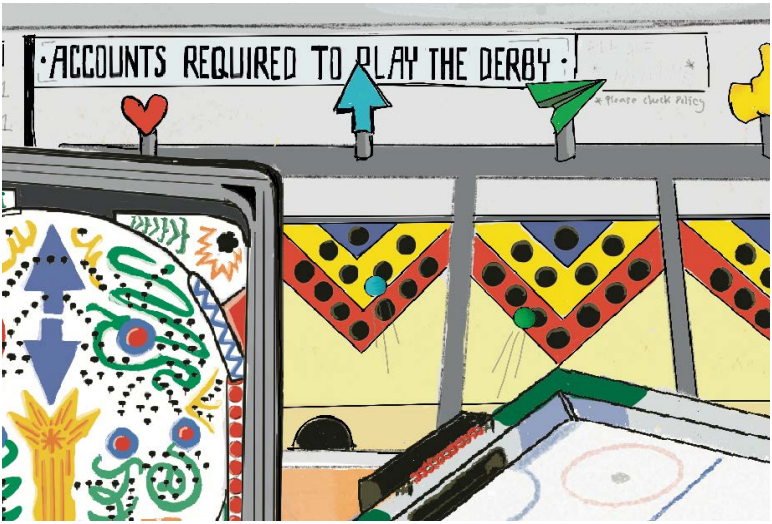
OPTIMIZING FOR WHAT?
ALGORITHMIC AMPLIFICATION AND SOCIETY



Teachable Agents for End-User Empowerment in Personalized Feed Curation

By Kevin Feng, David McDonald, and Amy Zhang

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In April 2023, the Knight Institute hosted a symposium, "Optimizing for What? Algorithmic Amplification and Society," to explore how law and policy regulates or should regulate the algorithms that power social media platforms' recommendation systems and the challenges and opportunities that arise from these systems. The symposium was organized in partnership with the Institute's 2022-2023 Visiting Senior Research Scientist Arvind Narayanan and took place at Columbia University.

The essays in this series were originally presented and discussed at this event. Written by leading scholars, nonprofit leaders, technologists, and professionals from a wide range of disciplines, including computer science, psychology, law, social science, philosophy and other fields, these essays offer more precise explanations of how algorithms are designed, deployed, and evaluated and propose interventions that would mitigate some of the harms caused by amplification. This series also engaged with normative questions about algorithmic recommenders: What should they optimize for? How do we design systems to promote healthier public discourse? Who makes these decisions?

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, executive director; Katy Glenn Bass, research director; Arvind Narayanan, visiting senior research scientist; Alex Abdo, litigation director; and Larry Siems, chief of staff. The essay series was edited by Glenn Bass and Narayanan with additional support from Lorraine Kenny, communications director; Victoria Tang, legal research fellow; Mia Speier, research coordinator; Kushal Dev, research fellow; Avian Muñoz, intern; and Stephen Dai, intern.

The full series is available at knightcolumbia.org/research/

ABSTRACT

A small handful of platforms function as the online social architects of our age, molding user norms and behaviors through algorithmically curated feeds and interface affordances alike. In their hands, the risks of curatorial centralization come into focus—top-down, platform-wide policies deprive users of their sense of agency, fail to consider nuanced experiences, and ultimately marginalize users and communities with unique values and customs. Previous research has demonstrated that users strive to reclaim their agency by developing "algorithmic folk theories" to probe black-box feed curation algorithms and by "teaching" algorithms to generate more satisfactory content through strategic interactions with their feeds. Given this, we ask: How can users' inherent teaching abilities be more explicitly utilized to empower personalized curation in online social settings? Drawing inspiration from the interactive machine teaching paradigm, we delve into user-teachable agents for feed curation. To achieve this, we initiated a formative study to comprehend how users would approach the task of explicitly teaching an algorithmic agent about their preferences in social media feeds, as opposed to the agent passively learning them. Based on our findings, we propose in-feed affordances that enable users to engage in a teaching loop by 1) articulating content preferences through examples of posts to a learnable agent, 2) evaluating the agent's effectiveness in learning, and 3) iteratively formulating a curriculum of teaching goals and examples. We conclude with a discussion of challenges and next steps, focusing on how our approach could be employed to better align the incentives of users and platforms within sociotechnical systems.

1. INTRODUCTION

ALGORITHMICALLY CURATED FEEDS lay the groundwork for our daily online social interactions. Their interface designs and affordances define the boundaries of our engagement with content, while their algorithms determine not only *what* we engage with but also with *whom*. Together, these factors influence our social norms, behaviors, and mental models within the online realm. The rise of algorithmic content curation and distribution coincided with a significant transformation in the ownership structure of social networks. Over the past two decades, social media platforms have shifted from being hosted on numerous distributed, independent servers to being controlled by a handful of private corporations that reach millions of users worldwide [1]. This process has distilled the intricacies and diversity of our social realities into a limited set of standardized parameters. We term this concentration of algorithmic curation capabilities within the hands of a select few as *curatorial centralization*.

The homogenization resulting from curatorial centralization does not necessarily occur at the content level. In fact, recommender systems have become remarkably skilled at swiftly and accurately personalizing content [2]. Instead, homogeneity becomes evident in *the approach* used for

personalizing feeds. In 2021, the Wall Street Journal conducted an investigation into the TikTok algorithm. They employed over 100 automated bot accounts, each assigned specific interests, to "watch" TikTok videos [3]. The investigation revealed that these bots were swiftly directed into specialized content rabbit holes based solely on one metric: watch time. Although social media recommendation algorithms consider an array of variables when determining what to suggest next [4, 5], this example empirically highlights the significant emphasis placed on engagement.

Curatorial centralization gives rise to substantial adverse effects on users' perception of agency and the broader freedom of expression. In the present landscape, platforms often employ a top-down, one-size-fits-all strategy in platform design and governance, providing limited opportunities for input from users themselves. This approach marginalizes individuals who do not conform to the prevailing platform-wide norms. For example, user groups that employ culturally significant language patterns might find their content wrongly labeled and "downranked" by platforms [6, 7], while users with neurodiverse traits could discover many posts overwhelming to consume [8]. Even users who align with the majority might become frustrated due to encountering irrelevant content, without an efficient means to control or eliminate such content from their feeds. The emergence of symptoms like doomscrolling, dissociation [9], and affective polarization [10] in the context of today's engagement-driven feeds serves as a clear indication of diminishing agency.

Reviglio and Agosti [11] coined the term *algorithmic sovereignty* to denote the ethical entitlement of individuals to exercise exclusive control over instances of algorithmic personalization that affect them. User agency, a concept extensively explored in human-computer interaction (HCI) literature as a "fundamental human need" [12] and "basic psychological need" [20], aligns closely with algorithmic sovereignty. Both these concepts encapsulate essential aspects of user empowerment. To ensure the fulfillment of this need, whether it's algorithmic sovereignty or user agency, a transition of curatorial power towards users is imperative. Empowering users through personalized curation allows them to tailor platform algorithms and experiences to their preferences, and to truly own their online social spaces. Nevertheless, a couple of pivotal questions remain: What specific forms of curatorial power

are users seeking, and how can we effectively grant them this power through in-feed interfaces? These questions constitute the focus of our research.

We start in Section 2 by examining current research that delves into user agency within the sphere of social media algorithms. Additionally, we introduce an interaction framework that shapes our design decisions moving forward. In Section 3, we present a comprehensive overview of insights gleaned from our empirical study involving 24 social media users spanning 4 platforms. These insights culminate in two provocative metaphors in Section 4, which serve as the foundation for the four in-feed design affordances we propose in Section 5. In Section 6, we engage in a discussion about the reasons for the current appropriateness of reconsidering feed design. Our focus is specifically directed towards the exploratory potential inherent in open-source alternative social media platforms, alongside an exploration of how to enhance the incentive alignment between major for-profit platforms and their users.

2. RECLAIMING USER AGENCY: WHERE WE STAND NOW

RESEARCHERS AND PRACTITIONERS commonly employ the term "black-box" to characterize algorithmic behaviors that are overly intricate and nontransparent for the average user to grasp. Within the HCI and social science literature, social media recommendation algorithms are frequently denoted as black-boxes [11, 14-17]. However, there exists an opposing argument suggesting that there is an ample existing understanding of these algorithms, and if this knowledge becomes widely accessible, they might cease to retain their black-box status [18]. Indeed, in late March 2023, Twitter released its For You feed algorithm on GitHub along with a high-level explanation [19]. Regardless of whether algorithms truly merit the label of black boxes, one aspect remains evident: the proliferation of engagement-driven content curation bereft of algorithmic choice can erode user agency.

But what precisely do we mean by user agency? To precisely define this concept, we draw upon the research of Bennett et al. [20], who conducted an

extensive review of over 30 years of HCI research on autonomy and agency (from 1990 to 2022) to discern contemporary interpretations of these terms. We concur with their identification of the key attributes of agency, which encompass 1) the availability of choices aligned with user preferences and values, 2) the capability to pursue those choices, and 3) the maintenance of a sense of control.1 It is worth noting that a black-box feed curation algorithm can undermine agency by obscuring the available choices and sense of control from users. On the other hand, a fully transparent "glass-box" algorithm can also diminish agency—despite offering a clear view of its internal mechanisms, users might still find themselves restricted to mere observation, lacking opportunities for choice or control.

2.1. Algorithmic folk theories on social media

The intersection of black-box algorithms and diminishing agency has prompted users to formulate "algorithmic folk theories" to comprehend their interactions and endeavor to regain some of the agency they have lost. Folk theories are informal conceptualizations about the world, often shaped through empirical experiences. For instance, a Twitter user might reason, *"if I 'like' this post, I'm telling the algorithm that I'm interested in similar content, and the algorithm will show me more of such content in the future."* While a user's folk theory might not precisely align with the technical intricacies of a system, it can still wield influence over their conduct and self-perception on the platform, thereby exerting an impact on the platform's behavior as a whole [21, 22].

The conceptualization of folk theories and the inherent nature of these theories offer valuable insights into how users approach algorithms within sociotechnical contexts. Siles et al. discovered that Spotify users either anthropomorphized the algorithm as a social entity that provides recommendations in exchange for user data or perceived it as a purely mechanical system that tailors experiences based on user training [23]. Examining tweets under the hashtag #RIPTwitter in response to Twitter's introduction of the engagement-based timeline in 2016, DeVito et al. [24] found strong negative reactions due to theories that the algorithm would transform the platform into a "popularity contest" saturated with "only popular Tweets, ads, and promoted accounts." Eslami et al. demonstrated that Facebook

users could easily formulate an array of folk theories from their News Feed, particularly if they were aware of the presence of an algorithmic curation agent [21]. However, this ability hinges on the "seams" in the interface that visibly expose facets of algorithmic operations. On the other hand, interfaces that are "seamless" and abstract or conceal the technology make it considerably more challenging, if not impossible, for folk theories to emerge [25]. Consequently, the concept of "seamful design" [26] has emerged as a critical factor in facilitating folk theories and cultivating user agency through the mechanism of "algorithmic resistance," [27] which arises when user and platform interests diverge.

Having crafted their theories, users then translated these into action by "teaching" the algorithm their preferences through strategic interactions within their feeds. On Facebook, users frequently visited the profile pages of individuals they wished to see more content from, actively tagging them in posts to capture the algorithm's attention [21]. On TikTok, users took measures to boost a video's visibility on others' For You pages; this involved repeatedly watching videos (even if they understood them fully on the first viewing), engaging more with videos featuring specific hashtags, leaving extended comments, and repeatedly hitting the like and share buttons [22]. On Twitter, users consistently liked or retweeted specific types of content while intentionally omitting or misspelling particular words in their posts to avoid the algorithm's detection [28]. Nonetheless, users retained an element of uncertainty regarding the effectiveness of these techniques [21, 28].

Although folk theories empowered users to develop teaching strategies, it is important to note that these strategies may not invariably lead to authentic or genuine interactions. Eslami et al. highlighted in their Facebook study that users perceived their own teaching behaviors as manipulative and coerced [21]. Teaching is an inherent human capacity that naturally transpires in various aspects of our lives; it should never come across as contrived. How can we effectively harness our familiar mental frameworks of teaching to convey desired concepts to a receptive algorithmic agent? To address this, we pivot towards the paradigm of interactive machine teaching (IMT).

2.2. Interactive machine teaching

IMT constitutes an interaction framework whereby domain

experts—individuals who might not possess expertise in machine learning (ML)—can utilize their own domain knowledge to instruct ML models [29]. Conventional ML focuses on developing algorithms that autonomously derive conceptual representations from training data. In contrast, IMT contends that the acquirable representations should stem directly from human knowledge. This approach imparts users with a more solid understanding of what the model is learning, resulting in more transparent and diagnosable models. It also diminishes the potential for the model to acquire irrelevant or undesirable representations, such as societal biases. IMT operates as an iterative process, encompassing three primary components:

1. **Planning:** The teacher identifies a teaching task along with a curriculum, which entails a collection of examples and representations designed to facilitate the model's instruction. These examples usually take the form of a compact dataset.

2. **Explaining:** The teacher shows the learning agent examples, explicitly pinpointing the concepts the agent should absorb.

3. **Reviewing:** The teacher offers the model the opportunity to predict previously unseen instances. Any erroneous predictions are corrected by the teacher, and the teaching strategy or curriculum is adjusted in response.

Fundamental to IMT is the concept of the *teaching language*, which serves as an interface for instructors to craft expressive representations capable of conveying desired concepts and being comprehensible by algorithmic agents. For example, in the context of Pearl [30], an IMT tool designed to help users analyze personal calendar data for workplace well-being, conceivable teaching languages encompass labels that users can attach to time blocks on their calendar and rules (e.g., using specific keyboard strokes after 6pm to signify working after regular hours). To engage with a teaching language, instructors must undertake a process known as *knowledge decomposition*. Ng et al. [31] define knowledge decomposition as the "process of identifying and expressing useful knowledge by breaking it down into its constituent parts

or relationships." In the context of workplace well-being, for instance, the teacher might hold an overarching objective of *improving work–life balance*, which can be decomposed into a relationship between *keyboard strokes after 6 p.m.* and the act of *working after hours* [30].

How can this perspective contribute to end-user empowerment in the curation of social media feeds? Although not explicitly detailed in previous literature, users might inadvertently carry out knowledge decomposition while formulating folk theories. However, a gap still persists between their disassembled preferences and the actual implementation of those preferences by an algorithm within the feed. We delve into how to bridge this gap in the subsequent sections.

3. EMPIRICAL TAKEAWAYS

While IMT demonstrates potential in personalized feed curation contexts, it would be hasty to embark on the design and development of a teachable curation agent without a clear understanding of what and how users would teach it. To gain a better understanding, we drew inspiration from techniques in knowledge decomposition to design an interactive study that utilized content from a participant's own feed.

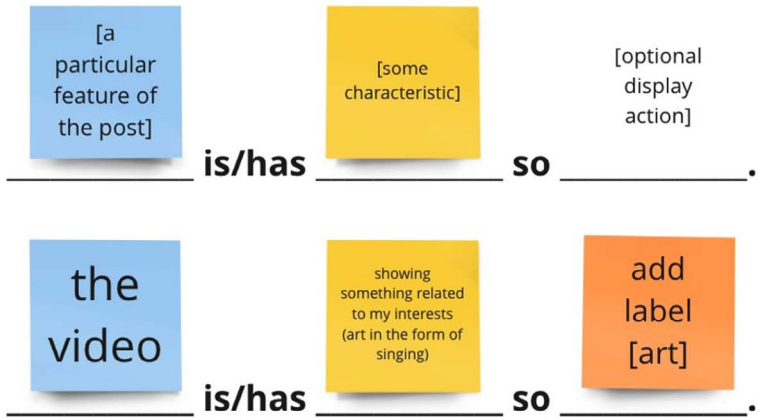
In this section, we present an overview of our study, along with emergent insights from an exploratory analysis of our data.

3.1. Study structure

To formalize the process of knowledge decomposition, we established an information unit, which we call a *signal* (see Figure 1). A signal consists of two primary components: a **feature** and a **characteristic**. A **feature** refers to an objective element present within a post. Illustrative examples of features include the post's author, textual content, image(s), number of likes, and the post's creation time. In contrast, a **characteristic** constitutes a subjective statement elucidating the significance of a feature. This subjectivity arises due to variations in significance across different participants. For instance, a characteristic like "someone I know from in-person interactions" describes

the feature "the post's author." A third, optional element known as an **action** represents a participant's potential response to a feature-characteristic pairing. For example, if the post author is an individual the participant recognizes through in-person interactions, they might opt to execute an action such as "trigger a notification for this author's posts."

Figure 1: Examples of signals from our study. On top is a default signal template that we provide to users, and on the bottom is one completed by a participant.



We conducted studies involving 24 social media users across four platforms: Instagram, Mastodon, TikTok, and Twitter. Our participants were evenly allocated to each of these platforms based on their indication of using the platform at least a few times a month. We requested them to provide screenshots of the first 10 posts they encountered in their "home" feed, including any encountered ads, while excluding posts they were uncomfortable sharing.[1]

1 On Instagram and Mastodon, this refers to the main feed users see when they log in. For TikTok, we equated the home feed with the For You page. For Twitter, we equated it with the Following feed, as Twitter was just starting to roll out its For You feed as we were conducting our study.

Subsequently, participants were invited to a Miro board,[2] where the 10 submitted screenshots were uploaded. They were then tasked with arranging these screenshots as they would prefer them to appear in an ideal, imaginary feed. Furthermore, we asked participants to compose at least one "signal" for each post, structurally assessing the content's value.[3] Many participants crafted multiple signals for a single post if they relied on more than one feature and/or characteristic. Optionally, they could also attach actions to their feature-characteristic pairs. Following this activity, we conducted brief interviews with participants. These interviews encompassed discussions about their most noteworthy signals, their objectives on the studied social media platform, and their general experiences with that platform's feed.

Our study underwent review and received approval from the University of Washington Institutional Review Board. All participants fell within the age range of 18 to 65 and were compensated with a $20 USD gift card upon completing the study. These studies were conducted virtually via Zoom between January and April 2023. Recordings were made of the sessions and subsequently transcribed.

In the subsequent section, we delve into several emerging themes that surfaced during the analysis of our study data.

3.2. Users prioritized people, platforms prioritized content

Social media, at least in its early forms, revolved around the concept of connecting with other individuals. However, over time, platforms have undergone a shift towards content-based recommendations, primarily due to their heightened efficacy in driving user engagement, particularly among younger demographics [32]. This evolution has not escaped the notice of participants in our study. Many of them articulated a longing to curate accounts belonging to individuals they have a genuine interest in, with a preference for featuring posts from these accounts prominently at the top of their feeds. Participant

2 Miro is an interactive and collaborative virtual whiteboarding tool.

3 We define value broadly here based on what the participant wants to see before others. Content that they want to see first has the highest value.

6 (P6), in particular, highlighted the idea of having a dedicated section that showcases *"my favorite people highlighted at the top,"* followed by a strictly chronological feed. Notably, the majority of participants underscored the significance of the post's author as the most crucial feature to them.

This people-centric mental framework, wherein individuals prioritize content from those they care about, often clashes with the content-centric strategies pursued by the platforms. The most striking illustration of this tension was observed on Instagram. Among the six Instagram users in our study, four of them encountered no content from close friends or acquaintances within the initial 10 posts of their feeds. Most notably, P9 became aware of this behavior *during* the course of the study, and this realization evoked a sense of aversion:

I don't have any posts here from, I just noticed, my friends which used to be a thing I really liked about social media, is just finding out what my friends are up to and feeling connected. Wow! What a horrible development. How have I only just noticed? (P9)

P9's quote sheds light on the occurrence of a frog-boiling effect within platforms and reveals that, despite the differing goals of users and platforms, users may not readily notice the gradual decline in content from individuals they deeply care about. Indeed, platforms can implement their changes in a gradual and strategic manner, often evading immediate detection.

3.3. Use of proactive curation mechanisms rarely persisted

Platforms offer existing mechanisms that provide users with some (albeit limited) degree of control over what they see on their feeds. These mechanisms can be proactive or reactive. We consider a mechanism *proactive* if users first explicitly specify their preferences and then allow those preferences to inform what types of content can appear in their feeds. An example of this is the ability to create lists on Mastodon and Twitter. On the other hand, *reactive* mechanisms allow users to take action *in response to* content they are not interested in. The mute function available on many platforms is an example of such a mechanism.

Some participants extensively utilized reactive mechanisms. For instance, P3 had muted approximately two-thirds of the accounts they follow

on Instagram. Similarly, on Instagram, P12 chose to snooze any suggested posts they encountered. Many participants also experimented with proactive mechanisms but eventually discontinued their use. When we probed participants who were once users of lists on Twitter and Mastodon, the main factors that emerged were the initial effort required for setup and subsequent maintenance. P2 expressed a wish that they had been aware of lists before following a larger number of people, as sifting through hundreds of accounts to construct a satisfactory list can become overwhelming. They remarked: *"some of these people don't post, some are muted, but I have to go through my follows [to find out]."* In contrast, P16 encountered a different issue when transitioning from Twitter to Mastodon: *"I don't really have a good idea of what to put in each list. I don't have enough people and content going through it."* P14 attempted to create a list for one of their fandoms on Twitter but soon realized that the creators either *"either moved on [from creating desirable content] or add nothing of what I followed them initially for."*

Despite their relatively shorter lifespans, proactive mechanisms were deemed valuable by participants. P15, for instance, utilized lists to organize "low-velocity posters" of interest—individuals who infrequently post but contribute pertinent content when they do. By placing them in lists, these individuals are prevented from being overshadowed by frequent posters in the standard feed. In essence, these mechanisms act as countermeasures against default algorithmic behavior when those behaviors do not align with user preferences.

3.4. Retrieving archived content was a common yet poorly-supported practice

Participants archived, or "saved," content from their platforms for various reasons: tracking articles to read after work, recipes to try cooking, job opportunities for the future, and more. Despite this, many platforms offer limited tools for users to curate their archived content. It's important to differentiate between *allowing a user to archive content* and allowing them to *curate their archive*, making it easy for them to retrieve specific content. Most platforms provide options for archiving content, such as bookmarks (Twitter and Mastodon) and collections (Instagram and TikTok). However, the process of retrieving archived content presents its own challenges. P8

mentioned that they would occasionally send content to another person, like their husband, as a way of archiving, thus avoiding the need to search for the content within their extensive archives. P7 vaguely recalled saving a post a few weeks ago but conceded to saving an excessive number of posts. They expressed a desire for the ability to search for a saved post using the author's name or specific keywords associated with the post.

We briefly introduced the concept of "curating archived content" with the perspective that an archive can essentially serve as an independent feed. P19 saw this potential, stating, *"it's like a separate news feed, and if an algorithm could help and start tagging [the content], that would be amazing."* While the precise role of algorithms in this context remains uncertain, we acknowledge that this curated "feed" of archived content constitutes a compilation that underwent thoughtful selection. This compilation can be employed to shape user preferences, functioning akin to a curriculum within the context of IMT. We delve into the implications of this design in Section 5.

3.5. Agency can be a double-edged sword

In the field of HCI, an increase in user agency is often regarded as a positive attribute [33]. Numerous participants conveyed a desire for greater agency, indicating a wish to wield more control in fine-tuning their feed algorithms to match their preferences or even eliminating engagement-based interventions entirely. Some harbored this sentiment due to a lack of trust in the algorithm. For instance, P16 expressed, *"I'm a bit of a perfectionist. [With chronological] order, I know that I'll get everything."* Regarding the manner in which they would customize their feeds, both P15 and P19 extensively employed labels to cultivate a more focused feed. P19 even applied labels to every piece of content during the activity, aiming to surface content for later use. P13 wished for a distinct method to explicitly dismiss a video on TikTok, rather than implicitly signaling dismissal by scrolling away. P4, who also used TikTok, proposed a hybrid approach involving both human and algorithmic efforts. In their concept, the platform would automatically label content with topic categories and prompt the user to confirm their continued interest in those categories.

However, numerous participants also suggested that enhancing agency might occasionally prove detrimental to the user experience. Such situations

emerge when users engage with social media purely for entertainment, a sentiment commonly expressed by TikTok participants. This aligns with prior research underscoring the arduous nature of personalized content moderation [34]. A handful of participants also indicated that social media serves as a much-needed respite or diversion during brief moments of idle time throughout the day. In these instances, instead of investing mental effort into a focused and purposeful social media encounter, participants are content with simply allowing the algorithm to amuse them. Imposing additional curation demands could potentially compromise the user experience by compelling users to actively participate in attentive tasks when they may prefer not to. In essence, despite the negative connotations associated with the term, "mindless scrolling" can occasionally manifest as a liberating and even cherished interaction.

Our key insight here is that enabling users to participate in personalized curation might not be suitable for every situation or consistently. We move forward with this consideration, recognizing that our designs are not intended to supplant every facet of the social media encounter. Rather, they could inhabit a "focus" mode, catering to deliberate interactions and valuing intentionality, while users retain the freedom to exit this mode at their discretion.

4. TWO PROVOCATIVE METAPHORS

OUR STUDY DREW INSPIRATION from knowledge decomposition and the broader framework of IMT. With our empirical insights in hand, we now pose the question: How can we effectively put this framework into practice to provide more tangible guidance for the design of teachable feed curation agents tailored for feed curation?

One potential initial approach could involve directly adapting IMT interfaces proposed in prior work (e.g., PICL [29]) to explore their potential application within the realm of social media. Nevertheless, our study served as a cautionary reminder that previous teaching languages may not be as

effective in this context.[4] Existing IMT systems typically demand substantial user effort to generate meaningful labels and descriptions from data examples. For instance, in the case of building a food recipe classifier using PICL, a user might designate the keyword "recipe" as a significant signal. However, since this term also appears in various non-food contexts (e.g., "recipe for success"), the user would need to locate and label other instances that encompass the diverse contexts in which "recipe" is used. While this effort-reward dynamic might be acceptable for tasks like recipe classification, it might not hold the same appeal for lower-effort, swift-paced content consumption commonly found on social media.

Before we outline the affordances that shape this teaching approach, it is crucial to recognize that contemporary feeds are well suited for implicit learning, favoring rapid consumption as mentioned earlier. This contrasts with the iterative teaching loop central to IMT, which encompasses essential teaching actions like curriculum development and learning assessment. The prevailing structure of today's feeds may restrict the potential impact of introducing new affordances, as these may not align with the existing flow.

In this context, we contend that reimagining metaphors for feeds becomes a pivotal step in navigating the balance between consumption speed and teaching efficacy. We regard these metaphors as thought-provoking and provocative lenses through which we can envision novel possibilities for the design of teachable feed curation agents. It is worth noting that these metaphors are intended to complement and enhance our proposed affordances, while the affordances themselves can still operate independently of the metaphors.

Here we introduce our two provocative metaphors for social media feeds: a notifications panel and a playlist.

4.1. The feed as a notifications panel

In a conventional feed, the primary interaction a user engages in with content is often one of disposal. Given the continuous influx of new content within

4 Recall that a teaching language is an interface by which the teacher can expressively communicate desired concepts for an algorithmic agent to learn.

the feed, any unarchived content is prone to swift burial at an unspecified depth, joining the multitude of other posts that may never resurface for human attention. This act of disposal is facilitated by a ubiquitous design pattern known as the infinite scroll. Setting aside the well-documented addictive nature of feeds that scroll infinitely, these feeds also elevate the opportunity cost of thoughtfully lingering on a specific piece of content, as there is perpetually more potentially captivating material waiting to be explored.

In stark contrast to the concept of an infinitely scrolling feed is that of a notifications panel. While notification panels take various forms in apps and operating systems, we specifically employ the notion of panels containing dismissible notifications in our metaphor. These notifications are akin to the ones visible at the top of Android or iOS interfaces. We underscore the ability to dismiss notifications, as it results in a deliberately finite stream of information. When notifications appear on our devices—whether they pertain to weather updates, text messages, or social media alerts—we are prompted to undertake one of two actions: either respond or dismiss.

The implicit objective is to eventually achieve an empty panel by addressing every notification. This intrinsic nature of notifications to induce action is so powerful that they have been declared to "hijack" our minds [35], prompting the development of various digital well-being tools aimed at managing and muting notifications [36, 37]. However, rather than allowing notifications to commandeer even more of our attention, we utilize this metaphor to redirect attention and facilitate *more intentional, goal-oriented interactions* with content in *purposefully terminating feeds*.

What might the mechanics of such a feed look like? Initially, to ensure the feed's definitively finite nature, the user designates a numerical value corresponding to the number of posts they wish to have concurrently present in the feed. As a fresh post is introduced into the feed, the user is presented with choices: They can opt to engage with it (such as archiving or liking), dismiss it, or let it remain untouched. Engagement serves as a positive signal, indicating interest or relevance; dismissal functions as a negative signal; and idleness signifies a neutral stance. Similar to a notifications panel, the implicit objective is to clear all content from the primary feed area. Furthermore, users might have the option, upon reaching the end of

the feed, to clear all remaining posts. This introduces what could arguably be considered the most pivotal attribute of this metaphor—an invaluable prospect to incorporate meaningful interactions at the end of a feed. We explore some sample interactions in Section 5.

4.2. The feed as a playlist

As discussed in Section 3.4, increasing user access to archives and refining archive curation represent interconnected yet distinct challenges. We now draw inspiration from the playlist feature found in recommendation-powered streaming services like Spotify to tackle the latter.

When curating a playlist on Spotify, users frequently alternate between leveraging existing preferences and exploring novel ones. For instance, they might include tracks they have previously heard and enjoyed, or directly add songs from albums of artists they know. Once a playlist is "seeded," Spotify begins suggesting song additions for the same playlist. Moreover, Spotify generates a "Discover Weekly" playlist [38] after users have spent additional time on the platform, enabling them to venture into new artists and genres beyond their immediate interests. Amidst this backdrop of recommendations, users maintain the ability to continually incorporate fresh tracks from familiar artists, often sorting content from an existing playlist into multiple smaller playlists.

Many of these playlist curation interactions can be adapted for the curation of a social media archive. Users commence by saving content they find appealing from their initial feed, which assists them in identifying authors or content themes they wish to follow more closely. Subsequently, they may partition their archive into meaningful categories and populate them accordingly. Recommendations for content that aligns with specific archive categories may then be provided, facilitated by a "mini-feed" at the conclusion of a category's content. The concept of a "main feed" parallels Spotify's Discover Weekly playlist: It empowers users to explore new authors or content topics they may not have otherwise encountered.

Overall, this metaphor lays the groundwork for a multiplicity of feeds to collectively function as dual sources of appreciation—achieved through the curation of their respective archives—and exploration.

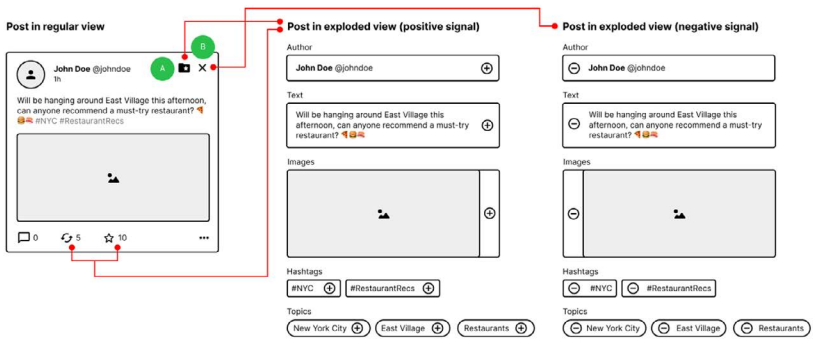
5. AFFORDANCES FOR IN-FEED TEACHING

WE NOW INTRODUCE DESIGN affordances within the feed that facilitate users' engagement in an iterative teaching loop encompassing three stages: 1) elucidating content preferences to an algorithmic agent using examples, 2) assessing the agent's proficiency in comprehending the elucidated concepts, and 3) managing a curriculum of objectives and instances for concept elucidation. These affordances have been devised with a focus on Twitter and Mastodon-style posts as a starting point. While we acknowledge that this might narrow their suitability to other platforms, we aspire for them to provide a conceptual illustration on a broader scale.

5.1. Explaining preferences via exploded views

In 3D diagramming, an exploded view is a technique that presents individual components of an object positioned slightly apart from each other. This arrangement aids the viewer in comprehending the relationships between these components and, when applicable, the sequence of assembly. We adopt a similar concept to function as a teaching language, allowing us to deconstruct the card user interface (UI) of a post to achieve a more

Figure 2: A user interface card for a social media post (left) along with exploded views showcasing positive (center) and negative (right) signals. A: option to send the post to the user's archive. B: option to dismiss the post from the feed. Note that the topic is automatically extracted from the post's textual content, leading to the rounded interface elements.



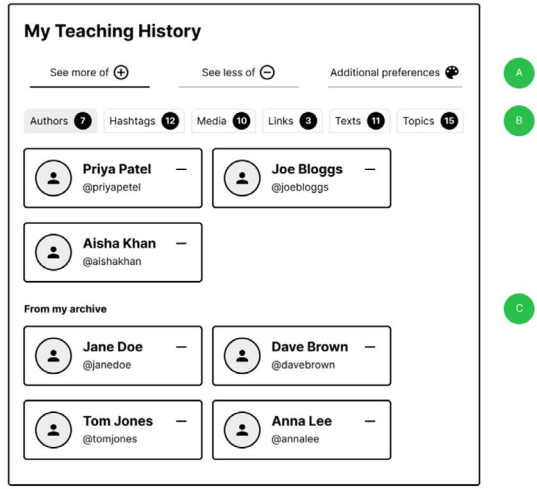
detailed expression of preferences. Using Tweets as an example, when a user expresses approval by "liking" a Tweet, the specific element(s) of the Tweet that prompted this positive response remain ambiguous within the existing framework. It remains uncertain whether they appreciated a humorous remark in the text, engaged with one of the mentioned hashtags, or simply endorsed all content from the particular author.

An exploded view of a card UI serves the purpose of unearthing this information. After a user registers either a positive (e.g., like) or negative (e.g., dismiss) signal in response to a post, the UI expands into distinct components, which may include the author, text, attached media, and hashtags.

Furthermore, supplementary components might be algorithmically extracted if they hold potential utility for the user-illustrated by the inclusion of "topics" as an example (see Figure 2). Subsequently, the user can teach the agent by selecting the components that are of interest or disinterest, as applicable. We devised the exploded views for positive and negative signals to exhibit subtle distinctions, ensuring users can readily discern between instances of imparting positive examples (those they wish to encounter more frequently) and negative ones (those they desire to encounter less frequently).

All of these interactions are to be documented within the user's teaching history. Maintaining a record of their instructional engagements is vital, given that user preferences—and subsequently, teaching objectives—can undergo frequent fluctuations. We present a method of reviewing teaching activities in Figure 3. We organize the teaching history based on the teaching method at the highest level (positive signal, negative signal, and any additional specified preferences). For positive and negative signals, we additionally organize the history according to the specific regions of the exploded post UI (see Figure 2) where teaching was conducted.

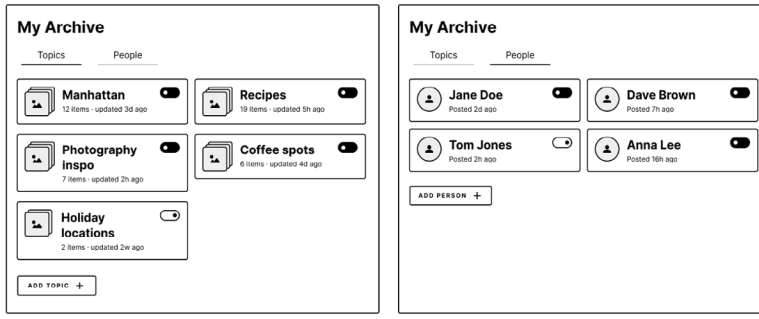
Figure 3: A view of a user's teaching history. Interactions with the exploded card view are recorded here. A: top-level tabs allow the user to review teaching based on positive or negative signals, along with any natural language teaching instructions. B: lower-level tabs offer filtering options based on specific areas of the exploded card. Items can be removed using (-) in the top right corner. C: relevant entries from the user's archive are automatically included as positive signals within the teaching history.



5.2. The archive as a teaching curriculum

We used our playlist metaphor to uncover the potential of archival in cultivating more meaningful engagement with social media. We now showcase what a user's archive might resemble in Figure 4. An archive consists of individual "collections" of content, which users can create and add to. Each collection can serve as a basis for instructing the algorithmic agent on whether analogous content should be featured in the user's primary feed as a set of instructional examples. However, the user retains the ability to make a collection "invisible" to the agent by toggling it off.

Figure 4: A user's archive. Tabs allow the user to navigate between topic-centric and people-centric views. Toggles indicate whether the collection is being used to influence the algorithmic agent.



We enable users to curate topic-based and people-based archives. However, these two archive types can serve distinct purposes. A topic-based archive might be employed to gather inspiration for future use, whereas a people-based archive is intended to function as a space for receiving updates from a limited circle of close friends or followers, akin to the "see friends first" concept discussed in Section 3.2.

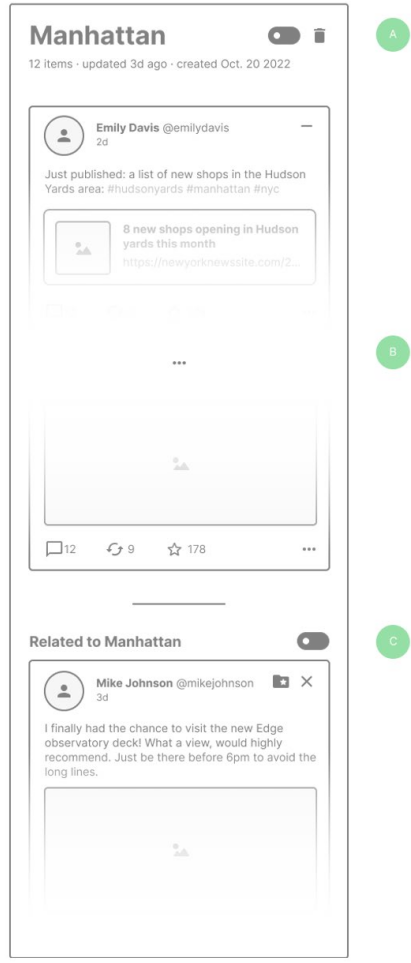
5.3. Evaluating learning performance with collections

How can a user ascertain whether a feed curation agent has effectively absorbed its instructional input? Unlike traditional ML assessments that involve scrutinizing F1 scores and confusion matrices, such metrics may not offer users substantial guidance. Instead, users are likely interested in determining if their collections lead to a heightened presentation of relevant content in their feed, and whether the agent comprehends the essence of a collection. Assessing the former is relatively straightforward: The user navigates to their main feed and observes whether the content becomes more aligned with their collections. Evaluating the latter can appear more challenging without the appropriate affordances. We return to our discussion of in-collection mini-feeds in Section 4.2, which aims to expand the realm of content from a specific collection. We further illustrate this concept in Figure 5.

Posts within the mini-feed retain many of the same interactions found in the main feed, including options like archiving (automatically placed into the current collection), dismissal, and the potential for additional instruction via exploded views.

Effective learning is attained when the content within the mini-feed closely aligns with the attributes of the collection itself. If users frequently find themselves dismissing mini-feed content, they have the option to include more examples and/or refine their teaching goals.

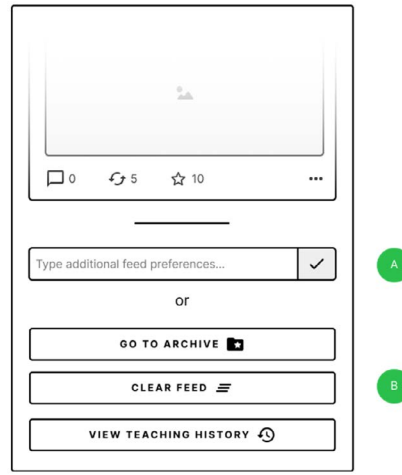
Figure 5: Inside of a collection. A: the collection's header allows the user to toggle algorithmic influence and delete the entire collection. B: the content within the collection itself, including an option to remove items using (-) in the top right corner. C: the mini-feed that surfaces content related to the collection. The user may toggle this feed on or off.



5.4. A control panel for purposefully terminating feeds

We consolidate the affordances we have presented thus far and revisit a question we raised during our discussion of the notification panel metaphor (Section 4.1): How can we foster intentional interactions in a purposefully terminating feed? We present some end-of-feed possibilities in Figure 6.

Figure 6: Possible end-of-feed interactions. A: natural language teaching input. B: call-to-action buttons for opening the archive, clearing the feed (eliminating all remaining posts), and viewing the teaching history.



The first interaction is natural language teaching, a teaching mechanism that supplements the teaching language of exploded views. If users find the teaching language overly restrictive or unsuitable for their intended use case, they can input their preferences via natural language for the agent to learn. We envision this teaching method leveraging the semantic richness and comprehension capabilities of large language models (e.g., models from the GPT family) to parse a free-form preference into a structured one. This structured preference can then be combined with existing teaching strategies to fine-tune the feed. In addition to natural language teaching, we offer quick navigation options to the archive (Figure 4) and teaching history (Figure 3). Collectively, these interactions constitute a central "control panel" that users can access at the end of their feed.

6. THE PROMISING ROAD AHEAD

OUR STUDY AND DESIGN proposals are not necessarily reliant on new, groundbreaking technologies. In fact, we draw upon the research efforts of numerous scholars who have been exploring matters related to algorithmic curation, agency loss, and folk theories since the mid-2010s [25]. The natural question that arises is, why now?

There has been a recent increase in curiosity about, as well as a willingness to experiment with, alternative social media platforms. Following Elon Musk's acquisition of Twitter in October 2022, hundreds of thousands of Twitter users migrated to open-source alternatives using the hashtag #TwitterMigration [39]. Many of these users joined Mastodon, while a smaller group also explored other options like the VR-focused CounterSocial and the decentralized Twitter-like social platform Bluesky. This migration sparked broader discussions within the mainstream about the decentralized and federated social network ecosystem.

6.1. Tapping into the potential of federated social media

Federated social media platforms are built upon decentralized social networking protocols. One such protocol is ActivityPub, which powers not only social media platforms like Mastodon but also various other platforms within the "Fediverse"—an extensive array of platforms catering to diverse use cases, from music sharing to book discussions [40]. In our discussion of federated platforms, we will primarily focus on Mastodon due to its popularity, growing user base, and developer communities. However, it's important to note that these discussions can be applicable to other federated platforms with a substantial user base and interconnected servers.

Mastodon, an open-source social media platform, enables users to post updates, share media, engage with other users through likes and reshares (referred to as "favorites" and "boosts"), comment on posts (known as "toots"), and follow other users, much like mainstream social media platforms. One key distinction, however, is that Mastodon feeds are chronological by default. Another is that *instances*—servers that users can join

or create—play a central role in shaping content distribution. This grants users greater control over their data, the ability to establish instance-specific moderation policies, and a more tailored online experience. Instances communicate with each other, facilitating interactions between users from different instances and forming a decentralized network of interconnected communities.

Mastodon provides an ideal testing ground for new social media tools and interfaces. Developers are drawn to Mastodon due to its open-source nature, well-documented API, and its existing exploratory culture, reflected in the wide variety of available mobile, desktop, and web clients [41]. The instance-based structure of Mastodon encourages a strong community-driven ethos. Developers and instance administrators can experiment with new features *within an instance*, gather user feedback, and then potentially expand those features to other instances or the Mastodon platform itself, if successful. This start-small, scale-later approach promotes experimentation while limiting potential negative impacts and allows developers to iterate and gather feedback more effectively than on today's mainstream platforms. Additionally, Mastodon's federated architecture facilitates the sharing of new interfaces and artifacts across users and instances.

As an illustrative scenario, imagine a user who creates a teachable agent on Mastodon using the metaphors and affordances described earlier. This user trains the agent to provide appealing restaurant recommendations in their local city, and the agent demonstrates effective performance. Another user from the same city, interested in similar restaurant recommendations, can leverage the agent's underlying algorithmic parameters and relevant collections. Enabling the sharing of these teachable agents can potentially cultivate a new ecosystem of robust tools and enhanced user experiences.

Looking further into the future, we envision that this sandbox environment could be bolstered by a technical infrastructure that empowers hobbyists to develop third-party plugins. Numerous previous tools have achieved remarkable success through community-driven extensibility. Examples include Google Chrome, and its open-source counterpart Chromium, as well as Figma. Once introduced, these plugins could be installed by individual users to tailor their personal social media interfaces or by instance

owners to customize experiences specific to their instances. Establishing this infrastructure will likely require a collaborative effort on a significant, multidisciplinary scale. Given the prevailing spirit of our times, now might be an opportune moment to embark on such an undertaking.

Up to this point, we have primarily delved into the potential of federated social media platforms. However, what about the dominant for-profit, corporate-owned platforms that continue to shape our social media landscape? These platforms undeniably still wield considerable influence over our social media experiences. We illustrate how such platforms could find the motivation to integrate teachable agents.

6.2. Aligning incentives on corporate-owned platforms: The case of ads

Ads have become an inescapable reality of the contemporary online landscape. Given that they constitute the primary financial foundation for large-scale social media platforms and no readily available alternative is evident, ads are poised to remain a fixture in the digital landscape for the foreseeable future. The profit-driven nature of corporate-owned platforms hinges on ad-generated incentives, which, unfortunately, often diverge from the more human-centric motivations of technological advancement. The pursuit of elevated ad revenue translates to a focus on prolonging user engagement on the platform, sometimes at the expense of users' extended screen time, declining mental well-being, and susceptibility to misinformation and disinformation.

In our user study, participants consistently downranked ads in their ideal feeds. While a couple of participants were insistent on removing all ads regardless of their content, most grievances related to ads stemmed from their irrelevance. P11 mentioned that the ads they encountered were *"really random"* and often repetitive: *"[The same ad] always comes up multiple times if I scroll long enough."* P18, a non-U.S. citizen, recounted seeing a U.S. military recruitment ad: *"The advertisement I was getting was about joining United States military or navy and I was like, why would you send me that? I could not even join this."* However, in some rare cases, ads turned out to be genuinely meaningful and valuable. P3, who does casual modeling

part-time, received an ad for a local "open call" modeling opportunity and not only gained more experience from open call modeling but also secured a contract from the agency.

One reason an ad might seem irrelevant is that the feed has failed to accurately grasp the user's interests and intentions through implicit means. In our proposed interfaces, users have the chance to explicitly indicate interests in their archive as a teaching method and arrange their content around those interests. Can platforms harness these teaching actions to deliver more relevant and higher-quality ads? In theory, if platforms could access user-taught preferences, they would be capable of delivering ads aligned with those preferences. This presents a substantial advantage in algorithmic transparency compared to current implicit learning methods. Because users had control over their taught concepts and curriculum, they can more easily deduce why a specific ad was presented to them, particularly if it closely aligns with their interests.

In stark contrast to this are today's non-teachable feeds, where a well-targeted ad may be perceived as privacy-invading and raise questions about non-consensual data retrieval and surveillance, while a poorly targeted ad leaves users wondering why the ad was even shown to them at all. However, in practice, directly accessing taught preferences without user consent may also be seen as a privacy infringement. Thus, it is crucial for users to be provided with opt-out controls for any preferences they wish to distance from ads.

In summary, teachable agents for feed curation offer to bridge the gap between user goals of agency and algorithmic control with the revenue goals of corporate-owned platforms. Currently, many platforms rely on a scattershot, quantity-over-quality approach in which they inundate users with ads, hoping that at least some will garner engagement. By reducing the required guesswork, platforms might be able to attain similar revenue goals with a smaller, more focused batch of higher-quality ads. The reduced quantity of ads can, in turn, enhance the platform's user experience.

7. CONCLUSION

TODAY, WE FIND OURSELVES at a turning point in the realm of social media feed algorithms. Platforms are increasingly shifting towards engagement-based approaches, departing from traditional social networking methods to distribute content [32]. At the same time, we acknowledge that these approaches can distort discourse, diminish user agency, and detrimentally affect our mental well-being in significant ways [35]. The emergence of open-source, federated social platforms offers an opportunity to return feeds to their chronological origins as a basis for exploring alternative curation experiences.

In this spirit, we explore the concept of user-teachable agents for personalized feed curation. Our exploration is informed by empirical insights gleaned from our user study, where we identify potential signals that users might employ to instruct a curation agent about their preferences. These insights guide the formulation of three design principles, and we propose two metaphors to reshape feeds for more effective teaching. By amalgamating our insights, principles, and metaphors, we present design affordances intended to facilitate in-feed teaching endeavors, with the aim of empowering users to regain agency in their interactions with their feeds.

However, the practical impact of our design implications is contingent upon empirical user evaluation. Future steps may involve gathering initial feedback from users on our affordances, refining these features iteratively, and embarking on a longitudinal field deployment by integrating them into established social media environments. It is important to note, though, that achieving the latter might necessitate specialized infrastructure. Looking ahead, we envision a substantial potential for a modular social media sandbox, where multidisciplinary minds—ranging from developers and designers to social scientists and policymakers—can converge to discuss and experiment with novel interfaces and algorithms. Through this collaborative effort, we can collectively take pivotal strides toward enhancing sociotechnical systems for society's benefit.

REFERENCES

[1] DeNardis, L., Hackl, A.M.: Internet governance by social media platforms. Telecommunications Policy 39 (9), 761–770 (2015)

[2] Bhandari, A., Bimo, S.: Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. Social Media + Society 8(1), 20563051221086241 (2022)

[3] Wall Street Journal: Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires (2021). https://www.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/ 6C0C2040-FF25-4827-8528-2BD6612E3796

[4] Boeker, M., Urman, A.: An empirical investigation of personalization factors on TikTok. In: Proceedings of the ACM Web Conference 2022. WWW '22, pp. 2298– 2309. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3485447.3512102

[5] Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X., Chi, E.: Recommending what video to watch next: A multitask ranking system. In: Proceedings of the 13th ACM Conference on Recommender Systems. RecSys '19, pp. 43–51. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3298689.3346997

[6] Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 1668–1678 (2019). http://dx.doi.org/10.18653/v1/P19-1163.

[7] Jia, C., Boltz, A., Zhang, A., Chen, A., Lee, M.K.: Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. Proceedings of the ACM on Human-Computer Interaction 6(CSCW2) (2022). https://doi.org/10.1145/3555096

[8] Pavlov, N.: User interface for people with autism spectrum disorders. Journal of Software Engineering and Applications 2014 (2014)

[9] Baughan, A., Zhang, M.R., Rao, R., Lukoff, K., Schaadhardt, A., Butler, L.D., Hiniker, : "I don't even remember what I read": How design influences dissociation on social media. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3491102.3501899

[10] Milli, S., Carroll, M., Pandey, S., Wang, Y., Dragan, A.D.: Twitter's algorithm: Amplifying anger, animosity, and affective polarization. arXiv preprint arXiv:2305.16941 (2023)

[11] Reviglio, U., Agosti, C.: Thinking outside the black-box: The case for "algorithmic sovereignty" in social media. Social Media+ Society 6 (2), 2056305120915613 (2020)

[12] Zhang, R., Lukoff, K., Rao, R., Baughan, A., Hiniker, A.: Monitoring screen time or redesigning it? Two approaches to supporting intentional social media use. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3491102.3517722

[13] Lukoff, K., Lyngs, U., Zade, H., Liao, J.V., Choi, J., Fan, K., Munson, S.A., Hiniker, A.: How the design of YouTube influences user sense of agency. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445467

[14] Christin, A.: The ethnographer and the algorithm: Beyond the black box. Theory and Society 49(5–6), 897–918 (2020)

[15] Pasquale, F: The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge (2015)

[16] Bartley, N., Abeliuk, A., Ferrara, E., Lerman, K.: Auditing algorithmic bias on Twitter. In: 13th ACM Web Science Conference 2021. WebSci '21, pp. 65–73. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447535.3462491

[17] Lustig, C., Pine, K., Nardi, B., Irani, L., Lee, M.K., Nafus, D., Sandvig, C.: Algorithmic authority: The ethics, politics, and economics of algorithms that interpret, decide, and manage. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. CHI EA '16, pp. 1057– Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2851581.2886426

[18] Narayanan, A.: Understanding Social Media Recommendation Algorithms (2023). https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

[19] Twitter: Twitter's Recommendation Algorithm (2023). https://blog.twitter.com/ engineering/en us/topics/open-source/2023/twitter-recommendation-algorithm

[20] Bennett, D., Metatla, O., Roudaut, A., Mekler, E.: How does HCI understand human autonomy and agency? Proceedings of the 2023 CHI Conference on Human Factors in Computing CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3580651

[21] Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., Kirlik, A.: First I "like" it, then I hide it: Folk theories of social feeds. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, pp. 2371–2382. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858494

[22] Karizat, N., Delmonaco, D., Eslami, M., Andalibi, N.: Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. In: Proceedings of the ACM on Human-Computer Interaction 5(CSCW2) (2021). https://doi.org/10.1145/3476046

[23] Siles, I., Segura-Castillo, A., Solís, , Sancho, M.: Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global south. Big Data & Society 7(1), 2053951720923377 (2020)

[24] DeVito, M.A., Gergle, D., Birnholtz, J.: "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. CHI '17, 3163–3174. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025659

[25] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, , Sandvig, C.: "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15, pp. 153–162. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702556

[26] Shen, H., DeVos, A., Eslami, M., Holstein, K.: Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. In: Proceedings of the ACM on Human-Computer Interaction 5(CSCW2) (2021). https://doi. org/10.1145/3479577

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

[28] Burrell, J., Kahn, Z., Jonas, A., Griffin, D.: When users control the algorithms: Values expressed in practices on Twitter. In: Proceedings of the ACM on Human-Computer Interaction 3 (CSCW) (2019). https://doi.org/10.1145/3359240

[29] Ramos, G., Meek, C., Simard, P., Suh, J., Ghorashi, S.: Interactive machine teaching: A human-centered approach to building machine-learned models. Human–Computer Interaction 35(5–6), 413–451 (2020)

[30] Jörke M., Sefidgar, Y.S., Massachi, T., Suh, J., Ramos, G.: Pearl: A technology probe for machine-assisted reflection on personal data. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23, pp. 902–918. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581641.3584054

[31] Ng, F., Suh, J., Ramos, G.: Understanding and supporting knowledge decomposition for machine teaching. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference. DIS '20, pp. 1183–1194. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3357236.3395454

[32] Heath, A.: Facebook's Lost Generation (2021). https://www.theverge.com/22743744/facebook-teen-usage-decline-frances-haugen-leaks

[33] Bennett, D., Metatla, O., Roudaut, A., Mekler, E.: How does HCI understand human autonomy and agency? arXiv preprint arXiv:2301.12490 (2023)

[34] Jhaver, S., Zhang, A.Q., Chen, Q., Natarajan, N., Wang, R., Zhang, A.: Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. arXiv preprint arXiv:2305.10374 (2023)

[35] Harris, T.: How Technology Is Hijacking Your Mind—From a Magician and Google Design Ethicist (2016). https://medium.com/thrive-global/how-technology -hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3

[36] Android: Digital Wellbeing (2023). https://www.android.com/digital-wellbeing/

[37] Apple Support: Turn on or Schedule a Focus on iPhone (2023). https://support.apple. com/guide/iphone/turn-a-focus-on-or-off-iph5c3f5b77b/ios

[38] Spotify: Spotify Users Have Spent Over 3 Billion Hours

Streaming Discover Weekly Playlists Since 2015 (2020). https://newsroom.spotify.com/2020-07-09/ spotify-users-have-spent-over-2-3-billion-hours-streaming-discover-weekly- playlists-since-2015/

[39] Ilyushina, N.: What Is Mastodon, the "Twitter Alternative" People Are Flocking To? Here's Everything You Need to Know (2022). https://theconversation.com/what-is-mastodon-the-twitter-alternative-people-are-flocking-to-heres-everything-you-need-to-know-194059

[40] Silberling, A.: A Beginner's Guide to Mastodon, the Open Source Twitter Alternative (2022). https://techcrunch.com/2022/11/08/what-is-mastodon/

[41] Mastodon: Apps (2023). https://joinmastodon.org/apps

About the Authors

Kevin Feng is a Ph.D. student in the Human Centered Design and Engineering Department at the University of Washington. His interests are at the intersection of social computing and human-centered machine learning. He is motivated by the observation that key stakeholders are increasingly at risk of agency loss from black-boxed but powerful technologies in sociotechnical systems. As such, his research strives to embed the reclamation of stakeholder agency as a vital and meaningful pillar of these systems. His work is supported by a 2022 Herbold Fellowship in data science and computation. Prior to UW, he graduated from Princeton University with a B.S.E. in computer science and a minor in visual arts in 2021, where he was affiliated with the Center for Information Technology Policy and worked on tools to democratize web advertisement research at scale.

David McDonald is a Professor in the Department of Human Centered Design & Engineering (HCDE) at University of Washington. Dr. McDonald has ongoing projects studying collaboration in Wikipedia, social matching, fostering healthy behavior all through systems that interleave computation with human activity. He has published research on ubiquitous sensing for behavior change, collaboration in distributed contributor systems, collaborative authoring, recommendation systems, and public use of large screen displays. His general research interests span Computer-Supported Cooperative Work (CSCW) and Human-Computer Interaction (HCI).

Dr. McDonald earned his Ph.D. in Information and Computer Science at the University of California, Irvine. At UC Irvine he was part of the Computing, Organizations, Policy and Society (CORPS) group. He worked at FX Palo Alto Laboratory in the Personal and Mobile technology group and at AT&T Labs, Human Computer Interaction group. Dr. McDonald recently finished serving as a Program Director for the Human Centered Computing (HCC), Network Science and Engineering (NetSE), and Social Computational Systems (SoCS) programs at the National Science Foundation.

Amy Zhang is an Assistant Professor at University of Washington's Allen

School of Computer Science and Engineering. Previously, she was a 2019-2020 postdoctoral researcher at Stanford University's Computer Science Department after completing her Ph.D. at MIT CSAIL in 2019, where she received the George Sprowls Best Ph.D. Thesis Award at MIT in computer science. During her Ph.D., she was an affiliate and 2018-2019 Fellow at the Berkman Klein Center at Harvard University, a Google Ph.D. Fellow, and an NSF Graduate Research Fellow. Her work has received a best paper award at ACM CSCW, a best paper honorable mention award at ACM CHI, and has been profiled on BBC's Click television program, CBC radio, and featured in articles by ABC News, The Verge, New Scientist, and Poynter. She is a founding member of the Credibility Coalition, a group dedicated to research and standards for information credibility online. She has interned at Microsoft Research and Google Research. She received an M.Phil. in Computer Science at the University of Cambridge on a Gates Fellowship and a B.S. in Computer Science at Rutgers University, where she was captain of the Division I Women's tennis team.

Acknowledgments

We thank Arvind Narayanan, Katy Glenn Bass, and the attendees of "Optimizing for What? Algorithmic Amplification and Society," a symposium hosted by the Knight First Amendment Institute at Columbia University, for feedback and insightful discussions on the first draft of this article.

We also thank the members of the Social Futures Lab at the University of Washington for conversations around new possibilities for social media, which influenced many of the ideas in this work.

Cite as: Kevin Feng, David McDonald, and Amy Zhang, *Teachable Agents for End-User Empowerment in Personalized Feed Curation*, 23-09 Knight First Amend. Inst. Oct. 10, 2023, https://knightcolumbia.org/content/teachable-agents-for-end-user-empowerment-in-personalized-feed-curation [https://perma.cc/TZC4-5A5M].

© 2023, Kevin Feng, David McDonald, and Amy Zhang.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Cover illustration: Emilie Flamme


KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY