A Browser Extension for in-place Signaling and Assessment of Misinformation

Farnaz Jahanbakhsh*
Stanford University
Stanford, USA

David R. Karger
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
Cambridge, USA

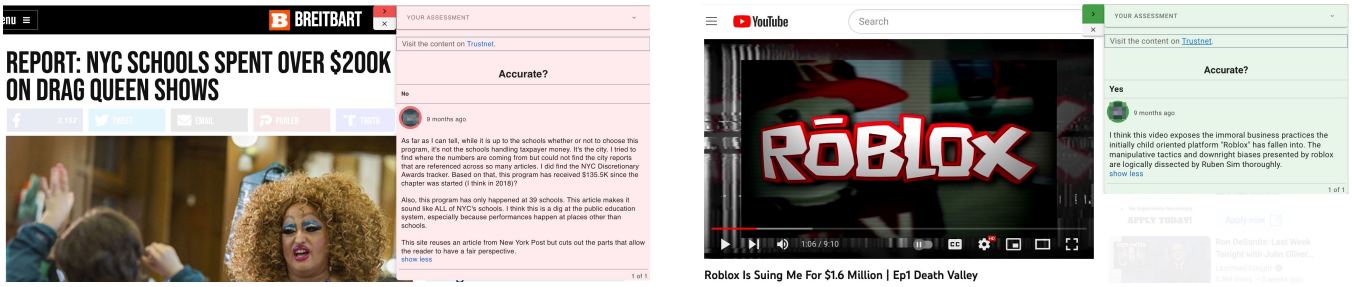


Figure 1: Upon visiting a webpage, the Trustnet extension opens a pane on the side showing the assessments of the page by those the user follows or trusts. The figure shows a news article (left) and a Youtube video (right) assessed by users of the study.

ABSTRACT

The status-quo of misinformation moderation is a central authority, usually social platforms, deciding what content constitutes misinformation and how it should be handled. However, to preserve users' autonomy, researchers have explored democratized misinformation moderation. One proposition is to enable users to assess content accuracy and specify whose assessments they trust. We explore how these affordances can be provided on the web, without cooperation from the platforms where users consume content. We present a browser extension that empowers users to assess the accuracy of any content on the web and shows the user assessments from their trusted sources in-situ. Through a two-week user study, we report on how users perceive such a tool, the kind of content users want to assess, and the rationales they use in their assessments. We identify implications for designing tools that enable users to moderate content for themselves with the help of those they trust.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**.

*Research conducted while at MIT.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05…$15.00
https://doi.org/10.1145/3613904.3642473

KEYWORDS

misinformation, democratized content moderation, fact-checking

ACM Reference Format:
Farnaz Jahanbakhsh and David R. Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642473

1 INTRODUCTION

The convenience of spreading and consuming misinformation on social media platforms [99, 100] and the media spotlight given to the havoc that misinformation has wrought in recent years [11, 16, 30, 96] have caused a surge of attention to this age-old problem. Researchers as well as platform operators are increasingly investing effort in combating misinformation. At present, the dominant approach to fighting misinformation is to place the onus for doing so on the large social platforms where many users spend their time. This approach places a critical social decision in the hands of for-profit companies, and limits users' autonomy in deciding who they trust to declare what is accurate [51]. It also does nothing to combat misinformation that users encounter off the social platforms, including the many news sites found on the web. The downsides of centralized misinformation moderation have motivated some researchers and practitioners to pursue decentralized solutions to the problem. For instance, Twitter Community Notes is an example of such an approach which involves all the users of the community in assessing the accuracy of content and rating each other's assessments. However, the final decision of what notes to eventually show to the public is taken in a centralized fashion, by Twitter's "bridging algorithm" [8].

In contrast, a body of work has advocated for giving users more autonomy while aiding them by providing the tools that they need for discerning truth from falsehoods. For instance, it is reported

that enabling users to assess the accuracy of content can reduce their likelihood to share misinformation [49], as it nudges users to have accuracy on top of their mind [81, 82]. If user assessments are captured as structured metadata, they can be displayed on content in structured form as well, similar to but separate from likes or comments, and have the potential to warn the assessor's social circle should they encounter the misinforming content [51]. Jahanbakhsh et al. argue that users should additionally be given the ability to specify whose assessments they would like to see—in essence, leaving it to each user to decide who they trust as their moderators. They report that even on the current platforms, users do ask those they trust for their assessments; however, they do this without platform support, and in fact despite the platforms getting in their way. In a user study of a prototype social news sharing platform that offers these affordances, the authors determine that users want them and are capable of using them [51]. Capturing assessments and trust relationships in structured form can be a first step in designing interventions that point users to fact-checking information from those that they trust.

To increase user autonomy, platform operators can be urged to incorporate the affordances of user generated assessments and specified trusted sources. However, this incorporation is unlikely to happen without activism or legislation, because changing the user experience drastically, particularly ceding too much control over content to their users, can impact platform revenues. Researchers could build a new social platform where they offer these affordances to users, for instance similar to the prototype platform presented by Jahanbakhsh et al. [51]. But such undertaking realistically may not succeed because first, it costs much development and maintenance for this new platform to have a chance to compete with the existing social platforms; and second, users are unlikely to abandon their social network on the existing platforms and join a new one without a critical mass of their connections already present there.

While social media platforms have been an important outlet for the circulation of misinformation, they are far from being the only one [34]. Efforts concentrated on getting *social platforms* to moderate content ignore that users consume content, by choice or by accident, outside these platforms as well. Users for instance, may encounter links to unverified or misinforming content on a news or media website that crosslinks to another, in an email newsletter, or on a news aggregator. It is unrealistic to expect that every website or platform on the web, some the very perpetrators of misinformation, offer content moderation and in a fair and rigorous way.

The pursuit for ubiquitous content moderation that preserves user autonomy is the motivation behind this work. What follows is a case study of how enabling users to moderate content, specifically in the context of misinformation, can be deployed *everywhere on the web*, in a platform-agnostic manner and without support from the social platforms or individual web sites. We use the term "moderation" as a concept that involves categorizing the status of content by applying labels, including those that are commonly used by platforms such as "misleading" or "explicit" as well as any other label that can help a user customize their web experience, such as "depressing". The other aspect of moderation is taking an action, if any, on content that has been positively or negatively labeled, e.g., whether to simply flag or remove it. These two aspects are often bundled together in the status-quo practices. In this work, we do not focus on the latter aspect and simply signal the status of content. If the categorization of content is customized according to users' needs and input, then a set of tools can exist that take the labels and act in different ways on them, or within the same tool, the choice of what to do with the content can also be deferred to the user [51].

In this work, we design, deploy, and evaluate a browser extension that lets *anyone* assess any content—news articles, Youtube videos, tweets, etc.—*anywhere* on the web, and lets each user independently decide whose assessments they *trust* and want to use to help protect them from misinformation. This tool is an extension of the Trustnet social platform previously offered by Jahanbakhsh et al. [51]. Upon visiting a page, the user will see assessments of the content by their trusted sources, if any have contributed one, as shown in Figure 1. Such assessments can aid the user in deciding whether the content they are viewing is credible. However, since links on social media are often not clicked, even when they are shared [35], showing the assessments of a page only when a user visits the page will not benefit those who do not navigate to the page. To address this issue, the extension also places any available trusted accuracy assessments next to outgoing links on the page the user is reading, as displayed in Figures 2 & 3.

To evaluate the potential of such a design and users' perception of it, we conducted a user study of the extension. We recruited 32 users and to use the extension for two weeks. Users were tasked with assessing the accuracy of two pieces of content daily. Users also could, but were not required to, share content with the other users via the extension. The shared content would be presented on each user's feed on the Trustnet platform instance that we asked them to visit on a daily basis. We surveyed our users after the user study to understand their fact-checking needs and desires and their perception of the tool. The user study serves as a technology probe [45] where users can interact with a tool that empowers them to have an active role in moderating misinformation for themselves and others, without needing to drastically change where and how they consume content. By offering the experience of a new alternative, our technology probe breaks our subjects free of the status quo, encouraging less constrained thinking about their needs and desires for misinformation moderation.

Our user study sets out to answer three main research questions:

RQ1: What are the facets of users' information needs that a democratized misinformation moderation, made possible through our proposed tool, can address?;

Our first research question aims to *empirically* demonstrate the potential for democratization of misinformation governance. Investigating the diversity of users' needs after their exposure to the tool helps us understand whether centralized fact-checking can sufficiently cover the breadth of user needs and desires, some of which users may be empowered to articulate after they realize other forms of misinformation moderation may be possible. We focus on multiple aspects of users' needs: how important they think it is to avoid misinformation on various topics, how easy they find it to fact-check content on various topics, whether they want to see assessments from others, and whether they want to contribute assessments for the benefit of others. We show that even in our small-scale study, the distribution of the types of content people deem important to assess differs across individuals and from

Figure 2: The extension places any available accuracy assessments from the user's trusted or followed sources next to outgoing links on the page the user is visiting. The screenshot shows an example of a red X marking such a link on the Twitter timeline. The link is visually faded because it has been assessed as inaccurate.

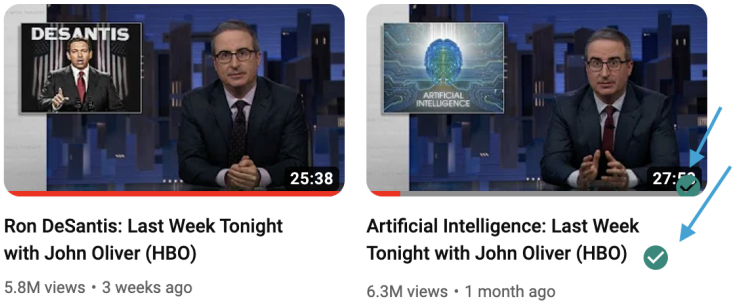


Figure 3: Outgoing links to a Youtube video on the page are displayed with a checkmark icon (marked with blue arrows), indicating that the video has been assessed as accurate by the user's assessors.

what is typically fact-checked by professional fact-checking initiatives. Moreover, depending on the topic, users wished for including viewpoints and assessments from than professional fact-checkers alone—e.g., journalists from other countries and immigrants on issues related to foreign affairs. These observations further bolster the argument that there is a need for decentralized approach to content moderation.

RQ2: What criteria for assessing content accuracy do users use when using our browser extension?

Examining the rationales that users use in their assessments helps us determine whether regular people can indeed assess content effectively. By extending a taxonomy of such rationales with new categories not previously reported in prior work, we contribute to the literature on user-centered indicators of content credibility [49, 105]. Additionally, these rationales can be built into tools to capture users' assessments of content credibility in structured form.

RQ3: How do users perceive a browser extension for in-situ signaling and assessment of misinformation from their customized sources?

- RQ3a. What do they perceive as the advantages and downsides of the approach?
- RQ3b. What are their ideas for improving it and how could they be incentivized to use it?

Investigating these questions helps us determine whether users find the tool of value to them and how it can be improved.

The main contributions of our work are: 1) the design of a system for democratized in-place signaling and assessment of misinformation anywhere on the web which allows users to moderate misinformation for themselves with the help of those they trust; 2) a taxonomy of criteria for assessing content credibility not previously reported in prior work emerged from our study, which can be used to extend the credibility indicators framework offered by prior work [49, 105]; and 3) An empirical understanding of how users use and perceive this system.

This tool provides users with the autonomy to moderate and filter misinformation in-situ anywhere on the web for themselves and to additionally help their social circle do so. We offer recommendations for extending the design and architecture behind this tool to domains beyond misinformation, and designing a class of tools that further empower users to moderate content for themselves and choose who they want as their moderator.

2 RELATED WORK

Throughout the paper, we use the term "misinformation" as an umbrella term encompassing the various types of false information that can be found on the web per Zannettou et al., including fabricated stories, propaganda, conspiracy theories, hoaxes, biased or one-sided content, rumors, clickbait, and more [102]. Here, we situate our work in the literature related to identifying and dealing with misinformation, tools that allow for personalizing user's web experience, and systems that aid in sensemaking.

2.1 Identifying and Dealing with Misinformation

Given the impact that misinformation on social platforms has had on people's lives and even democracy [11, 16, 21, 96], researchers as well as platform operators have been investigating and deploying approaches to detect and counter misinforming content. The space of these approaches is broadly contained within these two axes:

- Who gets to decide what is misinformation
- and whether the moderation decisions are contained within a single platform/website or applied ubiquitously

The arbiters of misinformation can range from a central authority to a body of a trusted few to a democratized process e.g., through majority votes from users to each individual.

2.1.1 Centralization. The dominant approach is centralization of the arbitration power—in most cases to the platforms—who decide on both the detection of misinformation as well as the decision on how to deal with the misinforming content. Centralized approaches to misinformation detection include using machine learning models for classifying whether content is accurate [13, 20, 85, 103], and leveraging human moderators and third party fact-checking and news organizations [7, 10]. Centralized approaches to dealing with misinformation include displaying warning labels next to, down-ranking, reduction (e.g., shadow-banning), or removal of the content or suspension of offender accounts [18, 24, 27, 38, 75, 77, 84].

A body of work has investigated the effectiveness of these centralized approaches. For instance, Bak-Coleman et al. show that content removal and account banning can be effective at stopping the growth of misinformation [15]. Researchers have also reported that warning labels can increase user discernment of content accuracy or reduce users' likelihood of sharing misinformation [17, 92, 101]. Other studies however, found warning labels have limited or adverse effects [25, 36, 79].

Nevertheless, the role of platforms as moderators of content and speech is contested by some researches, scholars of law, as well as users who believe the downsides of ceding such control to the platforms can be grave. One such downside is that by assuming the role of a truth arbiter, platforms can limit users' freedom of speech and listening and their autonomy in deciding what content to consume [41, 51, 60, 89]. Related is the concern that even platform-assigned labels or messages can be perceived as truth governance and a threat to freedom of speech [29, 54]. In fact, some users perceive platform labels as judgmental, paternalistic, and against the platform ethos [89]. Another concern is that any centralized decision, by platforms or otherwise, to for instance, down-rank or filter misinformation cannot address the needs of every user, as some users want to be aware of what misinforming content their social circle is exposed to, so that they can talk to them about it [14, 51, 69, 86]. Yet another concern is that for these centralized approaches to work, users have to rely on the trustworthiness, goodwill, and competence of the moderators [66, 71]; however, some users consider platforms profit-driven and politically biased [51, 89]. These concerns are valid given the history of platforms blocking content that arguably did not have potential to harm or content by activists or dissidents in certain autocratic countries [5, 37, 47, 52, 90].

While the centralized approaches mentioned above are generally contained within a single platform, some exist that have been applied ubiquotiously. These include classifiers to label clickbait or sensationalist headlines based on lexical features that were incorporated into browser extensions that warn users when they view such a headline [23, 88].

2.1.2 A Select Body of a Trusted Few. There exist approaches that put the power of arbitration in the hands of a trusted few (e.g., fact-checkers or journalists). ClaimBuster is a system of this nature, which checks claims in live discourses, news, and social media posts against a repository of fact-checked claims [43]. Another is ConsidertIt, a system for public dialog where participants discuss pros and cons of issues, integrated an on-demand fact-checking into the Living Voters Guide (LVG) deliberation space. Fact-checking was staffed by librarians and any LVG user could request that any pro or con point submitted by other users be fact-checked. The fact-checking information would then be placed in-situ [61]. The interactive fact-checking element in this system is similar to the one in our tool. Another example is Videolyzer that enables political bloggers and journalists to assess aspects of quality in videos including accuracy and bias [28]. Other systems have also been developed to help the body of the select few monitor and verify potentially critical content shared in social spaces [73].

While these initiatives are valuable, their effectiveness can be limited by being constrained to a platform and not incorporated into the systems where users encounter content. This issue is countered by tools such as NewsGuard, a browser extension that shows detailed transparency and credibility scores given by journalists to various sources (i.e., websites) [64]. These scores are placed alongside links that users encounter as they browse the web. A shortcoming of displaying credibility ratings at the granularity of a source is that not all content from an unreliable source is inaccurate; and conversely, not all content from a reliable source is accurate [51]. However, for approaches that confine the power to assess credibility to a limited set of individuals, e.g. journalists, scaling the investigation and signalling of the credibility at *content-level* would be impractical. Even at the source-level, such an approach fails to capture the reliability of sources that are not notable enough.

2.1.3 Democratization. Moving farther away from centralization of the arbitration power, a body of work has studied how to involve users in the process of content moderation. Most notably, Twitter Community Notes (previously known as Birdwatch) enables users to anonymously write "notes" critiquing or adding additional context to a tweet that they find misleading. Other community members can rate notes as helpful. Twitter's bridging algorithm then decides which notes are eventually shown to users by inspecting whether the raters seem to come from diverse perspectives [8]. Subreddits and Facebook groups also have moderators from the user body appointed to the role. In these cases however, the resulting decision is imposed on all users of the community, regardless of whether they agree with the outcome or the decision making process.

2.1.4 Individualization. At the end of the centralization/ decentralization axis lie approaches that do not impose a single assessment for content on all users and instead enable each individual to make more informed decisions about content credibility, e.g., by relying less on intuition and instead, engaging in critical thinking [74, 83]. These include nudges that ask users to assess the accuracy of content before sharing it [32, 49], subtly priming users to have accuracy on top of their mind [81, 82], or providing users with checklists based on recommendations from authoritative sources (e.g., the WHO) to guide users in evaluating content reliability for themselves [44]. Jahanbakhsh et al. proposed a set of design affordances, that they argue platforms should adopt, which democratize the power to not only determine what content is misinforming, but also decide what to do about such content. These affordances include enabling users to: 1) assess content as part of the data model 2)

specify whose assessments they consider trustworthy, and 3) filter misinformation from their feed as assessed by their trusted sources. These affordances are intended to empower users in their practices around fighting misinformation that the authors uncover through a study— users already ask those who they trust within their social circle to verify the information that they encounter. And they also provide fact-checking information to their friends and family [51]. Indeed, users are more receptive to correcting information from their friends than strangers [42, 70]. However, platforms do not support these user practices and sometimes even undermine them.

Jahanbakhsh et al. provided these user affordances through a prototype social media platform that they designed [51]. However, it is unrealistic to expect that the myriad of websites and platforms where users consume content will adopt these affordances. Therefore, in this work, we explore using a browser extension that acts as an overlay on the web, and allows users to assess the accuracy of any content that they encounter across platforms, as well as see the assessments of their trusted sources on the content. Our design deals with operationalizing trusted assessments ubiquitously and in-situ including where they should be placed considering the particulars of user's content consumption practices on the web; e.g., the fact that many users skim headlines and links to content, but click on them only sporadically [19]. In addition, our architecture broadens the coverage of content that can be assessed beyond articles.

There do exist other tools that both are individualized and whose arbitration decisions are applied ubiquitously. One is the Rehead-line browser extension that enables users to suggest headlines they deem better for news articles and specify whose suggested headlines they want to see on the web [50]. Another is Dispute Finder, a browser extension that highlights text snippets making disputed claims and shows articles arguing for and against the claim. The disputed claims in the system are restricted to those in certain fact-checking websites such as Snopes and Politifact. With centralized fact-checking often lagging behind the spread of false claims [93], the effectiveness of this tool can be limited when the user encounters content that is not old enough to have been fact-checked by fact-checkers of the two platforms. Dispute Finder is individualized as it is the user who specifies the sources of the articles a user wishes to see for or against the disputed claims, although users are restricted to websites that meet the Wikipedia criteria for being reliable [31].

2.2 Personalizing Users' Web Experience

Our browser extension belongs to the class of tools that personalize users' web experience across platforms. A classic example of such a tool was WebWatcher, which followed the user's actions on the web, learned the user's interests, and highlighted or added certain hyperlinks on the page for the user to follow [55, 56]. Within this class, web annotation tools exist that enable users to add content to webpages where they do not have authoring permission. One such example is Hypothesis, a browser extension that allows users to highlight and annotate text and for future visitors on the page to view and interact with the annotations. Hypothesis users can make their annotations visible to the public or only within a private group of users [2, 58]. NB is a similar social annotation tool for classroom settings [107]. Other tools have also existed that enable users to add comments on a page that are not necessarily tied to a particular piece of content on the page. Examples include Google Sidewiki and Eyebrowse [12, 104].

Our extension differs from these tools in two ways. One is that the assessments, i.e., the annotations or metadata, that are submitted via the extension are structured. The structure makes it possible to differentiate assessments arguing for the accuracy of the content and the counter arguments as well as the questions about the accuracy of the content. The differentiation makes it possible to show each of these categories differently in the UI. For helping the user determine content credibility as is the purpose in this domain, it is important to surface arguments as well as counter-arguments in a salient fashion. The other difference in this tool is the fine-grained control that the user has in determining their sources of assessments, i.e., their trusted sources after the model proposed in [51].

2.3 Systems that Aid in Sensemaking

By allowing users to share assessments with each other, and to contextualize their assessments in-situ, our browser extension is related to a class of systems that empower and aid users in the process of sensemaking. The sensemaking brought about by our tool is in fact distributed—i.e., it leverages other users' work without those users directly collaborating with each other [33]. Research has shown that empowering users to scaffold on each others' learning can increase the depth of their sensemaking [59]. Of the other work that similar to our case, aid users in making sense of information, Goyal et al. provided a sensemaking translucence interface for crime solving, which included a hypothesis window to promote idea exchange and a suspect visualization for automatic feedback on suspects discussed in the hypothesis window [39]. The role of visualization in improving sensemaking has been explored in other work as well, including for representing data links or analysis provenance [40, 65]. Other work has reported on the effectiveness of creating shared representations between domain experts and the crowd at increasing sensemaking [98]. Close to our scenario, Kuznetov et al. designed a browser extension with the goal of reducing the friction of collecting information while remaining in the flow of the sensemaking process. The extension allows users to access and organize external content on the web in-situ while preserving the provenance of and the context around the content [62].

3 THE SYSTEM

To build our system, we extended the open source Trustnet platform presented in [51]. The reason we used Trustnet was because it already provides an API for managing the data models that our envisioned extension would use. In this section, we provide a summary of the Trustnet platform and describe the design of the extension that we built.

3.1 The Trustnet Platform

Trustnet is an open source prototype social content sharing platform developed by Jahanbakhsh et al [51]. On this platform, users can post content or import content from other websites using their URL, *follow* other users or news media (all generally referred to as

sources) to see the content they post on their newsfeed, provide assessment of content, i.e., mark content as accurate or inaccurate and provide their reasoning, *trust* other users, and see the assessments of those they follow and trust on the content appearing in their newsfeed. The platform also offers a filter which a user can use to filter the posts in their newsfeed based on how the posts have been assessed by the user's trusted sources. Users can also inquire about the accuracy of posts, and optionally elaborate on their question. A user's question about content accuracy is relayed to their trusted sources by default; although the user can choose to specify from whom they would like assessments. Users can ask their question about content accuracy anonymously. On this platform, who a user trusts is kept private to the user. Follow relationships on the other hand, are public.

We extended the API of Trustnet's backend to serve the browser extension that we developed. In addition, we updated Trustnet's client to work with the extended backend. Using the extended system, users can sign up for an account on the Trustnet platform, and specify which other users they trust or want to follow, and then use the same account for logging into and working with the extension.

3.2 The Trustnet Extension

The design decisions pertaining to the Trustnet extension [1] were made through many rounds of discussion within the research team as well as pilot tests with a broader group of researchers who volunteered their time as alpha testers. Here we describe the design as well as the rationales behind the various decisions. The flow of the user interaction with the system is shown in Figure 4.

When a user is on a web page, the extension checks the URL (and its variations[2]) to see whether its associated content has been assessed by those the user trusts or follows. If there are such assessments for the page, a pane that contains the assessments pops open on the side of the page automatically. The pane stays open for a few seconds and then is minimized to a floating button on the top right corner of the viewport to avoid obstructing the user's view. The pane and the floating button color is determined by the assessed accuracy of the content, as we will later describe. Clicking on the floating button re-opens the pane. The reason we made the decision to show the pane automatically when assessments were present, as opposed to letting the user notice the colored floating button, was because we expected that with a small user base, users would only occasionally encounter pages which their sources had assessed and we wanted to prevent "change blindness". Change blindness is a phenomenon where changes in a visual scene go unnoticed because of visual inattention. The degree of change blindness is higher when the observer does not expect a change [87], e.g., because change happens infrequently. We deemed it important to successfully deliver the signal that assessments are present on a page. With a wider adoption and a higher likelihood of assessments being present on various pages, users could be given the option to have the pane minimized from the beginning and open automatically in certain situations, for instance, when the page has been evaluated as inaccurate. For ease of discernment, the color of the pane as well as the floating button signals the accuracy status of the content: green for accurate, red for inaccurate (Figure 1), and orange for split opinion—where some of the user's assessors have marked the content as accurate and others as inaccurate (See Figure 4).

The accuracy status of the page (and hence the pane's color) is determined by: 1. the assessment that the user has submitted for the page, if any. 2. In case the user has not assessed the page, the assessments from those that the user trusts if any exist. 3. If there are no such assessments, the assessments of those the user follows. A page is determined as accurate if *all* the assessments of the relevant assessors (i.e., the user, or those that the user trusts, or those that the user follows in absence of trusted assessments) have evaluated the page as accurate. The same is true for inaccurate. If there is disagreement among the relevant assessors, the page's accuracy is determined as "split opinion". If the page has no assessments from either the user or those that the user trusts or follows, the pane contains no assessments and its color is grey.

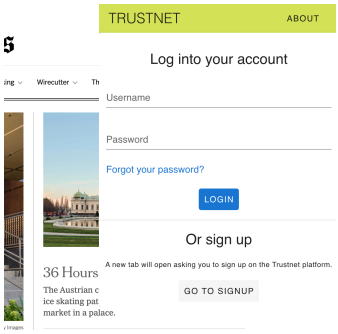
We chose the green for accurate and red for inaccurate color scheme to be externally consistent with fact-checking platforms such as Politifact and Snopes. We selected the (yellow-)orange color for split opinion because it is halfway between green and red on the color wheel. Another design that we considered for the split opinion status was for the color of the pane to be determined as a distance between red and green based on the proportion of assessments marking the content as accurate vs inaccurate. However, in cases when a minority assessment was outnumbered by assessments arguing the opposite, it was possible that the color would be so close to either red or green that it would be indistinguishable. We determined it was important to signal even a lone outlier disagreement amidst one-sided assessments, and that perhaps such contrary viewpoint can point the user to the specific trusted assessor who might help guide them out of an echo-chamber. Therefore, we decided against using a color gradient.

The same pane also shows inquiries about the accuracy of the content from those that either trust the user who is visiting the page or have specifically asked that their question be relayed to the user. The user can use the same pane to submit their own assessment of the page, update their previous assessment, or ask about the accuracy of the content. They can also use the pane to share the content into the Trustnet platform after they have assessed it or asked about its accuracy. By sharing the content, other users who follow the sharer will see the content in their Trustnet feed along with the sharer's assessment of it. To help users expand their network, the pane informs the user if there exist assessments on the page by someone the user does not follow or trust. The pane shows up to 10 such sources who are trusted by the most number of users platform-wide. The user can choose to follow any of these sources if they wish to see their assessments on the page. Users can expand, collapse, or close the extension pane.

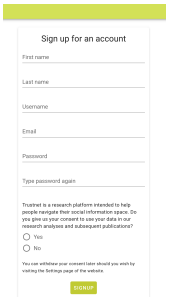
An advantage of using URLs to bind assessments to and find assessments of content is that they are universal. Any resource on the web including news articles, social media posts (e.g., Facebook posts and tweets) and even their comments, Youtube videos, etc.

[1] Code available at https://github.com/farnazj/Trustnet-Extension
Extension on the Chrome web store:
https://chromewebstore.google.com/detail/trustnet/nphapibbiamgbhamgmfgdeiiekddoejo

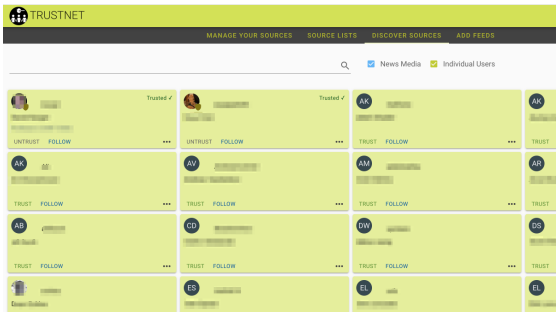
[2] Depending on how the page is served, the same resource can be accessed with slightly different URLs, for instance with or without index.html or / concatenated to the end, over both secure and plain HTTP, or using geolocation IP services to redirect requests to different domain names such as BBC.co.uk to BBC.com if the requester IP is in the U.S.



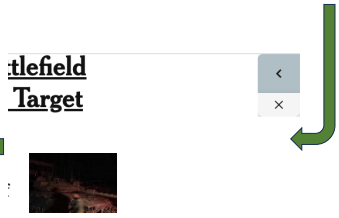
1. After the user installs the extension and opens any tab, the Trustnet extension pane appears on the corner, and asks the user to log in or sign up.



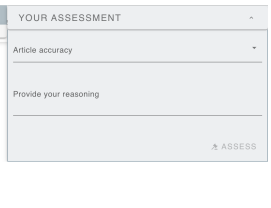
2. For sign up, the user is redirected to the Trustnet platform.



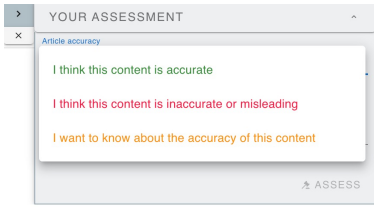
3. Once the user signs up on the Trustnet platform, they can log in. The platform informs them how they can browse through sources and select any they want to follow or Trust.
User can log into the Trustnet extension using the same credentials.



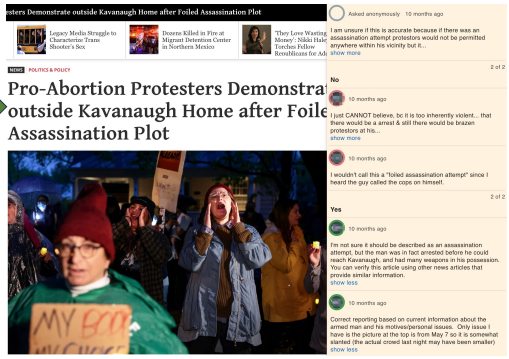
4. With the extension running, user will see a floating button in the corner of any page they visit. The color of the button signifies the accuracy status of the page as assessed by the user's sources. Grey means the page is unassessed.



5. By expanding the pane using the button, the user can input their assessment of the page.

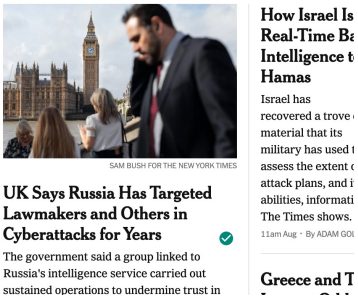


6. The user can assess the content as accurate or inaccurate, or they can ask about its accuracy.



7. If a user is on a page that has been assessed by their sources, the assessments side-pane automatically opens for a few seconds and shows the assessments. The color of the pane signifies the accuracy status of the page as assessed by the user's sources. This is a page that some of the user's assessors have marked as accurate and others as inaccurate. Questions asked about the accuracy of the content are shown at the top of the side-pane.

8. The color of the floating button in the corner matches the color of the pane. By clicking on the floating button, the user can bring the assessments side-pane into view.



9. If the user is on a page that has links to content assessed by their trusted sources, the extension places an icon indicating the accuracy status of their associated resource next to each link. Here, the green checkmark indicates that the content of this link has been assessed as true.

Figure 4: The flow of user interaction with the Trustnet extension, from signing up to assessing and seeing assessments from others.

has a unique URL and therefore the user can see its assessments by their social circle when navigating to the content's URL. However, many of these resources are usually embedded into a feed (e.g., social media posts) or are presented as links on an aggregator page (e.g., the front/home page of a news website). Prior work has reported that links, for instance on social media, are often not clicked, even when they are shared [35]. Therefore, showing the assessments of a page when a user visits the page will not benefit the multitude of those users who do not navigate to the page. To address this issue, the extension also checks the links on the page that the user is presently reading, and places an icon indicating the accuracy status of their associated resource next to

each (if there exists assessments for the resource from the user's network) by slightly modifying the rendered page (see Figure 3). The extension further places a question mark next to the links about the accuracy of whose content other users have inquired. We additionally made the design decision to reduce the salience of links assessed as inaccurate by making them look faded (somewhat transparent but still visible), as seen in Figures 2 & 3. The motivation for reducing the visual salience was to reduce the "illusory truth" effect, in which a single prior exposure to a headline increases subsequent perceptions of its accuracy [80]. By attempting to bias eye movements (a proxy for overt attention) through reducing the visual salience of content assessed as inaccurate, we aimed to make it easier for users to ignore links that are misinforming [46]. In effect, we tried to introduce a design friction as a disincentive—a small hurdle that users would need to overcome to read the content [78]. Nevertheless, users can still consume the content if they wish to as the content is still where it would otherwise be and is visible.

3.2.1 Dealing with Link Complications. A complication of finding assessments of outgoing links on a page is that these links cannot be used as they are found on the rendered page to check whether they have associated assessments in the Trustnet system's backend. The reason is that many links go through (sometimes multiple) redirections before finally redirecting to the link from which the resource can be fetched. For instance, a link shared on X (formerly known as Twitter) is automatically processed and shortened to another with the format of `https://t.co/xxx`. Links on other websites may also first redirect to an intermediary URL so that the website logs outgoing links (i.e., which link a user clicks to leave the page).

To address the issue of redirections, when the extension encounters links on a page, it recursively follows each [3] until it encounters the final link that contains the resource. The reason it is the client rather than the server that follows the redirects is three-fold: First, the number of links a user encounters on any one page can be large, and if all the links encountered by all the users are sent to the server so that it can follow their trail of redirects, the process can incur a substantial delay as there is a limit on the number of parallel requests the server can send and therefore, it would need to batch the requests sequentially. Second, it is conceivable that many of the requests across users are to the same websites and if a large number of requests are made to these websites by one or a few servers, the servers can get rate-limited. Third, the information to which the client has access allows it to visit some links that the server cannot, for instance, those that are behind paywalls. Nevertheless, the client cannot follow all the links as there are occasions when upon following the redirects, the browser encounters a Cross-Origin Resource Sharing (CORS) issue. The CORS issue typically occurs when a web application running in one domain (the "origin") tries to make an HTTP request to a resource located on a different domain. If the server hosting the resource does not explicitly allow the requesting domain to access it, the browser will block the request. This for instance, happens on Facebook or Twitter where the encountered links are shortened links from the social media domain but the target links belong to other domains which enforce the CORS policy [4]. Because CORS is only enforced in the browser, these links are sent to the server for following.

It is likely that the links which have been encountered and whose trails have been followed will be encountered again by the same or other users. Therefore, to boost future performance, after following the links and fetching their target links, the client creates a mapping of the (cleaned[5]) original to the target links and sends it to the server. The server then caches this mapping. When visiting a page, the extension first sends the encountered links on the page to the server and asks whether it holds a mapping for any of them, as fetching the targets is faster this way. With enough people using the extension, the load of following the link redirects on any one client becomes lower, as it will be more probable that the server already knows a mapping for at least some of the requested links. The server discards the mappings that have not been requested for some time.

Most redirects are either HTTP redirects (returning HTTP code 300) or HTML redirects (where the URL to follow can be found in a meta element). Another redirection method is doing so in the JavaScript running on the page which is harder to detect. We wanted the extension to be general-purpose and to work on as many platforms as possible. During development, we tested the extension on numerous news and content sharing websites where we anticipated users would want to post assessments. One where we found JavaScript redirection was Google News. We made specific accommodations to fetch redirect URLs from returned HTML of Google News URLs.

Many URLs have query parameters appended to them that are usually optional key value pairs customizing the query. For instance, clicking on an article link on Facebook will append a query parameter of the form `?fbclid=xxx` to the article's URL, which is used for tracking purposes. Before requesting assessments for URLs or requesting mappings of them from the server, the extension strips the URL of these query parameters because they do not play a part in identifying the resource and their inclusion would result in a false miss in the cache. However, we found a few websites where we anticipated users may want to use the extension and which unconventionally use query parameters to distinguish a resource. These websites include Youtube (where video URLs are of the form `https://www.youtube.com/watch?v=xxx`), Facebook photos, videos, and comments (but not for instance, Facebook posts), and Hacker News posts (where post URLs are of the form `https://news.ycombinator.com/item?id=xxx`). We made specific arrangements so that the extension treats query parameters from these websites as a resource.

The set of links on a page can change even as the URL of the page is constant, for instance, because more content is loaded as a result of an infinite scroll. The extension watches the page for changes and reacts by looking for new links for which to fetch assessments.

The extension additionally has an Options page where the user can specify their blacklisted domains, i.e., where they do not want the extention to run.

[3] In order to protect against broken looping links, there is a maximum depth after which the extension abandons the pursuit.

[4] For example: https://t.co/CUnCxRezGn (origin on Twitter) -> https://www.cnn.com/politics/live-news/election-live-updates-11-07-23/ (resource on CNN.com)

[5] The extension first cleans the links found on the page, for instance by adding missing protocol or hostname.

4 USER STUDY OF THE TOOL

We recruited participants to work with the Trustnet browser extension and the Trustnet platform. We advertised the study on a behavioral research platform managed by our institution with a pool of diverse participants. We asked those who were interested to fill out a survey which in addition to describing the study and providing the consent form, asked about the news sources participants consumed and the subreddits, Facebook groups, and Twitter accounts that they frequented or followed which discussed news. We collected this information to test the extension on these websites or subcommunities prior to the onset of the study and resolve any potential issues. In the survey, we also asked prospective users to refer others in their network who might be interested in participating because the tool can best be leveraged among members of a social circle with pre-existing trust relationships. However, we were unsuccessful in recruiting those that were referred. The survey additionally asked for demographic information, adopted from prior studies on misinformation [49, 51].

For onboarding, we asked that users watch a short video tutorial explaining the systems and their features, sign up for an account on the Trustnet platform (which would enable users to log into the Trustnet extension as well), and add the extension to their Chrome browser through the Chrome Web Store. To bootstrap the study, we configured every user to initially follow all other users of the study upon signup, so that they would be able to see each other's assessments. Users could later choose to unfollow one another. We asked that users perform the following tasks each day during the two-week period of the user study:

- Assess the accuracy of two pieces of content per day. These could be news articles, tweets, Youtube videos, or any other piece of content that the participant wanted. We encouraged our participants to assess as many more pieces of content as they liked.
- Check out the Trustnet feed to see what the users of the study share every day. The reason was because through simply using the extension, with such a small user base and varied interests, it was unlikely that users would serendipitously run into each other's assessments as they were browsing the web. We encouraged, but not required users to share content to the Trustnet feed through the Trustnet extension. The reason we did not require sharing was because a valid use of the system—a feature missing in existing social platforms—is assessing content as misinformation for the benefit of whoever comes across the content, but not wanting to further increase its visibility.

At the study's conclusion, participants completed two surveys asking about their experience with assessing, seeing assessments from others and their perceptions of the tool. The survey included questions such as "To what extent did you like being able to assess content" and "To what extent did you like to see the assessments of other users?" with 5-item likert scale responses as well as free-text elaboration. To better understand users' thought process in assessing content, the survey showed each user examples of content they had assessed as inaccurate and asked them to explain the characteristics of content they assessed as (in)accurate. The questions on the experience of assessing included if users believed their assessments helped others and what content they had difficulty assessing if any, among others. The questions on the experience of receiving assessments included how well-thought-out users found others users' assessments and their accounts of disagreements with others. Some of the survey questions were inspired by prior work on democratized misinformation moderation [51].

To compensate participants, we randomly selected 4 via a raffle and awarded each with a $100 gift card. After analyzing the results, we observed that users had assessed a wide range of content on different topics, not necessarily limited to political news. However, their responses in the post-study survey centered around assessment of political content, possibly because a few questions had primed them to have politics on top of their mind. Therefore, we sent our participants another follow-up survey to further understand their thoughts on how useful they would find assessments on different types of content and their current as well as desired fact-checking practices around different topics. This survey included questions such as "How important is it for you, if at all, to avoid inaccurate information related to each of these topics? (likert scale)", "Who (if anyone) would you like to see assessments from on each of these topics? Please elaborate", and "to what extent do you think you can contribute assessments that can help others on each of these topics". Participants were each compensated with a $10 gift card for completing this follow-up survey. The questionnaires are included in the Supplementary Materials. The study was approved by our Institutional Review Board.

4.1 Participants

A total of 32 users participated in the user study, of whom 25 submitted assessments for 3 or more days. View the full participation plot in Appendix Section A. Out of all the participants, 21 completed the post-study survey and 25 completed the follow-up survey asking about assessments on different topics. 66% were female and 25% were male. 28% identified as Democratic, 19% as Republican, and 53% as Independent. The median age was 26 (ranging from 19 to 67), the highest degree achieved Bachelor's degree (ranging from high school diploma to Doctoral degree or equivalent), and the median income $60,000 - $69,000 (ranging from less than $10,000 to $150,000 or more).

5 RESULTS

For categorizing the themes surfaced in users' qualitative data such as their assessments or answers to the surveys, one member of the research team made multiple passes over the data and used open coding to assign codes to the idea units and axial coding to consolidate the themes [95]. These codes were refined through subsequent passes over the data by constant comparison and theoretical sampling (i.e., building a temporary theory and then collecting more data to test that theory), resulting in the emergence of more generalizable concepts. Then through selective coding, the researcher focused more narrowly on codes of topical interest considering the question at hand [76], e.g., whether there is potential for democratization or personalization of misinformation moderation. During the analysis stage, we followed up with some participants via email to ask them to clarify some of their responses. As described, our use of grounded theory was interpretive and aimed to yield concepts and

themes; therefore, agreement or measures of inter-rater reliability were not appropriate [72].

5.1 The Breadth of User Needs Supports the Case for Democratized Misinformation Moderation (RQ1)

5.1.1 Users Differ with Each Other and with Fact-checkers on the Importance of the Accuracy of Various Topics. Responses in the follow-up survey revealed that users not only differ in the topics that they consume, but also to what extent they find it important to avoid misinformation related to each topic. For instance, while some said that the accuracy of any content matters to them, others considered certain topics of less relevance or importance to them; some believed certain topics are inherently non-factual or controversial, or that it is impossible to find unbiased news related to certain topics. Figure 5 shows the perceived importance of avoiding misinformation related to different topics averaged among our participants. The accuracy of health-related and political content was at least of moderate importance to most users. However, even in our small user sample, there was some variation in how important they found the accuracy of other topics compared to the topics that are often scrutinized by fact-checkers. For instance, one participant deemed it more important for them to avoid misinformation on spiritual content compared to political or scientific content. Another considered the accuracy of content on lifestyle, diet, nutrition, and beauty more important than that on foreign affairs. The status-quo of centralized fact-checking constrained to a limited set of professionals is unlikely to match with the breadth of the types of content for which different users want assessments.

Similarly, the ease of fact-checking content related to each topic, and therefore the need for the involvement of external fact-checkers, varied across users. Some participants explained that they fact-check by reading articles on a topic from different sources (N=6); others relied on their trusted news sources to have done their research (N=3); some said certain topics are harder to fact-check because they are rarely presented objectively or that they do not have objective standards by which they can be assessed (N=4), they are complex and the user does not have the required background knowledge to fully understand them or that the knowledge in the field is constantly evolving (N=5), or that it takes a great deal of effort to assess the accuracy of an article on the topic along various criteria such as missing information and statistics (N=4). Figure 5 shows how easy participants on average find it to fact-check content on various topics.

> *"The more dense topics such as medical questions or in-depth political issues require a lot of foreknowledge or quick education to understand what is being discussed before you can fact-check or make an opinion on something." (P25)*

5.1.2 There is a desire and potential for including others in providing assessments, than professional fact-checkers alone. The viability of an ecosystem that involves users beyond professional fact-checkers in misinformation moderation is dependent on two factors: 1) Whether users would want to view assessments from such assessors, and 2) whether they can contribute assessments that can help others.

To investigate the first aspect, in the follow-up survey, we asked users from whom they want to see assessments on content related to each topic in an extension like the Trustnet extension that allows them to see assessments from those they follow and trust. We have summarized the responses per topic in Table 1. For all of the topics, users wanted to see the assessments of experts and professionals as well as other users. However the notion of what qualifications the experts or what characteristics the users should have varied across participants. These responses corroborated prior work reporting that users want to and do seek the judgment of a variety of different sources they consider trustworthy to determine whether content is credible [51].

What we additionally uncovered was that depending on the topic, participants wished to also view assessments from a broader set of individuals. For instance, on political content, participants liked to see assessments from users with certain or diverse demographics or political leanings or those users who have been affected by certain policies. On issues related to foreign affairs, they wished for the assessments of users or journalists from other countries and immigrants.

To investigate the second aspect, we asked our participants to indicate on which topics they believe they can contribute assessments that can help others. In Section 5.2, we will further shed light on this aspect by reporting on the types of rationales they used in their assessments. Two outcomes that would pose challenges for the viability of such an ecosystem would be if users believed themselves incapable of assessing or if on the contrary, they were ready to profess their opinion about credibility of content without consideration. What we observed however, was that the extent to which participants perceived their own assessments on various topics helpful varied considerably and depended on whether they were an expert on, interested in, or well-read on a certain topic, their confidence in their knowledge, whether they could trust themselves to be unbiased, and whether they considered themselves capable of finding accurate and balanced information to pass their findings to others. Therefore, it was interesting that users exhibited discernment about when their assessments would contribute meaningfully, as demonstrated in this quote:

> *"I know more about certain topics because I work in that area or because I am interested and consume more about those topics. I feel that my assessments would be more useful in those areas rather than in areas I have no experience or interest in." (P20)*

5.1.3 Users Assessed Content from Diverse Sources . Our participants submitted a total of 641 assessments and asked 39 questions about the accuracy of various pieces of content. To understand what type of content users would want to assess if given the opportunity, we investigated the sources to which the content belonged. The sources of content that participants assessed or inquired about using the Trustnet extension included various local, national, and global news, hyper-partisan, or misinformation websites (e.g., BBC, CBS58 Milwaukee, National Review, Info Wars, Common Dreams), news aggregators (e.g., AllSides, Yahoo News), non-English news sources (e.g., El País), government websites (e.g., CDC), academic

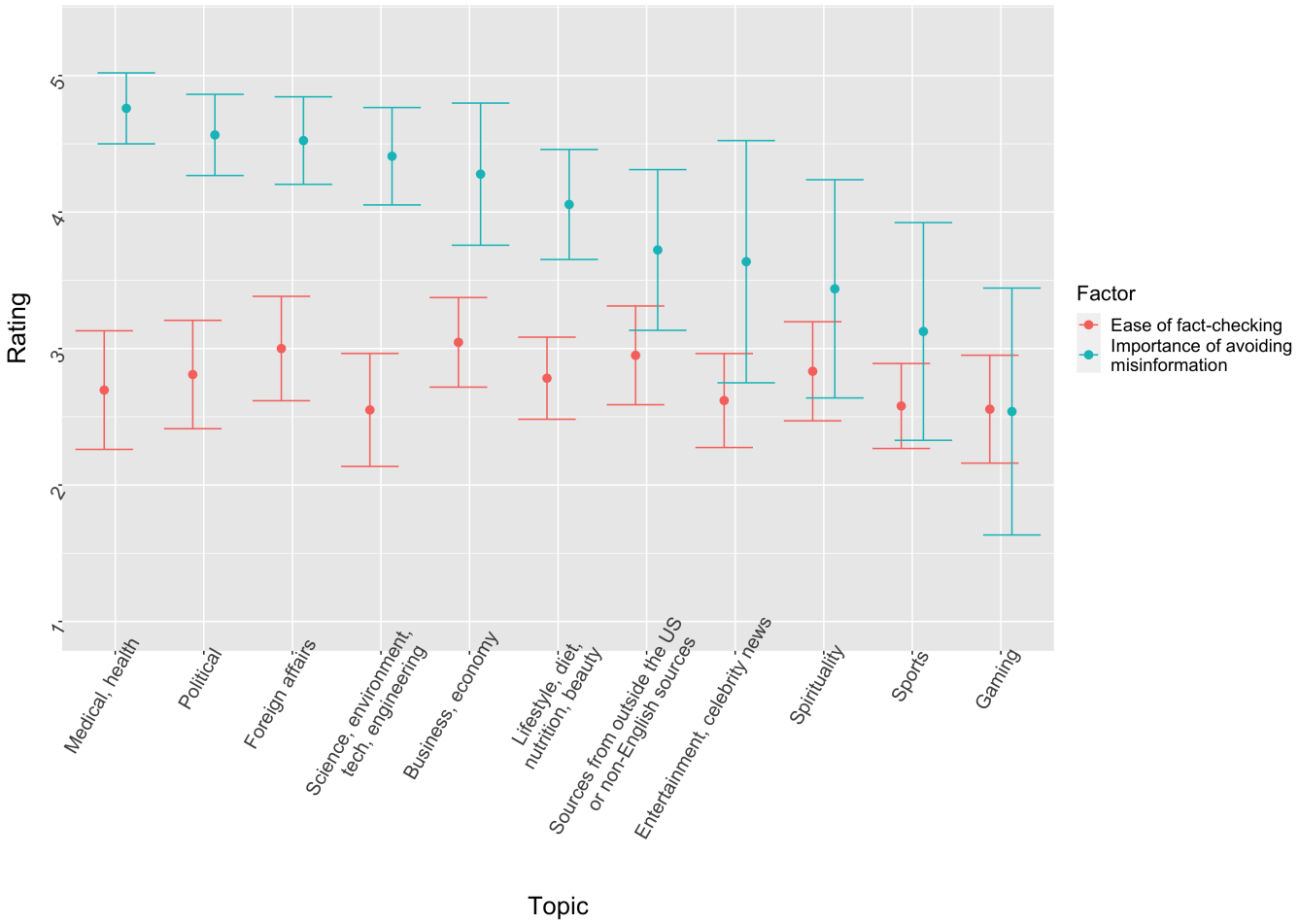


Figure 5: The extent to which participants found it important to avoid misinformation on various topics, and how easy they found it to fact-check information related to the topics they reported at least a little important to avoid misinformation on (rating >= 2), on a scale of 1-5. Each user answered these questions only on those topics that they reported they consume.

journals (e.g., National Library of Medicine), publications of various universities (e.g., NYU Law and University of Bath UK), posts on content sharing websites (e.g., Twitter, Youtube, and Hacker News), and specialty news and blogs (e.g., related to the economy, sports, entertainment and celebrity gossip, diet and nutrition, home improvement, etc.). Source of other content assessed included archives and personal content (e.g., email).

The tweets and Hacker News posts that participants had assessed had embedded links to outside news articles. The assessed Youtube videos were created by accounts that had between 200K and 10M (in the case of major media organizations) subscribers. The content that these videos discussed were diverse including some that touted certain misleading information and conspiracy theories, others that explained various scientific and historical events, some that disclosed the unethical practices of various corporations or the inauthenticity of certain claims such as those of psychics and mediums, and, because it was making news at the time, a number that discussed Johnny Depp vs Amber Heard trials, for instance, inviting body language and behavior experts to "reveal what she really thought".

5.2 Users' Criteria for Assessing Content (RQ2)

We investigated users' assessments to understand what kind of criteria and rationales they had used for evaluating accuracy. To analyze the themes in participants' rationales, we first consulted prior work that presented criteria for assessing content accuracy: the taxonomy of users' rationales for assessing news headlines [49], the context and content credibility indicators developed by the Credibility Coalition in an attempt to define a shared language for classifying misinformation and credibility [105], and problems with news headlines (e.g., clickbait, sensationalism, etc.) that users corrected when given a tool that enabled them to do so [50]. To analyze the rationales that emerged in the assessments (and questions—since some questions also contained rationales for why the content may or may not be accurate) of our users, we first partitioned their responses into idea units, resulting in 732 idea units, each being

Table 1: Participants' responses to who they would like to see assessments from on content related to various topics. Examples of assessors belonging to various categories that participants cited are shown in parentheses.

Topic	Professionals & experts	Academics & researchers	Users	Journalists	Friends & family	Other
Political	N=11 (political analysts, politicians, previous office holders, judges)	N=5 (Political scientists, historians, etc.)	N=9 (those from diverse demographics or political leanings)	N=5	N=2	N=4 (certain media outlets from all sides, political activists)
Medical, health	N=17 (medical specialists)	N=9	N=2 (patients)	N=1	N=1	
Science, environment, tech, engineering	N=9 (those in tech, designers, engineers)	N=9 (the researchers who conduct the studies being written about, peers in the field)	N=4	N=2	N=1	
Business, economy	N=9 (business analysts, CEOs)	N=4 (economists)	N=5 (small business owners)	N=3	N=1	
Foreign affairs	N=6 (diplomats, ambassadors, government workers)	N=3	N=7 (locals, people on both sides of issues)	N=8 (foreign correspondents	N=1	
Lifestyle, diet, nutrition, beauty	N=6 (medical professionals, life gurus, etc.)	N=3	N=6 (women, those interested in weight loss or skincare)		N=1	N=1 (influencers)
Sources outside the US or non-English sources	N=2	N=3	N=4 (immigrants, international audience, those who speak the language fluently)	N=5		N=1 (firsthand sources)
Sports	N=10 (athletes, coaches, retired sportspeople)	N=1	N=2 (fans)	N=3	N=1	
Gaming	N=6 (gamers, e-sports professionals)		N=2			
Entertainment, celebrity news	N=3		N=2	N=4	N=2	
Spirituality	N=6 (pastors and spiritual leaders from certain backgrounds or specializing in certain denominations)	N=2 (theologians)	N=5 (religious users or those with similar moral backgrounds as the user)			

a coherent unit of thought. Of these, a total of 204 idea units did not include rationales, e.g., because they were questions about the accuracy of the content, or they simply summarized the article. We then assigned codes from this body of related work to the rest of the idea units when the codes were appropriate, noting which rationales could not be captured by any of the proposed codes in the related work. The criteria presented in prior work and discussed by our users included the trustworthiness of the source, presenting evidence that corroborates the claim, presentation of all viewpoints, soundness of inference or logic, citation of studies, organizations, or experts, representative citations, title representativeness, among other factors. We then used grounded theory as described in Section 5 to build concepts that could encompass the user rationales that were not assigned labels. Some of these emergent concepts were generalizations of, and therefore encompassed, some of the criteria in the prior work. When reporting the number of citations of rationales based on these more general concepts, the count includes instances that were originally coded using the criteria from the earlier work. To test inter-rater reliability for this part of the results, another coder was trained on the criteria from the developed codebook as well as those in [49, 105] that the criteria in the codebook did not encompass. A sample of the idea units (about 12%) were labeled by the coder. Both the training and test samples were selected such that they covered all of the categories. Cohen's Kappa was 0.63, indicating substantial agreement [63].

Here, we elaborate on some of the criteria that are related to those reported in prior work. We do not elaborate on the rest of the criteria because the definitions accompanied by the examples convey the meaning well. We have summarized all criteria along with representative quotes in Table 2. These can be used as criteria to extend the list of credibility indicators for content and context developed in [105].

5.2.1 Citations are credible. Two indicators in [105] that are closely related to this category include citation of organization and studies and quotes from experts (in the field), as they can add context or support to the article. Our users however, found content credibly not only through researcher and expert citations but also from citations by others, including victim's family, eyewitnesses of various events, cancer patients (e.g., to attest to certain symptoms), and others. Therefore, the notion of who is a credible citation varies by topic.

5.2.2 The details or underlying data are disclosed adequately. This category is closely related to the rationale reported in [49] about why people *believe* news claims—that evidence presented in the article corroborates the claim. In [49], there was no counterpart to this rationale for why people would *disbelieve* a news claim. Our users however, when discussing the data or details presented in an article, sometimes reported inadequate or even more than adequate disclosure of details as a criteria that detracts from article credibility.

5.2.3 The Content Discusses a Scientifically Reproducible Process or a Claim that Can Be Verified. This category is related to a rationale reported in [49] for why people *disbelieve* claims—that the claim references something that is impossible to prove, e.g., using a

metric that cannot be measured. In [49], there was no counterpart to this rationale for why people would *believe* news claims. In the assessments that our users submitted, we observed that because a claim could be verified (e.g., through a scientific process or by cross-checking the evidence), it made users believe the claim as accurate. In addition, one use case of this rationale that we observed, and which was not reported in [49], was that at the moment when the article was published or being read, the general information or evidence at hand was or was not enough to deduce whether the article was accurate. Or, while the claim in the article was accurate at that moment, the situation was still evolving and future developments might render the claim inaccurate, or vice versa.

5.2.4 The Content Distinguishes Between Facts and Opinion. One of the criteria that our users used for assessing content credibility was whether the content was intended to be a factual piece as opposed to for instance, op-ed or anecdote, and if so, whether the piece exhibited any editorialization. Even if the content was editorialized, users considered author's transparency about separating the facts and the opinions as a signal of content credibility. This category is a generalization of a rationale reported in [49] for why users *disbelieve* news claims—that the claim appears motivated or biased.

While our users mostly used the tool to assess content accuracy, in some cases, they put it to other use cases which points to opportunities for enabling individuals to annotate content beyond assessment of factuality. For instance, they used the tool to explain that an article was not factual, but satirical or an opinion piece. Other use cases included expressing opinions or sentiments—that a particular story should not have been published (for instance because it distresses the subjects the story is about or the story is not news worthy) or conversely, that the story is insightful, important, or positive, or to express their (dis)agreement with the message of an article.

5.3 Participants' Perceptions of the Tool and the Affordances It Provided (RQ3)

5.3.1 Perceived Utility of the Tool and Its Advatanges and Downsides (RQ3a). Figures 6 and 7 show the extent to which participants liked being able to assess content and seeing the assessments of other users respectively. We investigated their free-text responses to understand their reasons. The reasons they liked being able to assess included that it helped them think about the news in a more analytical way or gauge their trust in a source (N=5), and that they liked being interactive with the news content they consumed and the ability to call out content they found biased or misleading (N=4). The cited downsides of assessing were that it took extra time and effort (N=4) and that sometimes they find it hard to assess a piece of content e.g., when there is no obvious cause to think the article might be inaccurate or worrying that one is biased in one's assessment (N=3). A frequently cited reason why participants liked seeing assessments from others included because users can get to know other people's perspectives and standards for evaluating content credibility or that it was fun (N=11):

> *"It was really interesting to see what people thought of different articles, and especially to read the reasons WHY they thought that. As opposed to people just blindly taking a side, I liked to see people actually articulate and think through their assessments. It was especially interesting to read when people had different views as to whether content was trustworthy or not." (P14)*

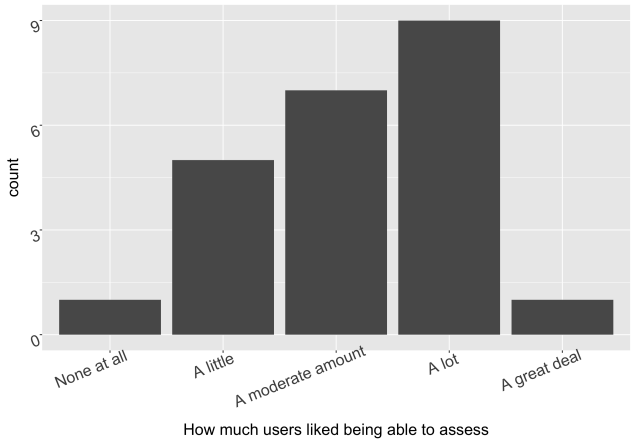


Figure 6: The extent to which users reported that they like the ability to assess content in the post-study survey.

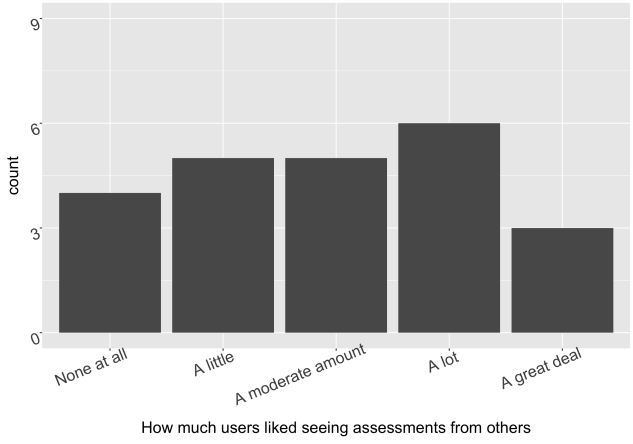


Figure 7: The extent to which users reported that they like seeing assessments from others in the post-study survey.

Other reasons included that by considering whether they agree or disagree with their assessments the user can think more critically about content (N=2), and assessments from others, especially trusted sources, help the user decide whether they should trust an article (N=2). Some participants however, disliked democratized assessments because they believed people are biased or incapable of assessing (N=2).

Figure 8 shows participants' post-study responses to how much they believe a tool like the Trustnet extension would be useful to them if adopted by many people ($\bar{X} = 2.57, s = 0.79$). In the follow-up survey, we asked participants a similar question of to what extent they find it useful to see the assessments of content right on the page. Responses to this question are shown in Figure 9 ($\bar{X} = 2.84, s = 1.07$).

Table 2: The criteria for content accuracy cited by the users of the extension which were not reported in prior work on indicators of content credibility [49, 105]. These criteria can be used to extend the list of credibility indicators.

Criterion	Yes	No
The content distinguishes between facts and opinion. (N=116)	*"It's a good theory; and it's presented as such: a theory. Nothing for sure! So that makes it accurately presented." (P14)*	*"Presents fact and personal opinion in the same setting without clarifying appropriately"(P24)*
The citations are credible. (N=93)	*"This article provides references and published articles/data." (P17)*	*"… the news source uses an unverified; anonymous Twitter account as its primary source." (P1)*
The details or underlying data are disclosed adequately (not less or more than needed). (N=71)	*'Provided quantitative data about the number and location of grocery stores w/in majority and non-majority Black neighborhoods." (P7)* *"I think the inclusion of surveys; statistics; and examples of what makes ads invasive and their prevalence on the internet through this article proves how detrimental they can be to a common user's experience on the internet." (P10)*	*"The writing style feels wrong. There are too many details that should have been withheld; and that gives me reason to think it might be sensationalized" (P33)* *"The article states over 2000 pre-existing brain scans of people with and without anorexia were used. There are many missing important factors that the article author did not address- 1. how many of the scans belonged to people with anorexia (proportions of data); 2. age groups looked at; 3. gender; 4. length of illness; or length of remission; 5. any other illnesses or pre-existing conditions* *Appropriate sourcing is used here and the author uses good quotes. I just wish they would have gone a bit more in depth about the actual data; rather than expecting readers to believe the conclusion right away." (P26)*
The content discusses a scientifically reproducible process or a claim that can be verified. (N=19)	*"Explains a process that is scientifically reproducible." (P32)*	*"Not entirely accurate. Reporting of a few unattributed quotes that businessfolks in China are against the govt's zero covid19 policy. There is no way of verifying the veracity of this..." (P4)*
The investigation appears to be of high quality. (N=17)	*"Well-researched; numerous sources including the original professor/director; explanations of both sides of the issues involved in lithium-sulfur battery production" (P27)*	*"This article is basically a series of tweets. Reading only a paragraph in will make it apparent that the title is absolute clickbait. There isn't a lot of substance here. This was a missed opportunity to report more about the 260 employees that got layed [sic] off." (P26)*
The relevant disclaimers are reported and are salient. (N=8)	*"This article does a good job of vetting the source and study (not peer-reviewed and only 1 study author). It also mentions the author's own; self-admitted weaknesses to the study. Overall this was a balanced take on an unconventional study (may be more accurate to call the study an analysis instead)." (P26)*	*"The caveats named in the story should have been placed higher in the content. This is unfortunately too common among viral health stories." (P1)*
The content answers the main question it poses coherently and stays on topic. (N=5)	—	*"The focus of the article is a question that the content does not answer or give any clarity to." (P1)*
The visual artifacts (esp the lede image) are appropriate and do not have a biased undertone. (N=4)	*"good use of graphs"* (P24)	*"I could already tell it was biased given the fact that it posted a picture of a protest instead of the actual pregnancy center that got torn down" (P11)*
Fabricating claims in the piece would not benefit anyone. (N=3)	*'There is nothing at stake for misinformation here. :)" (P32)*	*"The study was done by a group who actively promotes LGBTQ rights. Of COURSE it's going to be a flawed study!" (P33)*

Many of the rationales (N=14) people cited for why they (dis)liked seeing in-situ assessments were in alignment with reasons why they (dis)liked assessing or seeing assessments from others. Other pros that users discussed were that it in-situ assessments would

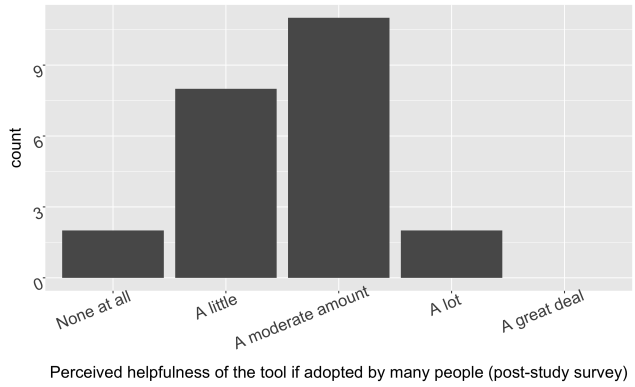


Figure 8: The extent to which participants reported the tool would be useful if adopted by many people on a 1-5 likert scale, in the post-study survey.

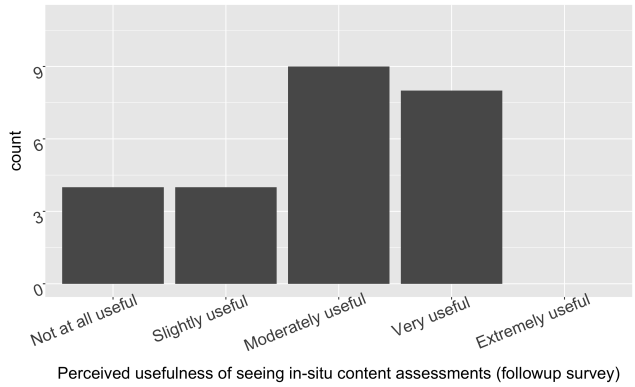


Figure 9: The extent to which participants found seeing in-situ assessments helpful on a 1-5 likert scale, in the followup survey.

save others the work of fact-checking that they may not have time for (N=2) and they provide helpful feedback to journalists (N=1):

"Everyone who has looked into certain topics in-depth should have a right to voice their opinion to others and I think this is a good way of doing so online. It can help improve journalism as they are getting more direct feedback/commenting on their sites instead of comments directed at them but on other social media." (P25)

Additional cons that they discussed included that they want to think for themselves, unassisted by anyone (N=4), a sentiment that has been reported in prior work as well [48], and a worry that users may become accustomed to taking all content with a green checkmark next to it as credible, and not fact-check content for themselves (N=1).

5.3.2 Incentives to Use the Tool, Perceived Problems with, and Ideas for Improving It (RQ3b). We asked our participants what would incentivize them to use such a tool, what they would consider as potential reasons they might want to avoid it, and their ideas for resolving those issues. We have summarized their responses in Table 3.

A common concern about the tool was that those who post assessments may abuse the ability, e.g., by pushing a certain narrative, using harmful language, engaging in coordinated attacks, etc. Related were concerns about not knowing an assessor's political leanings or agenda. In fact, the Trustnet extension gives users the ability to explicitly specify who they trust and follow and only shows a user assessments from their trusted and followed individuals. The user would not even be aware of assessments on a piece of content by others, even a multitude of bots, if the user does not follow or trust them. The reason why our participants were concerned about this point may have been due to the fact that we set up our study such that all our participants would follow each other by default. Although in our tutorials and our communications with them, we told them that they could unfollow others or choose certain individuals as trustworthy, they may not have paid attention to this point.

In their responses, some participants also mentioned their concerns about specific UI design decisions that we had made. For instance, while some participants found the green flag next to links assessed as accurate helpful, a few said that they would not care as much about if a piece of content is accurate compared to if it is inaccurate. In addition, although some users found the partial fading of content assessed as inaccurate helpful, others were against it or considered it a form of censorship. These differences in opinion across users point to a need to give users controls for customizing their experience with the actions taken on content once the content has been classified.

6 DISCUSSION AND FUTURE WORK

The universal in-place assessments displayed through the Trustnet browser extension would be in contrast to expecting the user to actively seek fact-checking information from external sources for every piece of content that they encounter. Furthermore, the extension enabling the user to provide assessment in-place can help others who find the user a trustworthy source.

We observed that our users used the extension to assess a variety of different types of content on various websites, blogs, and social media, and on various topics not limited to politics or health. Not all sources were widely known, but they still have enough consumers for the accuracy of the content they publish to be consequential. In the existing fact-checking model on today's web, the power to assess content is confined to a limited set of individuals and fact-checking initiatives who do not have the resources to cover every piece of content published by every source. For instance, a number of Youtube channels whose videos our participants had assessed as well as some specialty websites and blogs had enough subscribers for a wide reach but perhaps not enough to garner attention from the fact-checking organizations. To scale fact-checking in a world where everyone could be a publisher, we need all the help we can get, even from regular users and perhaps especially those who have interest in and knowledge about a particular topic, to sieve accurate from inaccurate information.

While assessments in our tool are democratized, how often a user views content with credibility assessments depends on who and how many sources they have marked as trustworthy. The issue of scaling such democratized trusted assessments so that a user

Table 3: Participants' reasons to avoid the tool and what would incentivize them to use it.

	Argument	Example
Reasons to avoid	User abuse including biased assessors, harmful language, assessment bombing of certain authors or pages, paid experts (N=9)	*"Assessors would have a lot of power. They would need incentives to not also become corrupted (push views of outside parties or their own)..." (P26)* *"I'm pretty shy online and I value my privacy and professional life, and wouldn't want anything traced back to me if it could be controversial or seen as too political." (P14)*
	Time consuming to contribute or read assessments (N=8)	
	Security and privacy concerns (the extension tracking user activity or the user posting assessments publicly) (N=3)	
	Lack of interest or knowledge in various topics (N=2)	
	Wanting to develop their own critical thinking, not wanting to be told what to believe (N=2)	
Incentives	User being knowledgeable and caring about a particular subject, wanting to avoid misinformation (N=10)	*"Some kind of points or a reward system could be an incentive for contributing. If the tool seems useful and valuable, that itself provides a good incentive too..."* (P20) *"Prior to signing up for the platform, if you are going to be someone who assesses, I think it should be displayed on what their political leanings are. If they have certain credentials to assess certain articles or topics, those should be displayed as well. This would give users a better idea on how the assessor is assessing information. It can instill more confidence if they have credentials on certain topics as well. "* (P17)
	Community engagement, having user's trusted sources use the tool (N=4)	
	Keeping the assessors in check, e.g., by detecting and removing or blocking bots and abusers, or detecting misinformation in assessments of other users (N=5)	
	Monetary compensation (N=4)	
	Earning reputation points or praise (N=3)	
	The tool should ask and display users' political leanings, biases, and credentials on different topics (N=2)	
	Having a pool of large and revolving base of experienced assessors (N=1)	
	Restricting the showing of assessments to topics of interest (N=1)	

sees assessments on content more often can be dealt with in two ways. One is using an AI that learns assessments from a select set of people (e.g., a user's trusted associates) and predicts how they would assess other similar content, such as the AI in [48]. Another is to explore whether a more extensive trust network can be built for each user by leveraging transitivity of trust. Prior work has also examined trust transitivity in social networks, in the context of e-commerce, recommender systems, and chat moderation [26, 57, 67]. Indeed, there is also the potential for a hybrid approach where the extension delivers professional fact-checking information to users in addition to that from regular users, but leaves it to the users to decide which fact checkers to trust.

6.1 Design Implications

6.1.1 User Criteria for Evaluating Content Accuracy. We reported on the criteria that our users used for assessing content credibility in Section 5.2. In contrast to other studies that have presented criteria for assessing content credibility [105] and [49], in our study, users were not restricted to a set of curated news claims or articles, but could use the tool to submit their assessment of any piece of content of their choosing. In this in-the-wild setting, some of the rationales that emerged out of the assessments encompassed, but were broader than, those previously reported in [49] and [105], and some had not been reported before.

Tools that allow users to evaluate content credibility, such as ours, can be augmented with a checklist of criteria surfaced in our study as a form of guideline to nudge the user to reflect on various aspects of credibility, when sharing content, similar to the proposal in [49], or even when reading content. While in [49], going through the checklist in addition to providing free-text reasoning at the time of sharing did not lead users to share less misinformation compared to providing free-text assessments alone, future work should study whether a different set of criteria such as ours could be effective. Even if the structured checklist does not reduce the sharing of misinformation, it can still help those who view the criteria that the assessors have highlighted in structured form, and perhaps even use them in filtering through content or assessments.

6.1.2 Incentivizing Participation. A hurdle in using the tool that participants mentioned is that it can be time consuming to contribute or read assessments. Future work can investigate how to further aid users in their sensemaking process, e.g., by providing a mechanism using which they can collectively cluster, synthesize, and surface unique information in the assessments, similar to Crowdlines or Wikum [68, 106]. The tool can then visualize the analysis provenance so that other users can audit the sensemaking pipeline [65]. The assessments pane can also be structured to include a section into which users can clip "evidence", including links to public hearing videos, studies, fact-checking articles, etc. that can help establish the veracity of claims in the content.

Our users also desired some form of compensation or recognition for their effort. The tool can be enhanced with a reputation system similar to the one in Stack Overflow where users accrue points based on how others rate their assessments or questions about content accuracy. To protect against coordinated attacks against an individual, the reputation score a user A views for user B would not be a global score, but rather a function of the scores that A's trusted sources (or those who have an extended trust path to A) have given to B's assessments. In fact, the score system can be used

to incentivize users to write assessments that can reach across the political divide or to read assessments from others that are not similar to them, e.g., by assigning a higher weight to the upvote on an assessment from an out-group user compared to those from the in-group users.

6.1.3 Democratizing the Design Space of Actions on Labeled Content. We scoped our approach and consequently our tool to democratizing the power to categorize the accuracy status of content. Informed by literature, we then made specific design decisions about what to do with the content that was labeled as accurate or inaccurate and how to signal its status. Our users however, were not in agreement about whether they liked these specific design decisions. For example, while the accuracy and inaccuracy of content were of equal importance to some, others wished to only be notified of inaccuracies. These differences make the case for democratizing the choice in the design space of actions as well. In fact, this is another aspect of moderation that can be democratized regardless of whether the arbitration itself is. Even if accuracy labels are generated by a select few, the decision of whether, how, and in what circumstances users want to view labels can be deferred to users.

6.2 Content Moderation on the Client Side

In the current information ecosystem and without much control over choosing what content they wish to see or avoid, users have submitted to the will of the platforms and their feed curation algorithms. The power of what information flows where is much tilted in the favor of the platforms. Our work is a case study of how we can give more agency to the users and democratize content moderation on the web, without support from the underlying platforms. While our tool was focused on enabling users to assess the accuracy of content and see assessments from their trusted sources, the system can be generalized, empowering users in a variety of other use cases. For instance, our users desired to see assessments from users or news outlets with diverse perspectives or from different sides. Taking this idea further, the architecture and design of our tool can be used to enable users to provide alternative sources for content—those with a different political leaning similar to an in-situ AllSides [1], those that a user thinks are more credible than the source at hand, or those that together with the source can paint a more complete picture of the story. The system can be adopted to label content that is hostile to various marginalized identities, similar to the Shinigami Eyes browser extension that colors trans-friendly and anti-trans social media groups, users, and search results [6]. The system can be used to label content with certain valence, for instance "depressing" or "angry", so that the user may choose to filter out labeled content if and when they are not in a state of mind to consume it. Future work can study different types of democratized governance on these tools, e.g., leaving the decision of categorization to customized trusted sources specified by the user as in our study or allowing the user to choose among various sets of moderator bodies with different standards for what constitutes a positive label. The choice of whether such content should removed from the user's view can be given to the user as well. Indeed, in our study, not every user was in agreement about whether visual fading of content assessed as inaccurate is an acceptable design decision.

The user-generated data in such a tool can further be augmented with machine learning or user-configurable word filters similar to [53] to block content on certain topics, with certain valence, or from certain people that appears on the user's feed. While users cannot stop the flow of unwanted and unasked for messages, for instance those that are forcefully boosted to them (e.g., Elon Musk's tweets receiving the "power user multiplier" special treatment [91]), they can leverage browser extensions such as ours to locally flag such content or remove it from the rendered page.

6.3 Broader Social Implications

6.3.1 Are regular users able to find trustworthy sources? The ecosystem that we propose relies on the users to find and vet others who should be trusted to provide assessments. But is this a challenge that could render the approach not feasible? We argue that on the contrary, this approach enables users to do online what they would do in natural settings, as finding trusted sources and heeding the information that they provide is how people have always determined the truth. Epistemically, social construction of knowledge is individuals receiving testimony from trusted sources about information that they cannot verify by direct observation [94]. On social platforms, users commenting about a post's accuracy contribute to the socially constructed knowledge of others who trust them [22]. Additionally, the presupposition of democracy is that we recognize our fellow citizens as having the capacity to participate in collective governance. However, social platforms, or the web, currently do not accommodate these age-old practices well.

6.3.2 Echo Chambers. An adverse effect of giving users autonomy to decide from whom they would like to see assessments can be the potential for echo chambers. But does this mean that instead, a governing body should make decisions that it decrees are in the best interest of users? Prior work has warned against relinquishing the power of content moderation to platforms and advocated for user autonomy in this space and argued that individuals have a moral right to freedom of listening and against compelled listening [48, 51]. These rights are delivered to users on some platforms through affordances that allow them to for instance, tag, rate, or upvote content and filter based on this metadata. Therefore, such affordances already have precedence on the web. Another advantage to structuring trust and assessments on the web is that such tools as ours can deliver features aimed at directing users to trustworthy contrary viewpoints. In fact, prior work has also reported that users are more receptive to fact-checking information from friends than strangers [42, 70]. In our tool, we made an initial attempt at encouraging users to expand their network by having the extension inform the user if there exist assessments on the page by someone that they do not follow or trust. Because of the sparsity of the data in our small-scale user study, the recommended sources on a page were those who were trusted by the most number of users platform-wide. However, with a larger adoption of the tool, the extension can recommend to a user contrary viewpoints by sources that for instance, have a history of assessments that align with the user's, or sources that are trusted by those that user deems trustworthy.

6.4 Challenges of Finding Link Redirects and Possible Solutions

To display user-generated content, for many use cases including ours, it is necessary to find embedded links on the page that the user is visiting as well as to which resource each link redirects. A potential complication of the client following the trail of redirects for many links from the same domain is that the client can get rate-limited. In our system architecture, because the server caches the redirects that the clients have found, with a large enough number of users, the number of links for which each client needs to find the redirected resource becomes fewer. To further protect against getting rate-limited, the client does not fetch all the links it finds on the page at once, but rather batches the requests and spreads them out. We set the delay between the request batches the same across all websites because many websites do not disclose their rate limits. This static delay results in suboptimal delay in fetching the links on some domains that would allow for more frequent requests and conversely, getting rate-limited on others that only allow for less frequent requests.

However, this issue can be dealt with in two ways. The first is to keep an estimate of the rate-limit of each domain on the server and dynamically change it until the estimate appropriately captures the domain's actual rate-limit. This can be done by for instance, additively increasing the rate of requests the client makes on the domain if it does not get rate-limited, and conversely, multiplicatively decreasing the rate of requests if it does, similar to congestion control schemes in networks. The second is to learn whether the links to internal and external resources on a website are likely to go through redirections. While the extension currently attempts to follow all links found on a webpage, the links on many websites in fact—at least those that direct to internal resources—do not go through redirections. Identifying these websites after a number of encounters can prevent the extension from getting rate-limited, decrease the delay of fetching user-generated content on the links, and reduce the load of caching mappings of links on the server.

7 LIMITATIONS

A limitation of the empowerment brought about by our tool is the inherent limitation of browser extensions—that they only work on desktop and not on mobile browsers or within mobile applications. This is unfortunate especially because certain messaging apps such as WhatsApp have been a hotbed of misinformation [97]. To partially address this issue, we can provide databases or platforms such as [51] where users can copy and paste in-app messages or links to assess them or see their assessments. However, this solution relies on users adopting this extra step into their content consumption process.

Another limitation of this tool concerns how Facebook deals with links. When displaying posts or external links embedded in posts (e.g., posts from news media), the rendered DOM does not include their URLs, unless as a result of user action such as the user hovering their mouse over the elements. Therefore, without this extra action from the user, the extension cannot know the URLs of the displayed posts or content to fetch their assessments. Interestingly, embedded links used to have their URL in the DOM in the early days of the Trustnet extension, but no longer do. The rationale behind this change in design is unclear. However, it points to the need for active maintenance of browser extensions such as Trustnet to keep working on hostile platforms.

While it is conceivable to view a source as generally well-informed in their assessments, trust in a source may not always be a simple binary choice of yes or no. Future work should study how users reason about their degree of trust in various sources and how to incorporate that into such a tool as ours.

8 CONCLUSION

In this work, we explore democratizing misinformation moderation on the web by providing users with a browser extension that empowers them to assess any content including news articles, social media posts, or videos, and to see assessments from the sources they have marked as trustworthy in-situ. We evaluated the potentials and the challenges of such a design through a two-week user study where we asked users to use the tool to assess any content that they wanted. We observed a gap between the content that users want to assess and whose accuracy they consider important and the type of content that is usually assessed by fact-checking initiatives, pointing to a need for giving users autonomy in this space. Our users perceived value in the ability to assess content or see assessments from others right on the content, for instance, because they believed the process helped them think critically about news. They also had ideas for improving the process such as incorporating a reputation or a point system to incentive users to contribute assessments. We also investigated user's rationales for their assessments and believe they can extend credibility indicators reported in prior work and can be used for guiding users in how to evaluate content accuracy. Our work suggests there is potential for democratizing content moderation on the web through tool such as ours.

ACKNOWLEDGMENTS

We thank all the alpha testers of our tool as well as the users of our study. In addition, we thank Professor Michael Bernstein for his feedback on the manuscript.

REFERENCES

[1] 2012. AllSides | Balanced news via media bias ratings for an unbiased news perspective. https://www.allsides.com/unbiased-balanced-news
[2] 2012. Annotate the web, with anyone, anywhere. https://web.hypothes.is/
[3] 2015. Media Bias / Fact Check. https://mediabiasfactcheck.com/
[4] 2018. Ad Fontes Media. https://adfontesmedia.com/
[5] 2018. *Facebook apologises for blocking Prager University's videos.* Retrieved July 9, 2022 from https://www.bbc.com/news/technology-45247302
[6] 2018. Shinigami Eyes. https://shinigami-eyes.github.io/
[7] 2021. *How Facebook's third-party fact-checking program works.* Retrieved August 25, 2022 from https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works
[8] 2022. Diversity of Perspectives | Community Notes. https://communitynotes.twitter.com/guide/en/contributing/diversity-of-perspectives.html
[9] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–19.
[10] Mike Ananny. 2018. The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation. (2018).
[11] Marc-André Argentino. 2021. QAnon and the storm of the US Capitol: The offline effect of online conspiracy theories. *The Conversation* 7 (2021).
[12] Charles Arthur. 2009. *Google Sidewiki: the idea that won't die, but never lives.* Retrieved July 6, 2022 from https://www.theguardian.com/technology/blog/2009/sep/24/google-sidewiki-commenting

[13] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–27.

[14] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2022. What is the Will of the People? Moderation Preferences for Misinformation. *arXiv preprint arXiv:2202.00799* (2022).

[15] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* 6, 10 (2022), 1372–1380.

[16] Shashank Bengali. 2019. How WhatsApp is battling misinformation in India, where "fake news is part of our culture.". *Los Angeles Times. https://www. latimes. com/world/la-fg-india-whatsapp-2019-story. html* (2019).

[17] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. Nudgecred: supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[18] Monika Bickert. 2019. *Combatting Vaccine Misinformation - About Facebook.* Retrieved August 25, 2022 from https://about.fb.com/news/2019/03/combatting-vaccine-misinformation/

[19] Pablo Boczkowski, Eugenia Mitchelstein, and Mora Matassi. 2017. Incidental news: How young people consume news on social media. (2017).

[20] Adrian MP Braşoveanu and Răzvan Andonie. 2021. Integrating machine learning techniques in semantic fake news detection. *Neural Processing Letters* 53, 5 (2021), 3055–3072.

[21] J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation.* Ph. D. Dissertation. University of Oxford.

[22] Amy S Bruckman. 2022. *Should you believe Wikipedia?: online communities and the construction of knowledge.* Cambridge University Press.

[23] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM).* IEEE, 9–16.

[24] Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha, and Abdullah Mueen. 2021. Examining factors associated with twitter account suspension following the 2020 us presidential election. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* 607–612.

[25] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. 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* 42 (2020), 1073–1095.

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

[27] Josh Constine. 2017. *Facebook puts link to 10 tips for spotting 'false news' atop feed.* Retrieved August 25, 2022 from https://techcrunch.com/2017/04/06/facebook-puts-link-to-10-tips-for-spotting-false-news-atop-feed

[28] Nicholas Diakopoulos and Irfan Essa. 2008. An annotation model for making sense of information quality in online video. In *Proceedings of the 3rd International Conference on the Pragmatic Web: Innovating the Interactive Society.* 31–34.

[29] James Price Dillard and Lijiang Shen. 2005. On the nature of reactance and its role in persuasive health communication. *Communication monographs* 72, 2 (2005), 144–168.

[30] Pranav Dixit and Ryan Mac. 2018. How WhatsApp destroyed a village. *Buzzfeed News* (2018).

[31] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web.* 341–350.

[32] Ziv Epstein, Adam J Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G Rand. 2021. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review* (2021).

[33] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 247–256.

[34] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election.. In *CIKM*, Vol. 17. 6–10.

[35] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter?. In *Proceedings of the 2016 ACM SIGMETRICS int'l conf. on measurement and modeling of computer science.* 179–192.

[36] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–16.

[37] Parham Ghobadi. 2022. *Instagram moderators say Iran offered them bribes to remove accounts.* Retrieved November 16, 2022 from https://www.bbc.com/news/world-middle-east-61516126

[38] Tarleton Gillespie. 2022. Do not recommend? Reduction as a form of content moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

[39] Nitesh Goyal and Susan R Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* 288–302.

[40] Nitesh Goyal, Gilly Leshed, and Susan R Fussell. 2013. Effects of Visualization and Note-taking on Sensemaking and Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2721–2724.

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

[42] 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 *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. 187–196.

[43] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.* 1803–1812.

[44] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–21.

[45] 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.

[46] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience* 2, 3 (2001), 194–203.

[47] Amanda Silberling Ivan Mehta. 2022. *Twitter bans posting of handles and links to Facebook, Instagram, Mastodon and more (Updated).* Retrieved March 27, 2023 from https://techcrunch.com/2022/12/18/twitter-wont-let-you-post-your-facebook-instagram-and-mastodon-handles

[48] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. *To appear in the Proceedings of the ACM on Human-Computer Interaction* CHI3 (2023).

[49] 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.

[50] 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.

[51] Farnaz Jahanbakhsh, Amy X Zhang, and David R Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–40.

[52] Tracy Jan and Elizabeth Dwoskin. 2017. *A white man called her kids the n-word. Facebook stopped her from sharing it.* Retrieved March 29, 2023 from https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html

[53] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–21.

[54] Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeff Hancock, and Michael S Bernstein. 2023. Embedding democratic values into social media AIs via societal objective functions. *arXiv preprint arXiv:2307.13912* (2023).

[55] Thorsten Joachims, Dayne Freitag, Tom Mitchell, et al. 1997. Webwatcher: A tour guide for the world wide web. In *IJCAI (1).* Citeseer, 770–777.

[56] Thorsten Joachims, Tom Mitchell, Dayne Freitag, and Robert Armstrong. 1995. *Webwatcher: Machine learning and hypertext.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

[57] 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).* Citeseer, 9–22.

[58] Jeremiah H Kalir. 2019. Open web annotation as collaborative learning. *First Monday* (2019).

[59] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In

Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. 999–1010.

[60] András Koltay. 2021. The Protection of Freedom of Expression from Social Media Platforms. *Mercer L. Rev.* 73 (2021), 523.

[61] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1188–1199.

[62] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[63] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[64] Issie Lapowski. 2018. Newsguard wants to fight fake news with humans, not algorithms. *Wired, August* 23 (2018).

[65] Tianyi Li, Yasmine Belghith, Chris North, and Kurt Luther. 2020. Crowdtrace: Visualizing provenance in distributed sensemaking. In *2020 IEEE Visualization Conference (VIS)*. IEEE, 191–195.

[66] 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.

[67] 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. 1222–1229.

[68] Kurt Luther, Nathan Hahn, Steven Dow, and Aniket Kittur. 2015. Crowdlines: Supporting synthesis of diverse information sources through crowdsourced outlines. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 3. 110–119.

[69] Pranav Malhotra. 2020. <? covid19?> A Relationship-Centered and Culturally Informed Approach to Studying Misinformation on COVID-19. *Social Media+ Society* 6, 3 (2020), 2056305120948224.

[70] 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.

[71] James C McCroskey and Jason J Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* 66, 1 (1999), 90–103.

[72] 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.

[73] Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. 2019. Whatsapp monitor: A fact-checking system for whatsapp. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 676–677.

[74] Mohsen Mosleh, Gordon Pennycook, Antonio A Arechar, and David G Rand. 2021. Cognitive reflection correlates with behavior on Twitter. *Nature communications* 12, 1 (2021), 921.

[75] Adam Mosseri. 2016. News feed fyi: Addressing hoaxes and fake news. *Facebook newsroom* 15 (2016), 12.

[76] Michael J Muller and Sandra Kogan. 2010. Grounded theory method in HCI and CSCW. *Cambridge: IBM Center for Social Software* 28, 2 (2010), 1–46.

[77] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[78] Paul Ohm and Jonathan Frankle. 2018. Desirable inefficiency. *Fla. L. Rev.* 70 (2018), 777.

[79] 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* 66, 11 (2020), 4944–4957.

[80] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.

[81] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[82] 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.

[83] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.

[84] Sarah Perez. 2019. *Facebook News Feed changes downrank misleading health info and dangerous 'cures'.* Retrieved August 25, 2022 from https://techcrunch.com/2019/07/02/facebook-news-feed-changes-downrank-misleading-health-info-and-dangerous-cures/

[85] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 153–162.

[86] 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.

[87] Ronald A Rensink. 2005. Change blindness. In *Neurobiology of attention*. Elsevier, 76–81.

[88] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2018. BaitBuster: a clickbait identification framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[89] 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.

[90] Aaron Sankin. 2017. *How activists of color lose battles against Facebook's moderator army.* Retrieved March 29, 2023 from https://revealnews.org/article/how-activists-of-color-lose-battles-against-facebooks-moderator-army/

[91] Zoe Schiffer and Casey Newton. 2023. *Yes, Elon Musk created a special system for showing you all his tweets first.* Retrieved March 27, 2023 from https://www.theverge.com/2023/2/14/23600358/elon-musk-tweets-algorithm-changes-twitter

[92] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*. 265–274.

[93] 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).

[94] Matthias Steup and R Neta. 2005. Stanford Encyclopedia of Philosophy. Epistemology.

[95] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).

[96] Naomi Thomas. 2022. Doctors worry that online misinformation will push abortion-seekers toward ineffective, dangerous methods. *CNN* (2022).

[97] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. 2022. Accost, Accede, or Amplify: Attitudes towards COVID-19 Misinformation on WhatsApp in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[98] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting expert image geolocation with crowdsourcing and shared representations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[99] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[100] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine* 240 (2019), 112552.

[101] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–14.

[102] 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.

[103] Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass* 15, 10 (2021), e12438.

[104] Amy X Zhang, Joshua Blum, and David R Karger. 2016. Opportunities and challenges around a tool for social and public web activity tracking. In *Proceedings of the 19th ACM Conf. on Computer-Supported Cooperative Work & Social Computing*. 913–925.

[105] 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.

[106] 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.

[107] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the sigchi conf. on human factors in computing systems*. 1883–1892.

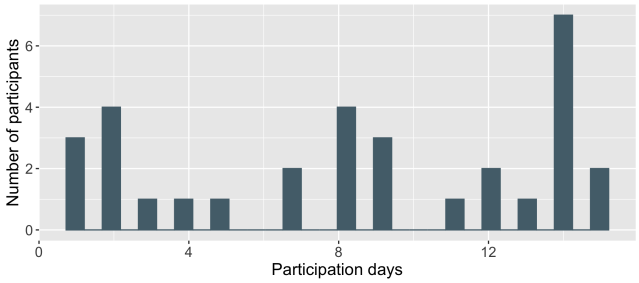


Figure 10: The distribution of how many participants completed the daily tasks for how many days. Not all the ones that completed the tasks for fewer than 14 days dropped out before the study's conclusion. Rather, they did not complete the tasks in certain days.

A PARTICIPATION

Figure 10 shows the participation distribution across our study participants.

B PARTISANSHIP

Prior work has reported that users' assessments of content accuracy correlates with whether the their partisanship aligns with that of the content [9, 49]. We wanted to further investigate whether partisanship plays a role in the context of our study where users were not restricted in their choice of which content to assess. Therefore, we consulted the media bias ratings provided by 3 sources (AllSides [1], Ad Fontes Media [4], and Media Bias/Fact Check (MBFC) [3]) and developed a normalized numeric political leaning value averaged from the ratings given by these sources. The value ranged from -1 to 1, with negative values indicating a left-leaning ideology, and positive values indicating a right-leaning ideology. A number of domains in our dataset were aggregator websites such as Hacker News, Allsides, and Yahoo News that present articles from various media. For retrieving the bias of the articles assessed by our users on these aggregators, we considered the original source of the articles, for instance LA Times for an article assessed on Yahoo News.

The political leaning information on a number of sources was not available in these repositories. Some of these sources for instance, were not major news websites or covered specialty news. In fact, the lack of information about bias or credibility of a number of sources that our small user sample consumed points to the gap—perhaps resulting from limited resources—between the content fact-checked by professionals and the content that the general public consumes and about whose accuracy they inquire. The number of datapoints (i.e., user assessments) for which we had the political leaning of the media source was 588.

To measure partisanship concordance between a user and the content that the user assessed, we made the simplifying assumption that the leaning of the content is the same as the leaning of its media source. We labeled each user as either a Democratic or a Republican. We were able to place all participants including those who identified as Independent or other in one of the two Democrat or Republican categories because in addition to party, we had asked participants about their political preference (strongly Republican, lean Republican, Republican, Democrat, lean Democrat, strongly Democrat). If the numeric political leaning value of the media source was negative and the user was labeled as a Democratic, or if the media value was positive and the user was a Republican, partisanship concordance had a value of 1. Otherwise, partisanship concordance was 0.

To understand if partisanship concordance has an effect on users' accuracy judgment, we fit a generalized linear model to the data. To do so, we used the function "glmer" from the R package "lme4" and used the family function "Binomial" with the link "logit". We included user and content identifiers as random effects to account for variation in the outcome as a result of the unobserved characteristics of a particular user or a piece of content. The model was as follows:

$$\begin{aligned}\text{accuracy assessment}_{uc} = \beta 1(\text{ partisanship concordance})_{uc} \\ + b_u Z_u + b_c Z_c + \epsilon_{uc}\end{aligned} \quad (1)$$

where the accuracy assessment$_{uc}$ is user u's assessment of content c. The indicator variable 1(partisanship concordance)$_{uc}$ denotes whether user u's political leaning is aligned with content c. Z_u is the matrix for the random effects for observations for participant u, and Z_c for content c. ϵ_{uc} is the error term. β is the coefficient of interest.

Interestingly, we observed that partisanship concordance did not have a significant effect on users' assessment [$\beta = 0.45, z = 0.56, p = 0.58$], suggesting that the effect of partisanship on users' assessments when users are not restricted in what content to assess warrants more research.