Leveraging Structured Trusted-Peer Assessments to Combat Misinformation

FARNAZ JAHANBAKHSH, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA
AMY X. ZHANG, Allen School of Computer Science & Engineering, University of Washington, USA
DAVID R. KARGER, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA

Platform operators have devoted significant effort to combating misinformation on behalf of their users. Users are also stakeholders in this battle, but their efforts to combat misinformation go unsupported by the platforms. In this work, we consider three new user affordances that give social media users greater power in their fight against misinformation: (1) the ability to provide structured accuracy assessments of posts, (2) user-specified indication of trust in other users, and (3) and user configuration of social feed filters according to assessed accuracy. To understand the potential of these designs, we conducted a need-finding survey of 192 people who share and discuss news on social media, finding that many already act to limit or combat misinformation, albeit by repurposing existing platform affordances that lack customized structure for information assessment. We then conducted a field study of a prototype social media platform that implements these user affordances as structured inputs to directly impact how and whether posts are shown. The study involved 14 participants who used the platform for a week to share news while collectively assessing their accuracy. We report on users' perception and use of these affordances. We also provide design implications for platforms and researchers based on our empirical observations.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Misinformation, Social Media, Fact-checking, News Reading and Sharing Platform, Trust

ACM Reference Format:
Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (November 2022), 40 pages. https://doi.org/10.1145/3555637

1 Introduction

Modern social media platforms invest substantial effort in detecting and addressing misinformation [5, 10, 101]. However, these efforts treat users as mere consumers of that information, without agency.

In reality, people have always worked together to determine truth. Epistemologists have characterized social construction of knowledge which is based on individuals receiving testimony from *trusted* sources about information that they cannot verify by direct observation [100]. On social

Authors' addresses: Farnaz Jahanbakhsh, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA; Amy X. Zhang, Allen School of Computer Science & Engineering, University of Washington, Seattle, USA; David R. Karger, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA.

2573-0142/2022/11-ART524
https://doi.org/10.1145/3555637

platforms, any user posting a comment about the accuracy or inaccuracy of a post is contributing to the socially constructed knowledge of others who trust them [14]. As we will argue below, while social platforms *permit* the interactions needed for such knowledge construction, they do not *support* them particularly well and indeed sometimes interfere with them— for example, when a user's comment asserting that a post is false is treated as a signal of "engagement" that spreads the post more widely.

In this work, we explore how social platforms might be modified to provide more explicit support for the social construction of knowledge, in particular for distinguishing accurate from inaccurate information. We propose to do so by incorporating some of the key concepts from the social construction of knowledge—trust and testimony—*explicitly* into the data model and user interface of a social platform, to make it easier for users to work with these components. In particular, we propose:

(1) accuracy assessment of posts by users as part of the data model;
(2) user-specified indications of which sources and other users they trust to assess posts; and
(3) filters that users can configure to block from their feed posts assessed as inaccurate by other users they trust.

As we elaborate below, these user affordances can provide support *within* a system for tactics that users currently must (and often do) apply *without* system support. Support within the system makes these tactics easier to apply and more effective. In contrast to centralized systems of accuracy assessment currently employed by many platforms, such as content labels applied by fact checkers or platform moderators [3, 5], we propose a *decentralized* approach led by users. Such approaches can be more scalable then centralized ones, marshalling more resources than platforms can allocate to combat information. In addition, a centralized approach where platforms make a determination of truth can be at tension with freedom of speech and autonomy of individuals deciding what content to consume [101]. While empowering users to make these choices may carry risks for certain users whose social networks have succumbed to misinformation, it might also benefit those who mistrust platforms or find labels paternalistic [91], and for whom existing interventions backfire. We assess these novel user affordances in two ways: first, a survey of 192 participants that aims to assess whether and how users would value the affordances we propose, and second, a one-week field study of fourteen participants using a prototype social media platform that implements all three ideas as system features. Finally, we discuss the benefits (and risks) of providing greater autonomy to users towards combating misinformation collectively.

The main contributions of our work are: 1) A broader understanding of how social media users deal with misinformation in their feed, how they seek the help of each other in this effort, and what is lacking in the platforms to support user needs 2) The design of user affordances that give a social platform's users greater agency to protect themselves and their social circle from misinformation 3) empirical understanding of how users would perceive and use these affordances; and 4) design implications for platforms and researchers based on our empirical observations.

1.1 Approach

This work was motivated by observations of how users currently attempt to help each other cope with misinformation, and how their efforts are obstructed by current social platforms. Years ago, the third author noticed a number of anti-vaccine posts from an individual they did not follow appearing in their Facebook feed. Further investigation revealed that a doctor friend had been commenting on these posts, explaining why they were false. Because Facebook surfaces commented posts to friends of the commenter, the doctor's efforts led to *increased* visibility of this misinformation.

More generally, as platforms track engagement to decide what to prioritize in their algorithmic feeds, engagement to repudiate a post can make it spread.

Imagine if instead the doctor could mark the anti-vaccine post as *inaccurate*, similar to how posts are currently "liked" or "upvoted" on platforms. Unlike an ambiguous human-language comment or reaction—e.g., a "laugh" that could signify disdain at an incorrect post or conversely, approval of its remarks—an "inaccurate" flag provides an unequivocal signal informing the platform that the post should *not* be spread more widely. This scenario motivates the first user affordance that we examine: affordances for users to rate the *accuracy* of posts they see, for the benefit of other users.

The second user affordance we investigate is providing to users the ability to indicate which other accounts they *trust* when it comes to assessing information accuracy. In existing data models for capturing relationships, users can "follow" or "friend" other accounts; however, a user may choose to follow or be friends with others that they nevertheless do not trust when it comes to assessing information accuracy. If users have the ability to assign these more nuanced relationships, then they can delegate to their trusted connections—such as the doctor in the above example—the power to flag or even remove misinformation from the user's feed. As part of this design idea, we also examine allowing users to post articles with a *request for assessment* from trusted associates; when the associates respond with assessments, the original poster can be notified and their information need resolved.

Integrating accuracy assessments and trust relationships into the platform's data model creates an additional opportunity to *filter* posts from a user's feed based on their accuracy status as assessed by trusted associates. Thus, the third new user affordance we investigate is to provide *user-configurable filters* for blocking certain information from their feed, for example, information assessed false by someone they trust. This design idea is informed by prior work which reports that users are more satisfied when given controls for manipulating their feed [104]; and that in the absence of such control over their algorithmically curated feed, users try to find ways to see or hide certain content, although with uncertain efficacy [27]. Not only can accuracy assessments be used individually, but because assessments are structured, they can also be automatically aggregated and compared. For example, if a user assesses a post as true, and someone they trust assesses it as false, then the discrepancy indicates a disagreement. The participating assessors' attention can be drawn to the disagreement, and they can attempt to resolve it, leading either to a change in assessment, or perhaps a change in trust relationships.

1.2 Evaluation

To determine whether these ideas might help users, we designed two studies. The first phase was aimed at understanding whether there are indeed needs for the affordances we were envisioning, by investigating users' current practices with reading and sharing online content. This investigation surfaced themes that are largely missing from the body of research on misinformation, related to how users use platform features to take collective action against misinformation and what is lacking in platforms to support users' needs and processes.

Through a survey of a diverse group of 192 people, we found that users already provide and receive corrections in their social circle, and they ask and are asked about the accuracy of content. To do so, they repurpose the features that are offered by platforms, such as comment, like, and share. These ambiguous signals can be misinterpreted by the platform or different users, with undesirable consequences. In addition, participants follow both those sources they do and do not trust for a variety of reasons, and would find benefit in being able to curate their feed based on these trust relationships. We also found that while some users wanted to automatically filter out content assessed as inaccurate or unverified, others preferred to view that content with a flag or be able to access it separately from their main feed.

To test how users might actually use these new affordances if given the opportunity, we built and conducted a field study on a social content sharing platform prototype called Trustnet [1] that offers the paradigm of user empowerment described above, where users can assess news, specify the sources they trust, and filter their news feed based on accuracy assessments provided by their trusted sources. The platform then displays structured accuracy assessments next to posts. The user study involved 14 users who joined the study with at least one other person from their social circle and used the system for a week. We saw that participants used the system as intended, were capable of assessing posts, and saw value in the norms that the platform established, such as vetting posts before sharing. They also saw value in facilitating inquiries about the accuracy of posts and in seeing assessments from others. We found that participants used the filters in different ways, some configuring their feed to show only articles confirmed by trusted sources, others keeping refuted or disputed articles in their main feed, and some using the filters to seek out such articles. The user study also revealed the need for enhancing the assessment functionality to cover a variety of complications such as true-but-misleading statements, slant, or a mismatch between content and headline.

2 Related Work

We draw from related work to motivate each of our design affordances and explain how they can help users in the fight against misinformation.

2.1 Accuracy Assessment of Content By Users

Below, we discuss related work to explain why we chose a decentralized approach to labeling of misinformation and that our affordances help users have rich indications of credibility upon encountering content and even after exposure.

2.1.1 Informing Users of Inaccuracies at Reading Time. The prevalence of misinformation in online spaces has led researchers to examine how to identify misleading or inaccurate information using strategies such as machine learning algorithms [17, 86, 88, 94, 108] or crowdsourcing credibility annotations of articles or sources [4, 9, 26, 39, 52, 85]. Platforms have also invested in misinformation detection by using automation, paid human moderators, and partnerships with third party fact-checkers [3, 5].

Platforms take action against misinformation by flagging or labeling it, reducing its visibility, and even removing it [10, 21, 76]. While some degree of platform intervention is necessary for harmful behaviors such as toxic language or child abuse [58, 81], a broader platform goal of governing *truth* can be at tension with freedom of speech and autonomy of individuals deciding what content to consume [101]. Related is a study which reports that while some users find platform assigned labels helpful and even wish for stronger labels for content that has a stronger perceived harm, others find the labels judgmental, paternalistic, and against the platform ethos [91]. The decentralized approach that we pursue in assigning accuracy labels to content allows for developing labels that are not governed by the platforms.

Rather than determining what information users should or should not see, a body of work has examined attempts to warn users about misinformation or bad actors. These attempts include compiling a list of credibility indicators that news media can use to differentiate themselves from alternative or low credibility media [112] and flagging the accuracy of posts by platforms and content moderators [76]. Among the studies that examine the effectiveness of these approaches, Clayton et al. report that tagging news stories as false or misleading reduces their perceived accuracy [19]. This approach can have unintended consequences as the perceived accuracy of news

[1] http://trustnet.csail.mit.edu/

stories without warnings may increase, or users may be enticed to click on the tagged inaccurate content [57, 82]. Informed by this prior work, in our prototype system, we distinguish content that was assessed (either true or false) distinctly from content that was not assessed and is thus unverified, in order to discourage users from ascribing any default accuracy to content. Our design makes this possible because assessments are captured and displayed in structured form.

Research is mixed about whether fact-checking leads to correcting misperceptions, and the contentions seem to suggest the circumstances play an important role in when corrections succeed. On one hand, in some studies involving commonly held misperceptions in the domain of news, such as Iraq's possession of WMD before the U.S. invasion and a link between vaccines and autism, researchers find that presenting corrections alongside articles touting misperceptions can in fact strengthen those misperceptions among people who are more prone to believe them [78, 79]. Related are studies reporting that exposure to opposing views can exacerbate polarization [8, 18, 109]. On the other hand, Bode & Vraga report that presenting correcting information through debunking articles suggested by Facebook's algorithm as well as through comments left by other users can in fact change prior misperceptions [11, 12]. Other studies have examined the role that corrections from sources such as governmental agencies, research institutions, news media, and other social media users have on improving the accuracy of users' belief in scientific issues in various domains [13, 105, 106].

Unlike this prior work, we focus on accuracy assessments performed only by sources a user has specified they trust. We take inspiration from studies finding that social media users are more likely to attend to and accept fact-checking information from friends compared to strangers [38, 68]. Related is a Twitter field experiment that found being corrected by a stranger significantly reduced the quality of content that users subsequently shared [75].

2.1.2 Correcting Misinformation after Exposure. Another body of work has probed how posts making corrections fare on social media compared to the posts they refute. These corrections typically appear only after an inaccuracy has already spread and been viewed by users. For instance, Starbird et al. report that tweets that correct rumors on social media often propagate with a delay and their number is much lower than those that spread the misinformation [98]. Zubiaga et al. find that tweets supporting unverified rumors are retweeted more than tweets denying the rumor or those that are posted after the rumor is resolved. Therefore, once a rumor has been debunked, users do not make the same effort to inform others about its falsehood [114]. Shin et al. analyze rumors spread on Twitter during the 2012 election and find that they mostly continued to propagate even after professional fact-checking organizations had published debunking information [92]. As can be seen, a major drawback of such corrections is that they may be slow to arrive while a piece of misinformation is spreading widely. Thus, we design our new affordances to mitigate this issue in a few ways. First, we prompt would-be-sharers for an accuracy assessment before they share any content. This results in assessments that match in scale and time to the shared content. Second, any content that is missing an assessment is shown as such and does not have the same appearance as content assessed as true. Finally, we enable users to inquire about the accuracy of posts from others they trust, thus speeding up the process of getting important content assessed.

2.2 User-Specified Indication of Trust in Others

Our affordances confer the power of content moderation to users. This is similar to peer production systems like Wikipedia or Reddit [41] where governance is determined by the community. However, while users in peer production communities are expected to collaboratively develop a single source of "truth", we propose that we also empower users to choose whose "truth" they want to heed. We present a model where users make their own decision about who they trust to provide sensible and

useful assessments and whose content they wish to see based on these assessments. This model is rooted in prior work on the importance of trust in building knowledge.

2.2.1 Situating Our Work in Epistemology. Our work relates to epistemology in that we consider structured ways to capture and report on how knowledge is constructed in a social network. According to theories relating to epistemology, to have knowledge about something is to have a justified belief in it. A belief is justified if the believer possesses evidence for it (Evidentialism theory) or if it originates in a reliable source (Reliabilism theory). Examples of reliable sources are perception, reason, and the testimony of someone reliable [100]. Believing someone's assessment about a post online is an instance of obtaining knowledge through testimony. Bruckman asserts that the reliability of such a testimonial exchange online is determined by the features of the platform where this exchange happens as well as social norms for people contributing content there [14]. The affordances that we propose for adoption by platforms in this work are aimed at making users more aware of what content is credible by virtue of its assessor and thus, enhancing the reliability of a testimonial exchange.

2.2.2 Source Credibility and Trust. Because of the sheer volume of information available online, users rely on cognitive processing of heuristics and cues to determine which content is credible and worth their further attention [72, 73]. Although these heuristic judgements are triggered prior to more systematic cognitive analyses, they can also aid subsequent systematic information seeking [103]. Research has called for using these cues to guide users toward more accurate credibility judgements [61, 103]. One such heuristic is the credibility of the information source [30, 45, 74], which is not an objective property but rather the source's believability as perceived by a user [30]. Research characterizes source credibility as comprised of three dimensions: expertise/competence (i.e., the degree to which a perceiver believes a sender to know the truth), trustworthiness (i.e., the degree to which a perceiver believes a sender would tell the truth as they know it), and goodwill (i.e., the degree to which a perceiver believes a sender has the receiver's best interests at heart) [61, 70]. Users have also self-reported using their trust in the poster of a content and their perception of the poster's influence, topical expertise, and reputation as heuristics in determining the content's credibility [64, 74, 99].

Our design idea of allowing users to explicitly mark certain sources as trustworthy can conceivably help users reflect on the credibility of their information sources. It can also increase attention to the perceived credibility of sources of content appearing on users' feeds. Although one's trust relationship to a source may vary from topic to topic, in our proposed design, we start with the simpler case of enabling users to mark who they generally trust irrespective of topic. In the future, this could be enhanced to allow for topic-specific trust relationships. However, it is conceivable that this may not be necessary if users self-select towards only assessing on topics where they have expertise, as others have the power to stop trusting their assessments in the system.

This approach of democratizing trust and decision-making rather than ceding these powers to centralized platform-assigned moderators has been implemented on platforms that permit subcommunities or groups to appoint moderators, such as Reddit or Facebook. The community moderators on these platforms develop their rules and norms independently from the platforms and are given the power of deciding which content to allow or remove. The space of users' involvement in the decision-making is constrained to joining these communities or leaving them. Our proposed affordance is a natural extension of this idea—rather than limiting the assessment decision to a small number of moderators, we explore the idea of allowing users to appoint any arbitrary set of other users to serve as moderators for them, by heeding their assessments of content accuracy. Other areas in prior work that have explored decentralized decision-making include subjective

moderation in chat, against email harassment, for allowing or removing content, or for developing policies for platform governance [20, 29, 33, 48, 65, 111].

2.3 User-Configured Filter of Accuracy Over Feed

While structured assessments and customized trust relationships could be used as input for algorithmic curation of feeds, we made the decision to instead give the power of feed filtering to users. This decision was rooted in the body of research on how users interact with and perceive curation algorithms and users' desire for more agency.

On platforms, users are given a limited set of actions such as like, follow, or mute to tune their algorithmically curated feeds. However, because of the complexity of these algorithms, the large number of features that they take as inputs besides the ones consciously controlled by the user, and the fact that their inner-workings are hidden [24], users' actions do not necessarily procure users' desired effects or the algorithms end up violating user expectations [23, 60]. The absence of a direct translation from user inputs to effects has led users to conjure folk theories of how algorithms operate [22, 28, 89]. Research has found that users are dissatisfied with algorithms making all decisions on their behalf [15] and more satisfied when given control over curation of their feed [104]. In the absence of controls, user attempt to find ways around the algorithm, albeit with uncertain efficacy [27, 80].

Kerr & Earle have warned against the dangers of opacity in algorithms that employ preemptive prediction which delimit a user's access without providing a choice, and often without their knowledge [51]. An example of such an algorithm could be one that presents a feed of only the content that is deemed as reliable or accurate to the user. If given the power, a user may wish to seek unverified or inaccurate content for a variety of reasons, e.g., to assess it for the benefit of their friends and family [66] or to be aware of what content they share [89]. Therefore, in our design, we provide users with an accuracy filter and leave it to them to decide what content they wish to see based on their settings of accuracy (posts assessed as accurate, inaccurate, those with disagreeing assessments, or those whose accuracy others want to know about) and according to which assessors.

3 Method

To evaluate our design affordances, we designed two studies. *Study 1* was a survey that aimed at evaluating whether there were indeed needs for the design affordances that we were envisioning. *Study 2* evaluated the design affordances through a field-study of a platform that implemented them. The reason for the first study is that the effort and time that goes into developing a prototype that offered the affordances was substantial, and therefore we needed to determine whether the pursuit was worth the effort.

The survey of Study 1 had two main parts: First, it inquired about people's current practices of reading and sharing online content and if and how they ask for assessments or provide assessments to others. This part allowed us to ascertain that users do indeed engage in providing and seeking assessments but without system support. It also informed us about the diversity of behaviors that users manifest as they try to repurpose existing system functionalities in their quest for the truth. The second part was intended as an early evaluation of our design ideas, by capturing users' perceptions of a hypothetical system that offered affordances that were aimed at facilitating their current practices.

Study 2 was a field study of a prototype system that complemented our earlier evaluation Study 1 and served as a *technology probe* into the problem space. It allowed us to understand user needs and observe user behaviors in a real setting, where instead of imagining the technology, users could interact with it and with each other by using it [42].

4 Study 1: A Survey of Practices, Laments, and Wishes

This section describes our Survey study. We began this study with our general hypothesis about user needs and behaviors, and about new affordances that would help users meet these needs more effectively. We designed questions aimed at determining (i) whether there is evidence that our hypothesized needs are real, and (ii) whether there is evidence that our affordances would help address those needs. We found evidence supporting these hypotheses, both explicitly articulated needs *for* the affordances, as well as signs that user behavior and effort were shifted in sub-optimal directions by the *absence* of these affordances.

4.1 Procedure

The survey comprised four sections. The first concerned users' online content reading and sharing practices including what they like about the platforms they use and their reasons for sharing news or debating with friends. The second targeted the sources within users' social circles including whether they are trustworthy, whether and how users cope with sources that post inaccurate information, whether and how they tweak their feed to see fewer posts by untrustworthy sources given the controls offered by platforms, and whether and how they provide fact-checking information to their friends or seek their friends' judgement to assess content accuracy. The third section asked about users' trust practices including how they decide who to trust to share reliable content. The fourth investigated how users would engage with a social content sharing tool that provides various kinds of accuracy filters and whether and how they would help friends filter out false or unreliable information from their feeds. The research team held multiple meetings to discuss and iterate on the questions for clarity and comprehensiveness.

Following the main section, participants then answered a demographics questionnaire that involved questions on political preference and theistic ideologies, among others. We adopted the questionnaire from previous work on misinformation [45]. The full questionnaire is included in Supplementary Materials.

We advertised the study on a mailing list and through Facebook. We asked that the Facebook post be reshared so that we would reach a broad and diverse population. Indeed, the responses indicated that some users had encountered the survey in private Facebook groups of which they were part. At the end of the study, we raffled off $50 Amazon gift cards to 3 random participants. Our study was approved by our Institutional Review Board.

Once we had obtained answers, a member of the research team performed an inductive thematic analysis on the responses to each of the questions and assigned codes to them. Sometimes a response to a question contained answers to other questions in which case those idea units were also given the relevant codes. Through subsequent passes over the data, the codes that had too much similarity were merged and others showing distinct changes were split. We did not calculate inter-rater reliability because we developed the codes as part of a thematic analysis to yield concepts and themes [71].

4.2 Participants

A total of 154 participants completed the survey of whom 149 provided at least partial demographic information. The number of responses for some of the questions exceeds 154 because some participants abandoned the survey before its completion. The total number of participants who provided at least partial answers was 192. Participants who provided demographic information were distributed across a wide range of age with a median of 50 years old. 53% of participants identified as female. The median for income was $100,000 to $149,999 (ranging from less than $10,000 to $150,000 or more); and for highest degree received, Master's degree (ranging from high

school diploma to Doctoral degree or M.D./J.D.). 31% identified as Democrat, 22% as Republican, 29% as Independent, and 19% as Other including Libertarian. With respect to ethnicity, 104 identified as White/Caucasian, 15 as Asian, 8 as mixed (e.g., White, and/or African American, and/or American Indian or Alaska Native, and/or Native Hawaiian or Pacific Islander, and/or Asian), and 18 as Other.

4.3 Results

4.3.1 Users Share Content for Multiple Conflicting Reasons that Are Not Transparent to Readers. The responses to this question uncovered the often conflicting reasons for sharing content. This inconsistency was mostly due to the absence of explicit channels for providing or asking the others for credibility assessments. This led users to share content sometimes to endorse it, sometimes to declare it as inaccurate, or other times to ask other users for assessments. This conflation can have the unintended consequence of surfacing a false or unverified post to other users as well as the platform who may have their own interpretation of why the content is shared. Because of this conflation, some users refrained from sharing content, thus missing the opportunity of warning others about content they knew to be false.

For the majority of participants (N=152 out of 192 respondents who provided answer to this question), sharing was a sign of implicit endorsement, as they stated that they share news to inform others.

"I fairly often share news stories in comment threads that are on topic either to inform, to correct prior misstatements, or because there was a request for information."

Many participants also shared content to analyze it and provide assessment of its legitimacy, accuracy, bias, etc (N=88 or 46%). The occasions when participants engage in sharing for assessing a post include when its content or headline is bogus, out of date, or misleading, they want to provide relevant disclaimers and context related to one's area of knowledge that may not make it into popular coverage for news, they wish to offer commentary, or they share a story from multiple different sources to demonstrate bias:

"I sometimes offer comical commentary to accompany high-level summaries that appear in the news. For example, comparing local news coverage of how universities initially responded to the COVID-19 outbreaks to the statements that were shared internally. Or, for news that is relevant to my research field I try to provide the relevant disclaimers and context that don't necessarily make it into popular coverage."

A number of participants (11% or N=22) share to invite assessments and opinions from their social circle, to encourage conversation, and to challenge other people's beliefs:

"I always like to see what other people think about the same news story I've read. Sometimes getting other people's input helps me educate myself more about the topic. Especially if they offer different links and views. I like to look at ALL sides before making my final judgement"

Conversely, some (5% or N=9) specifically *refrain* from sharing articles that they know are not accurate because they believe sharing is an endorsement of the accuracy of the content, they do not want to invest their energy, they believe journalists and advocacy groups are already providing fact-checking information, others would not change their mind about their misperceptions, or that the biases of different sources are already known.

"Explicitly, I don't think I do this (in the sense that i'm not likely to share a story with a comment like "wow, how biased towards X" or "wow, this is wrong"). But implicitly, I don't share stories that i don't think are a) interesting and b) either accurate factually or transparent in their viewpoint/opinion/bias if that's relevant. So my sharing is implicitly an endorsement of merit in some dimension."

This observation points to a missed opportunity that although these users possess the knowledge to assess certain content, because they choose not to conflate their assessment with sharing, their

social circle who are likely to come across the misinforming content by other means may fall prey to misinformation.

53% of participants (N=102) said they share a news post to convey their feeling about it. Other reasons include to advocate on an issue or ask for an action (e.g., signing a petition), for entertainment, because someone would find it funny or relevant, or for social grooming and to maintain connectivity with one's social circle. These use cases are all other examples of how the share functionality is overloaded with multiple and often incompatible intentions and consequences [44].

Some participants (24% or N=47) however, stated that they rarely share news posts. Their reasons included because they want to keep their online presence minimal, do not want to overwhelm the others, assume others are more informed or can find their own sources, do not want to start arguments between those who are of different opinions or to alienate people, believe their views are presumed not welcome, cannot fact-check every piece and are wary of spreading misinformation, or fear that their sharing news can be deemed inappropriate based on the sharing platform and the user's role within it, for instance, when their colleagues are on the platform.

The fear that one might contribute to the spread of misinformation when sharing posts can be addressed with a clear signal at posting time inquiring about the accuracy of the post. In addition, introducing controls for filtering their feed to see what they wish could mitigate users' concerns of potentially overwhelming recipients.

Some of the themes that our participants discussed are aligned with prior work that report information sharing and seeking, changing minds, and maintaining and extending one's social network as drivers of sharing news on social media [40, 59, 63, 64, 107]. Prior studies also mention reputation seeking and being part of the crowd for instance, for viral posts, as other motivations for sharing news [59, 107], which our participants did not discuss.

All these observations indicate the need for the explicit separation of accuracy assessment from sharing or, more broadly, the functionalities that platforms are already providing. With accuracy captured and displayed in structured form, as our proposed affordances offer, the intended message of a share will not be confused with endorsement, by either the platform or other users.

4.3.2 Users Encounter Content that They Know to be Unreliable. The responses to this question indicated that regular users should not be treated as mere passive consumers of misinformation who do not realize that they have fallen prey to it and who should be saved. Rather, many users view online content with scrutiny and at the time of exposure to a piece of content or later judge the content to be false.

82% of participants (N=146) stated that they at least occasionally see people or organizations they follow online share content that they know to be false, biased, or of unverified accuracy, or that they later find to be false. These unreliable posts take the form of articles from biased press or networks with propaganda, outdated, sensationalized, misleading, or inaccurate articles, misleading headlines or translations, misattributed quotes, misplaced image captions, claims with no supporting proof, and claims that are not in agreement with one another:

"Even though I'm right of center, I don't trust or value everything that comes from PragerU, for example. Though I probably agree with them on a lot, I think their content is not well explained or documented and know that some of it is designed to manipulate rather than educate. I would not share their content with friends. I feel the same about Daily Wire, Breitbart, and Daily Caller, for example."

Participants believe these posts to be false based on their general knowledge of the matter, domain expertise, firsthand knowledge—*"I used to have a role in government, and I would see news stories that I knew were inaccurate because they were about meetings where I had been present."*, presentation of the claim, and their knowledge of the bias of the source—*"...many individual journalists see themselves as crusaders against Trump, and they allow those biases to override any professional standards they*

once had. Anything that cites an expert with a PhD is likely biased to the left, like academia in general is 80+% leftist...". These rationales are in alignment with the reasons why people disbelieve news, reported by Jahanbakhsh et. al [45].

4.3.3 Many users Rely on Trusted Sources to Provide Assessments. We investigated whether participants consider certain people or organizations whom they follow online to have informed opinions and whether and how they seek the judgement of these sources on posts that they encounter. We found that users already engage in seeking assessments from various trusted sources. They do so by repurposing various existing platform features or, because there is no explicit feature for seeking assessments, by passively waiting for information from those sources.

79% of respondents (N=142 out of 180 respondents who provided answer to this question) stated that they indeed follow certain sources they believe are worthy of trust. Examples included scientists and people with domain-knowledge, certain commentators, diarists, analysts, authors, journalists, or politicians that they deemed well-informed, certain news organizations, activist groups, Facebook groups, subreddits, friends, and family. Therefore, the traditional practice of knowledge building by testimony from someone reliable is also exercised on online platforms [100].

Participants indicated they use a variety of methods to *seek the judgement* of these sources. Some tag these sources in a comment or on the post where their assessment is needed, or share the post with them on social media. Others seek their opinion privately, via direct messages or email. Some participants specified that they can obtain the opinion of their trusted sources only by reading their posts because they are not personally acquainted with these sources or because they are concerned the sources might be too busy. Other approaches included writing comments on a trusted blogger's blog post or visiting one's trusted websites such as Snopes that have already published a piece on the topic.

This observation provides evidence for the value of our affordances. In a system that captures and displays assessments on content, a trusted source might have already assessed the content when the user encounters it. In this scenario, displaying the assessment would save the user the effort of (and fear of burdening others by) requesting that assessment. Such a system would also provide the assessment at the time the information is encountered, as the effect of post-hoc corrections can be too little too late [97, 98, 114]. Our affordances facilitate requesting assessments if none yet exists and lower the burden of providing one, both of which are likely to increase this desired behavior.

4.3.4 Users Provide Assessments to Their Social Circle. We found that many users are assigned a fact-checking role by their social circle. With users already embracing this role even without system support, there is potential for streamlining the process, through structured requests for assessments and structured ways to capture them.

51% of the participants (N=92) said that their friends rely on them to verify or refute information found online. They receive these requests by having their names tagged on posts or through private messages. In response, participants research the topic, provide explanations grounded in their expertise or experience, translate scientific articles into a more approachable language, direct their friends to better sources or to supporting articles on trusted sources, point out misleading language in the article, ask questions and allow people to think and come to their own conclusions, or explain that they are not well-informed about the subject.

"My parents often ask me about medical information they see or health recommendations. I do research and try to distill conclusions from peer-reviewed articles into everyday terms."

Participants mostly felt positive about helping provide information to their social circle, deemed the request for their judgement as a display of trust and pleased to be considered an expert and have their opinion respected and valued.

The responsibility of providing fact-checking information for family and friends has been discussed in the context of private and encrypted chat spaces, in which it is difficult for platforms to moderate misinformation [66]. However, even on public spaces where platform moderation is possible, users do still happily fulfill the role of on-demand fact-checkers for their social circle.

These observations provide evidence that people value the ability to provide assessments to their social circle, which also suggests that our affordances for providing simpler and more visible assessments would be valued.

4.3.5 Users Invest Effort in Increasing the Reliability of Their Own and Their Social Circle's Feeds. We found that many participants already use the small set of features offered by the platforms in an attempt to curate a reliable feed for themselves, suggesting that our proposed accuracy filter affordance could be well received by them. In addition, in their pursuit of a reliable feed, many act as vigilantes against misinformation, actively fighting inaccuracies. We envision that by leveraging their assessments captured through our proposed structured accuracy affordance and input into our proposed accuracy filter, other users can benefit from a more reliable feed as well.

54% of participants (N=75 out of 140) employ a variety of methods to tweak their news feed to see fewer posts by the people or organizations who post inaccurate content. These methods include completely taking the content out of their feed (N=75) by unfollowing, unfriending, muting, blocking, or snoozing the untrustworthy sources. Other methods are intended to train the feed curation algorithm to show content from the unreliable sources less often (N=4):

"I remove the offending articles from my feed in hope that the algorithm picks up my hints."

"Liking other things such that these show up less"

The phenomenon of users finding ways around a curation algorithm that chooses content for them has also been reported by Eslami et al. [27].

Sometimes participants assume a more active role in fighting the unreliable content that they encounter (59% or N=82 out of 138) by debating the credibility of the post with the poster, reaching out to news publishing entities in an effort to notify them of inaccuracies in their content— *"There are numerous news agency channels in social media, whom i have seen put false news in them. I even tried informing them of the mistake and they said they will investigate but they didn't remove it."*, reporting false content or the users who post inflammatory content to platforms, pointing the unreliable content out to others in an effort to inform them— *"Occasionally I'll point out bad information to people I know who are both affected by and it and might not be aware of it. e.g. mentioning to my parents that "facemasks do nothing to stop corona" is probably wrong"*, or expressing support for those who correct the unreliable content— *"...if something is particularly egregious, I will often "like" someone's comment that corrects/refutes/clarifies."*

In order to protect themselves from content that appears unreliable, participants seek fact-checking information either actively, or through more passive means– *"I have a 3-day rule – if something "outrageous" is reported, wait three days and see what the real fall-out is. Usually, it disappears quietly, after it's been discounted and an apology has been issued on page"*.

4.3.6 Users Do Not Always Want a Feed of Completely Reliable Content. Some participants (62% or N=84 out of 136) believed that there were benefits to keeping unreliable content in their feed or that the consequences of removing them outweighed the benefits. This is because in the absence of clear accuracy signals, the separation of the concepts of *follow* and *trust*, or fine-grained accuracy filters for feed customization, the only certain way to maintain a reliable feed on platforms is to unfollow the sources of unreliable content and participants had various reasons for not doing so.

Therefore, not all participants engaged with unreliable content. Some indicated that they allow it to remain in their feed because the unreliable content is infrequent, they believe all information is to some extent unreliable, the sources of the unreliable content are friends, family, or people

with whom they have shared interests, sometimes mistakes happen—*"Doesn't everyone make a mistake now and then? While a news organization is the sum of its parts, individual contributors and editors can still make bad judgement calls or find additional info after initial reporting."*, they want to be aware of other perspectives and where different narratives overlap, they want to be aware of what their friends and family think so they know how to talk to them or correct their misconceptions—*"Because I can then point out their fallacies and kindly show them where they are wrong. And if I happen to share something unreliable, I expect them to do the same."*, unreliable sources sometimes offer valid arguments—*"Even though someone or [an] organization publishes false info, they still have an absolute right to do so. I'm interested to see the whacked out POV of a conspiracy theorist, like Alex Jones or David Remnick. Not for the novelty but for the grain of truth that may be in there that I never would have seen if a craft algorithm scrubbed it from view..."*, they are concerned they might miss otherwise reliable content from these sources, they rely on major news publishing media to correct inaccuracies in their content—*"I do a lot of passive information consumption from major sources, and they correct stories, flag warnings and try to be clear about their editorial policies"*, they deem the situation entertaining, or they do not know how to tweak their news feed— *"...I have looked into trying to mute them but I wasn't sure how..."*. Prior work has also reported on users' interaction with unreliable content in an attempt to learn more about different viewpoints and to build counter arguments [64].

While, as indicated, many attempt to provide correcting information to the sources of unreliable information in their feed, some believe such an engagement infringes on the source's freedom of speech rights—*"Because it's her [my mother-in-law's] damn page to be annoying on if she wants to. Nobody is making me follow it, read it, or believe it. If I don't like it I can keep scrolling or unfollow her. She has freedom of speech and expression, no matter how her chosen content makes anyone feel."*

Although many platforms already provide methods for filtering content from specific sources out of one's feed, participants' responses indicated that these methods do not sufficiently address the diversity of users' needs. These needs motivate our design affordance of separating the concepts of follow and trust, as well as allowing for accuracy assessments at the granularity of content, rather than source.

Table 1. Participants' ideas for a filter that can help users separate accurate from inaccurate content.

Filtering Ideas	Quotes
Block content originating from certain sources	*"I'd rather be able to block a source (i.e., news channels I know are unreliable) so that I don't see misinformation from news site A, newspaper B, pundit C, etc."*
Block information that has not yet been verified only on certain topics	*"it's also domain-dependent. in the medical domain, i would like to see only the accurate news; but in the politics domain, it's hard to tell what accuracy means.")*
Block information that has not yet been verified only from certain sources	*"I would appreciate this feature as well because it places the burden on me to think hard about whom I trust."*
Flag unreliable or biased content but keep it in the user's feed	*"I don't think I'd use the filter if it removed it entirely- I think I'd want to see news that had been flagged as false so if my parents or friends brought it up later, I'd be able to refute it as fake news."*
Provide reasons why certain posts are flagged	*"I would want a tool that allowed me to view "blocked" articles separately along with the specifics as to why they had been blocked. This would allow me to verify what is being filtered and why which I believe is very important in a proposed filter."*
Display the provenance of the content	*"Show the trail of who passed it on to whom!"*

4.3.7 Users Are Receptive to a Filter for Blocking Misinformation Unless Operated By Platforms. We asked participants to imagine they had a filter to block out all the misinformation on their feed, but

not humor or complaints or any post that is not news. We then inquired whether they would use such a filter and what features they desired in it.

Some participants (54% or N=83 out of 154) saw value in such a filter—*"I think this kind of filter is very useful for me to save my time and my attention."*. However, many respondents (48% or N=74) including a number of those who perceived value in the filter interpreted this filter as one that would be managed by the platforms, and raised doubts about or were vehemently opposed to relinquishing the power of content filtering to an algorithm that executes the decision or carries the biases of the people who have made it and whose judgement users do not necessarily trust.

"How could I rely on the filter? Who would be setting the parameters? As it is what Facebook and Twitter do to "fix the problems" on their platforms is reprehensible. I would NEVER use ANYthing they had to offer! It's like 1984 coming to life! Or Big Brother, Altho from Big Business, not government. Semitism is rampant on their sites, terrorists can communicate with no blowback, pedophiles, too. But Twitter and Facebook check speech that might hurt someone's feelings? Or political points of view that counter the mainstream culture? Or points of view that don't believe in global warming? It is outright fascism to block those things. Especially while true bad stuff happens on their platforms."

"I need to evaluate all sources. Good, bad, ugly. I don't want some fuzzy-faced tech whiz or algorithm written by a university-indoctrinated Progressive doing it for me. Of a MAGA partisan, for that matter."

The participants who believed that this filter would be applied universally by the platforms feared that its application could result in censorship of certain ideas.

These ideas mirror concerns raised by some scholars of law who argue that unregulated algorithmic filtering undermines the notion of agency in the selection and consumption of content, and that it can inhibit the development of a free market of ideas needed for citizens in a democratic society to perform their civic duties [35, 54].

Other users interpreted the question as describing a single-source-of-truth filter that would be enabled or disabled by the users, and were wary of creating a divide between those who chose to apply the filter and those who did not:

"I also wonder what the bigger implications are – would such a filter only further the divides between what people see based on algorithms vs what other people don't see?"

Other reasons why participants did not wish to have their feed always filtered were to stay abreast of what other people are thinking (N=25)—*"I want to know what people are hearing/believing. Can't address it if you don't know it...educated friend was starting to buy covid-is-a-bio-weapon... seeing it told me to send him the 2007 research paper on mutations of SARS/Covid in horseshoe bats in Wuhan... that concluded by describing the risk of transfer to humans as a "time bomb""* and to cultivate their own critical thinking (N=3).

The participants who saw value in such a filter volunteered several ideas that they imagined would help them separate accurate from inaccurate content, summarized in Table 1. Almost all of these ideas are supported in the prototype platform that we built by virtue of our affordances.

This section validate users' need for our affordance of *personally configurable* filters.

4.4 Summary

In summary, our survey provided evidence for the usefulness and desirability of our proposed affordances for end users. Users passively consume accuracy assessments and actively seek them out from sources they trust. Some are eager to share accuracy assessments and some are assigned a fact-checker role by their friends and family. The lack of visible accuracy assessments can confuse users as to the intended message of posting, as many take sharing a post as a signal of endorsement. The fear of this confusion sometimes deters people from sharing their (in)accuracy assessments. Users would like to have filters that block unreliable content, but do *not* trust those filters to be

operated by the platforms. Instead, they want to be able to configure those filters themselves, including to sometimes show inaccurate information so they can understand or refute it.

5 Study 2: Field Study of User Affordances for Truth Assessment on a Social Media Platform

To complement Study 1's survey of users' current practices and desires, we turn towards collecting empirical evidence of what users would actually do in a naturalistic setting if given user affordances for truth assessment. We developed a prototype social content sharing platform (Trustnet) that implements our three design ideas. It allows users to follow different *sources*—other users or news publishing entities—to view the content that they post. In addition, users can specify which of their sources they trust, and can filter their feed based on accuracy assessments provided on posts by their trusted sources. This platform gives users enhanced control over their feed, allowing them to block misinformation and leverage their social circle to highlight quality and accurate content to them.

Due to the major changes to the data model and user experience that our design ideas necessitate, we were compelled to design and build a system from scratch as opposed to piggybacking on an existing platform [36], which permits only surface-level changes to user interactions.

Our platform's client is written in Vue.js[2] and connects to a Node.js server that we developed[3]. The server interfaces with a MySQL database and a Redis server.

5.1 Features

Trustnet is modeled on the popular content-sharing platforms such as Facebook and Twitter. Users, who we refer to more broadly as *sources* since we also represent entities such as news organizations that publish content, are able to *follow* other sources. Users can *post* content, both content they author themselves within the platform, as well as content on other sites specified by a URL. The platform constructs a *news feed* for each user, consisting of content posted by any source that the user follows.

On top of this typical social network structure, we added our special affordances as described below.

5.1.1 Sources and Relationships. Accounts managed by individual users and proxy accounts for news publishing entities are both considered sources in the platform. In addition to the usual *follow* relationship, the platform provides a *trust* relationship. Both relationships have binary values. Following another source lets a user see what the source posts, as usual. In contrast, trusted sources can be leveraged for filtering articles in one's news feed. For instance, a user can choose to see only the articles that have been confirmed as accurate by their trusted sources. Trust and follow relationships are independent from each other; therefore, it is possible for one source to follow but not trust or to trust but not follow another source.

Both following and trusting are asymmetric relationships, meaning that a source can follow or trust a source without being followed or trusted by that source. A source's followers are publicly displayed on the source's profile. However, who a source trusts is kept private. Even the trusted sources will not know whether or by whom they are trusted. While public trust relationships could serve as a form of endorsement, which may especially be useful for news publishing entities or journalists, we made the design decision to keep them private to the user to ease any social pressure on users of marking sources they do not actually trust as trustworthy. This decision is tied to Study

[2] https://github.com/farnazj/Trustnet-Client
[3] https://github.com/farnazj/Trustnet-Backend

1 in which many participants reported that a source of unreliable content in their feed is their friends and family.

The platform provides the ability for each source to construct arbitrary *source lists*, or collections of sources. A user can use accuracy assessments given by the sources within each list in a fashion similar to trusted sources when filtering among the articles in their news feed. For instance, a user may use source lists to group sources according to the topic on which they have expertise (e.g., healthcare) and filter their feed using accuracy status and article tags (described in 5.1.5) to view healthcare related articles confirmed as accurate by those sources of theirs who are expert in healthcare. Source lists are private to the source that has created them.

Figure 1 shows the Discover Sources view of the platform where users can browse among the sources they are not following. Blue source cards are sources that are managed by the system, for instance, those that have RSS feeds associated with them, and the lime cards represent individual users. The user can edit their trust or follow relationship with the sources they follow in the Manage Sources tab. Figure 2 shows a source's profile.

5.1.2 RSS Feeds. To simplify the process of integrating assessment into normal news-reading activities, the system makes it possible to "wrap" *RSS feeds*—machine readable listings that many news sites publish to announce news articles as they are published—as sources in the system. Each source account can have multiple linked RSS feeds. The system periodically checks the feeds for updates using an estimate of each feed's update rate which is regularly adjusted. Once the contents of a feed are fetched, the system treats the announced articles as posts by the source.

The platform provides a UI for users to add RSS feeds to the system. When an RSS feed is submitted, the system searches among the existing sources to find if there is a source that may own the feed. If no such source is found, it creates a source with the feed credentials and associates the feed to the source. A user may then follow this source to view the contents of the feed.

5.1.3 Accuracy Assessments and Sharing. Our platform helps support users' need to provide and receive fact-checking information by enabling them to provide accuracy assessments of posts. In assessing a post, a source may assert that the post is accurate or inaccurate and provide their reasoning. Figure 3 shows the UI for submitting an assessment of a post. Any post on the platform is shown with indications of whether it has been assessed as accurate, inaccurate, whether there is disagreement on its accuracy, whether its accuracy has been questioned (discussed more in 5.1.4), or with no credibility indications at all if the post has not received any assessments. See figures 2 and 6 for examples. The absence of indicators by itself is a clear signal that the credibility of the post is yet unknown.

For our prototype we made the aggressive design choice of *requiring* a user to assess any post before sharing it. This decision was motivated by related work that reports that requiring assessment reduces the frequency of inaccurate posts [45]; however, optional assessment is another interesting part of the design space that could be explored.

Users may choose to assess posts *without* sharing them. This model contrasts with that of the existing social media where any engagement with a post, including a comment that refutes it, can disseminate the post further. When sharing posts, users can choose the sources or source lists to which they wish to limit their audience. The default target audience for a shared post is any source that follows the sharer. When a user shares a post, their status as a sharer of the post is added to it as meta-data. This is akin to a retweet in which the original post appears along with the original poster, with an indication of who retweeted it. Therefore, the context of who had originally posted the content is never lost.

Users can import articles from other websites into the system via their URL. If a source corresponding to the originating website does not already exist in the system, the system creates that

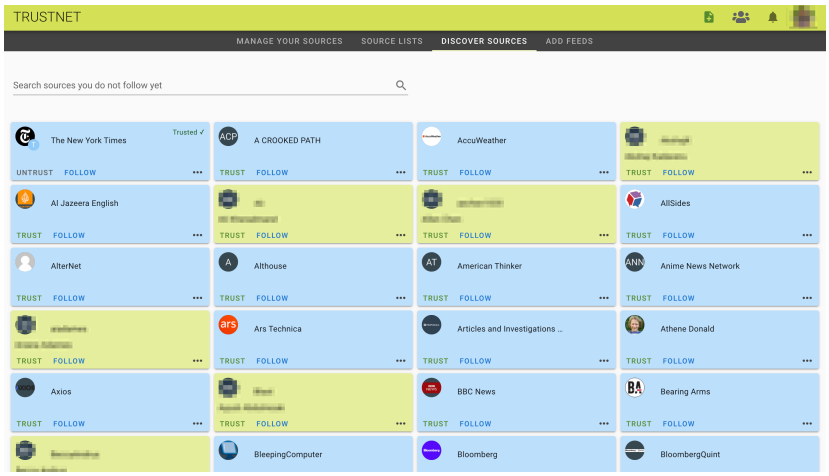

Fig. 1. The Discover Sources view where the user can see the sources they are not following. The sources that are managed by the system (e.g., those that have RSS feeds linked to them) are differentiated with a blue color. The sources with lime cards are managed by individual users. The avatars of the sources that the user trusts are marked with a "T" badge–e.g., The New York Times in this screenshot. Clicking on a source's avatar redirects the user to the source's profile.

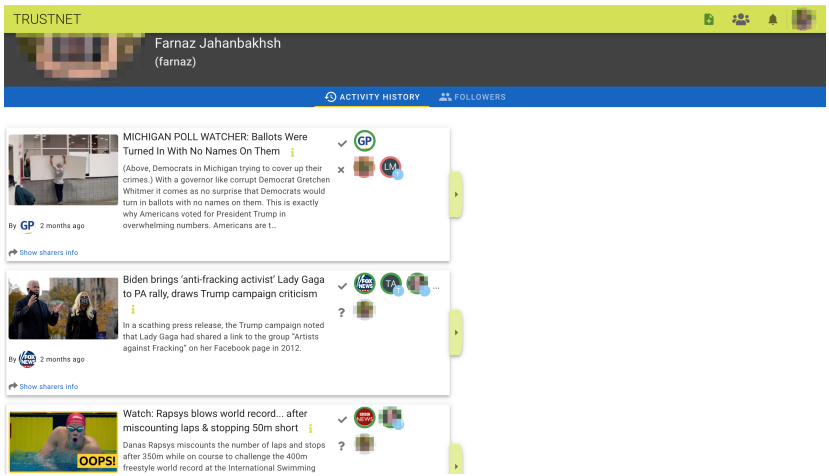

Fig. 2. The profile view of a source which contains all the posts they have assessed, asked a question about, or shared. Each post tile contains the post's metadata in addition to the assessments given to the post by the source whose profile is being viewed as well as the followed and trusted sources of the logged in user, and the post's original author. The source's followers can be viewed in the Followers tab.

source and associates the post to it. By importing an article into the system, the user is in fact sharing it and therefore, needs to assess the article.

Additionally, users can write posts of their own and share them with others. A post written by a source is automatically assessed as accurate by the source. Any assessment however, can be changed. Earlier versions of an assessment are kept and are available to the public.

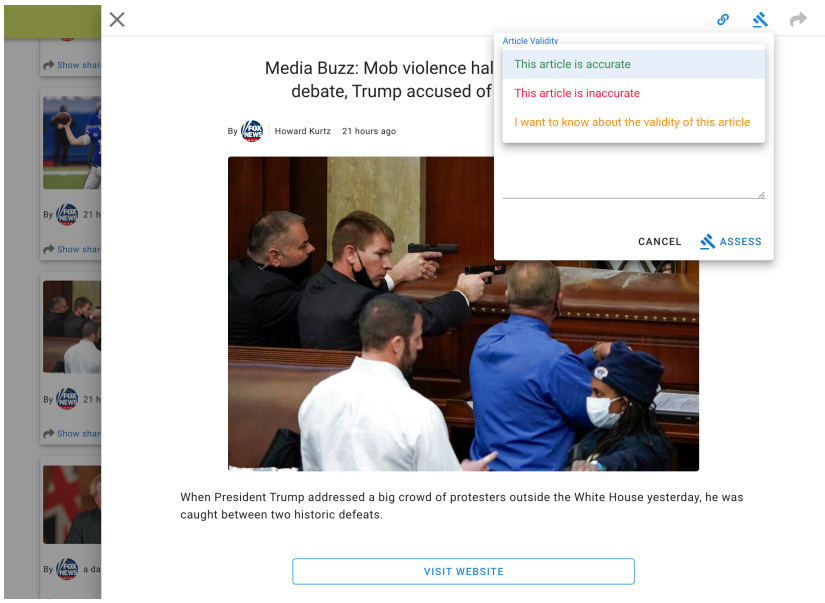

Fig. 3. The UI for assessing a post. When asserting that the post is accurate or inaccurate, a rationale is required. Assessing is required before sharing is enabled.

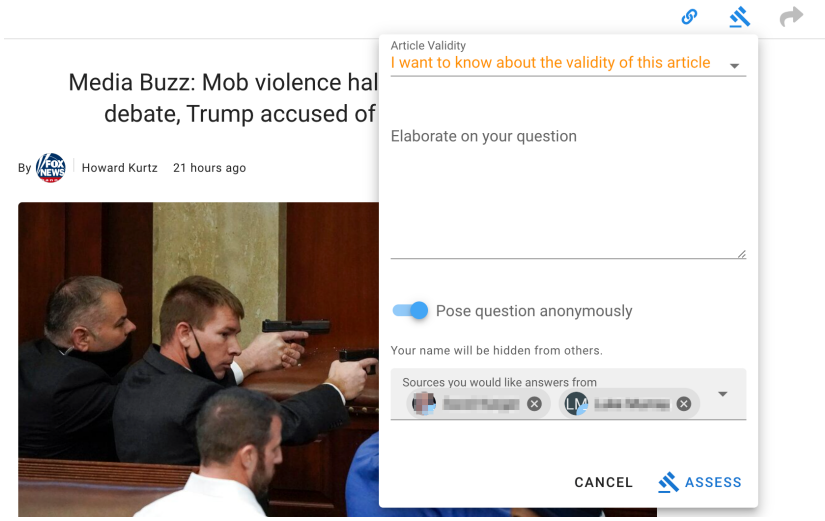

Fig. 4. The UI for inquiring about the validity of a post. The question by default is routed to the user's trusted sources. However, the user can customize to which sources their question should surface.

5.1.4 Inquiring about Validity of Posts. As discussed in Study 1, users may wish to share an article to learn whether or not it is accurate. Thus, a third option in accuracy assessment is to *inquire* about the validity of the post. [4]

[4]In a later update of the platform, we changed the word "validity" to "accuracy" to avoid user confusion.

Questions about a post's validity can be posed anonymously or with one's name attached to the question. By default, a request for assessment is surfaced to all sources that user trusts. However, a user can also choose to notify specific sources of their question. A source who has been asked a question receives a notification about it. A source is also notified when their assessment request is answered. The requester can then view the post in order to see any assessments by sources they follow or trust. The reason we show a user the assessments of those they follow as well as trust is so that the user can decide over time if they want to trust any of their followed sources and expand their trust network.

A user can view questions about the validity of a post posed by any source regardless of whether the user follows or trusts the source, provided that the source has not specified that their question be asked of specific sources.

Figure 4 displays the UI for submitting one's question about a post.

5.1.5 Filters. Users view content shared by all the sources that they follow. The platform provides a number of filters that users can use to narrow down the articles presented to them. The filters work in conjunction with each other. The *Validity* filter is used to determine the validity (accuracy) status of the articles that the user wishes to be shown and the *Assessors* filter determines whose assessments the filtering process should take into account.

Articles' validity status includes *Confirmed*—articles that have been marked as accurate in all the assessments they have received from the specified assessors, *Refuted*—those that have been marked as inaccurate in the assessments they have been given by the specified assessors, *Split opinion*—articles that some of the specified assessors have marked as inaccurate and some as accurate, and *Questioned*—articles about whose accuracy the specified assessors have inquired. These validity filters meet the needs discussed in Study 1 (see e.g. Table 1): some participants wished to see only fact-checked information, others wanted to peek at information that was deemed inaccurate and understand why, and others wanted to help their friends in their quest for the truth when they ask for fact-checking information.

The *Assessors* filter lets the user choose whose assessments to consider: the sources that the user follows (*Followed*), those that the user has marked as trusted (*Trusted*), the user themselves (*Me*), and a checklist set of individual sources or source lists that the user can construct. Any number of these assessor groups can be selected to use for filtering.

The *View Status* determines whether to filter among the articles that the user has yet not seen, those that the user has seen, or both. When a user scrolls past an article, the platform considers the article as seen by the user and therefore, the article will no longer be in the user's *Not Seen* feed the next time they visit their feed. However, if a source that the user follows or trusts posts a new assessment on or question about the validity of the post, the post will be returned to the user's *Not Seen* feed.

Additionally, users can select tags (e.g., Politics, Coronavirus, or Donald Trump) so that the filtering process filters only among the articles with the selected tags. These are the tags that are fetched from an article's originating website if the website has provided them; they are associated with the article at the time of its insertion into the system.

Figure 5 shows the UI of a user's homepage on the platform. Filters are selected from the sidebar on the left and previews of articles satisfying the filters are shown in the middle. Figure 6 displays the expanded assessments pane for an article which contains the assessments given by the sources the user follows or trusts.

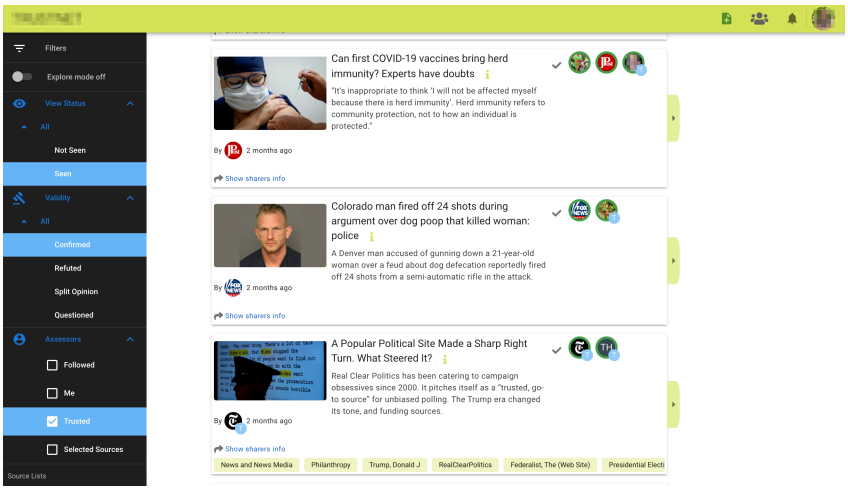

Fig. 5. The homepage view with articles filtered according to the filters on the left sidebar. Articles can additionally be filtered using tags (e.g., those on the New York Time's article in this screenshot).

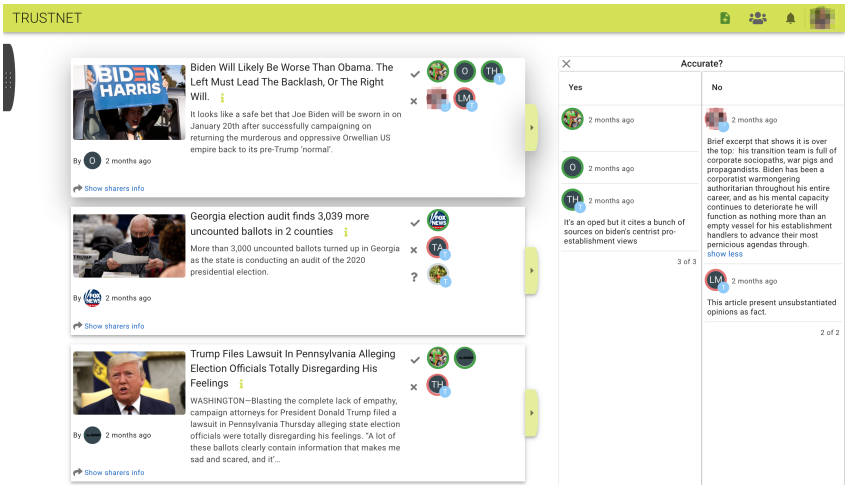

Fig. 6. An article tile shows a preview of which of the user's followed or trusted sources has assessed the article as accurate, inaccurate, and who has inquired about its validity. The assessment pane can be expanded to show their full assessments.

5.2 Field Study

We conducted a field study to understand users' perceptions of our social content sharing platform and the features that it offers relate to content accuracy. Because fact checking and the recommendation of posts through sharing can best be leveraged among members of a social circle, in recruiting participants, we asked that users join the study with at least one member from their social circle. We advertised the study by reaching out to those participants from Study 1 who had indicated interest in joining a follow-up study of a social content sharing tool. We also advertised through word-of-mouth, mailing lists, and on Facebook and asked that our Facebook post be reshared. Some of our participants joined the study in pairs and others belonged to larger groups some of whom

joined together and others with a chain of referrals. Our participants' social groups consisted of siblings, spouses, parents and children, friends, and more distant relationships as a result of a chain of referrals (e.g., a friend of the spouse of the child of a friend).

We asked participants to view a short video tutorial of the core features of the Trustnet platform before the study and to each sign up for an account on the platform. Their daily tasks for the one week period of the study consisted of spending at least 15 minutes a day on the platform, sharing at least 2 posts every day, and checking if someone has asked for their assessments. We encouraged them to help by providing their assessment. Every few hours at random points in time, participants would receive a prompt on the platform reminding them about how many posts they had shared and assessed in the past 24 hours if they had not yet accomplished their daily goals. The reasons we set daily goals for participants were twofold: to ensure that they would have enough interactions with the system and possibly other users' posts to form an opinion about the utility of the tool and because per IRB requirements we needed to be transparent about what amount of contribution would be required for receiving the compensation.

We wished to investigate how participants use the system to engage with false, misleading, or biased posts from sources with a political orientation similar to or opposing theirs as well as accurate posts from sources with an opposing viewpoint. Such posts are likely to make their way into participants' feed on social media where their extended social circle is present. However, having only a limited subset of their social circle on the platform, we needed to increase their likelihood of encountering such posts via other means. We therefore created a source named "Trending News" that we asked our study participants to follow. The Trending News source was managed by one member of our research team and each day imported and shared a number of news articles ($\mu = 9.29$, $\sigma = 1.36$) from a variety of news sources across the political spectrum and with different degrees of credibility (bias and factual accuracy in reporting).

The articles presented to participants in their feed were ordered by a combination of how recently they were published and how many sources in the whole system had recently shared the articles. In ordering articles, we took the shares from all the sources in the system, rather than only the sources a participant followed or trusted, into account because the activities of the limited social circle of the participant would be too sparse to yield a useful ordering. However, for a post to actually appear in a user's feed, the user needs to be following a sharer or the original source of the post in the first place.

After the study, participants completed a survey that inquired about participants' use and the usefulness of the filters, how participants decided what sources to follow or to mark as trustworthy, whether and what changed their mind about trusting a source to share reliable content, whether they encountered unreliable content and whether and how they tweaked their feed or relationship with other sources to see fewer such posts, their sharing practices, how assessments of other sources helped them, and how they believed their assessments helped other sources. In addition, some questions captured their perception of the tool, its strengths and weaknesses, and their ideas for improving it. The full set of survey questions is included in the Supplementary Materials.

5.3 Participants

A total of 14 users participated in and completed our user study. To sign up, participants completed a survey which asked what sources they consumed so that we could add their RSS feeds to the system, with whom they were planning to join the study and their relationship to their study partners, and demographics questions similar to those in Study 1. Another 7 participants signed up but dropped out either at the onset or in the midst of the user study. Of the 14 participants who completed the study, 8 were male and 6 female. One identified as Republican, 2 as Independent, and the rest were Democratic. The median for age was 33.5 (ranging from 24 to 66); for highest degree

received, Master's Degree (ranging from some college no degree to Doctoral degree or M.D./J.D.); and for income, $100,000 to $149,999 (ranging from $20,000–$29,999 to $150,000 or more). With respect to ethnicity, 9 identified as White/Caucasian, 2 as Asian, 2 as both White and Black, and one as other. The demographics of the 7 participants who dropped out were similar to those who stayed, except that the median for age in that group was 61. They included one who identified as Independent and 6 Democrats.

Participants were each compensated with a $30 Amazon gift card at the end of the study.

5.4 Results

Throughout the user study, participants revisited who they wished to follow and trust. The final snapshot of their relationships is shown in Figures 7 and 8 in the Appendix. The 14 participants did not belong to the same network of trust. We also observed that trust between pairs of users in our study were often (but not always) bidirectional.

Participants posted content to the system, which is not surprising since this was required by the study conditions. However, participants also made interesting use of many of our implemented affordances, which provides some evidence that they considered them both usable and useful (or at least interesting) as we discuss below.

In the sections that follow, where we present quotes from participants, we identify them with strings of the form "p-participant identifier".

5.4.1 Participants Were Capable of Assessing Posts. Over the one week period of the study, participants produced 235 assessments, 66 questions, and 209 shares (213 assessments, 59 questions, and 184 shares from those who completed the study).

Although we did not train participants in how to assess articles, the majority of the assessments they posted (N=212) indeed evaluated the veracity of the article based on various rationales. The rationales that they cited in support of their accuracy assessments indicated that users are capable of assessing content and can be recruited to help out each other. Sometimes participants based their assessment on their expertise or firsthand knowledge of the subject. The source of the article was also cited as a signal of its credibility. Sometimes in their assessment, participants referred to other trusted sources that had reported on the topic. In some instances, participants investigated the sources, for instance, authoritative government websites or scientific journals, mentioned in an article and reported on whether the content of the article accurately represented the content of the sources: *"The research credentials of WHOI are substantial and the study to which this article refers has been published online via peer-review (https://www.annalsofglobalhealth.org/articles/10.5334/aogh.2831/)" (p-18)*—Headline: New study takes comprehensive look at marine pollution (Woods Hole Oceanographic Institution)

In some assessments, participants evaluated the presentation, the logic, or the language of the article. Another rationale that participants cited in arguing for or against an article's accuracy was that it was (in)consistent with their past experience and observations. All of these rationales are in alignment with those reported by Jahanbakhsh et al. in [45] for reasons why people believe or disbelieve news claims. We have included some of the examples of the rationales that we observed in our study and the types of claims our participants chose to assess in Table 2 in the Appendix.

We also observed that participants used the question feature as intended, to inquire about the veracity of articles and receive answers on them. For instance:

"Yipes – I have a half dozen oysters in the refrigerator picked by a neighbor directly off the beaches of Cape Cod. Planned to slurp tonight (as I have done for many years with no ill effect). Always a risk with raw seafood, of course, but I'd really like to know how MUCH of a risk it is, and if there's any

way at all to avoid it (since I REALLY like raw oysters...)." (p-18)—Headline: Is it dangerous to slurp oysters straight from the shell? Japan's Ministry of Health weighs in (SoraNews24 -Japan News)

A participant with an M.D. degree assessed the article as inaccurate because: *""The reason is because what causes food poisoning isn't bacteria but actually a virus — the norovirus to be specific." This article makes it seem like food poisoning from oysters is only caused by norovirus which isn't true. Claims from Japan Ministry of Health is taken out of context." (p-14)*

5.4.2 Different Participants Used the Filters Differently. Participants reported using the homepage filters in different ways. One participant mentioned that they configured their default feed to consist of only articles confirmed by their trusted sources, protecting themselves from misinformation, but sometimes changed it to explore—*"[my filters setting was] trusted, unseen. But sometimes I turned off all filtering to see what was around." (p-18).* Others set their Validity filter to show *all* articles, including those that were refuted, out of curiosity—*"Validity: All, since I read the news out of curiosity, and that curiosity extends to what people found that they didn't think was valid." (p-3).* Others indicated they specifically sought out controversial articles—*"I used the confirmed, refuted, and questioned filters sporadically. I found refuted of most help to find content that I normally don't see. The questioned filter was more for seeing what interests some of the people I followed had." (p-7).* A caveat however, was that since users had only a few trusted sources, such articles were few—*"Specific source filters were useful when I wanted to see what trusted sources were sharing. I was curious to see what articles had controversial opinions and I used the "Split Opinion" filter sometimes, but I ended up not finding it as useful mostly because I just didn't see many articles there. This may change with increased activity though." (p-14).*

In summary, when given the opportunity, users chose to configure filters to control the delivery of information into their feeds. While no one filter setting for accuracy appealed to all users, many users found particular filter settings of value to them—supporting our design idea of placing more filtering control in users' hands.

5.4.3 Learning to Trust. The structured trust affordance in the platform captures the trust relationships that users have already developed. However, it is interesting to understand whether the emphasis on accuracy and credibility and the available assessment history from everyone encourages users to rethink their trust relationships, either by doubting someone they previously deemed trustworthy or by adding someone new to their list of trusted sources.

We observed that although our participants set their trust relationships with various sources according to their pre-existing familiarity with them, they examined assessments from unknown sources to determine whether they should extend their trust to these sources. In the post-study survey, some participants stated that they followed certain sources to see content and assessments from them and later adjust their network of trusted sources: *"[I followed sources I did not trust because] I was curious to see whether I could assess the trustworthiness of a source previously unknown to me just by reading its coverage of current events and trying to apply the criterion of asking whether they are reasoning based on evidence." (p-20).* Conversely, in one case, a validity question from one participant led another to decide the first was not trustworthy:

"Knowing that COVID numbers can/have been exaggerated, it makes me wonder about the numbers stated here. Also, if some people are stepping forward simply for the money." (p-19)

*"In response to Anonymous: * How do the # of positive Covid tests come into play here? Different process, different situation entirely. * Exactly how many of the people publicly claiming to have been sexually abused as a boy are you accusing of lying? Would it make the number seem less horrific if 40,000 men are telling the "truth" and 50,000 are "lying"? btw, reveal your name so I know who I shouldn't trust." (p-8)*

Headline: About 90K sex abuse claims filed in Boy Scouts bankruptcy (The Associated Press)

Our prototype captures the full history of a user's assessments and does not permit them to be deleted. This record will be of value in determining whether to trust someone. Someone who rushes to (wrong) judgements about content accuracy cannot disguise this by updating their assessments over time; their early mistakes will be visible to users who wish to determine who to trust.

5.4.4 Reported Strengths. The most cited strength of the tool in the post-study survey was that the tool brought the veracity of articles to the foreground (N=8). Participants valued that articles had to be assessed before they could be shared: *"News vetting in order to share was the primary strength (and the exposing of the reasoning). I spent a lot of time researching stories and I feel that the system really makes that worth it." (p-11)*, that they could view other people's assessments: *"The most interesting articles to examine were those that had differing judgements from the Trustnet users. I didn't notice many of them, but it was helpful to read what had been written." (p-18)*, and that they could ask others about an article's validity. A number of participants (N=6) appreciated the ability to make the inquiry anonymously—*"Asking about the validity of an article is something I'd rarely consider doing publicly, but being able to do so anonymously as part of sharing the article was great.(p-3)"*; however, one participant believed that *"Anonymous is used for trolling mostly." (p-8).*

Users also appreciated being able to curate their feeds using their trust relationships with other sources (N=2): *"The obvious features were the trust-based personal and media feeds. I don't think you're leveraging this enough right now, but they could theoretically produce a more focused news feed. It would be nice to have an alternative to Facebook for sharing stories with friends." (p-4)*,

These comments provide additional evidence for the usability and usefulness of our proposed design affordances.

5.4.5 Reported Weaknesses and Areas for Improvement. Many of the cited weaknesses of the tool (N=8) concerned the particulars of the UI for instance, how it appeared on mobile or the ordering of articles. A number of participants (N=4) wished for the platform to have a channel for discussions that were not limited to assessing articles:*"I also felt like it didn't need to be always so assessment-focused as people sometimes just want to exchange ideas/inspirations from articles" (p-14).*

Two of the cited weaknesses were unavoidable consequences of our small research deployment. First, the small user base meant that there was not enough activity to curate interesting feeds for users (N=7)—*"[One weakness is] lack of critical mass in the user base. Critical mass would be most effective if the user base included the journalists whose reports we are reading in the media. It would be especially helpful and interesting to see their assessments of each other." (p-20).* A second problem was that our system lacked the sophisticated tracking and profiling history that lets the large platforms choose content matching a user's interests—*"It's harder to find interesting stories due to the lack of aggregate peer vetting. I can add sources as RSS feeds but there's no easy way to filter those feeds to only show interesting things. This may be fixed by having a critical mass of people with similar interests that I followed, but compared to reddit or twitter I felt like I was struggling to find diamonds in the rough. That being said, the idea of vetting before sharing was super interesting and it's something that all platforms should adopt." (p-11).* A combination of our design elements with the automated feed-ranking of the large social platforms could address this complaint, with users manually filtering for accuracy *after* the platform has filtered for a user's interests. Another weakness was that accuracy assessment in its current form is not generalizable to articles with multiple claims, satire, opeds, or those that have slants, which we elaborate more on in Section 5.4.6.

One participant raised skepticism related to the approach and not the design. One concern they had was that accuracy assessments may not make a difference to the people who fall for misinformation—*"The things I saw marked as "not true" were all extremist pap. When an untrustworthy source gives oxygen to conspiracy theories, it goes without saying it's trash. I'm not sure how to capture a more nuanced reaction, but if there has to be factual errors in an article for it to be marked "wrong,"*

then the sources who chronically misreport will fall into that category very often without the tagging making any difference." (p-18). They also speculated that assessments may not be helpful in the context of published editorial material—*"it didn't seem helpful for published editorial material. There seem (to me) to be enough 'trusted' sources in my life already (sources I've come to trust over years of reading), so I'm less needful of having someone I trust tell me they also trust something from a source I already believe" (p-18).* Interestingly however, while none of our participants refuted any post from the news publishing sources they had marked as trustworthy, we found a few instances of them asking questions about the validity of such articles—*"I ask about validity because I'm always frustrated reading about a source described as a "U.S. official who was not authorized to speak publicly." What are we to make of this? I believe the reporting is true, but how do we rate confidence in a source so described?" (p-18).* Interestingly, this quote is from the same participant who had concerns about the usefulness of assessments for content from the sources they already trust.

5.4.6 Assessment Complications. To understand the problem points of our design, we also looked for instances of conflicted or unexpected uses of the different features that we offered. Understanding the shortcomings or the existence of needs that were not addressed can help designers provision concepts that fulfill them.

5.4.6.1 Lack of Discussion. We observed that participants sometimes used the assessment functionality not to assess content accuracy, but rather to convey their sentiments about an article to to paraphrase the contents of it (N=24). In a few other instances, the validity question feature was used to respond to an assessment or a question raised by someone else. The repurposing of assessments or questions for other types of discussion may have been due to the absence of a salient placeholder for such discussions as we had prioritized incorporating assessments into the tool. Indeed, in the post-study survey some participants mentioned the need for supporting discussions: *"For conversational threads, I found that having discussions only possible under "Questioned" assessment a bit limiting." (p-8).*

5.4.6.2 Headlines versus Content. Another occasion that gave rise to conflicted usage of assessments was when an article's headline misrepresented its content (N=9). The disconnect between the article's content and its headline resulted in some participants basing their accuracy evaluation on the content, others on the headline, and some explained the inconsistency in the guise of a question about the article's accuracy. The participant assessments below demonstrate this inconsistency, with the first one being based on the content, and the second, on the headline:

"Despite the bait-clicky [sic] headline; this article does not fall into the trap of claiming that space aliens have sent out signals that have been detected on earth. Instead; it is a compellingly written description of what scientists who look for such signals have been detecting and what those findings might mean." (p-6)—Headline: Alien hunters detect mysterious radio signal from nearby star (National Geographic)

"The article's title is worded to evoke a strong emotional response. The contents of the article contradict the assertion that there is no Thanksgiving surge. The article cherry-picks a few states where there was an active surge leading up to Thanksgiving and have seen a slight decline afterwards. But tbh the article also presents states that have seen a higher surge since Thanksgiving than the 2 weeks preceding. This contradicts the title. btw; they have a survey embedded about people's support of a mask mandate. The results show 68% strong support mask mandates with only 19% strongly opposing them. " (p-8)—Headline: After Weeks Of Holier Than Thou Howling, The 'Thanksgiving Surge' Is A No Show (The Daily Caller)

5.4.6.3 Un-Assessable Posts. Sometimes participants appropriated veracity assessments or questions on some articles to explain that the articles were in fact not factual, but rather op-eds or

satire, suggesting the need for such posts to be identified and characterized separately from factual news pieces, e.g., by enabling users to label them as such: *"Pieces of this article are largely based on opinion. Not sure if that makes them true; or if that piece is irrelevant." (p-19)*—Headline: 4 Takeaways From Biden's Electoral College Victory (The New York Times)

The inconsistency in participants' accuracy assessments of non-factual content pieces also surfaced in participant responses in the post-study survey: *"There did not seem to be consensus about what validity is. Someone posted a fictional webcomic he liked and said it was valid. Someone posted an Onion article and some people marked it as valid because its falsehood was consistent with The Onion's mission. I thought it was enough to check the validity of the article's main claim, but [Friend's Name] seemed to think you had to ascertain the validity of every claim. I don't think either interpretation would make sense for long articles that make multiple complicated claims." (p-3)*

Another point of confusion about how to assess the accuracy of articles was in cases where participants deemed the content of an article accurate but its intended message misleading or its language biased or its publishing source not generally trusted by them:

"This article reports on statements supposedly made by Tulsi Gabbard; a member of Congress from Hawaii. The tone of the article certainly sounds like reasonable journalism. But this appears in Breitbart; which gets very low ratings by the organization NewsGuard. So I'm inclined [to] doubt it." (p-6)–posted as a question

"Is it true that Gabbard made misleading and inflammatory statements? Yes; it is. Should we hold Breitbart culpable for reporting this conspiracy theory as if it has any credibility? Yes; we should; but editorial slant is hard to capture in a yes/no" vote on an article.""(p-4)–posted as an assessment marking the article as accurate

Headline: Tulsi Gabbard: "Heartless, Arrogant, Unelected CDC Bureaucrats" Giving Vaccine to Healthy Americans Before Elderly (Breitbart)

Some of these issues, dealing with humor and opinion, may have been avoided if we had made accuracy assessment optional instead of required.

5.5 Summary

In summary, participants understood the new affordances that we designed into the system, made effective use of them, and spoke positively of their value. They made proper accuracy assessments of content they posted and of some they encountered from others. They valued the opportunity to make these assessments and appreciated receiving them from others. They curated a trust network based on their familiarity with other users as well as other users' assessments. They leveraged filters, choosing a variety of different settings, to control what kinds of accurate, inaccurate, or questioned information was allowed into their feeds. Many of them spoke positively about having the power to do these things. We also discovered certain assessment challenges that demand further design for certain difficult-to-assess types of posts. In the Discussion, we draw from our findings and the related work to propose design changes that can address these challenges.

6 Discussion & Future Work

Motivated by prior work on misinformation [7, 38, 68], information seeking [100], source credibility and trust [61, 64, 99, 103], and users' wish for more agency over their social media feeds [27, 104], in this work we designed user affordances that can empower social media users in their fight against misinformation: (1) allowing accuracy assessment of posts in structured form as part of the data model, (2) enabling users to indicate which users or sources they trust to assess posts, and (3) providing filters that users can configure to block posts from their feed based on the accuracy status of posts assessed by their trusted sources.

We wanted to understand the potentials of these affordances for facilitating users' current practices in dealing with misinformation online and whether user are receptive to them. Therefore, we first conducted a survey study of 192 people. This study revealed that many users view online content with caution and ask their social circle or are asked by them to provide assessments of content accuracy. Some even expect their social network to proactively correct them should they post inaccuracies. However, because social media platforms do not have designated metadata for assessing accuracy of posts, participants use a variety of features intended for other purposes, such as likes, comments, and shares that can be picked up by the platforms as signals of engagement or misinterpreted by other users. We also uncovered a diversity of preferences in whether and how users want unreliable content to be presented in their feed. Some users would like to take the misinformation out of their feed but continue following unreliable sources, and others would like to keep the unreliable content in their feed along with signals of the content's inaccuracies. These wishes suggest the need for filters that empower users to take control over their feed rather than simply surfacing content chosen by opaque social algorithms [15, 104].

To gain more insight into how users would use these affordances if given the chance, we developed a social content sharing and reading platform that offered the affordances and we then conducted a user study on the platform. The platform enables any user to assess news, specify which sources they trust, and filter their news feed based on accuracy assessments provided by their trusted sources. Consulting prior work that reports providing accuracy assessment as well as rationales before sharing can shift users' attention to accuracy and away from social feedback [45, 83, 84], the platform allows users to share a post only after they have assessed it. The filters that the platform presents enable users to narrow their feed down to posts with a certain accuracy status, for instance, verified or refuted, assessed by their trusted sources. The platform gives accuracy status of posts salience to counteract the effects that social engagement metrics can have on perceptions of content credibility [7, 73]. Our prototype shows how assessments could be structured, aggregated, and directly incorporated into the UI so that users can view and understand them at a glance and filter their feed based on them.

In our user study, we asked a set of users to use the platform to read and share news over the course of a week. We found that participants were capable of providing assessments. In addition, they perceived value in the salience of accuracy signals and the ability to ask questions about the accuracy of a post. We also observed that different users found different filter settings useful. Most of the weaknesses they saw in the tool concerned the particulars of the UI and the lack of a large user base. These grievances were expected as our tool was a research prototype with limited scope and scale, for which recruitment was especially difficult because participants needed to sign up as part of a group that already had connections and commit to performing tasks over an extended period of time.

In summary, this work contributes: 1) A broader understanding of how social media users deal with misinformation in their feed, how they seek the help of each other in this effort, and what is lacking in the platforms to support user needs 2) The design of user affordances that give a social platform's users greater agency to protect themselves and their social circle from misinformation 3) empirical understanding of how users would perceive and use these affordances; and 4) design implications for platforms and researchers based on our empirical observations.

We devote the rest of the Discussion to implications for deploying these affordances on the platforms, or more broadly, on the web.

6.1 Untrained Users Are Capable of Accuracy Assessment

We observed that although we did not train participants in how to assess articles, they assessed posts based on various rationales including firsthand knowledge, expertise in the subject, referring

to other trusted sources that had reported on the topic, examining the logic or the language of the article, the credibility record of the source, etc. A question that may arise is whether our participants were able to provide assessment while being untrained because they came from a particular user population (with rather high income and high education). This question has implications for whether our results would generalize to a setting where platforms made these affordances available to the public. The rationales that our participants reported were indeed in alignment with those reported by prior work [45] in which a large sample of users were asked to provide explanations for why they believed or disbelieved various news claims. The sample in the said study had a lower median income and education compared to the one in our study. Yet while being untrained, those participants were able to not only provide rationales for their accuracy assessments but also reliably label their own rationales according to a taxonomy of rationales provided to them by the researchers. Drawing from the similarity of rationales across the two studies, we believe a platform such as ours that incorporates our proposed design affordances can be used by and help at least some part of the general public. Nevertheless, it is possible that the distribution of rationales, and therefore their strength, differs with demographics factors such as income or education. To help users better evaluate the strength of their assessments as well as those from their social circle, when asking for assessments, platforms can augment the process of capturing assessments by presenting the checklist of rationales reported by Jahanbakhsh et al. and signaling their strength, as described in 6.2.3.

Additionally, the affordances we propose do not require everyone to be assessors. It suffices that enough people perform assessments to support all the users who want them. Indeed, just a small number of widely trusted sources—journalists, fact checkers, etc.—could provide sufficient assessments for a large group of news consumers. Nevertheless, our affordances allow any user to ignore those popular assessors in favor of others they trust more.

6.2 Design Recommendations Related to Assessment

Most of the assessment-related complications in the user study arose because users had purposes without concepts provisioned in the system to support them [44]. Therefore, they overloaded the assessment or the request for assessment functionality to fulfill these purposes, resulting in sometimes inconsistent usage. Below, we describe recommendations for platform designers on how to design mechanisms that fulfill these concepts and issues that future work should consider.

6.2.1 Non-factual Content Pieces. While labels of accuracy may not be appropriate for op-ed and satirical pieces, their status as opinion or satire needs to be signaled to users as these pieces are often shared on social media where their non-factuality status is obfuscated or ignored by or unknown to users [110]. A mechanism to deal with this issue could be to crowdsource the labeling of articles as factual, op-ed, or satire. However, because as mentioned, recognizing the status of these articles may sometimes require a deeper knowledge of the issues at work and a high engagement on the part of the reader [90], not all users may agree on the labels assigned to a piece. The labeling of factuality status of content therefore, can leverage the use of trust networks similar to accuracy assessments in this work. Potential disagreements on the factuality status of a post can be presented in a similar way to how disagreements of assessments are currently shown. Future work could investigate the implications of this approach.

6.2.2 Articles with Multiple Claims. To deal with content that contains multiple claims, designers can allow for individual claims to be identified and assessed by users, inspired by annotation tools such as [1, 115]. For assigning an accuracy label to an entire article or a part that is composed of several claims, crowd-sourced summarization techniques such as the one in [113] can be used. Future work however, should study how to display and resolve conflicts in circumstances when

a user's trusted sources have disagreements on the boundaries of individual claims, the labels assigned to each claim, or the accuracy status of a collection of claims as a whole.

6.2.3 A Richer Taxonomy of Assessments. Our prototype explored a ternary categorization of accuracy: accurate, inaccurate, or questioned. Our users encountered situations orthogonal to this categorization: political slant, the article being misleading, and the (un)trustworthiness of the publishing source. Some of these were discussed in prior work as other dimensions of credibility [45]. As an attempt to split up the overloaded concept of accuracy [44], future work can explore using the broader, but still quite small, accuracy scale developed in [45] while adding the other dimensions of credibility encountered in our user study. The taxonomy in [45] can also be used on platforms as a checklist set of rationales that participants can select upon assessing a post as it may serve as useful metadata based on which users could filter their news feed. Using this taxonomy, users could choose to view for instance, only those articles that have been verified by a trusted source who declares they have firsthand knowledge or specific expertise.

6.2.4 Inconsistency Between an Article and Its Headline. As it is the headlines that users first encounter and what can entice them to click through articles, manipulation techniques such as clickbait, sensationalism, and even incongruence between the headline and the message of the article are increasingly applied to headlines [34]. Such headlines have the potential to harm because even if their associated article ultimately presents accurate information, users often simply skim headlines or do not read articles in full [67]. Even if they do read the article, they are more likely to retain the message of the headline even if it disagrees with the article [55, 56].

Our users encountered examples of inaccurate or misleading headlines that linked to accurate articles but were unsure of how to assess them. To address this issue, platforms can augment the accuracy assessment categories with a "misleading headline" category so that the headline's veracity, and even its sensationalism can be captured independently from the article. Another approach, employed by the Reheadline browser extension, is to allow assessors to fix the problem by proposing more accurate headlines for such articles [46].

6.3 Filter Bubbles

Empowering users to block information they do not trust could conceivably lead to stronger "filter bubbles" in which a closed group of like-minded users mutually reinforce each others' perspectives while their trust filters shield them from any divergent views [53, 95, 102]. However, there exist other possibilities. As we observed in our user study, users do seek information that has been questioned or refuted, or shared by untrusted sources. This observation is in alignment with prior work that reports on individual differences in people's preference of and receptiveness to collections of news articles that confirm their views vs those that challenge their views [77]. Our design does not prevent users from seeking counter-attitudinal content; it simply puts them more in control. They can decide when they wish to remain safe in their bubble and when they wish to explore. This control and the safe space of a community of like-minded people that users can enter and exit whenever they wish, or *epistemic respite*, can help prepare users to re-engage with differing views [6]. Platforms could also implement features aimed specifically at puncturing filter bubbles. Consider a situation where the majority of a user's friends have assessed an article as true, and one as false. In today's networks, that outlier friend's disagreement would likely be invisible in a sea of one-sided comments. However, if structured assessments were captured, then a system could surface the fact that the article is disputed, and could highlight the outlier assessment and the fact of its being in disagreement with the rest. A user could then be directed to the specific trusted friend who might help break their filter bubble. It however, needs to be explored whether

friend-sourced credibility assessments can exacerbate inaccurate beliefs in certain communities, for instance, those characterized by extreme polarization.

The use of structured assessments could also empower a user to seek out *trustworthy* contrary perspectives. When a user encounters another who opposes their view on a particular article, an examination of that other user's (public) history of assessments on other articles would help provide a broader picture of this other user's judgement. Seeing that the users are usually in agreement could provide confidence that the other user is generally reliable, which could add weight to their rare disagreement. Indeed, given structured assessments the system could automatically recommend such opponents who would be deemed trustworthy. The hope that users will be receptive to such quality suggestions is rooted in prior work which argues that people exclude opposing viewpoints not out of aversion to other opinions but because they perceive less benefit in them. This thread of work reports on factors that lead people to favor counter-attitudinal information in particular settings [31, 32].

6.4 Burdening Users or Empowering Them?

Questions may arise about whether our proposed affordances place the burden of moderating content for a user's friends on the user. This burden could come in different forms including time, effort, and affective [47, 69]. Today's internet ecosystem is full of participants performing labor of high value without compensation, for instance on platforms like Wikipedia, because they view the work as a form of civic participation [16], or because it provides them with indirect value such a sense of accomplishment, community, or creativity [87]. Similarly, any comments made by a user on social media can be seen as both uncompensated labor and self-expression. Our studies revealed that many users do already enjoy using comments or other available means to assess posts for the benefit of their social circle. Our affordances simply empower users, in ways previously not possible, to choose to help their friends should they wish to help.

This empowerment comes as an alternative to ceding truth governance to platforms. Allowing commercial platforms running on ads and benefiting from user clicks to have an absolute power over what content can make its way to the information space and what becomes censored is perhaps not in the best interest of the users [37, 50]. Indeed, social media platforms have time and again made decisions to block content that arguably did not have potential to harm or content by dissidents in certain countries [2, 25, 43, 93, 96].

Although we argue that platforms should enable end-user content moderation, we believe they are not absolved of responsibility when it comes to content delivery. For instance, in addition to intervening in contexts that contain harmful behaviors such as toxic language or child abuse [58, 81], they need to be wary of increasing the visibility of content that is false or misleading, e.g., through "trending" or "recommending" it.

6.5 Leveraging Assessments without Compliance from Platforms

The adoption of our proposed affordances requires persuading platforms, perhaps through activism or legislation, to adopt a content curation model that is not as focused on increasing user engagement as the present. Until that goal is realized, we can explore ways to leverage the affordances without compliance from the platforms. For instance, the credibility cues and assessments implemented into the tool in our study can be offered via a browser extension that recognizes content via its URL as users browse various news websites. The extension could then allow users to assess and see the assessments of their trusted sources in-situ on the article. As social media posts also have their unique URLs, the extension could support the proposed features on social media platforms as well. Some of these platforms have made it notoriously hard for third party extensions to reliably read or edit content on them. However, extensions such as FB Purity that modify the Facebook

feed demonstrate that with sufficient effort it is possible to compete in this arms race. Future work can investigate the feasibility and implications of this approach.

7 Limitations & Future Work

7.1 User Sample

A potential limitation of both Study 1 and Study 2 is the demographics of our participant sample who were rather highly educated and had high income. The skewed demographics may have been a consequence of recruitment via snowball sampling. We also acknowledge that the scale of the user study was relatively small which is due to the fact that recruiting for longitudinal studies in which participants are required to sign up as part of a group that already has connections and commit to performing tasks over an extended period of time is difficult. However, while we cannot generalize that all potential users will find our designs useful, the subset of users that we studied did in fact find them valuable. Even if our affordances help a user population similar to the one we studied and their social circle—and do not impact other populations—they still have played a positive role in the information war. Our work can serve as a stepping stone for additional research that further investigates how our proposed design decisions are received in other user populations.

7.2 Topic-based Trust

While it is possible to believe a source generally well-informed and sensible in the assessments they provide, it is not always the case that a source's trust in the expertise of another in one domain, for instance healthcare, translates into a universal trust of that source in any domain, e.g., politics [70]. Related was one comment from one of our participants in the user study of the tool: *"Making the idea of trust explicit in one's assessment is valuable. However, there were only 2 levels of trust: yes or no. Trust is not usually so binary. I would like to see more levels of trust." (p-8).* Trust on our platform is not granular at the topic level, however, it can be approximated by using source lists the assessments of whose members can be used in a manner similar to a source's trusted list. Therefore, to view health-related articles that are assessed as accurate by a user's friends who are knowledgeable in healthcare, the user need only compile a source list of such friends, and choose to see articles confirmed by the members of the list. With trust differentiated at the topic level, posts on the platform should also be classified according to their topics. News publishing media often present tags along with their article, and these tags could be used as indicators of an article's topic. Our platform already allows users to filter posts based on such tags in addition to accuracy signals. However, a limitation is that these tags may be at finer or coarser levels of granularity than the topics about which users deem their trusted sources well-informed.

7.3 Propagating Assessment through the Trust Network

One limitation of a customized trust-based approach in filtering misinformation is that the set of a user's trusted sources may not be large enough or have the expertise to assess content on a variety of topics. To address this issue, future work can leverage trust transitivity to build a more extensive trust network for each user. Prior work has examined transitivity of trust in social networks, for instance in the context of e-commerce, recommender systems, or chat moderation [20, 49, 62]. With transitive trust relationships, when a user leaves a credibility assessment on a post, the platform can propagate that information to all the users that either immediately trust or have an indirect trust path to the assessor, while maintaining the assessor's anonymity. This chain of trust could help users benefit from more extensive assessments even though they may not immediately know the benefactors. Future work can investigate several interesting research questions related to the incorporation of trust networks. One question for instance, is how fast trust decays as the distance

of two sources in the network who are connected by an implicitly inferred trust relationship increases. Another is how assessments from different sources, some not immediately connected to the user, should be weighted, aggregated, and presented to the user in an interprettable manner. One scenario could be that each trusted source is given an equal weight in deciding the accuracy of an article. In another scenario, the user could decide that a particular source or rationale be given priority over others.

7.4 Misplaced Trust

Finally, we acknowledge that our proposed design is not a panacea for the misinformation problem. While we found in both the Study 1 and Study 2 that many of our participants would like to receive assessments from the people they trust, we acknowledge that some users may place their trust in sources that are not credible, such as media manipulators. Nevertheless, as mentioned in the Discussion, if such users trust at least one person or entity with credible assessments, our design could highlight the disputed posts to them, and encourage them to explore further.

8 Conclusion

In this work, we explored the potentials of incorporating three user affordances into social media in an effort to give users more agency in combating misinformation: accuracy assessments of posts by users as part of the data model, users indicating whose assessments they trust, and filters that users can configure to display or hide posts based on assessments provided by sources they trust. We validated the need for these affordances through a survey of a diverse set of 192 people. These users indicated that they already ask for assessments from their social circle and provide assessments to them. They do so by using the affordances offered by platforms for other purposes, such as like, comment, or share which can be interpreted by platforms as signals of engagement, or misinterpreted by other users, propagating inaccurate content further. In addition, users follow sources that they deem trustworthy as well as those they consider untrustworthy for a variety of reasons and have difficulty adjusting their feed to show or hide certain content to them, as it is a curation algorithm that is in charge. These findings all confirm the need for facilitating users' practices through the integration of our proposed affordances onto platforms.

We then evaluated our design affordances through a user study of 14 users on a social content sharing platform we prototyped that provides the affordances. The user study gained us an empirical understanding of how users would perceive and use these affordances in a realistic setting. Users perceived value in accuracy signals and providing and seeking accuracy assessments. They were capable of providing assessments based on various rationales, suggesting that they should not be treated as passive consumers who need to be saved but rather be given the right tools to help themselves and their social circle against misinformation. Additionally, we observed that they configured their filters in different ways, some setting a default to only view posts confirmed by their trusted sources occasionally peeking at refuted or disputed posts, and others choosing to include articles of all accuracy statuses in their feed. These various preferences indicate the appeal of being in control of one's feed. We draw from the results of the user study to offer recommendations to tool designers on how to incorporate our proposed affordances into platforms.

9 Acknowledgments

This work was supported by the National Science Foundation Award 1915724.

References

[1] [n.d.]. *Annotate the web, with anyone, anywhere.* https://web.hypothes.is/

[2] 2018. *Facebook apologises for blocking Prager University's videos.* Retrieved July 9, 2022 from https://www.bbc.com/news/technology-45247302

[3] 2021. *How Facebook's third-party fact-checking program works.* Retrieved July 9, 2022 from https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works

[4] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2020. Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi. org/10.31234/osf. io/9qdza* (2020).

[5] Mike Ananny. 2018. The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation. (2018).

[6] Natalie Ashton. 2020. *Why Twitter is (Epistemically) Better Than Facebook.* Retrieved July 9, 2022 from https://www.logically.ai/articles/why-twitter-is-epistemically-better-than-facebook

[7] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review* 1 (07 2020). https://doi.org/10.37016/mr-2020-033

[8] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[9] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[10] Monika Bickert. 2019. *Combatting Vaccine Misinformation - About Facebook.* Retrieved July 9, 2022 from https://about.fb.com/news/2019/03/combatting-vaccine-misinformation/

[11] Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

[12] Leticia Bode and Emily K Vraga. 2018. See something, say something: Correction of global health misinformation on social media. *Health communication* 33, 9 (2018), 1131–1140.

[13] Leticia Bode, Emily K Vraga, and Melissa Tully. 2020. Correcting Misperceptions About Genetically Modified Food on Social Media: Examining the Impact of Experts, Social Media Heuristics, and the Gateway Belief Model. *Science Communication* (2020), 1075547020981375.

[14] Amy S Bruckman. 2022. *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge.* Cambridge University Press.

[15] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, communication & society* 20, 1 (2017), 30–44.

[16] Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. 2002. Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work* 1 (2002), 171–194.

[17] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web.* 675–684.

[18] Rhia Catapano, Zakary L Tormala, and Derek D Rucker. 2019. Perspective taking and self-persuasion: Why "putting yourself in their shoes" reduces openness to attitude change. *Psychological science* 30, 3 (2019), 424–435.

[19] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2019. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* (2019), 1–23.

[20] Alexander Cobleigh. 2020. TrustNet: Trust-based Moderation Using Distributed Chat Systems for Transitive Trust Propagation. (2020).

[21] Josh Constine. 2017. *Facebook puts link to 10 tips for spotting 'false news' atop feed.* Retrieved July 9, 2022 from https://techcrunch.com/2017/04/06/facebook-puts-link-to-10-tips-for-spotting-false-news-atop-feed

[22] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems.* 1–12.

[23] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems.* 3163–3174.

[24] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).

[25] Tory Newmyer Elizabeth Dwoskin and Shibani Mahtani. 2021. *The case against Mark Zuckerberg: Insiders say Facebook's CEO chose growth over safety.* Retrieved July 9, 2022 from https://www.washingtonpost.com/technology/2021/10/25/mark-zuckerberg-facebook-whistleblower/

[26] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[27] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. 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*. 2371–2382.

[28] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 153–162.

[29] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[30] Brian J Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 80–87.

[31] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

[32] R Kelly Garrett and Paul Resnick. 2011. Resisting political fragmentation on the Internet. *Daedalus* 140, 4 (2011), 108–120.

[33] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[34] Joshua Gillin. 2017. The more outrageous, the better: How clickbait ads make money for fake news sites. *PolitiFact, October* 4 (2017).

[35] Sofia Grafanaki. 2018. Platforms, the First Amendment and Online Speech Regulating the Filters. *Pace L. Rev.* 39 (2018), 111.

[36] Catherine Grevet and Eric Gilbert. 2015. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4047–4056.

[37] Jennifer Grygiel and Nina Brown. 2019. Are social media companies motivated to be good corporate citizens? Examination of the connection between corporate social responsibility and social media safety. *Telecommunications Policy* 43, 5 (2019), 445–460.

[38] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations.. In *ICWSM*.

[39] Alfred Hermida. 2012. Tweets and truth: Journalism as a discipline of collaborative verification. *Journalism Practice* 6, 5-6 (2012), 659–668.

[40] Avery E Holton, Kang Baek, Mark Coddington, and Carolyn Yaschur. 2014. Seeking and sharing: Motivations for linking on Twitter. *Communication Research Reports* 31, 1 (2014), 33–40.

[41] Benjamin Horne and Sibel Adali. 2017. The impact of crowds on news engagement: A reddit case study. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 751–758.

[42] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[43] Yasmin Ibrahim. 2017. Facebook and the Napalm Girl: reframing the iconic as pornographic. *Social Media+ Society* 3, 4 (2017), 2056305117743140.

[44] Daniel Jackson. 2021. *The Essence of Software: Why Concepts Matter for Great Design*. Princeton University Press.

[45] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–42.

[46] Farnaz Jahanbakhsh, Amy X Zhang, Karrie Karahalios, and David R Karger. 2022. Our Browser Extension Lets Readers Change the Headlines on News Articles, and You Won't Believe What They Did! *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–33.

[47] Rae Jereza. 2021. Corporeal moderation: digital labour as affective good. *Social Anthropology* 29, 4 (2021), 928–943.

[48] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[49] Audun Jøsang, Elizabeth Gray, and Michael Kinateder. 2003. Analysing topologies of transitive trust. In *Proceedings of the First International Workshop on Formal Aspects in Security & Trust (FAST2003)*. Pisa, Italy, 9–22.

[50] Makena Kelly. 2019. *Facebook proves Elizabeth Warren's point by deleting her ads about breaking up Facebook*. Retrieved July 9, 2022 from https://www.theverge.com/2019/3/11/18260857/facebook-senator-elizabeth-warren-campaign-ads-

removal-tech-break-up-regulation
[51] Ian Kerr and Jessica Earle. 2013. Prediction, preemption, presumption: How big data threatens big picture privacy. *Stan. L. Rev. Online* 66 (2013), 65.
[52] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 324–332.
[53] Silvia Knobloch-Westerwick and Jingbo Meng. 2011. Reinforcement of the political self through selective exposure to political messages. *Journal of Communication* 61, 2 (2011), 349–368.
[54] András Koltay. 2022. The Protection of Freedom of Expression from Social Media Platforms. *Mercer Law Review* 73, 2 (2022), 6.
[55] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[56] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[57] Hanna Kozlowska. 2017. *Facebook is ditching its own solution to fake news because it didn't work*. Retrieved July 9, 2022 from https://qz.com/1162973/to-fight-fake-news-facebook-is-replacing-flagging-posts-as-disputed-with-related-articles/
[58] Nicholas Kristof. 2020. *The Children of Pornhub*. Retrieved July 9, 2022 from https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html
[59] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society* 1, 2 (2015), 2056305115610141.
[60] Francis LF Lee, Michael Che-ming Chan, Hsuan-Ting Chen, Rasmus Nielsen, and Richard Fletcher. 2019. Consumptive news feed curation on social media as proactive personalization: a study of six East Asian markets. *Journalism Studies* 20, 15 (2019), 2277–2292.
[61] Xialing Lin, Patric R Spence, and Kenneth A Lachlan. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in human behavior* 63 (2016), 264–271.
[62] Guanfeng Liu, Yan Wang, and Mehmet Orgun. 2011. Trust transitivity in complex social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25.
[63] Long Ma, Chei Sian Lee, and Dion Hoe-Lian Goh. 2011. That's news to me: the influence of perceived gratifications and personal experience on news sharing in social media. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 141–144.
[64] Long Ma, Chei Sian Lee, and Dion Hoe-Lian Goh. 2013. Investigating influential factors influencing users to share news in social media: A diffusion of innovations perspective. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. 403–404.
[65] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friend-sourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[66] Pranav Malhotra. 2020. <? covid19?> A Relationship-Centered and Culturally Informed Approach to Studying Misinformation on COVID-19. *Social Media+ Society* 6, 3 (2020), 2056305120948224.
[67] Farhad Manjoo. 2013. You won't finish this article. *Why people online don't read to the end: Slate* (2013).
[68] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* 35, 2 (2018), 196–219.
[69] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.
[70] James C McCroskey and Jason J Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* 66, 1 (1999), 90–103.
[71] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[72] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.
[73] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.
[74] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 441–450.

[75] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[76] Adam Mosseri. 2016. News feed fyi: Addressing hoaxes and fake news. *Facebook newsroom* 15 (2016), 12.
[77] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
[78] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
[79] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L Freed. 2014. Effective messages in vaccine promotion: a randomized trial. *Pediatrics* 133, 4 (2014), e835–e842.
[80] Megan O'Neill. 2012. Youtube responds to reply girls, changes related & recommended videos algorithm.
[81] Kari Paul. 2020. *Pornhub removes millions of videos after investigation finds child abuse content.* Retrieved July 9, 2022 from https://www.theguardian.com/technology/2020/dec/14/pornhub-purge-removes-unverified-videos-investigation-child-abuse
[82] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* (2020).
[83] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, Dean Eckles, and David Rand. 2020. Understanding and reducing the spread of misinformation online. *ACR North American Advances* (2020).
[84] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.
[85] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.
[86] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://aclanthology.org/C18-1287
[87] Hector Postigo. 2009. America Online volunteers: Lessons from an early co-production community. *International Journal of Cultural Studies* 12, 5 (2009), 451–469.
[88] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European Conference on Information Retrieval*. Springer, 810–817.
[89] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 173–182.
[90] Ian Reilly. 2012. Satirical fake news and/as American political discourse. *The Journal of American Culture* 35, 3 (2012), 258–275.
[91] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
[92] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *new media & society* 19, 8 (2017), 1214–1235.
[93] Robert Shrimsley. 2016. *Facebook photos: snap judgments.* Retrieved July 9, 2022 from https://www.ft.com/content/dbcdf744-7ac6-11e6-b837-eb4b4333ee43
[94] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[95] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160.
[96] Sara Spray. 2016. *Facebook Is Embroiled In A Row With Activists Over "Censorship".* Retrieved July 9, 2022 from https://www.buzzfeed.com/saraspary/facebook-in-dispute-with-pro-kurdish-activists-over-deleted
[97] Kate Starbird. 2021. Online Rumors, Misinformation and Disinformation: The Perfect Storm of COVID-19 and Election2020. (2021).
[98] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings* (2014).
[99] David Sterret, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, Kevin Loker, and Emily Swanson. 2018. Who shared it?: How Americans decide what news to trust on social media. *NORC Working Paper Series WP-2018-001* (2018).
[100] Matthias Steup. 2005. Stanford Encyclopedia of Philosophy. Epistemology.

[101] Frederik Stjernfelt and Anne Mette Lauritzen. 2020. *Your Post Has Been Removed: Tech Giants and Freedom of Speech*. Springer Nature.
[102] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576.
[103] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.
[104] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[105] Toni GLA van der Meer and Yan Jin. 2020. Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication* 35, 5 (2020), 560–575.
[106] Emily K Vraga and Leticia Bode. 2017. Using expert sources to correct health misinformation in social media. *Science Communication* 39, 5 (2017), 621–645.
[107] Lorraine YC Wong and Jacquelyn Burkell. 2017. Motivations for sharing news on social media. In *Proceedings of the 8th International conference on social media & society*. 1–5.
[108] Liang Wu, Jundong Li, Xia Hu, and Huan Liu. 2017. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 99–107.
[109] Sarita Yardi and Danah Boyd. 2010. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society* 30, 5 (2010), 316–327.
[110] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–37.
[111] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.
[112] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*. 603–612.
[113] Amy X Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2082–2096.
[114] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.
[115] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the sigchi conference on human factors in computing systems*. 1883–1892.

A Network of Participants' Relationship with Other Sources

B Users' Rationales In Support of Their Assessments

Received January 2022; revised April 2022; accepted August 2022

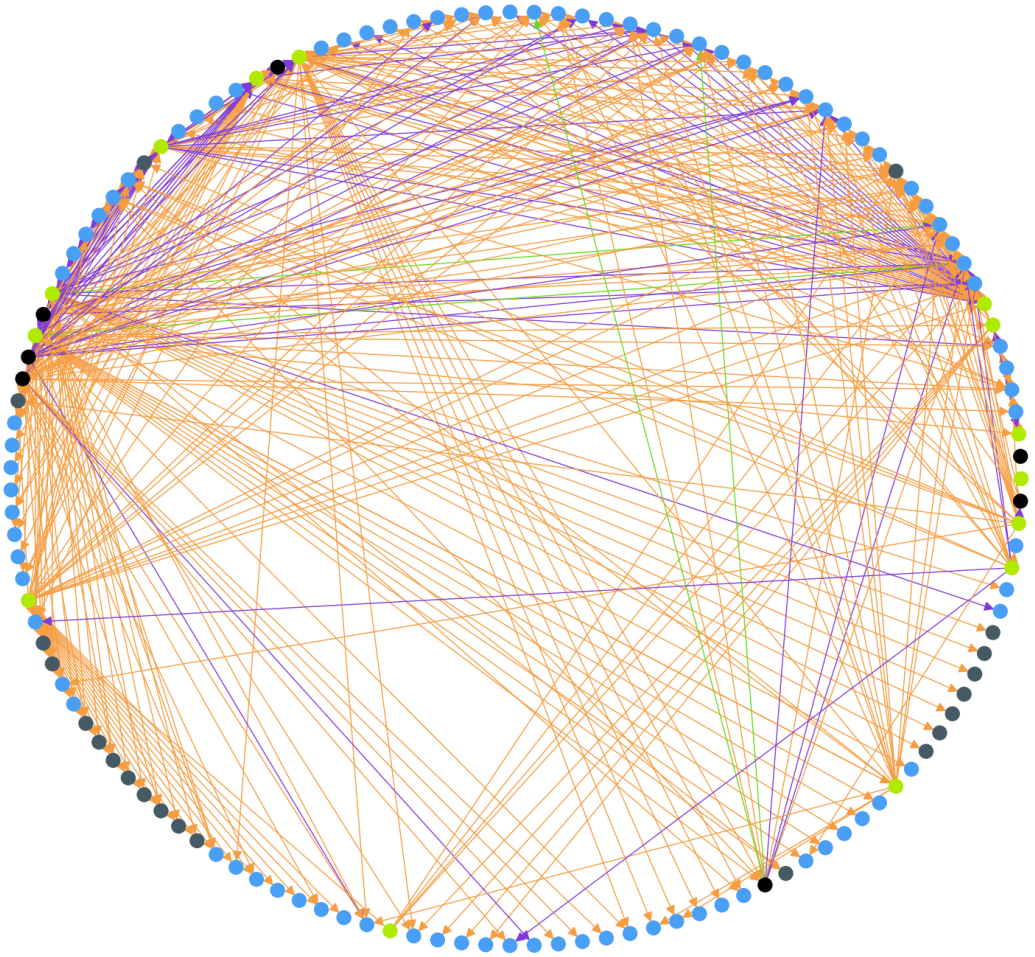

Fig. 7. Figure shows the graph of participants' relationship with other sources. Four types of nodes are shown in the figure: 1. Lime colored nodes depict participants who completed the user study; 2. Black nodes depict participants who abandoned the study before its completion; 3. Blue nodes show proxy accounts for news publishing entities (e.g., CNN); 4. Grey nodes show other individual users who had an account on the platform but that did not participate in the user study. Some of these accounts belonged to members of our research team or the broader research lab. The reason that these nodes are included in the graph is that they were trusted or followed by study participants. The nodes (individual or proxy accounts for news publishing entities) that none of the participants followed or trusted are not shown in the graph. The graph edges denote follow and trust relationships and all originate from user study participants; i.e., the figure does not show the relationship between grey and blue nodes. The orange edges denote follow and the green edges denote trust relationships. A purple edge between a pair of nodes (outgoing from node A to B) marks that A both trusts and follows B.

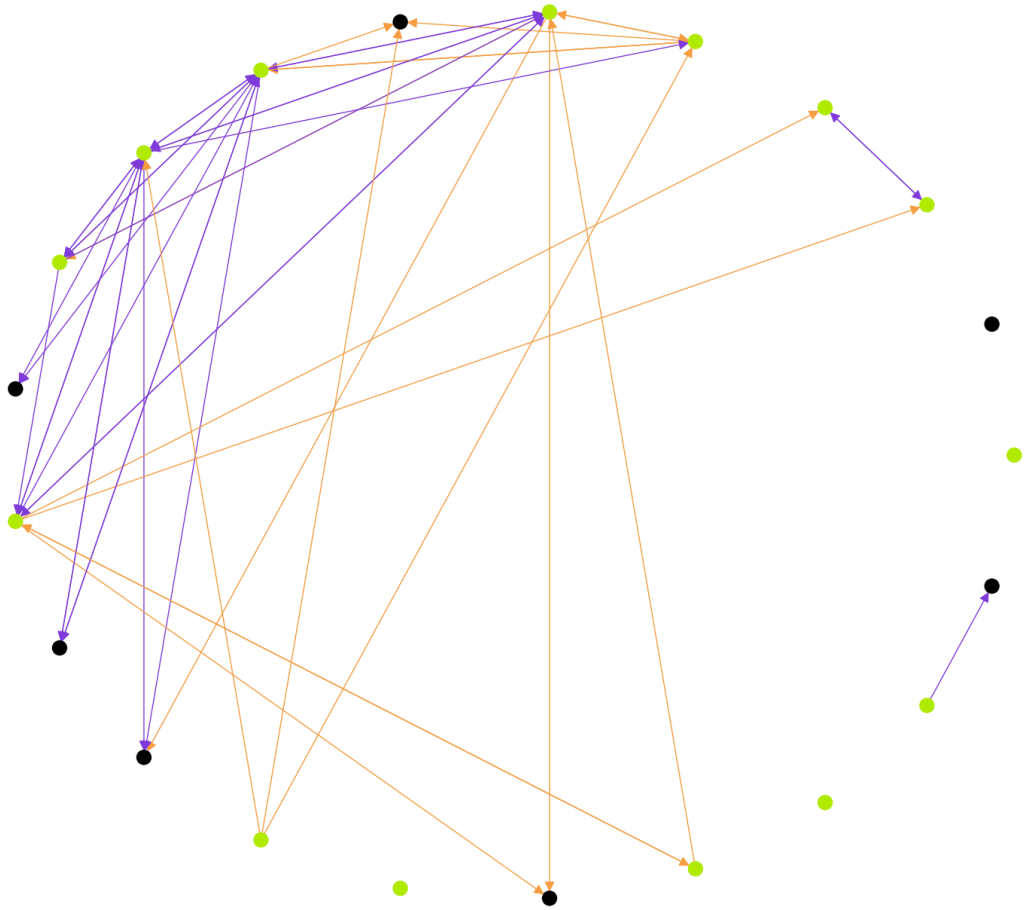

Fig. 8. Figure shows the graph of participants' relationship with other users of the study including those who dropped out. Two types of nodes are shown in the figure: 1. Lime colored nodes depict participants who completed the user study; 2. Black nodes depict participants who abandoned the study before its completion. The graph edges show follow and trust relationships. The orange edges denote follow relationship. An orange edge outgoing from node A to B denotes that node A follows B. If both A and B follow each other, the edge between them is bidirectional. A purple edge between a pair of nodes (outgoing from node A to B) marks that A both trusts and follows B. We observe that trust between pairs of users in our study were often (but not always) bidirectional.

Table 2. Examples of some of the rationales that we observed in our user study of the social content sharing platform and the types of claims our participants chose to assess.

Headline	Assessment	Rationale Type
Why The Apple M1 Chip Is So Fast - A Developer Explains (Production Expert)	*"This is a technical area I understand and I find the article plausible given my knowledge of the state of the art." (p-20)*	High degree of knowledge on the topic
U.S. Orthodox leadership congratulated Biden, but many in community stick to Trump (Haaretz)	*"I am amazed by the small number of orthodox folks I know who think the President should challenge the results" (p-12)*	Firsthand knowledge
Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic (Nature)	*"Nature is an esteemed journal with a reputation for sound scientific research" (p-10)*	The claim is from a trusted source.
Fox News forced to issue retraction on election fraud claims after legal challenge (Occupy Democrats)	*"I verified on other sites and watched the Lou Dobbs segment itself to verify this article is correct." (p-8)*	Confirmed by other trusted sources
'Very low' rates of coronavirus in schools, British study finds (Yahoo News)	*"Although published in Dec 2020; this article reports on a study conducted in the UK from June - July. During that period; COVID was controlled in the UK with cases decreasing. The article notes that the study found that for every 5 cases in 100k people; the odds of a school outbreak increased by 72%. Assuming a linear; (rather than exponential) trend; I would estimate school infections are more than 700% more likely to occur in the UK In December 2020; when this article was published. Therefore; I view this article as inaccurate at best." (p-13).*	The claim appears to be inaccurate based on its presentation (its language, flawed logic, etc.).
Where is the kit to protect NHS workers? (The Guardian)	*"Sounds plausible from what I've been hearing from UK medics and from my medical experience in the UK. I remember running around the hospital almost everyday looking for resources (that should've been well stocked) at one of the top hospitals in the country." (p-14)*	Consistent with past experiences and observations