweets which reply to each other. This was to ensure that filtering would not break up a thread and render it incoherent. Each of the seven ticks of the slider filtered the feed to only show threads in one quantile of popularity. The popularity of a thread was defined as the number of likes in the first tweet of the thread. The placebo version of the slider instead shuffled the full feed (i.e. the set of all threads) and divided it into seven equally-sized partitions. Each tick of the slider showed one of the partitions.

3.3.2 The Swap Button. Under the Clicking Mechanism, the mere act of clicking a button increases satisfaction, regardless of whether clicking leads to meaningful changes in the feed. To test this mechanism directly, we created a setting that afforded clicking but arguably caused minimal changes to objective feed quality. The result was a blue button labeled "Swap first two threads" that caused the first two threads in the feed to swap positions.

3.4 Measures

Our primary outcome measures were feed satisfaction and feelings of control over the feed. For feed satisfaction participants rated their agreement with "I enjoyed browsing the feed" and "I was satisfied with the final feed I saw." For feelings of control the statements were "I felt in control" and "The feed was the result of forces I couldn't control" (reverse-coded). Agreement ratings were on 5-point Likert scales from Strongly Disagree to Strongly Agree, which were then coded as integers and summed. Spearman correlations of the items within these two constructs were 0.66 and 0.36, respectively.

Besides these outcome measures we collected whether participants used a setting – crucial for testing the Clicking and Sovereignty Mechanisms. We deliberately avoided measuring user expectations for testing the Expectations Mechanism, as measurement can influence expectations [54].

Lastly, a process of reasoning, reflection, and literature review identified three potential confounds that could affect two or more of our primary measures simultaneously. We statistically controlled for these during analysis.

Before feed browsing happened, we measured the covariate of Locus of control (LoC). LoC describes the degree to which people attribute events in their lives to themselves ("internal locus") versus external causes ("external locus") [23]. It is predictive of people's tendency to seek and exert active control, as well as mental well-being [46]. We hypothesized that LoC could jointly influence both setting usage and feelings of control, as well as moderate the effect of control setting use on feed satisfaction. The survey assessed LoC using three items selected from Levenson [39]. Levenson's instrument encompasses three sub-scales validated through confirmatory factor analysis. From each sub-scale, we selected the one item with the highest loading on its corresponding factor.

Second among our covariates, we measured *perceived feed quality prior to setting use* (PFQ). People who expect a higher-quality feed might feel less need to improve the feed and thus use the setting less. At the same time, perceived feed quality likely causes feed satisfaction. The measurement of this variable depended on whether a participant expressed their willingness to use their Twitter account. If they did, prior to feed browsing they completed two questions about their general satisfaction and enjoyment towards their Twitter feed. Responses were on 5-point Likert scales. If a participant did not use their personal Twitter account, we instead used two proxy variables for PFQ: the feed topics that they selected and the age of the sampled tweets at the time they viewed the feed.

[4]https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-reverse-chronological

Lastly, after feed browsing, we measured the covariate of *familiarity with social media*. Prior work in recommender systems suggests that domain expertise, like knowledge of music when using a music recommender system, positively influences perceived recommendation quality [32, 36]. Therefore, social media or Twitter familiarity could predict feed satisfaction. At the same time, we hypothesized that social media expertise could impact a user's understanding of a setting and therefore their likelihood to use the setting. To report their familiarity with social media, participants rated their frequency of use of several popular social media services, including Twitter, Facebook, YouTube, Snapchat, Instagram, Reddit, and more. Frequency ratings were on a 6-point Likert scales ranging from Never to Multiple Times Per Hour.

3.5 Falsifiable Predictions

Given our experimental setup, each proposed mechanism predicts a different outcome. The Expectations Mechanism proposes that users' expectations positively bias their evaluations of the feed *after a setting makes changes*. It thus implies that using a setting will cause higher feelings of control, and subsequently feed satisfaction. The Sovereignty Mechanism predicts greater feelings of control and feed satisfaction in the presence of a setting, even when participants do not use it. The Clicking Mechanism predicts greater feed satisfaction, but not feelings of control, among those that click a setting. This prediction should hold true especially for those assigned to the Swap Button condition. The Randomness Mechanism predicts greater feed satisfaction when the feed is randomized, compared to chronological sort. The first two experimental conditions in Table 1 effectively model of the effect of random sorting under the constant condition of having no setting.

Figure 2 illustrates these predicted relationships in the form of a Directed Acyclic Graph (DAG). Note that while the DAG suggests path analysis, we tested the DAG's prediction using Bayesian linear regression (see Section 3.7).

3.6 Data Cleaning

First, we removed 203 participants who had experienced an older version of the Popularity Slider. Then, we removed participants showing evidence of low-quality responses considering behaviors across the entire survey, and not just the current study. Behaviors justifying removal included failing multiple attention checks, clear evidence of straightlining, completing fewer than 60% of the questions in the survey, failing to respond to our key outcome measures, or recalling a feed setting that the participant could not have experienced. This step removed 459 participants.

Of the remaining 802 participants, 40 used their own Twitter feeds. So few respondents used their own Twitter feeds that we opted to exclude them from the Main Experiment's dataset, and analyze them in the dataset for the Follow-Up Experiments which only contained people using their own Twitter feeds. This led to a final data set size of 762 participants for the main experiment. Tables 2 and 3 display various summary statistics of this dataset.

3.7 Data Analysis

Based on the DAG (Figure 2), we derived a series of Bayesian ordinal regressions, which Table 4 summarizes. Model 1 addresses RQ1 (placebo effect strength) by estimating the effect of experimental treatment on feed satisfaction. Models 2 and 3 each test a subset of the DAG's edges in a piecewise fashion, which addresses RQ2 (explanation of the placebo effect). This piecewise approach shares conceptual similarity to Baron and Kenny's method of using several regressions to test statistical mediation [6].

Treatment	Num. assigned	Num. using setting
(1) No Setting	66	0
(2) No Setting, randomized feed	66	0
(3) Popularity Slider	254	80 (32%)
(4) Placebo Slider	226	64 (28%)
(5) Swap Button	150	43 (29%)

Table 2: Number of participants assigned to each experimental treatment and setting usage statistics for the Main Experiment, after data cleaning.

Attribute	Distribution
Age	Mean = 45 years, median = 42 years, standard deviation = 17 years.
Gender	374 male, 376 female, 4 nonbinary, 5 no response
Race	535 (70%) White, 82 (11%) Hispanic, 62 (8%) Black, 46 (6%) multiracial, 22 (3%) Asian, 9 (1%) other, 4 (1%) American Indian, 2 Middle Eastern

Table 3: Participant demographics for the Main Experiment after data cleaning.

We used Gaussian priors for all regression coefficients, with means and variances encoding skepticism but not impossibility towards large effect sizes. For instance, before seeing any data, our models assigned under a 1% probability to setting use tripling the likelihood of the highest possible feed satisfaction score of 8

Ordinal regressions predict how the likelihoods of each possible DV value change as the IVs change. To distill this into one summary statistic, we selected the DV value of 6 as a pivot point, as obtaining a 6 requires at least one rating of agreement (a 3 or a 4) among the two items that make up each scale. We thus summarize many effect sizes as an odds ratio expressing *the predicted change in the likelihood of selecting a 6 or higher* due to a one-unit change in an IV. For IVs modeled as categorical (feed randomization, using a setting, UI type), the unit of change is moving from one category to another category. For IVs modeled as continuous, the unit of change is one sample standard deviation.[5] As an example, consider the effect of using a setting vs not using a setting on feed satisfaction. An effect size (odds ratio) of 2 indicates that using a setting doubled the likelihood of rating 6 or higher, 0.5 indicates a halving of the likelihood, etc.

Bayesian analysis provides posterior distributions of the most likely effect sizes from changing an IV (or more precisely speaking,

[5]Despite modeling feelings of control as an ordinal DV in Regression 2, we modeled it as a continuous IV in Regression 3. A desire to reduce the number of parameters in Regression 3 and observing a normally-shaped distribution for feelings of control justified this decision.

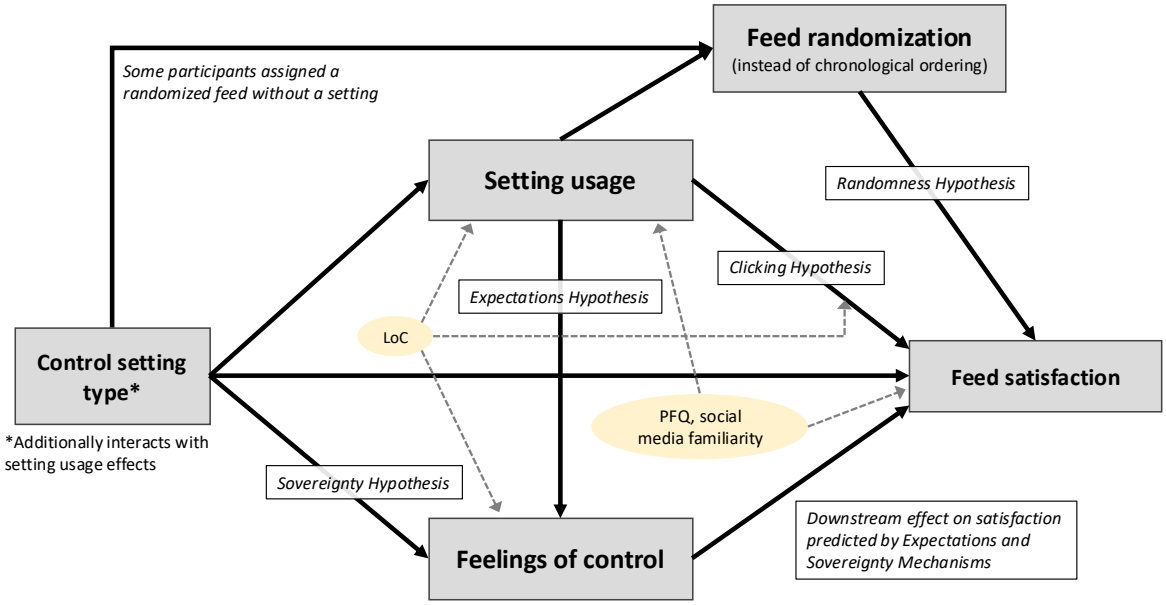

Figure 2: Diagram of how settings improve feed satisfaction. This diagram contains pathways for all mechanisms under RQ2. Control setting type (leftmost node) refers to the control setting that a participant was assigned; see the Control Setting values in Table 1. Our mechanisms predict the existence of certain edges, which are annotated. The direct arrow from "Control setting type" to "Feed satisfaction" captures any remaining effect of control settings on feed satisfaction that our four mechanisms do not explain. Lastly, confounding variables appear in light bubbles, which include Locus of Control (LoC), perceived feed quality prior to setting use (PFQ), and social media familiarity. The arrow from LoC to the Setting usage → Feed satisfaction relationship indicates our hypothesis that LoC mediates this relationship.

#	Independent Variables	DV	Controlled Confounds
1	Experimental treatment	Feed satisfaction	none
2	Control setting type, Used setting, Control setting type x Used setting	Feelings of control	LoC, LoC x Used setting
3	Control setting type, Used setting, Control setting type x Used setting, Feelings of control, Feed randomization	Feed satisfaction	LoC, LoC x Used setting, Perceived feed quality prior to setting use, Familiarity with social media

Table 4: Piece-wise modeling of the causal diagram through a series of regression models. DV stands for Dependent Variable or the response variable of each regression. Controlled Confounds are added as additional predictors for each DV.

the odds ratios most compatible with the parameters and assumptions of ordinal regression). We report several statistics regarding effect size distributions. First, we report the total posterior likelihood of an odds ratio exceeding 1.1, which can be interpreted as the likelihood of a non-negligible positive effect. Second, we report the mean of the posterior distribution. And third, we report the boundaries of the 94% Highest Posterior Density Interval (HPDI) – the narrowest interval containing 94% of the posterior's probability mass.[6]

3.7.1 Original data and models. The supplementary materials include survey questions, formal model specifications, and code to run the models. We did not gain consent from participants to share our data in a public repository, but researchers may contact silash2@illinois.edu to request access to the anonymized dataset.

[6]We select 94% as a relatively strict yet arbitrary threshold, just like the relatively strict yet arbitrary convention of p=0.05.

4 Main Experiment Results

4.1 Similar Average Satisfaction Across All Experimental Treatments

We first examine Regression 1, which analyzed how assignment to each of our five experiment conditions affected satisfaction scores. Our regression model predicts largest difference in feed satisfaction occurs between the "No Setting, random" and the "Popularity Slider" treatments. Compared to the No Setting, random treatment, assignment to the Popularity Slider treatment increases the likelihood of a 6-or-higher satisfaction rating by 2.7% on average, with a 96%

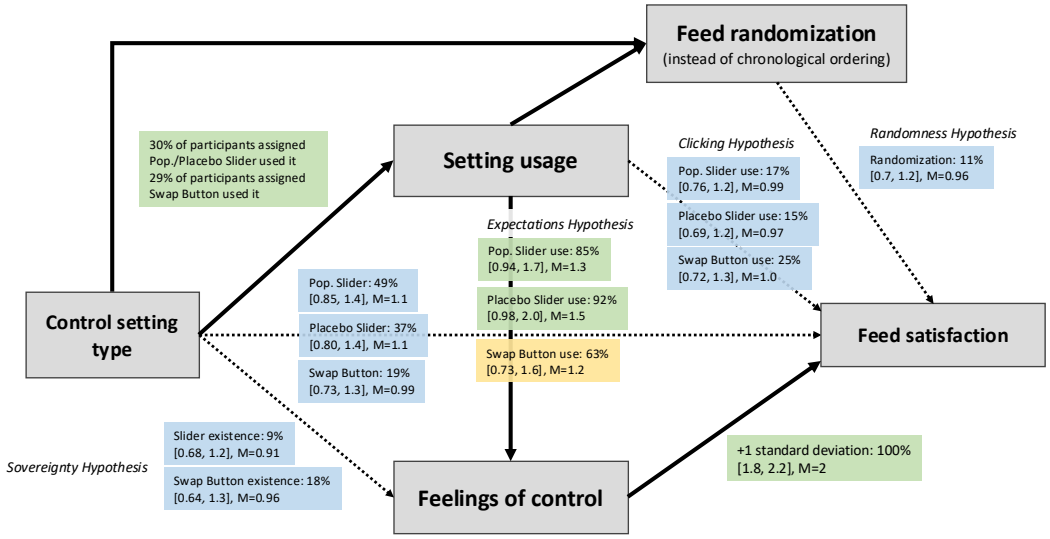

Figure 3: Causal diagram annotated with predicted effect sizes for the main experiment. Here, we define "effect size" as the odds ratio representing the change in the likelihood of a 6 or higher on an outcome scale, where the outcome scales are feelings of control and feed satisfaction. For each predicted effect size, we report three statistics. Percentages represent the posterior probability of an effect size greater than 1.1. The numbers in brackets represent the boundaries of the 94% HPDI (i.e. the range of the most plausible effect sizes). Finally, after the brackets, we report the mean of the posterior distribution. Green and amber backgrounds emphasize posterior probabilities of an effect size greater than 1.1 exceeding 80% and 60%, respectively.

Figure 4: Estimated probabilities of each satisfaction score for the "No setting, random" and "Popularity slider" treatments in the Main Experiment. This comparison has the largest effect size among all pairs of conditions. Each vertical line indicates the 94% Highest Posterior Density Interval (HPDI) containing the most plausible estimates, with the center marker (x for No setting, filled circle for Slider) indicating the mean of the posterior distribution. Takeaway: on average, assignment to the Popularity slider treatment very slightly increases the probability of a higher satisfaction score. However, since all the HPDIs overlap, a null effect is still very plausible.

probability of an increase less than 8.5% (94% HPDI boundaries = [0.9, 1.5]). Figure 4 presents exact details on predicted differences in satisfaction ratings.

Vaccaro et al. found that compared to no setting, a placebo slider increased satisfaction by an average of approximately 1 point on a 7-point Likert scale [58]. If we rescale our model's predictions to a 7-point Likert scale, it estimates an increase of 0.13 Likert points on average, with a 96% probability of an increase less than 0.45 Likert points. In other words, providing a setting affected feed satisfaction ratings much less than expected.

4.2 Support for Expectations Mechanism

Despite Regression 1 not finding significant differences among the treatments, we still conducted the remaining regressions to test our proposed placebo effect mechanisms and to determine if the settings satisfied any subset of participants. Figure 3 summarizes the findings of these regressions. Our models collectively predict that participants who used a slider, working or placebo, felt slightly more feed satisfaction, with a probability of at least 85%. This happens through the route of Slider use → Feelings of control → Feed satisfaction, which aligns with the predictions of the Expectations Mechanism. Using the Placebo Slider increases the estimated probability of feeling in control from 8% to 12%, with downstream effects on feed satisfaction sharing the same magnitude. Meanwhile, a sizable proportion of participants did not use a setting. This explains why assignment to experience a setting had such a small total effect on satisfaction. Supplemental Figures 8 and 9 illustrate these effect sizes with the raw response distributions.

Treatment	Num. instances	Num. instances w/ setting use
(1) No Setting	49	0
(2) No Setting, randomized feed	5	0
(3) Popularity Slider	59	38 (64%)
(4) Placebo Slider	40	24 (60%)
(5) Vague Popularity Slider	24	16 (67%)

Table 5: Experimental treatment distribution and setting usage statistics in the follow-up experiments. This table compiles statistics of 177 feed interactions from 123 participants. 27 participants from within-subjects Follow-Up W contributed three interactions each.

Attribute	Distribution
Age	Mean = 35 years, median = 32 years, standard deviation = 13 years.
Gender	69 male, 52 female, 2 nonbinary
Race	75 (61%) White, 16 (13%) multiracial, 15 (12%) Asian, 7 (6%) Black, 6 (5%) Hispanic, 2 (2%) Middle Eastern, 1 American Indian

Table 6: Demographics of the 123 participants in the follow-up experiments.

Our models assigned probabilities of at most 25% and effect sizes of at most 1.3 to the predictions of the Clicking, Sovereignty, and Randomness Mechanisms (see Figure 3). Notably, higher probabilities of positive effects are assigned to the Swap Button than the sliders. We attribute this to decreased model confidence from having fewer participants assigned to Swap Button.

5 Follow-Up Experiments with Twitter Users Only

We hypothesized that differences between our and Vaccaro et al's experiments may have caused a smaller effect size. In response we conducted two follow-up experiments that replicated additional aspects of Vaccaro et al's methodology: (1) a convenience sample of Twitter users, (2) a within-subjects design, and (3) a "Vague Popularity Slider" that omitted all explanation of functionality, leaving only a label that said "Popularity." Moreover, it omitted the activation switch and always actively filtered the feed.

All follow-up experiments occurred on the crowdsourcing platform Prolific in July 2023. We screened for participants located in the U.S. who reported using Twitter at least once per month. We recruited as many participants as our Twitter API limits would allow – at the time of the experiments, Twitter had reduced these limits significantly. Participants received a flat payment to match the rate of $13 USD per hour.

Follow-Up Experiment B compared the Vague Popularity Slider to No Setting and Popularity Slider in **B**etween-subjects fashion. Follow-Up Experiment W exposed subjects to No Setting, Popularity Slider, and Placebo Slider in counterbalanced **W**ithin-subjects fashion. Each participant completed feed browsing and responding to post-browsing measures three times in a row, once for each treatment. 74 and 29 participants completed Follow-up Experiments B and W, respectively.

5.1 Combining Datasets and Data Cleaning

The follow-up experiments used the same measures and feed interface as the Main Experiment, and many of the same experimental treatments. This justified combining the data from the follow-up and Main Experiment participants who used their own Twitter feeds. We start with 143 participants – 40, 74, and 29 from Main, Follow-Up B, and Follow-Up W respectively. We excluded 8 participants from the Main Experiment who experienced the Swap Button treatment because we wanted to avoid modeling a control setting type with only 8 people. We additionally removed 12 participants because the Twitter API fetched 20 or fewer tweets for their feeds (5, 5, and 2 removals in Main, Follow-Up B, and Follow-Up W respectively). After removals 123 participants remained collectively providing 177 feed interactions. Tables 5 and 6 compile summary statistics for this dataset. Bayesian ordinal regression proceeded in the same way as the Main Experiment, except with additional ordering effect terms.

6 Follow-Up Experiment Results

6.1 Similar Average Satisfaction Across All Experimental Treatments

Like the main experiment, the regressions indicate small effects across all conditions. Assignment to Popularity Slider produces the same satisfaction ratings on average as No Setting. The 94% probability upper bound of effect sizes for this comparison is 1.2, equivalent to an increase in probability of a 6-or-higher rating of at most 10%, or at most 0.43 Likert points on a 7-point scale. Figure 6 presents exact details on predicted differences in satisfaction ratings; Supplemental Figure 10 presents raw response distributions.

In addition we estimated ordering effects. Regression 3 predicts that compared to experiencing No Setting for the first time, experiencing No Setting after Slider multiplies the probability of a 6-or-higher satisfaction rating by 0.97 on average (94% HPDI boundaries = [0.79, 1.1]). We find a similar effect size for the other ordering of Slider after No Setting. That is, ordering effects do not explain the discrepancy between our findings and Vaccaro et al.'s.

6.2 Support for Expectations Mechanism

Again, we conducted further regressions to identify the mechanism(s) of the placebo effect, the results of which Figure 5 summarizes. The models predict that participants who used a slider, working or placebo, felt slightly more feed satisfaction with probabilities of 90% or higher. Again, we found alignment with the predictions of the Expectations Mechanism through the route of Slider use → Feelings of control → Feed satisfaction.

Due to the smaller sample size, the models indicate wider plausible ranges of effect sizes, making it harder to rule out placebo mechanisms. Regression 3 assigns a 25% posterior probability to the prediction of the Randomness Mechanism, and probabilities between 33% and 60% to the prediction of the Clicking Mechanism.

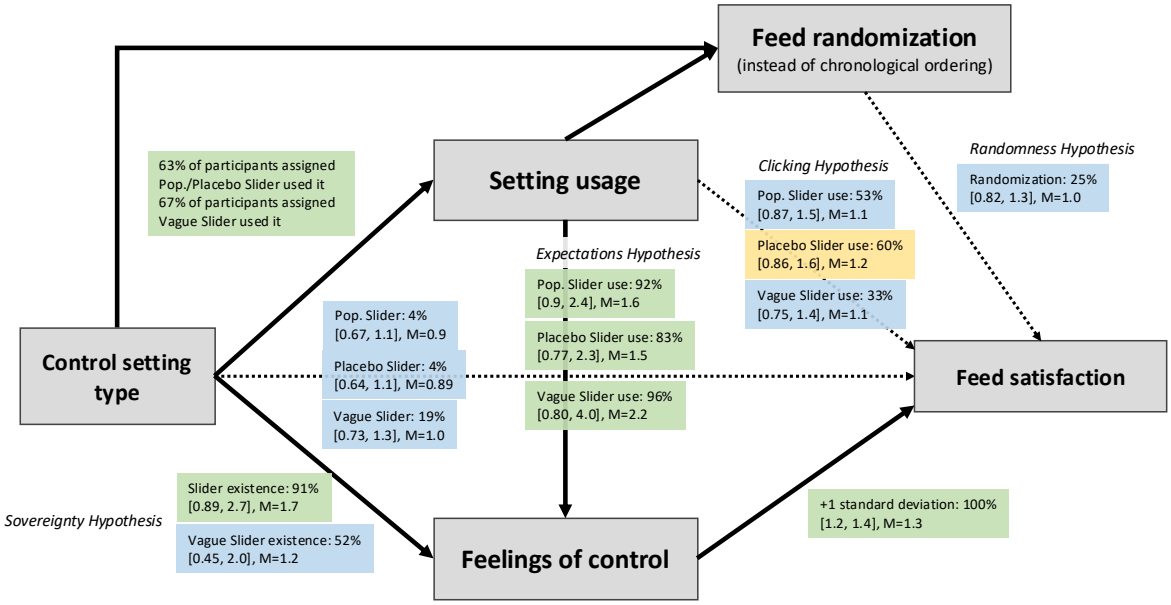

Figure 5: DAG annotated with predicted effect sizes for the follow-up experiment. Here, we define "effect size" as the odds ratio representing the change in the likelihood of a 6 or higher on an outcome scale, where the outcome scales are feelings of control and feed satisfaction. For each predicted effect size, we report three statistics. Percentages represent the posterior probability of an effect size greater than 1.1. The numbers in brackets represent the boundaries of the 94% HPDI (i.e. the range of the most plausible effect sizes). Finally, after the brackets, we report the mean of the posterior distribution. Green and amber backgrounds emphasize posterior probabilities of an effect size greater than 1.1 exceeding 80% and 60%, respectively.

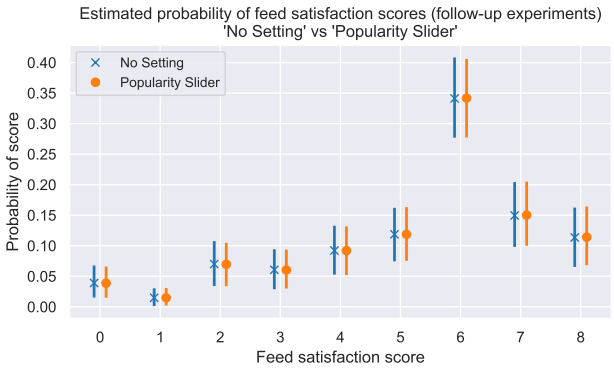

Figure 6: Estimated probabilities of each satisfaction score for the "No setting" and "Popularity slider" treatments in the follow-up experiments with Twitter users only. Each vertical line indicates the 94% Highest Posterior Density Interval (HPDI) containing the most plausible estimates, with the center marker (x for No setting, filled circle for Slider) indicating the mean of the posterior distribution. Takeaway: Our models estimated the No setting and Popularity slider treatments to have about the same level of feed satisfaction.

Our models at first glance support the Sovereignty Mechanism, with confident predictions that slider existence increases feelings of control, and feelings of control increase feed satisfaction. This is partially visible in the raw response data (Figure 7): participants had higher ratings of control when provided a slider, even when not using it. But contradicting the models, we did not observe a downstream effect of higher satisfaction among the non-users. Due to this apparent contradiction we consider support for the predictions made by the Sovereignty Mechanism inconclusive for this dataset, and discuss further implications in Discussion section 7.1.

7 Discussion

In a series of experiments with Twitter feeds, we scrutinized mechanisms underlying a placebo effect for control settings on social media. In this section, we discuss how we interpreted the data in light of the four mechanisms we tested, and we propose explanations for why we found a smaller effect size than expected. Lastly, we state recommendations for design and regulation, and discuss limitations and possibilities for future work.

7.1 Which Placebo Effect Mechanisms Can We Rule Out?

7.1.1 Evidence Against Clicking, Randomness, and Sovereignty Mechanisms. Neither the the Swap Button nor the randomly-ordered feed with no setting increased feed satisfaction to any practical

Figure 7: Feelings of control and satisfaction ratings in the follow-up experiments, grouped by slider usage. Ratings are as follows – 1: Strongly disagree, 2: Somewhat disagree, 3: Neither agree nor disagree, 4: Somewhat agree, 5: Strongly agree. Takeaway: Participants with a slider had higher feelings of control on average, but only those that *used* the sliders displayed increased feed satisfaction.

degree. Thus the data supports ruling out the Clicking and Randomness Mechanisms. However we do not rule out interaction effects, such as a conditioned response that only happens when users click on a setting they expect to be useful (i.e. interaction with the Expectations Mechanism). Fully disentangling such effects would require providing a way to assert control without clicking, such as via a proxy that clicks on a person's behalf.

The Sovereignty Mechanism predicted that participants provided a slider but not using it would feel more feed satisfaction, a prediction the data did not reflect. However, recall that in our follow-up experiments our models displayed confidence that setting existence increases feelings of control, with a downstream effect on feed satisfaction. This inconsistency could indicate issues in our measurement of feelings of control (see Section 3.4). Participants who declined to use a control setting may have interpreted our questions to mean that they felt agency, or the ability to act to control the feed. In comparison, participants who used the settings may have interpreted our questionnaire to mean that they had, in fact, successfully changed the content of the feed. Only this second meaning of control might predict feed satisfaction. In any case, our DAG and models indicate that *some* component of feeling in control predicts feed satisfaction. Future work should investigate more precise ways of conceptualizing feelings of control, not only through measurement but experimental manipulation.

7.1.2 Evidence Favors the Expectations Mechanism. The Expectations Mechanism's predictions consistently came true. However, our experiments only represent one test of this mechanism. Measuring user expectations could provide more definitive evidence, but only with careful implementation. Marketing research has shown the act of measurement causes users to reflect on their expectations, which amplifies expectation effects [54]. Therefore experiments that measure expectations should include a condition where expectations are not measured. Alternatively, we could use experimental treatments that alter system behavior, but *not* users' expectations, such as our treatment of No Setting with a randomized feed.

Additionally, our experiments provide a foundation for further study of *how* expectations lead to more positive ratings of the feed. Future work could attempt to separate the effects of assumptions about the quality of popular content, confirmation bias, or other factors. To do so, controlled studies could manipulate user expectations in more nuanced ways, such as by changing people's attitudes towards the quality of popular content before exposure to a popularity slider. This type of work could not only clarify how user expectations matter, but also give us more ways to manage those expectations.

7.2 Small Effect Sizes and the Need to Separate Placebo Mechanisms in HCI

Despite strong positive effects in prior work [59], the setting in our study slightly increased satisfaction ratings. Our follow-up experiments ruled out several explanations, but we did not and could not test every possibility. One remaining explanation is that our study's participants viewed a large continuously-scrolling feed, but Vaccaro et al.'s showed 10 posts at a time. If a participant never used a control setting, these were the only 10 tweets they saw during the study. Their slider design, placebo or not, enabled the viewing of more tweets, possibly providing significant objective utility. In our study, the benefits of the filters may have been more muted, as scrolling could already surface tweets participants wanted to see.

This alternative explanation highlights the need in HCI to separate "placebo" effects from non-placebo, "objective" effects. It also highlights the insufficiency of placebo terminology. HCI researchers borrowing the idea of a placebo from medicine may take for granted differences between the two contexts. Placebo effects in medicine occur when a treatment with no plausible biological mechanism improves patient outcomes, allowing us to conclude that a psychological mechanism must be the cause. But in HCI, the study of deceptive or placebo interactions involves not the separation of the biological and psychological, but rather different psychological effects. "Placebo effect" fails to capture the richness behind users' subjective experiences. Terms like "Expectations Effect" or "Sovereignty Effect" give designers and regulators a clearer picture of what interventions to adopt.

7.3 Implications for Design and Regulation

In summary, a deceptive "placebo" popularity slider increased short-term feed satisfaction, but less than previously thought; and user expectations towards the slider explains this placebo effect. An unethical designer might use these findings to implement placebo settings that target user expectations more strongly. But designers should strongly consider that placebo settings might not cause a large effect and that expectations can quickly deteriorate. User trust is fragile, especially with violated expectations [34] or investigations exposing misleading interfaces, such as what happened with Facebook's privacy settings [22]. Low trust towards settings may even cause a "nocebo effect" that undermines their ability to improve user experiences. Already-documented symptoms of nocebo effects may include users' belief in the inevitability of online tracking [44], or the belief in seeing personalized ads even after using a setting to turn them off [30]. Future research should examine nocebo effects in more detail.

Our findings underscore that not only overt deception can influence subjective experiences, but also inflation of user expectations. Many examples of potential harms from mismanaged expectations have been documented. Facebook users thought that turning off "ads based on data from partners" would stop Facebook from tracking them, when in fact the setting did not [30]. A court recently ordered Google to dispel misconceptions of privacy and clarify to users that it could still track users even with Incognito web browsing [10]. People provided more granular control over information sharing decided to share more information [7]. That is, even subtle cues like control setting granularity might influence people's expectations of security and cause them to share more.

Our work complements these examples. When user expectations shape subjective experiences and evaluations of a system, as it did in our study, it potentially becomes a tool for manipulating user behavior. This would clearly violate EU regulations which prohibit practices that "materially distorts or impairs the ability of [consumers] to make free and informed decisions" [49].

Mismanagement of expectations may also run afoul of FTC regulations as well. The FTC restricts deceptive practices, defined as any "representation, omission, or practice" that is "material" and "likely to mislead consumers who are acting reasonably under the circumstances" [12]. "Material" means relating to "information that is important to consumers and, hence, likely to affect their choice of, or conduct regarding, a product" [12]. Our findings suggest a stronger connection between user expectations and their subjective evaluations of feeds, and therefore their behaviors surrounding feeds. In other words, there is now stronger evidence that user expectations are "material" and worth it for auditors to monitor. As mentioned before, directly querying expectations via asking users how "good" a product will be might change expectations [54]. Auditors should alternatively consider interviews that ask users why they use a feature or how a feature works, for example [30]. If users' reasons do not align with reality, it clearly indicates that something has gone wrong.

7.4 Limitations

Our study used a Twitter feed, a "popularity" slider with instant feedback, and short-term evaluation of feed satisfaction. Future work should interrogate how our findings generalize to different contexts, including (1) different user goals, (2) settings with different types and timing of feedback, and (3) longitudinal use of systems.

First, users' goals may affect the strength of placebo effects and introduce new mechanisms of user satisfaction. In our experiments, because participants had 60 seconds to browse the feed, they likely defaulted to the browsing mode of "passing time" and general entertainment with no specific information-seeking goal. But people use news feed systems with a variety of goal-directed and undirected manners [17, 43]. As an example, instead of for entertainment, a user might want to use a popularity slider to understand what is popular and unpopular. This might introduce another placebo mechanism, where users' (untrue) understanding of what is popular and unpopular increases system satisfaction. As another example, if participants had much more than 60 seconds, our sliders may have gained large objective utility in their ability to shuffle the feed to reduce boredom.

Differences in user goals become even more pronounced with different social media platforms, as well as different content curation systems like news feeds vs shopping recommendations. For instance, some participants in Vaccaro et al. stated they used certain platforms for keeping up with close friends and other platforms for news [59]. In part, the variety of goals highlights the limitations of satisfaction of an outcome measure. We urge future researchers to measure users' experiences of accomplishing specific goals, like feeling informed vs feeling entertained.

Lastly, future work should study placebo effects in other social media UI elements, and longitudinally. This study used sliders with instant feedback, but many social media control settings do not function in this way. For example, the "See less like this" button does not give instant feedback; it alters a feed for an unknown period of time. Moreover, some settings on social media *never* provide feedback, such as a setting that deletes personal data. This raises questions as to how different amounts and delays of feedback affect users' feelings of control and their overall evaluation of their feeds, and how this evaluation changes over multiple sessions.

8 Conclusion

In this work we found that a deceptive "placebo" slider slightly increases users' evaluation of a social media feed only when they use it. User expectations that the slider will improve the feed are a promising explanation. The smaller-than-expected effect size demonstrates the need to carefully design experiments of placebo interfaces in HCI, and we have provided a scaffold for doing so. In addition, we recommend monitoring of expectations to provide additional evidence that a suspicious design unethically manipulates users. This will give regulators a clearer picture of what deceptive practices to prohibit.

Acknowledgments

A big thanks to everybody that participated in our study. Additional thanks to the people in the Social Spaces group that gave feedback on this paper. This work was supported by NSF grant CHS-1564041 and The Center for Just Infrastructures at UIUC.

References

[1] Eytan Adar, Desney S. Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1863–1872. https://doi.org/10.1145/2470654.2466246

[2] Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research* 24, 4 (2013), 956–975.

[3] Saleema Amershi et al. 2019. Guidelines for Human-AI Interaction. In *CHI '19*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Marco Annoni. 2018. Chapter Eighteen - The Ethics of Placebo Effects in Clinical Practice and Research. In *Neurobiology of the Placebo Effect Part II*, Luana Colloca (Ed.). International Review of Neurobiology, Vol. 139. Academic Press, Cambridge, MA, USA, 463–484. https://doi.org/10.1016/bs.irn.2018.07.031

[5] American Medical Association. 2001. Code of medical ethics, opinion 2.1.4: Use of Placebo in Clinical Practice. https://code-medical-ethics.ama-assn.org/ethics-opinions/use-placebo-clinical-practice Accessed June 9, 2024.

[6] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

[7] Laura Brandimarte, Alessandro Acquisti, and George Loewenstein. 2013. Misplaced Confidences: Privacy and the Control Paradox. *Social Psychological and Personality Science* 4, 3 (2013), 340–347. https://doi.org/10.1177/1948550612455931

[8] H Brignull. 2010. Dark Patterns. https://old.deceptive.design/main_page/index.html Retrieved April 25, 2023.

[9] H. Brignull, M. Leiser, C. Santos, and K. Doshi. 2023. Deceptive patterns – user interfaces designed to trick you. https://www.deceptive.design/ Retrieved April 25, 2023.

[10] Dell Cameron and Andrew Couts. 2024. The Incognito Mode Myth Has Fully Unraveled. *Wired* (2024). https://www.wired.com/story/google-chrome-incognito-mode-data-deletion-settlement/ Accessed December 11, 2024.

[11] X Help Center. 2024. Topics on X. https://help.twitter.com/en/using-x/follow-and-unfollow-topics Accessed January 25, 2024.

[12] Federal Trade Commission. 1984. In the Matter of Cliffdale Associates, Inc., et al. https://www.ftc.gov/system/files/ftc_gov/pdf/Cliffdale-Assocs-103-FTC-110.pdf 1984 WL 565319, decided March 23, 1984.

[13] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 585–592. https://doi.org/10.1145/642611.642713

[14] Frobrukerrådet (Norwegian Consumer Council). 2019. Deceived by Design: How tech companies use dark patterns to discourage us from exercising our rights to privacy. https://storage02.forbrukerradet.no/media/2018/06/2018-06-27-deceived-by-design-final.pdf

[15] Alena Denisova and Paul Cairns. 2015. The Placebo Effect in Digital Games: Phantom Perception of Adaptive Artificial Intelligence. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 23–33. https://doi.org/10.1145/2793107.2793109

[16] Alena Denisova and Eliott Cook. 2019. Power-Ups in Digital Games: The Rewarding Effect of Phantom Game Elements on Player Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY '19)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/3311350.3347173

[17] K. J. Kevin Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W. McDonald, and Amy X. Zhang. 2024. Mapping the Design Space of Teachable Social Media Feed Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 733, 20 pages. https://doi.org/10.1145/3613904.3642120

[18] Damien G Finniss, Ted J Kaptchuk, Franklin Miller, and Fabrizio Benedetti. 2010. Biological, clinical, and ethical advances of placebo effects. *The Lancet* 375, 9715 (2010), 686–695. https://doi.org/10.1016/S0140-6736(09)61706-2

[19] Jacob Frantz. 2015. Updated Controls for News Feed. https://newsroom.fb.com/news/2015/07/updated-controls-for-news-feed/ Accessed September 1, 2023.

[20] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[21] Frobrukerrådet. 2018. Deceived by design: How tech companies use dark patterns to discourage us from exercising our rights to privacy. https://storage02.forbrukerradet.no/media/2018/06/2018-06-27-deceived-by-design-final.pdf

[22] FTC. 2019. FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook. https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook Accessed June 9, 2024.

[23] Benjamin M. Galvin, Amy E. Randel, Brian J. Collins, and Russell E. Johnson. 2018. Changing the focus of locus (of control): A targeted review of the locus of control literature and agenda for future research. *Journal of Organizational Behavior* 39, 7 (2018), 820–833. https://doi.org/10.1002/job.2275

[24] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 61–73. https://doi.org/10.1145/3357236.3395486

[25] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[26] Scott Griffiths, Emily A. Harris, Grace Whitehead, Felicity Angelopoulos, Ben Stone, Wesley Grey, and Simon Dennis. 2024. Does TikTok contribute to eating disorders? A comparison of the TikTok algorithms belonging to individuals with eating disorders versus healthy controls. *Body Image* 51 (2024), 101807. https://doi.org/10.1016/j.bodyim.2024.101807

[27] Jessica Guynn. 2023. Tired of what you see in your Facebook feed? Facebook says it's going to hand you more control. https://www.usatoday.com/story/tech/2023/04/05/facebook-giving-more-control-news-feed/11608792002/ Accessed September 1, 2023.

[28] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376511

[29] David Halbhuber, Maximilian Schlenczek, Johanna Bogon, and Niels Henze. 2022. Better be quiet about it! The Effects of Phantom Latency on Experienced First-Person Shooter Players. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia* (Lisbon, Portugal) *(MUM '22)*. Association for Computing Machinery, New York, NY, USA, 172–181. https://doi.org/10.1145/3568444.3568448

[30] Silas Hsu, Kristen Vaccaro, Yin Yue, Aimee Rickman, and Karrie Karahalios. 2020. Awareness, Navigation, and Use of Feed Control Settings Online. In *Proceedings*

of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376583

[31] Sara Chandros Hull, Luana Colloca, Andrew Avins, Nancy P Gordon, Carol P Somkin, Ted J Kaptchuk, and Franklin G Miller. 2013. Patients' attitudes about the use of placebo treatments: telephone survey. *BMJ* 347 (2013). https://doi.org/10.1136/bmj.f3757

[32] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 13–21. https://doi.org/10.1145/3240323.3240358

[33] Julia Kirchhof, Liubov Petrakova, Alexandra Brinkhoff, Sven Benson, Justine Schmidt, Maike Unteroberdörster, Benjamin Wilde, Ted J. Kaptchuk, Oliver Witzke, and Manfred Schedlowski. 2018. Learned immunosuppressive placebo responses in renal transplant patients. *Proceedings of the National Academy of Sciences* 115, 16 (2018), 4223–4227. https://doi.org/10.1073/pnas.1720548115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1720548115

[34] René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[35] Agnes Mercedes Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. "AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 299, 24 pages. https://doi.org/10.1145/3613904.3642633

[36] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22 (2012), 441–504.

[37] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6, Article 56 (jan 2023), 32 pages. https://doi.org/10.1145/3529225

[38] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111 (2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[39] Hanna Levenson. 1973. Multidimensional locus of control in psychiatric patients. *Journal of consulting and clinical psychology* 41, 3 (1973), 397.

[40] Jacob Liedke and Luxuan Wang. 2023. Social Media and News Fact Sheet. *Pew Research Center* (15 11 2023). https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ Accessed April 21, 2024.

[41] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* 13, 1 (03 2021), 43–109. https://doi.org/10.1093/jla/laaa006 arXiv:https://academic.oup.com/jla/article-pdf/13/1/43/36669915/laaa006.pdf

[42] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[43] Colleen McClain, Regina Widjaya, Gonzalo Rivero, and Aaron Smith. 2021. The views and experiences of U.S. adult Twitter users. *Pew Research Center* (2021). https://www.pewresearch.org/internet/2021/11/15/1-the-views-and-experiences-of-u-s-adult-twitter-users/

[44] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. 2016. (Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking. *Proceedings on Privacy Enhancing Technologies* 2016, 2 (2016), 1–20. https://doi.org/10.1515/popets-2016-0009

[45] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 192, 15 pages. https://doi.org/10.1145/3544548.3580695

[46] Thomas W. H. Ng, Kelly L. Sorensen, and Lillian T. Eby. 2006. Locus of control at work: a meta-analysis. *Journal of Organizational Behavior* 27, 8 (2006), 1057–1087. https://doi.org/10.1002/job.416

[47] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[48] NORC. 2022. Panel Design. https://amerispeak.norc.org/us/en/amerispeak/about-amerispeak/panel-design.html Accessed April 9, 2024.

[49] Council of the European Union. 2022. The Digital Services Act (DSA) - Regulation (EU) 2022/2065, Article 25. https://www.eu-digital-services-act.com/Digital_Services_Act_Article_25.html

[50] Keith J. Petrie and Winfried Rief. 2019. Psychobiological Mechanisms of Placebo and Nocebo Effects: Pathways to Improve Treatments and Reduce Side Effects. *Annual Review of Psychology* 70, 1 (2019), 599–625. https://doi.org/10.1146/annurev-psych-010418-102907 arXiv:https://doi.org/10.1146/annurev-psych-010418-102907 PMID: 30110575.

[51] Jacopo Prisco. 2018. Illusion of control: Why the world is full of buttons that don't work. *CNN* (03 09 2018). https://www.cnn.com/style/article/placebo-buttons-design/index.html Accessed June 27, 2024.

[52] Urbano Reviglio and Claudio Agosti. 2020. Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media. *Social Media + Society* 6, 2 (2020), 2056305120915613. https://doi.org/10.1177/2056305120915613

[53] Elisa Shearer and Elizabeth Grieco. 2019. Americans are wary of the role social media sites play in delivering the news. *Pew Research Center* (2 10 2019). https://www.pewresearch.org/journalism/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news/ Accessed May 24, 2024.

[54] Baba Shiv, Ziv Carmon, and Dan Ariely. 2005. Placebo Effects of Marketing Actions: Consumers May Get What They Pay For. *Journal of Marketing Research* 42, 4 (2005), 383–393. https://doi.org/10.1509/jmkr.2005.42.4.383

[55] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Pearson, New York, NY, USA.

[56] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (feb 2018), 39–41. https://doi.org/10.1145/3180492

[57] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI Spring Symposium Series.*

[58] Kristen Vaccaro. 2024. Personal communication.

[59] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *CHI '18*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173590

[60] Barbara H. Wixom and Peter A. Todd. 2005. A Theoretical Integration of User Satisfaction and Technology Acceptance. *Information Systems Research* 16, 1 (2005), 85–102. https://doi.org/10.1287/isre.1050.0042 arXiv:https://pubsonline.informs.org/doi/pdf/10.1287/isre.1050.0042

[61] Molly Wood. 2014. OKCupid Plays With Love in User Experiments. *The New York Times* (28 07 2014). https://www.nytimes.com/2014/07/29/technology/okcupid-publishes-findings-of-user-experiments.html

[62] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark patterns in the design of games. In *Foundations of Digital Games 2013.* Society for the Advancement of the Science of Digital Games, Chania, Crete, Greece, 39–46. http://www.fdg2013.org/program/papers.html

A Feed Generation for Non-Twitter Users

To designate accounts for each topic we relied on Twitter's topic-following feature [11], which displays a feed of tweets relevant only to a selected topic. The first author manually inspected accounts appearing in these topic-restricted feeds, adding accounts that consistently tweeted on-topic content to the corresponding account lists. They browsed each feed until they deemed it evident that further browsing would add few new accounts. In the weeks before the study they occasionally revisited the topic-restricted feeds to capture accounts that tweeted less frequently.

For each topic, we attempted a download of no more than 680 tweets, a number chosen to exceed the feed sizes used during the study by a reasonable margin but also to avoid consuming too much of our Twitter API download limits. Each account within a topic was allocated an equal proportion of the 680 tweet limit; i.e. the download limit for each account was $680/N$ where N equaled the total number of accounts in its topic. We downloaded each account's most recent tweets up to its limit. The Twitter API often returned fewer tweets than the download limit, especially for accounts that tweeted infrequently.

During the study, the feed software sampled a total of 400 tweets from the participant's selected topics to build their feed. The sampling process iterated through each topic, and then through each

account in that topic, selecting a random tweet without replacement from that account. This strategy ensured that accounts that tweeted more than others would not dominate the feed, and that topics as well as accounts within a topic were represented in roughly equal proportions.

Although we initially set up a server to automatically update the tweet pool once per day, the server crashed six days into the study. This remained unnoticed until after the study's end, five weeks later. Thus, approximately 90% of participants that viewed our generated feeds were viewing tweets sampled from the same tweet pool.

Algorithm 1 Sampling process for building a participant's feed

```
loop 400 times
    select the next topic among the user's selected topics
    select the next account within the selected topic
    randomly sample a tweet from the selected account without replacement
    add the tweet to the feed
end loop
sort the feed according to experimental treatment
```

B Additional Figures

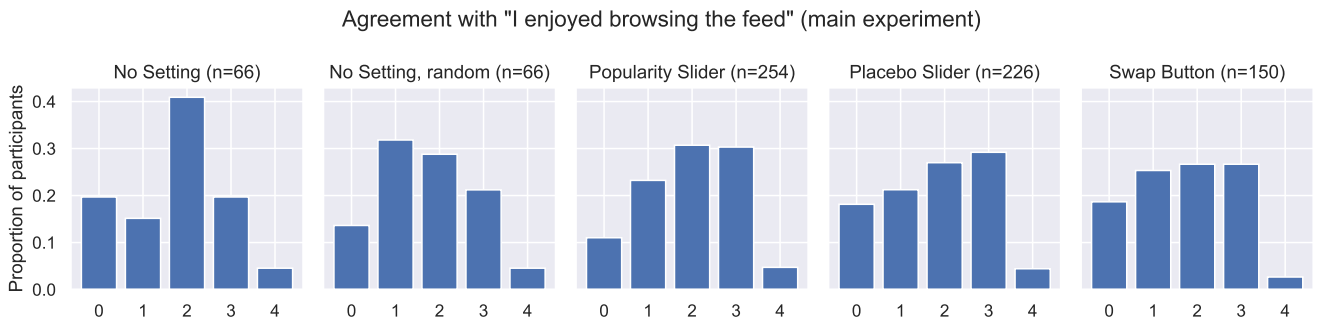

Figure 8: Participants' satisfaction ratings grouped by experimental treatment. Ratings are as follows – 1: Strongly disagree, 2: Somewhat disagree, 3: Neither agree nor disagree, 4: Somewhat agree, 5: Strongly agree.

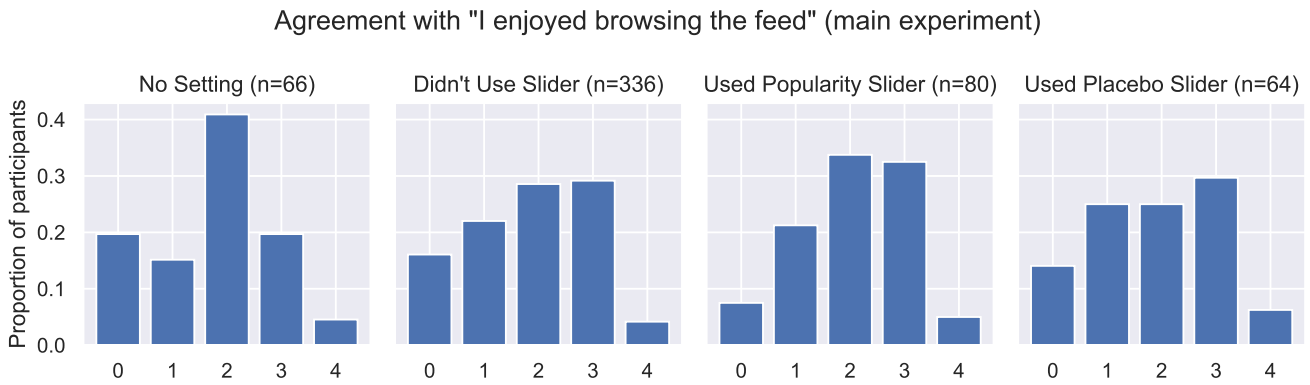

Figure 9: Participants' satisfaction ratings by setting usage. Ratings are as follows – 1: Strongly disagree, 2: Somewhat disagree, 3: Neither agree nor disagree, 4: Somewhat agree, 5: Strongly agree.

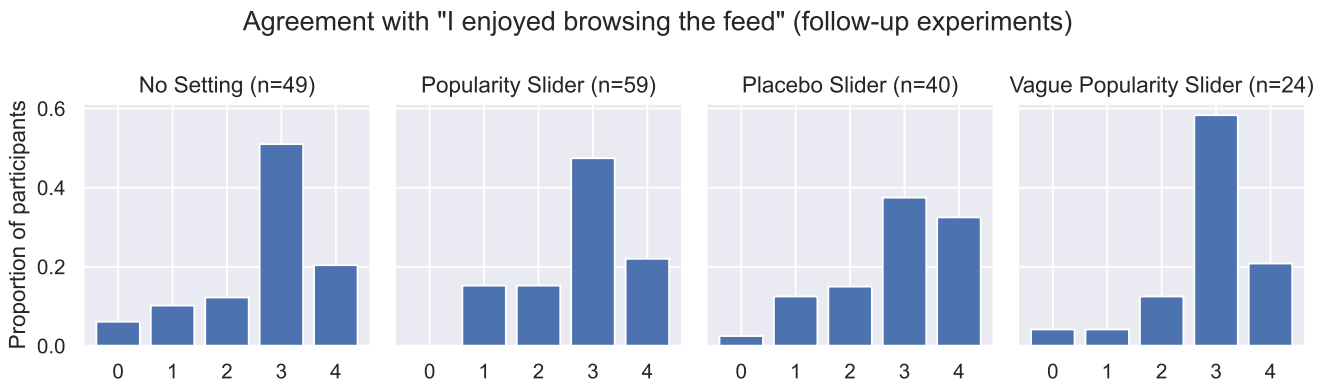

Figure 10: Participants' satisfaction ratings grouped by experimental treatment. The No Setting, randomized feed treatment is omitted because it only had 5 ratings in this dataset. Ratings are as follows – 1: Strongly disagree, 2: Somewhat disagree, 3: Neither agree nor disagree, 4: Somewhat agree, 5: Strongly agree.