Exploring the Use of Personalized AI for Identifying Misinformation on Social Media

Farnaz Jahanbakhsh*
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
Cambridge, USA

Yannis Katsis
IBM Research
Almaden, USA

Dakuo Wang
Northeastern University
Boston, USA

Lucian Popa
IBM Research
Almaden, USA

Michael Muller
IBM Research
Cambridge, USA

ABSTRACT

This work aims to explore how human assessments and AI predictions can be combined to identify misinformation on social media. To do so, we design a personalized AI which iteratively takes as training data a single user's assessment of content and predicts how the same user would assess other content. We conduct a user study in which participants interact with a personalized AI that **learns** their assessments of a feed of tweets, **shows** its predictions of whether a user would find other tweets (in)accurate, and **evolves** according to the user feedback. We study how users perceive such an AI, and whether the AI predictions influence users' judgment. We find that this influence does exist and it grows larger over time, but it is reduced when users provide reasoning for their assessment. We draw from our empirical observations to identify design implications and directions for future work.

CCS CONCEPTS

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

KEYWORDS

Misinformation, Artificial Intelligence, Social Media, Fact Checking, Democratized Content Moderation

ACM Reference Format:
Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3544548.3581219

*Research performed while interning at IBM Research.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581219

1 INTRODUCTION

With the rise of misinformation on online social spaces and the grievances that it has caused especially in recent years [9, 13, 28, 81], researchers as well as social media platforms have been investigating how to identify misinformation and what to do with the misinforming content once it has been identified. The approaches currently deployed on social media platforms treat the platforms (their AI, human moderators, and partnered third-party fact-checkers) as the authorities on the truth. However, centralized content moderation by the platforms is contested by some scholars as well as platform users who believe that it can inhibit users' freedom of speech rights and autonomy in deciding what content to consume [41, 54, 60, 82]. In addition, the centralized decision to, for instance, block misinforming content from the users' view may not be aligned with what users want or need, as some users want to see such content nonetheless, so that they can assess it for the benefit of their friends and family who are otherwise exposed to it or to be aware of what content their social circle share [54, 64, 80].

Rather than deciding what users should or should not consume, another body of work has studied how to empower users to determine content credibility for themselves. An approach studied in this corpus of work is to enable users to assess the accuracy of content which has been shown to reduce the sharing of misinformation [52], as it primes users to have accuracy on top of their mind [75, 76]. With user assessments captured as structured metadata, they can be displayed on content in a structured form as well, and have the potential to warn the assessor's social circle should they come across the misinforming content [54].

While it is feasible to ask users to assess content whenever they are about to share it, the intervention cannot be employed on every piece of content that users see as they scroll through their feed. Although a user can rely on assessments from their trusted sources or social circle [54], the assessments from this limited set of users are also unlikely to match in scale to the amount of content that the user encounters. Additionally, with many different sources publishing the same story or posting about similar claims, there is the missed opportunity that a user's assessment on one such post is only tied to the post and is not displayed on similar content.

Therefore, in this work, we attempt to tackle these issues by amplifying democratized assessments through the design of a **Personalized AI** system. The personalized AI can train on the data from a single user and does not need to share that data (or the

trained model) with other users, ensuring data privacy. For each user, the personalized AI takes as training data the assessments that the user has provided so far, and makes predictions on how the user is likely to assess other content that they have not assessed. The AI iteratively retrains itself as the user provides more feedback.

A personalized AI that predicts a user's assessment of content can potentially benefit the user by serving as an aide — a first-pass inspection of the content that the user encounters. It can save the user time by directing their attention to content that they would likely find credible, or conversely, to scope the user's attention to items that they would likely find inaccurate to demand the user's explicit assessment on them. Such a personalized AI can also act as a guard — when the user is about to share a post that the AI predicts the user may assess as inaccurate, the AI can nudge the user to have accuracy on top of their mind. This could be an alternative to asking users to assess the accuracy of each item that they are about to share, as proposed in [52]. If the accuracy predictions of a user's personalized AI are displayed publicly, they also have the potential to benefit the user's social circle. In this scenario, a user's warning of misinformation would have a wider reach if it is also displayed on content similar to the one that the user has assessed. This approach can address the scale problem mentioned before.

In this work, we begin to explore the potentials and the challenges of incorporating a personalized AI for determining content accuracy on social media. We have no intention to suggest such a personalized AI should or could replace a user's own judgment. In addition, we do not argue that such an approach is better than the status quo of leaving content moderation to the platforms; rather, we perform an initial exploration of this setting and identify needs for such an approach, reservations about it, and potential problems that need to be dealt with before a similar personalized AI technology for content moderation can be deployed in the wild in the foreseeable future. We investigate users' perception of such an approach through a **user study** where rather than asking users to imagine the technology, we have participants interact with it in a setting similar to the one we envision, in effect conducting a technology probe [49]. We present participants with a feed of tweets that they need to assess and train a personalized AI model for each participant in real time based on the assessments that the participant provides. To more closely mimic our envisioned AI if it were deployed on social media, the AI that we train for each participant evolves and updates its predictions as the participant provides more assessments.

A potential challenge that could arise if such a tool were deployed on social media is that displaying the AI predictions of how a user would assess content may end up influencing the user's own assessment. The existence of a somewhat similar effect has been reported in contexts where users collaborate with a system to make a decision [22, 39, 69]. This potential influence can be problematic in cases where the AI mispredicts the user's assessments, causing the user to either believe content that they would assess as inaccurate if they were nudged to think about it; or conversely, to disregard as inaccurate content that they would otherwise assess as credible (i.e., Overreliance on AI [11]). The influence of the AI can manifest differently when the AI does accurately predict how the user would assess a piece of content. In such cases, by seeing that the AI agrees with their assessment, the user's confidence in their assessment may grow, potentially making their stance on the issue more extreme.

To understand the possibility of the existence of such effects, we designed our user study such that we would be able to compare users' assessments when the predictions of their personalized AI were shown and when the predictions were withheld from the users' view. Our results suggest that users' decisions in deciding whether a piece of content is accurate were swayed by seeing the predictions of their personalized AI. Additionally, the influence of AI over users grew over time. However, this effect disappeared when users followed up on their assessment by providing justifications for their choices. Nevertheless, users' agreement with their AI did not affect their confidence about their assessment.

This work contributes 1) insight into how users perceive a personalized AI for assessing the accuracy of online content that takes as training data a user's assessments, 2) an empirical understanding of whether and how users are influenced by seeing the accuracy predictions of the AI, 3) identifying an intervention that could mitigate this influence, and 4) identifying design implications and research directions for the use of a personalized AI for content moderation based on our study observations, as well as ethical considerations around such an approach.

2 RELATED WORK

We situate our work in the literature related to detecting and dealing with misinformation on social media platforms, as well as the influence of AI on human decision making.

2.1 Approaches to Detect and Deal with Misinformation

Given widespread concerns about misinformation on social media, platforms as well as researchers have been investigating how to address the misinformation problem. The design space of the approaches against misinformation is generally contained within three dimensions:

- who gets to decide what is misinformation (i.e., if it is centralized or democratized),
- to what extent automation is used in the detection of false content,
- and what is done to the content detected as misinformation.

For instance, many machine learning algorithms have been proposed to detect misinformation [5, 8, 44, 63], with the general assumption that the (large) dataset of ground truth that is fed to them as training data is determined by an authority, e.g., the platforms. Fact-checking initiatives serve as another centralized approach, albeit one that employs human moderators. Platforms have also reported using a combination of AI and third-party moderators and human fact-checkers to detect misinformation [4, 7]. Once the misinforming content has been identified, platforms take action against it by flagging, downranking, or even removing it [17, 25, 73, 79]. Researchers have also investigated the use of warning labels or placing fact-checking information alongside articles by the platforms. Some of these studies tested different types of warning labels (e.g., content disputed or source evaluated as unreliable by the crowd, news media, fact-checkers, or a centralized AI) and reported that they increase user discernment of content accuracy or reduce users'

intention of sharing misinformation, albeit different types have varying degrees of efficacy [15, 84, 98]. Epstein et al. further found that explanations for how a hypothetical hybrid crowd-AI labels misinformation increase the effectiveness of warning labels [30]. Others studies however, found warning labels have limited or negative effects [23, 36, 74].

Content moderation by social media platforms as a central authority has been a point of contention among scholars as well as social media users. Relinquishing the power of truth governance to the platforms can be at tension with freedom of speech and the autonomy of individuals in deciding what content to consume and can inhibit the development of a free market of ideas needed for citizens in a democratic society to perform their civic duties [41, 60]. Some users describe fact-checking labels assigned to content by social media platforms as judgmental, paternalistic, and against platform ethos and distrust platforms as fact-checkers because they consider them profit-driven and politically biased [82]. Other users however, call for stronger labels for content with stronger perceived harm, more obvious and striking labels, or even removal of completely inaccurate content [57, 82]. Other studies have also reported users' distrust of the platforms as arbiters of truth [54].

The concerns about the role of platforms as content moderators are legitimate as there have been cases when the platforms have blocked content that arguably did not have potential to harm or content by dissidents in certain autocratic countries [3, 29, 50, 87, 92]. In addition, there may be reasons why a user wishes for inaccurate information not to be filtered from their view, e.g., to assess it for the benefit of their friends and family who are nonetheless exposed to it or to be aware of what their social circle share or think [54, 64, 80].

The approaches that democratize the power to decide what is misinformation fall into two categories. Those that belong to the first delegate the decision of what content is false or harmful to the crowd or a set of users appointed as moderators [6, 16, 31, 56, 77], but the resulting decision is nevertheless imposed on all other users of the community as well. Content moderation on subreddits or Facebook groups is an example of such an approach. On the other hand, the approaches in the second category do not impose a single source of truth on all users and instead study how to enable individuals to make more informed decisions about the credibility of content they encounter online. For instance, Zhang et al. compiled a list of credibility indicators that news publishing media can use to differentiate themselves from low quality publishers or bad actors [100]. Jahanbakhsh et al. studied interventions that could be deployed on social media to nudge people to think about the accuracy of a post before sharing it and reported that they result in a reduction in the likelihood of sharing falsehoods. These interventions include asking people to assess the accuracy of the post, and to explain their rationale for their assessment [52]. These findings are corroborated by Pennycook et al. who found that subtly priming users to have accuracy on top of their mind reduces the likelihood that they share misinformation [75, 76]. They have argued that the reason these nudges work is that users are generally discerning of content accuracy; however, at the time of sharing, their attention is directed away from accuracy, to the social feedback that they would receive. This argument aligns with prior work that found that people who rely more on their intuition and engage in less critical thinking are more susceptible to believing political fake news in controlled survey experiments [78] and in fact share news from lower quality sources on Twitter [72]. In a similar vein, Heuer et al. studied the effectiveness of an interactive checklist based on recommendations of the World Health Organization placed alongside news articles to nudge people to investigate the reliability of articles. The checklist reminds people to follow certain practices such as examining the article's author and the supporting evidence [48].

An example of democratizing the power to not only determine what content is misinforming, but also of what to do about it was proposed by Jahanbakhsh et al. in [54]. Through a survey, the authors uncovered how users attempt to help each other avoid misinformation. They then proposed a set of design affordances which, if incorporated into social media platforms, can support users' practices in collectively fighting against misinformation: 1) enabling all users to assess content and for their assessments to be captured as structured metadata, 2) enabling users to specify a set of trusted sources—those whose assessments they deem trustworthy, and 3) giving users filters that they can use to block out misinformation from their feed as assessed by those they trust. These affordances are rooted in the observation that users already seek fact checking information about the content that they encounter online from those they trust within their social circle and provide fact-checking information to their friends and family, albeit without platform support, and that they have different preferences for whether misinformation should be kept in or out of their feed [54]. Therefore, enabling users to assess content can not only help them engage more in critical thinking and reduce the likelihood that they share misinformation, but also help their social circle receive assessments from them, which they are more likely to heed than assessments from strangers [45, 65]. Decentralized decision making or delegating content moderation to personalized moderators has been explored by prior work in the context of modifying news headlines perceived as misleading or inaccurate, subjective moderation in chat, email harassment, allowing or removing content, or developing policies for platform governance [24, 34, 37, 53, 55, 62, 99].

While users can be asked to assess or think critically about every piece of content that they are about to share, and therefore reduce the spread of misinformation at the source, misinformation will still make its way to users' feeds and users still need to be wary of it. Deploying the same intervention for every post that comes into the users' view is not feasible and although displaying structured assessments from users' trusted sources and social circle on content can help users avoid misinformation, such assessments are limited in scale and may not cover the entirety of posts from a user's feed. In this work, we aim to tackle this issue by exploring a part of the design space that that has not been studied, i.e., the use of automation for democratized detection of misinformation. We investigate how a personalized AI trained on a user's assessments can be used to predict how the user is likely to assess other content that they have not assessed, and with that widen the reach of the user's limited set of assessments.

2.2 AI's Influence on Human Decision Making

A potential negative consequence of deploying such AI as the one we envision is that its predictions about how a user is likely to assess content may end up influencing the user's judgment on the content's accuracy, hence creating a self-fulfilling prophecy. Users' blindly following the AI can undermine one of the motivations for deploying democratized assessments in the first place—to encourage users to engage more in critical thinking and pay attention to accuracy [52, 75]. Research investigating the use of automated system has reported the existence of automation bias, which describes errors that occur when decision makers rely on automated cues rather than engaging in vigilant information seeking and processing [22, 39, 69]. Prior work reports lack of vigilance and cognitive laziness as some reasons behind this phenomenon [88].

A thread of work has investigated positive or negative machine heuristic as another type of bias that users have toward automated systems. Individuals with a positive machine heuristic believe that machines are more accurate and precise than humans. For instance, Sundar et al. reported that users are more likely to reveal personal information if they believe that a machine, rather than a human, is handling their information [94]. Molina et al. found that users who distrust other humans favor content moderation by AI over moderation by human judges. Moreover, those who fear AI believe that AI is unable to make nuanced subjective judgments [67]. Wang et al. compared the influence of AI vs human experts on users in a task of rating the quality of profile photos and found no difference [96]. Related are studies that investigate the factors that influence users' trust in AI models. Such factors include opportunity for user to provide feedback to the AI, the presence of explanations or confidence scores provided by the AI [12, 90, 101].

Researchers have investigated how to reduce over-reliance on AI in systems that aid in decision making. For instance, prior work has studied whether explanations provided by AI models can give users insight into when AI's reasoning is incorrect and found that explanations have limited success in reducing users' over-reliance on AI [12, 20, 101]. Buçina et al. proposed cognitive forcing interventions to disrupt heuristic reasoning such as users waiting before they are shown AI's suggestion as a measure to counter the over-reliance [21]. In the context of robots, Wagner et al. suggest avoiding features that nudge users towards anthropomorphizing robots, as it can give users a false sense of familiarity [95].

A difference between our scenario and the context of automation bias in decision support systems or negative machine heuristic, is that those systems either are knowledgeable beings of their own who could complement the user's knowledge in decision making or replace another individual's expertise on which the user would otherwise rely. Our envisioned personalized AI, however, does not offer insights beyond what the user already knows and instead, attempts to capture what the user's decision would be (and remind the user of the decision when they appear not to be thinking critically). In that, the context of our envisioned AI resembles personalized recommender systems that attempt to predict user interests by finding the interests of similar users [18, 19, 89]. Prior work has found evidence that such recommender systems can manipulate users into following the action that is recommended to them or agreeing with the recommender's predicted ratings [27, 42]. Users' satisfaction with a recommender system has been found to be correlated with the soundness of their mental model of how the system operates [61]. Although these recommender systems and our envisioned AI are both personalized, the former draw from the data of others in addition to that of the user, and the latter learns from the user only. Therefore, it remains to be seen whether even by knowing that a (personalized) AI's knowledge is an imperfect version of a user's knowledge, the user would still be influenced by seeing the predictions of the AI.

The bulk of the body of work on automation bias focuses on decision support systems that help users decide whether to take a particular action, for instance, for a medical professional to decide whether to pursue a particular treatment for a patient or aviation aids for pilots. In such conditions, two types of errors can happen as a result of over-reliance on automated systems: *commission errors* that are made when decision makers take inappropriate action because they over-attend to automated information or directives and *omission errors* that occur when decision makers do not take appropriate action because they are not informed of a problem or situation by automated aids [71]. Similarly, the studies investigating whether recommender systems create a self-fulfilling prophecy gauge an observable outcome, such as the rating a user gives to a recommended item or whether the user engages with the recommended item [27].

In these scenarios, the existence of the influence of the AI can be ascertained by inspecting the user's actions or inactions. However, in the context of the personalized AI that we imagine, over-reliance on AI will not result in an immediate observable action or outcome. Rather, it may result in, for instance, a shift in users' (latent) beliefs over time or possibly a higher likelihood to share misinformation in the long run. The lack of visible and immediate consequences therefore, makes it all the more important to study and elicit, through a careful experiment design, what the potential consequences would be before such an approach can be deployed in the wild.

2.3 Research Questions

Motivated by prior work, our work explores the following research questions:

- RQ1: How do users perceive a personalized AI for determining content accuracy on social media?
- RQ2: Does showing the predictions of a personalized AI about the accuracy of content affect how users would assess the content?

3 METHOD

To answer our research questions, we designed and developed a platform that made the human-AI interaction we envisioned possible. On this platform, our study participants could assess a feed of tweets and receive their personalized AI's predictions of how they would assess other tweets that they had not already assessed. A participant's personalized AI would retrain and its predictions would update as the participant provided more assessments. The purpose of the task was twofold: first, to expose participants to the experience of interacting with a personalized AI for determining content accuracy on social media and gauge their perceptions; and

second, to understand whether a user's decision of how to assess a tweet would be influenced by an AI that attempts to predict how the user is likely to assess the tweet.

To study the influence of the AI on user's decisions, we could not ask that users assess the accuracy of tweets before and after seeing AI's predictions; because once primed, users would be likely to retain their initial assessment of a tweet even when the AI prediction was displayed to them later. Instead, we needed to compare user's agreement with the AI across tweets, with some tweets displayed with the AI predictions, and some where the AI predictions were withheld from the users' view. However, because our system incorporated an AI that evolved as the user assessed more tweets, we could not use only one AI and display its predictions on some tweets and withhold them on others. Otherwise, the order of assessing tweets with and without predictions displayed and the AI's accuracy at various points in time would create a confounding effect on user decision. We decided to train two separate AI models evolving in isolation from each other. Then we could display to the user the predictions of one on some tweets and withhold the predictions of the other AI on other tweets and compare how often the user agreed with the predictions. The prerequisite for this comparison was that the *expected* performances (the expected user-AI agreement) of the two AI models be similar so that any difference in the *observed* AI-user agreement would be attributed to the users seeing or not seeing the AI predictions, and not the performance of the AI models. Below, we explain how we set up the study to achieve this goal.

3.1 Task

The task involved participants interacting with 3 feeds that we had curated for them, each containing 26 tweets related to COVID-19. Each participant was required to:

- assess all the tweets on each feed as accurate or inaccurate,
- indicate how confident they were in their assessment (on a 5 point likert scale),
- and provide their reasoning for why they believed the tweet was (in)accurate for at least 3 tweets in each feed.

The accuracy ratings and confidence scores would help us ascertain whether participants would be influenced by seeing AI predictions, either by agreeing with the accuracy ratings generated by the AI, or by gaining more confidence about their assessments on the items they agree on with the AI. The free-text reasons could help us gain insight into users' reasons for (dis)believing social media posts as well as distinguish between spammers and workers who were performing the task legitimately. The 78 tweets of the study were all different, meaning that each participant assessed each tweet only once.

In deciding how to capture the accuracy assessments from participants, we consulted the prior studies on misinformation that ask users to assess the accuracy of news claims as an accuracy nudge, which we determined were most similar to the context of our study [52, 54, 77]. We adopted the two-item measurement of accuracy from these studies.

Each of the 3 feeds marks a step in our experimental setup. The purpose of the first, which we refer to as the Seeding Step, was to collect some initial data from the participant to bootstrap the first version of a personalized AI model that predicted how the participant was likely to assess other tweets. The tweets belonging to the Seeding Step were randomly sampled at the onset of the study and were the same across all users. In this step, for each participant, 2 identical models were trained on the same data (i.e., the tweets from the Seeding step) in real time, which we refer to as models Hidden and Visible. Section A in the Appendix describes the details of the personalized AI system that we used for training the models in the study.

After the Seeding step, the participant was prompted that we needed more data from them, and was redirected to the second step, which we refer to as condition Unassisted. In this step, the participant viewed a feed much similar to that of the Seeding step and was asked to assess each tweet on this feed. However, the experimental platform had under the hood used the user's personalized model Hidden to predict the user's assessment of each tweet on this feed. As the user assessed more tweets on this feed, model Hidden evolved and so did its predictions. These predictions were all recorded in the backend but were not displayed to the user.

After completing condition Unassisted, the participant was redirected to the third step, condition Assisted. In this step, next to each tweet, the user saw their personalized AI's prediction of how they would assess that tweet. The user was asked to guide the AI to become better at learning their assessments by indicating whether they agree or disagree with the AI's predictions, which they could do by assessing the tweets as (in)accurate. These predictions in this step were in fact generated by the user's personalized model Visible, which was kept untainted by the user's assessments in the Unassisted condition. As the user assessed more tweets on the feed of this third step, model Visible evolved and so did its predictions. To give users a better understanding of how their personalized AI evolved, a list of updated predictions was displayed on the side pane. After each update of the model, the pane specified the number of assessment prediction changes broken down by whether they were changed from inaccurate to accurate or vice versa. The participants could further interact with the pane to bring the tweets with the updated predictions into view and decide whether their AI was improving.

Figure 1 shows a schematic of the flow of the study steps. Figures 2 and 3 show the views that participants saw when completing the study.

Models Hidden and Visible start as being identical, as both are initially trained on the exact same data from the Seeding step. Then they evolve independently from each other, with model Hidden updating based on the user's assessments on the 2nd feed, and model Visible based on the user's assessments on the 3rd feed. If we can ensure that the *expected* performances of the two models across the 2 feeds (the Unassisted and the Assisted conditions) are similar, then any statistically significant difference in how often users agree with the AI predictions (*observed* performance) can be attributed to whether or not the AI predictions were displayed to users.

A model's performance in this setting depends on two factors: (1) on what set of tweets it is trained and tested, and (2) what is the order according to which the tweets labels are fed to the model. The order plays a role because the model updates in iterations after receiving a small set of assessments. If the model is initially given

Figure 1: Flow of the steps of the user study.

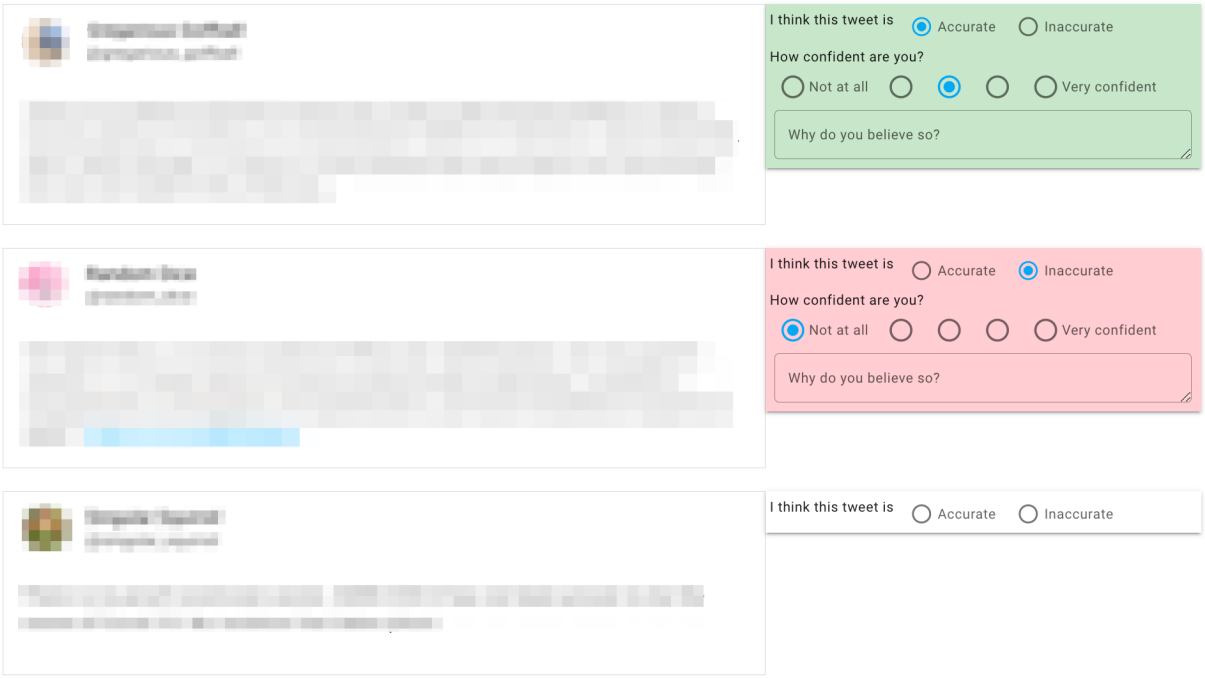


Figure 2: Assess tweets without AI assistance (UI used in Steps 1 and 2). Participants assess each tweet in their feed, rate their level of confidence in their assessments, and provide reasoning for at least 3 of the assessments without seeing any AI predictions. In Step 2, the order of assessing tweets was predetermined, with the user being able to only assess the next few tweets marked with a blue arrow (not shown in the Figure).

datapoints that result in a higher information gain for the model, the model improves faster and its performance in future iterations is higher.

To minimize the effect that the selection of tweets across steps 2 and 3 has on model performance, for each user, of the 52 tweets that did not belong to the Seeding step, we randomly sampled half to assign to step 2 and the other half, to step 3. This sampling led to each user having a different set of tweets in their condition Assisted compared to other users; and similarly, for condition Unassisted. To minimize the effect of order, we restrained the participants to only be able to assess 4 random tweets in their feed at a time. Although participants could see the rest of the tweets in their feed, those tweets were locked for assessment. The set of 4 tweets that a participant could assess would start from the top of the feed and would advance to the next four tweets on the feed sequentially once the participant assessed all the tweets in the previous set. Because the selection of tweets in each feed and the order of tweet placement in a feed was determined randomly, sequentially unlocking tweets for assessment ensures random ordering.

If we had not restrained the user in their choice of which tweets to assess first and which to defer to later, it was possible that in condition Assisted, where users could see the AI's predictions, they would attempt to first correct the cases where the AI had mispredicted their assessments. This could lead to (1) a higher disagreement in user and AI labels in the beginning compared to a random selection, and (2) a better, more accurate model in future iterations. This phenomenon would confound the results because while a better model would be desired if such a tool were deployed in the wild, if the performance of the model in condition Assisted is higher compared to the model used in condition Unassisted, a statistically significant higher user and model agreement in condition Assisted could not be attributed to the users seeing the model predictions in this step, because the model itself would also be more accurate compared to the other condition's model.

At the end of the task, participants were directed to a survey that asked about their perceptions of the AI, their views of different content moderation approaches, and their demographics. The full

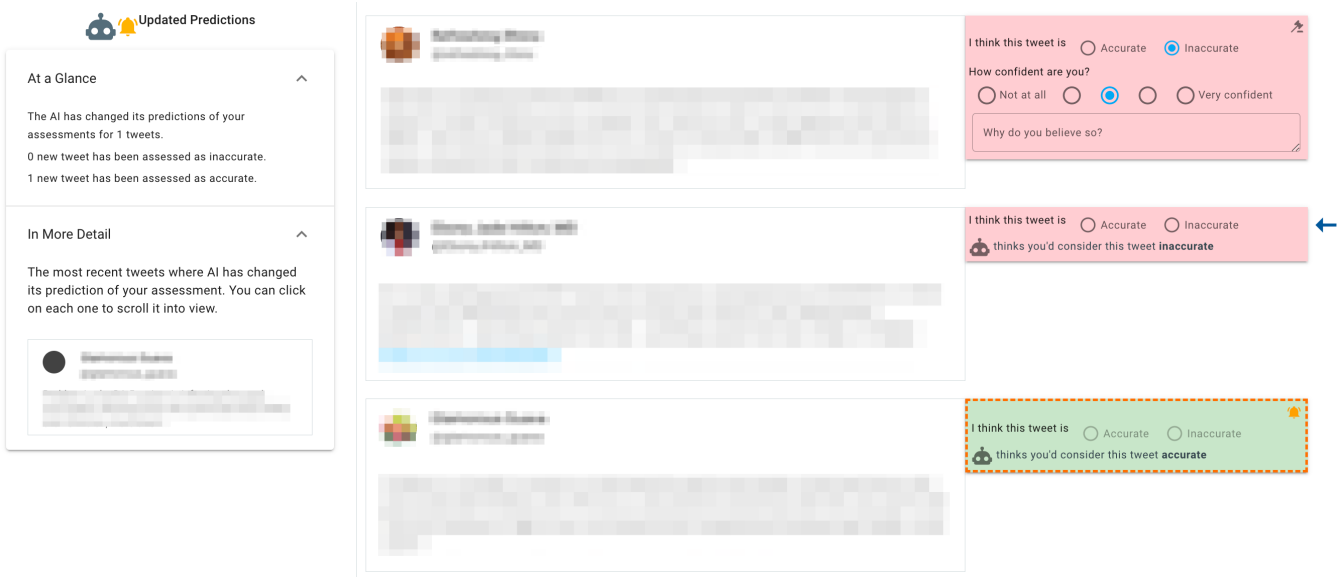


Figure 3: Assess tweets with AI assistance (UI used in Step 3). Participants assess each tweet in their feed while seeing AI model predictions. As shown in this screenshot, the user is in the middle of an experiment and has assessed the top tweet, while the rest of the tweets in this feed (for which the AI model predictions are shown) have not been assessed yet. The AI model predictions may change in response to user feedback. The tweets for which the AI predictions have recently changed are shown on the left pane. Newly changed predictions are differentiated visually with a border and a notification icon, similar to the bottom tweet in the image.

questionnaire from the post-study survey is included in the Supplementary Materials. At the end of the survey, the participants were notified that some of the tweets that they viewed during the study have been assessed as misinformation according to fact-checking initiatives and the scientific and medical communities.

3.2 Tweets

The vision for a personalized AI for determining content accuracy is a tool that is general purpose enough to work on any social media post, or perhaps even on any content on the web including news articles. However, such a versatile tool would require different components to handle different content types such as texts of various lengths, photos, and videos. As a first venture into this space, we decided to limit our scope to tweets, which are mostly text based and of a relatively short length. Nevertheless, tweets cover a wide range of domains including news or otherwise verifiable claims, opinions, life events, jokes or satire, advertisements, etc. A binary label of accuracy may not be appropriate for some of these types of tweets. For a binary accuracy classifier to work on an unrestrained set of tweets, first another classifier needs to determine which tweets are of the type "assessable with an (in)accurate label". Therefore, to make the problem tractable for the purpose of this work, we decided to manually curate a collection of tweets that we determined contained verifiable claims.

In curating the collection, we decided to limit the topic of the tweets to one, rather than including a range of topics e.g., related to healthcare, sports, or politics. This decision was because the text accuracy classifiers in the Machine Learning literature have mostly limited their scope to one topic, e.g., COVID-19 and therefore it was uncertain whether the knowledge that an accuracy classifier would learn about one topic can transfer to another. Especially, in our experimental task that contained only a rather small set of tweets, if the set consisted of multiple domains, there could be little chance for a model to learn any one domain.

We chose COVID-19 as the topic of our collection of tweets because it is an ongoing concern and topic of discussion on social media, various instances of misinformation have been propagated about this topic online [35], and it contains several subtopics (e.g., related to vaccines or masking) that are points of contention among social media users.

We needed a collection of tweets that contained both accurate and inaccurate tweets related to COVID-19. We examined several tweet datasets previously published by researchers to determine whether any would be appropriate for our task. We eventually decided that we needed to develop a new dataset that contained recent tweets. The reason was because per Twitter policy, any published dataset should only consist of Tweet IDs and any future user of such a dataset would need to "rehydrate" the dataset using the Twitter API and retrieving the tweets associated with the IDs. However, many of the misinforming tweets from previously published datasets are not retrievable anymore because either access to them has been restricted by Twitter or the poster accounts have been suspended. Another problem was that even the tweets labeled as accurate in such datasets were not relevant anymore and were difficult to assess, as they mostly concerned certain statistics such as the number of deaths in a certain week of a certain month of the

previous year or how many doses of vaccine had been administered by the time of the tweet.

3.2.1 Developing the Tweet Dataset. We used the Twitter API to fetch recent tweets related to COVID-19. The API allows for searching among tweets that have been posted within the previous week prior to making the query. To query for tweets related COVID-19 and its various subtopics, we used certain keywords including: long covid, covid vaccine, covid vaccine children, covid vaccine pregnancy, vaccine side effect, vaccine blood clots, vaccine sperm, vaccine infertility, vaccine fetus, vaccine fetal matter, covid wuhan, covid lab, covid origin, covid death, covid exaggerated death tolls, covid inflammation, paxlovid rebound, bill gates covid, pfizer raw information, ba4 ba5, booster, new variant, N95, masking, covid hoax, etc. Most of the tweets were posted around mid July 2022.

A member of the research team examined 2100 tweets and developed an initial set of criteria for including tweets in our experimental task. The research team held multiple meetings to discuss and iterate on the criteria for clarity. The final set of criteria for inclusion of tweets were:

(1) The tweets have to contain verifiable claims, rather than for instance arguments that cannot be verified, opinions, or life stories.
(2) The tweets have to be in the form of statements or evident rhetorical questions, rather than questions. This was because a binary accuracy label may not apply to questions.
(3) The topic of the tweets should not be news related to outside the United States. This was because we intended to recruit our users from within the US and we were concerned that they may not find such content relevant.
(4) The tweets have to be self-contained, i.e., for interpreting the tweet, other tweets (e.g., one that the tweet replies to, or other parts of a thread of tweets) should not be needed. Although many of the tweets that the research team assessed as misinformation were replies to other tweets, some were included because they were nonetheless self-contained.
(5) The main topic of the tweet should be about COVID-19, rather than have COVID-19 as a sideline, such as the impact of the pandemic on the economy. This was to limit the number of subtopics in our small experimental dataset.
(6) The topic of the tweet should not be about statistics related to COVID-19, such as death tolls, because we were concerned that these numbers may change before we could start the user study.

This set of criteria yielded 103 tweets out of our pool of 2100. We wanted the 76 tweets of our experimental task (3 feeds, each containing 26 tweets) to represent the different viewpoints related to each of the subtopics of our tweet set. A member of the research team inspected the 103 tweets, developed the set of subtopics that the tweets discussed, and categorized each tweet according to its stance on the two sides of the argument in the subtopic (e.g., believing in or doubting the efficacy of masks in the prevention of COVID-19). Table 1 lists the opposing claims of each subtopic. Many tweets contained multiple claims, each taking a stance on a different subtopic. For instance, it was possible for a tweet to assert that COVID exists and that death tolls are on the rise, but that masking is not effective against it.

To curate the 76 tweets of our experimental task, for each subtopic and among all the 103 tweets that discussed the subtopic, we randomly selected a sample such that it would be balanced with respect to how many of its tweets supported each side of the argument in the subtopic. We did this by iterating over the subtopics and for each one, drawing a sample from the pool of tweets that had not yet been sampled. The number of tweets that we sampled for each subtopic was the lesser of the following two numbers: the tweets that supported and those that were against the argument (in the pool of unselected tweets). Some subtopics had more tweets discussing one side of the argument either in the larger pool or in the pool of yet unselected tweets. Therefore, this method yielded a sample consisting of 50 tweets which was smaller than our desired sample size of 76. Then, we completed the sample by randomly drawing from the rest of the tweets that had not been selected.

3.2.2 Tweet Authors. Research has reported that content source is an important factor influencing a user's perception of content credibility [52, 68]. Therefore, for our experiment to be ecologically valid, we needed to show our participants the source of each tweet. Otherwise, by creating an artificial setting where the tweet sources were not shown, it was possible that users would scrutinize tweets from certain authors more than or less than they would if they knew who the authors were, impeding the generalizability of our findings. Nevertheless, we also wanted to be respectful of the privacy of tweet authors. Therefore, we decided to keep the author credentials for those tweet authors who were either institutions (such as "American Psychiatric Association") or individuals who had been given the status Verified by Twitter and replace the rest with fictitious credentials. A Verified badge by Twitter is given to Twitter users who have an account of public interest and are authentic, notable, and active [1]. The reason the unverified individual users were also given names, albeit fictitious, was to keep the appearance of all the tweets consistent in our experimental task. We decided that the absence of real names for the unverified regular users would be unlikely to affect the results substantially because the participants were unlikely to know those tweet authors even if we displayed their original names. Additionally, it was common in our original dataset for the accounts to have arbitrary names or pictures that did not represent the human behind the account. To create fictitious credentials, we developed a pool of photos that were owned by or licensed to our institution to serve as profile pictures. These were photos of animals, plants, and objects. Next we assigned names to these photos that were composed of either two part nouns or and adjective and a noun (such as Iconic Iguana or Gregarious Golfball). The nouns referred to the most salient entity in the photo. We avoided using negative adjectives or adjectives that expressed a state of credibility (such as "credible" or "reliable"). We created a mapping of original names of unverified non-institution tweet authors to fictitious names so that any tweet author from whom we had multiple tweets would always be replaced by the same fictitious name. We did not use human names and pictures to avoid the confounding effect that demographic factors such as gender, race, or age can have on perceptions of credibility [10, 47, 66, 85]. In the experimental task, each tweet was displayed with its text as well as author (real or fictitious) username, name, and profile picture. To

Table 1: The two sides of the arguments in COVID-19 related subtopics of our pool of tweets. Note that most of the claims on the right-hand side have been proven false according to the scientific community.

COVID-19 exists	COVID-19 does not exist (is completely made-up or is simply a strain of flu)
COVID-19 is dangerous	COVID-19 is not dangerous or is a mild disease
Masking is effective against COVID-19	Masking is not effective
Vaccines are effective against COVID-19	Vaccines are not effective
Vaccines are safe (e.g., safe during pregnancy, do not affect fertility)	Vaccines are not safe (e.g., unsafe during pregnancy, cause infertility, blood clots, death, or other side effects, pharmaceutical companies are not disclosing side effects)
-	Vaccines are bio-weapons made with malicious intent
COVID-19 is not man-made	COVID-19 is man-made
COVID-19 death tolls are high or on the rise	Death tolls are exaggerated
Information about new cures, vaccine discoveries, new variants, medicinal effects	New (false) cures

further protect the privacy of the tweet authors, we programmatically disabled copying text from the experimental platform, so that participants would not be able to copy and search a tweet's text. At the end of the study, we notified the participants that we had changed the usernames, names, and pictures of some tweet authors in order to protect their privacy.

3.3 Participants

We recruited participants from Amazon Mechanical Turk. The eligibility criteria for the Turkers were that they needed to have more than 500 HITs approved, an approval rate of higher than 98%, and US as their location. In our task, we further specified that they must be 18 years of age or older, at least occasionally read news online, be fluent in English, and be a US citizen or US permanent resident. From our pilot studies with our research group, we determined that the average time for completing the task was approximately an hour. Therefore, we set a compensation of $17 for the task.

A total of 65 (non spammer) workers participated in our study. We determined low quality response by investigating participants' free-text responses to the reasoning questions as well as the post-study survey. The responses from the spammers were unrelated to the question (e.g., responding "GOOD" to all the questions). There were 4 participants who we determined had attempted to perform the task in good faith, and therefore were paid, but whose submitted texts was unintelligible due to language errors. We removed these cases from our dataset as well because we determined that these participants had not received the treatment (e.g., did not fully understand the tweets or the questions).

Among the rest of the participants (N = 61), the median for age was 35-44 (ranging from 18-24 to 65-74). The median for income was $50,000 - $59,999 (ranging from Less than $10,000 to more than $150,000), and for highest education achieved Associate degree in college (ranging from High school graduate to Bachelor's degree). Our institution's data privacy policy prevented us from collecting other demographics information, such as those related to gender, race/ethnicity, or political leaning.

3.4 Analysis Procedure

When investigating the datapoints, we realized that for 11 participants, due to a bug in the open source AI system that we were using, the AI models Hidden or Visible or both had failed to retrain after a few initial iterations [1]. Because in these cases the model performances across conditions Assisted and Unassisted would be different, we removed the datapoints of these users from our analyses of comparing assessments across the two conditions. Therefore, the data that we include in these analyses is from 50 users. In our analysis of the answers to the post-study survey however, we retained the responses of those participants for whom model Visible had not experienced any problem in training, regardless of whether the training of model Hidden had failed. This was because the performances of the two models were isolated and these participants' experience with the AI which happened in condition Assisted was not affected by the failure in condition Unassisted. The data that we include in our analyses of the post-study survey is from 54 users.

Throughout the results, whenever we performed a statistical test to predict a binary dependent variable, 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" to accommodate the assumption of linear models that the residuals are normally distributed. For continuous outcomes, we fit a linear model to the the data using the function "lmer". In all our models, we included the tweet and participant identifiers as random effects to account for the variation in the outcome attributed to (unobserved) characteristics of a particular tweet or a participant, rather than the variation attributed to the independent variable of interest. In the tables where we present regression estimates, if the dependent variable is binomial and hence the fitted model is a generalized linear model (logistic regression), we present the exponentiated coefficient (i.e., the odds ratio) as a measure of effect size. We present partial η^2 as a measure of effect size for linear regressions to predict continuous outcomes. Additionally, we present the marginal R^2 (the variance in the outcome explained by the fixed effect), as well as the conditional R^2 (the variance explained by the entire model including both fixed and random effects) in all our analyses.

4 RESULTS

4.1 Users' Perception of a Personalized AI for Identifying Misinformation (RQ1)

4.1.1 Users' Perceived Accuracy of the AI and Why It Errs. To gauge participants' perception of a personalized AI in this domain, we first needed to understand whether they in fact found the AI capable of learning their assessments and responsive to their feedback. Figure 4 shows that many participants did find the AI good at predicting their assessments, with 36 out of 54 users (67%) in the post-study survey reporting that they found the AI "somewhat

[1] We reported the bug and it was immediately fixed.

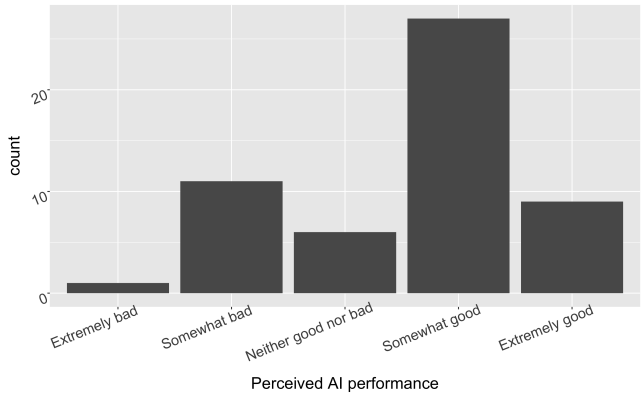


Figure 4: The distribution of participants' responses to the question "In general, how good was the AI at predicting your assessments?".

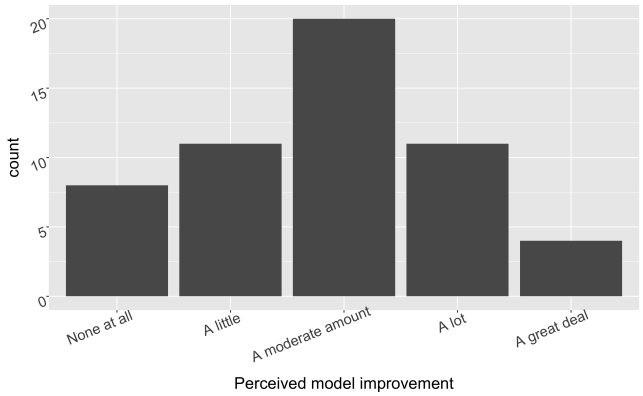


Figure 5: The distribution of participants' responses to the question "How better did the AI get at predicting your assessments over time as you provided more assessments?".

good" or better ($\bar{X} = 3.59$, $s = 1.06$). Figure 5 shows that many participants believed that the AI became better at predicting their assessments over time as they assessed more tweets, with 35 out of 54 users (65%) reporting that the AI improved at least a moderate amount ($\bar{X} = 2.85$, $s = 1.14$).

Participants understood that the AI was personalized. We explained in the task instructions that the AI was personalized and would learn and predict the user's assessments (see Figures 13, 14, & 15 in Appendix). Furthermore, to ascertain that participants understood the personalized aspect of this AI, we investigated their free-text responses and confirmed that this was indeed the case. For instance, in response to the questions of how good the AI was or how well it improved over time, many had explained that the AI had heeded their assessments. Those who were dissatisfied with the performance of the AI also indicated that they understood the purpose of the AI but that it had not delivered its promise well.

> *"Overall AI did a good job predicting whether I judged a Tweet as accurate or inaccurate. I was a little surprised that the AI thought I would assess a Tweet as inaccurate when I had previously labeled several similar Tweets as accurate."*

Factors influencing the AI's perceived performance. To understand what affected users' perception of the AI's performance, one member of the research team used open coding to assign labels to participant explanations related to how good they perceived the AI to be at predicting their assessments, to what extent it improved, and what cases they perceived as difficult for the AI. We have summarized the surfaced themes in Table 2.

We expected that participants' mental models of how the AI works would affect their perception of the AI [61]. Therefore, we investigated their responses to understand their mental models and report the results in the Appendix Section C.

The Effect of Users' Confidence in Their Assessment on Their Agreement with the AI. Some users justified the mistakes of their personalized AI by speculating that some of their disagreements with the AI were due to the user not being confident about their assessment of the tweets on which they disagreed or previous similar tweets that had been used for AI's training. We wanted to see if this interpretation is in fact valid or if users were inventing excuses for the AI's mistakes, especially since users' ratings of their confidence in their assessment were not taken into account when training the AI, contrary to what the users may have assumed. Therefore, we performed an exploratory analysis of the dataset containing the accuracy labels given by users and their personalized AI models in condition Unassisted. The reason we scoped our analysis to the data from condition Unassisted only was to avoid the confounding effect of users seeing the AI predictions. We performed the regression in the Table 3 on this data, which consisted of 1300 datapoints (50 users, each assessing 26 tweets).

The result of the regression demonstrates that the higher a user's confidence is about their assessment of a tweet, the higher their likelihood of agreeing with the AI is on whether the tweet is accurate. Figure 6 also depicts this relationship. Therefore, in line with what some participants hypothesized, it was indeed the case that some cases of "mistakes" by a user's personalized AI's were because the user was not confident about how to assess the tweets either.

4.1.2 Users' Perceived Usefulness of the AI. In the post-study survey, we had asked our participants to what extent they believed a tool like the one with which they interacted, which predicts their assessment of content accuracy based on the assessments that they provide, would help them if it were deployed on social media. See Figure 7 for participants' stance on this question ($\bar{X} = 2.69$, $s = 1.33$). The figure shows that different users have widely varied opinions about the usefulness of such an AI.

We examined participants' free-text responses through open coding to understand their views on the pros and cons of incorporating such an AI on social media platforms. Table 4 summarizes these views. Here, we elaborate more on some of the surfaced themes.

Specifics of the AI's implementation or deployment. Of these concerns, one was that users would not necessarily trust the AI's predictions. This concern was sometimes rooted in the users' observation that the performance of the experimental personalized AI did not

Table 2: Factors influencing AI's perceived performance

Factor	Example
AI mispredicted "obvious" tweets (N=11)	*"It felt like the AI got a couple of predictions wildly wrong, like on ones that felt very obviously correct or incorrect.".*
AI mispredicted tweets that were similar to those the user had assessed before (N=2)	*"...[The AI] did miss out on some I felt like it should have obviously knew [sic] since I'd rated similar tweets a certain way beforehand."*
User was conflicted about how they would assess the tweet (N=5)	*"...there were cases in the first two sets where I wasn't sure if some were accurate or not - they seemed reasonable, but them being false could also have been reasonable and I just hadn't encountered that claim/info before and couldn't know. SO I might have been "wishy-washy" in my choices on those, which lead to the AI sometimes picking the wrong way as well.".*
User believed AI's mistake was the user's fault—because the user did not understand the tweets they assessed previously or the ones on which they disagreed with the AI (N=3)	*"Most of the predictions were correct except for a few and that might have been my fault from not understanding a tweet or 2 and just plain confusing the AI.".*

Table 3: Are a user and their personalized AI more likely to agree on their assessment of a tweet if the user is more confident about their assessment? The regression outlined in the Table fit to the data from condition Unassisted confirms this hypothesis.

Independent Variable	$exp(\beta)$	CI	z test
user's confidence in their assessment*	1.25	[1.13-1.39]	$z = 4.14$, **p<0.001*****

Dependent variable: Whether a user's assessment of a tweet agrees with the AI's — Marginal R^2/Conditional R^2 = 0.019/0.176

Table 4: Participants' views in favor of and against adopting a personalized AI for identifying misinformation on social media.

	Theme	Argument	Example
In favor	AI could serve as a first-pass filter, filtering out blatantly false content so that finding accurate content among the rest could be easier (N=25)		*"I think it could sort out blatant untrue content and statements while leaving the rest for me to sort through. Most I could get rid of and this would save me time."* *"It could provide a good starting place or guide. Something marked as inaccurate would get more scrutiny."*
	AI could be used as a guideline for inspection of content (N=3)		
	AI could point out alternative opinions in a structured form (N=2)		
	AI is less biased than those employed by social media platforms (N=1)		
	AI is not fallible to human errors and misinformation (N=1)		
Against	Specifics of the AI's implementation or deployment (N=11)	AI's predictions may be wrong (N=8)	*"AI did improve over time, but the main problem is that AI doesn't know the reasons behind why I judged a tweet accurate or inaccurate. The AI is smart enough to recognize which Tweets generally get an accurate/inaccurate rating from me, but can't pick up on the nuances that make me decide how I judge a Tweet."*
		AI does not provide reasoning for its predictions (N=2)	
		AI, if operated by the platforms, will be manipulated by them (N=1)	
	AI would not help user's current practices with content reading on social media (N=12)	User would like to think for themselves—unassisted (N=4)	*"... I don't like being told what to think, even if it's correct. I find in-your-face fact-checking on social media to be intrusive and annoying."*
		User does not visit social media to find trustworthy content (N=3)	
		User fact-checks social media posts themselves (N=3)	
		It would take user longer to process information with such a tool (N=1)	
		User does not encounter misinformation (N=1)	
	Broader implications of democratized assessments (N=9)	There is one objective truth, or assessments should come from experts, not regular users (N=6)	*"It's catered to what I feel might be right or wrong. I'm going off of personal experience, but I can't always decipher whether it is right or wrong on my own as I am not fully aware of everything. If it is predicated to what I feel is right, then the AI will cater to that, and I don't necessarily want that. I would prefer the truth, just finding it is hard when there's so many opinions."*
		Democratized assessments would result in echo chambers or the AI would filter opposing viewpoints that are important to consume (N=3)	

meet their expectations and that they assumed that the envisioned AI would have a similar performance. However, the performance of an AI model in practice may vary based on the exact implementation details, including how many tweets are used to train the model and the type of model used. Other times, users were concerned that the envisioned AI may run into difficult cases if deployed in the wild, such as small nuances in wording that determine whether a piece of content is accurate or misleading, resulting in mispredictions. Prior work has reported that users have a similar concern about centralized AIs for content moderation used by social media

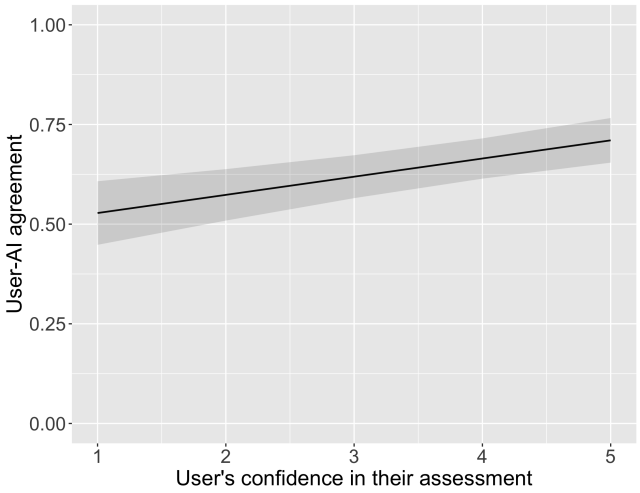


Figure 6: The predicted values of user-AI agreement by the user's confidence in their assessment. The higher the confidence of a user in their assessment, the more likely their AI is to correctly predict their accuracy rating.

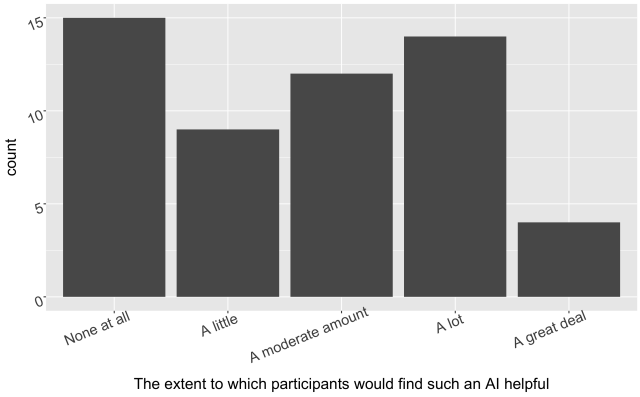


Figure 7: The distribution of participants' responses to the question "To what extent do you believe a tool like this would help you if it were deployed on social media?".

platforms [40]. Another concern in this class was that the AI that we offered did not provide reasoning for its predictions, and that participants believed that the AI that would ultimately be deployed on social media would also be lacking in this aspect. This issue can be addressed by drawing from research in text generation and using a user's set of provided reasons for their assessments as training data for generating rationales for such an AI's predictions [51]. In fact, explanations for how such an AI works can potentially increase the effectiveness of its labels as reported by Epstein et al. for the case of a hybrid crowd-AI misinformation detector [30]. Yet another concern was rooted in the deep distrust of social media platforms that some users harbored, believing that if such a tool were operated by the social media platforms, it would eventually be manipulated by them. Such a tool however, does not need to be operated by the platforms, and can be offered for instance, as a browser extension directly managed by the user. This would be similar to the Reheadline browser extension that users can install to edit misinforming headlines across all platforms without needing support or compliance from them [53].

The AI would not help the user's current practices with content reading on social media. One of the reasons for the AI's perceived lack of usefulness was that the user configured their social media feeds to be composed entirely of posts by those they trust and therefore do not encounter misinformation. Prior work however, has shown that many users have difficulty taking control over their news feed [32] or that they follow sources they consider untrustworthy for a variety of reasons, for instance, to be aware of to what content their friends and family are exposed [54].

We additionally analyzed how users perceive the usefulness of a personalized AI in comparison with other content moderation strategies. We present those results in the Appendix Section D.

4.2 The Impact of Personalized AI on User Assessments (RQ2)

In this section we report on the results of our analyses to understand the impact of showing AI predictions on users' assessment decisions, as well as how this effect can be mitigated.

4.2.1 Showing AI Predictions Sways User's Accuracy Ratings. To understand whether displaying the predictions of users' personalized AI models affects how users would assess content, we created a dataset of accuracy assessments generated by each participant and their two AI models, with the assessments belonging to the tweets in the participant's Unassisted and Assisted feeds. Each datapoint in this dataset was a pair of accuracy assessments given to a tweet t, one by a participant p, and the other, p's personalized AI— model Hidden if the tweet t for the participant p belonged to condition Unassisted, and model Visible if the tweet t for the participant p belonged to condition Assisted. Because each model and its predictions evolved, there could be multiple accuracy predictions from a participant's AI for a single tweet at various points in time. In constructing the pairing of user-AI assessments, we considered the last (i.e., the most recent) assessment that the AI had generated before the participant had assessed the tweet. For condition Assisted, this AI assessment would be the one that was displayed to the participant when the participant assessed the tweet. This dataset consisted of 2600 datapoints (50 participants, 2 conditions, each condition having 26 tweets).

Across all users and all tweets, the average user-AI agreement on the accuracy of tweets in condition Unassisted was 65.6% and in condition Assisted was 74.7%.

The analysis from Table 5 performed on this dataset shows that users have a higher agreement with the AI when they see the AI's predictions compared to when the predictions are withheld from them and that this difference is statistically significant. This result provides a partial answer to RQ2, indicating that users do in fact shift their accuracy rating of content to match that of their personalized AI, when exposed to the AI's predictions.

The Impact of Seeing the Model Prediction on Users' Accuracy Ratings Increases Over Time. A particular characteristic of our study

Table 5: Are users' assessments swayed by seeing AI's predictions? We fit the regression described in the Table to pairs of AI and user assessments for each tweet across the Assisted and Unassisted conditions. Condition (i.e., whether AI predictions were shown to users) did have a statistically significant effect on whether users agreed with the AI.

Independent Variable	$exp(\beta)$	CI	z test
condition (whether the AI's prediction for a tweet was shown to the user)*	1.60	[1.33,1.92]	$z = 5.03$, **p<0.001*****

*Dependent variable: Whether the user's assessment of a tweet agrees with the AI's Marginal R^2/Conditional R^2 = 0.014/0.182

that made it resemble real world scenarios was that a user's personalized models would evolve in response to the feedback that they received from the user. We wanted to examine whether there was a difference in the user-model agreement across the 2 models for each user over time. Each model updated in iterations—a model was in iteration i until it received 4 more assessments from the user and then it would start retraining. While the model was being retrained, a user could still continue assessing. Note that for each user, the model Hidden was used to train on and predict the accuracy of the tweets in the Seeding step as well as condition Unassisted, and model Visible was used to train on and predict the accuracy of the tweets in the Seeding step as well as condition Assisted. Therefore, while a user was in step 2, only the iteration of model Hidden would advance; and similarly, while a user was in step 3, only the iteration of model Visible would advance.

To compute the user-model agreement for a model at a particular iteration, we first examined what the model predictions were at that iteration. We excluded the tweets belonging to the Seeding step because the two models for a user would perform identically on those tweets. In other words, the *test set* for calculating the agreement between user p and the model Hidden for user p, is the tweets that the user p saw when they were in condition Unassisted. And similarly, the *test set* for calculating the agreement between user p and the model Visible for user p, is the tweets that the user p saw when they were in condition Assisted.

We then developed a dataset with each datapoint being a model's prediction of a tweet at a particular iteration paired with the assessment that the user would eventually submit for the tweet. This could be thought of as a post-mortem analysis—that although by the time the user assessed a certain tweet the model may have advanced to another iteration and so its prediction may have changed, we want to understand what the user's agreement would have been with a prior version of the model. For model Hidden, where the confounding effect of users' seeing model predictions does not exist, this agreement ratio would indicate how accurate the models from the prior iterations had been. We performed the regression described in the Table 6 on this dataset.

Note that for calculating the user-model agreement for the iterations of the model when the user was still in the Seeding step, the test set is constant. However, when the user exists the Seeding step and advances to the next steps, the user starts submitting accuracy ratings for the tweets in the test set. Therefore, for calculating the user-model agreement at a particular iteration i, the test set would be the tweets that the user had not yet assessed before iteration i. As the iteration for a model advances after the user has exited the Seeding step, the test set keeps shrinking since the user is providing "the ground truth" for more and more of the tweets. The dataset consists of 22263 datapoints.

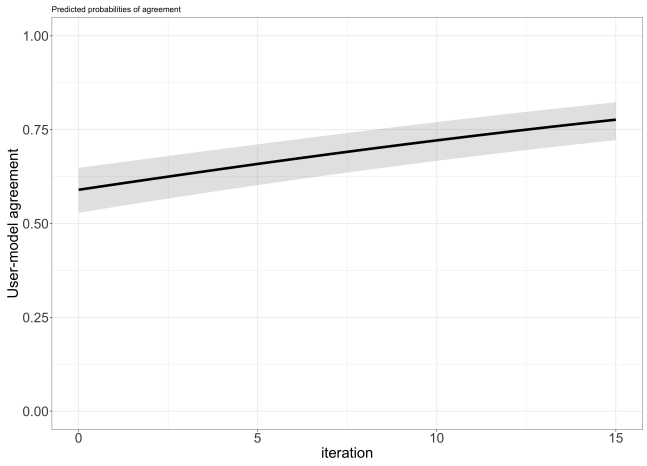


Figure 8: The predicted values of user-model agreement by iteration. Model iteration is positively correlated with user-model agreement, independent of which model. This shows that models become better at predicting users' assessments over time.

The regression in Table 6 on this dataset revealed that iteration had a positive and statistically significant correlation with user-AI agreement. Figure 8 displays this correlation. This result suggests that regardless of model, user-AI agreement increased over time, suggesting that the models improved and became better at predicting users' assessments.

Interestingly, we also observed that the interaction effect between model and iteration was significant. Figure 9 shows the effect of the interaction between model and iteration on the predicted user-model agreement. The figure demonstrates that over time (i.e., as the iteration increases), the difference in user-model agreement between models Hidden and Visible grows larger. This observations suggests that when users see AI predictions, they become more reliant on the AI over time.

4.2.2 Agreement with AI's Accuracy Rating Does Not Impact User's Confidence in Their Assessment. We wanted to understand whether by seeing that their assessment agrees with the AI's, users gain more confidence in their assessment; and conversely, whether they lose confidence in their assessment if the AI disagrees with them. Therefore, we split the dataset from Section 4.2.1 into 2 partitions: the partition where the users had agreed with their AI's predictions (N=1824), and the partition where the users had disagreed with their AI's predictions (N=776). This partitioning allowed us to test whether conditioned on the fact that users agreed (or disagreed)

Table 6: Does users' agreement with their personalized AI change over time? We used the regression described in the Table to compare AI's predictions of the accuracy of a particular tweet at the different iterations of the AI's evolution with the assessment that the user eventually gave to the tweet. The analyses revealed that over time, the user-AI agreement increases. However, this increase is higher in condition Assisted where users could see AI's predictions, compared to condition Unassisted.

Independent Variable*	$exp(\beta)$	CI	z test
model (whether the model that assessed a tweet for a user was Hidden or Visible)	0.98	[0.89, 1.09]	$z = -0.30$, **p=0.76**
iteration of the model	1.06	[1.05, 1.08]	$z = 7.98$, **p<0.001*****
model × iteration	1.05	[1.03, 1.07]	$z = 4.40$, **p<0.001*****

*Dependent variable: Whether a user's assessment of a tweet agrees with the AI's Marginal R^2/Conditional R^2 = 0.014/0.182

Table 7: Do users gain more confidence in their assessment if they see that the AI agrees with them? We fit the regression described in the Table to 2 data partitions: where users had agreed with the AI's assessments, and where they had disagreed. Seeing their (dis)agreement with the AI did not have a statistically significant effect on their confidence in their assessment.

Independent Variable	Partition	β	CI	t test	partial η^2
condition (whether the AI's prediction for a tweet was shown to the user)*	User and AI **agree** in assessment	0.08	[0.00, 0.17]	$t(1745) = 1.87$, $p = 0.062$	0.002
		Marginal R^2/Conditional R^2 = 0.001/0.342			
	User and AI **disagree** in assessment	0.07	[-0.09,0.23]	$t(725.64) = 0.83$, $p = 0.41$	0.001
		Marginal R^2/Conditional R^2 = 0.001/0.333			

*Dependent variable: confidence of the user in their assessment of a tweet

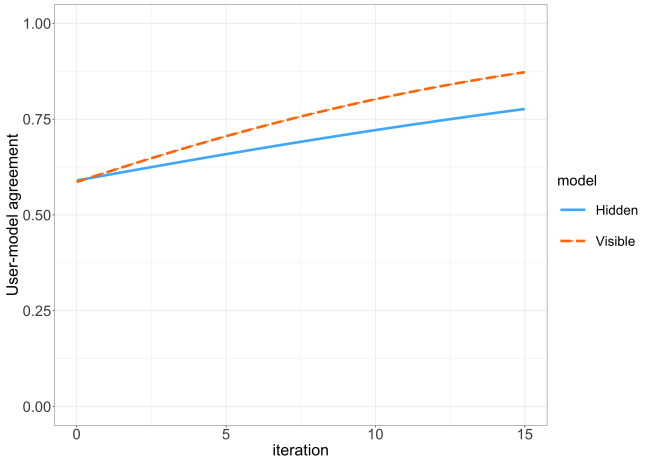


Figure 9: The effect of the interaction between model and iteration on the predicted user-model agreement. Over time, the difference is user-model agreement across the two models increases.

with their AI, seeing vs not seeing AI's predictions affected their confidence in their assessments. The regression model we fit to these partitions and the statistical results are displayed in Table 7.

The results show that seeing or not seeing AI's predictions did not have a statistically significant effect at $\alpha = 0.05$ regardless of whether the participants agreed or disagreed with the AI in their accuracy ratings. This result paints a more complete picture of the answer to RQ2, indicating that when users are exposed to AI predictions, their agreement with the AI on the accuracy of content does not increase their confidence in their assessment.

4.2.3 Providing Reasoning Mitigates the Influence of Seeing AI's Predictions. Since we observed that showing AI's predictions to users biases users' judgment in how they assess content, we looked into ways this bias could be mitigated. Previous work has reported that many cognitive biases can be mitigated by asking people to provide justifications for their choices [52, 70]. We performed exploratory analyses to understand if users' providing reasoning for their assessment of content could act as an intervention against their blindly accepting the displayed prediction of the AI. Although the requirements for the task was that participants provide reasoning for at least 3 of their assessments in each feed, many had provided more assessments than what was required. In total, participants had provided 608 reasons across the feeds of the 2 experimental conditions (average of 12.16 per user), with 310 belonging to condition Unassisted, and 298 belonging to condition Assisted.

To determine whether providing reasoning can mitigate the over-reliance on AI when AI predictions are shown, we first examined whether the users' likelihood of agreeing with their AI's predictions in condition Unassisted (which we would consider as the baseline agreement—untainted by the effect of seeing AI's predictions) was (statistically) different from their likelihood of agreeing on the items for which they provided reasoning in condition Assisted. Therefore, we created a partition from the dataset in Section 4.2.1 consisting of assessments from condition Unassisted, as well as those from condition Assisted on which the users had provided their reasoning (in total 1598 datapoints).

The regression in Table 8 revealed that when AI predictions were shown, if users followed up on their assessments with reasoning, they were as likely to agree with the AI as when the AI predictions were not shown to them. This result suggests that providing reasoning may mitigate the over-reliance of users on their AI.

If providing reasoning mitigates the influence of seeing AI's predictions, then in condition Assisted, there should be a (statistically significant) difference in user-AI agreement across the tweets for

Table 8: Is there a difference in user-AI agreement across tweets for which AI predictions were shown but for which users also provided reasoning in support of their assessments and those tweets for which AI predictions were not shown? We fit the regression described in the Table to this data, and found that there is no statistically significant difference.

Independent Variable	$exp(\beta)$	CI	z test
condition (whether the AI's prediction for a tweet was shown to the user)*	1.06	[0.79, 1.43]	$z = 0.39$, $p = 0.69$

*Dependent variable: Whether the user's assessment of a tweet agrees with the AI's Marginal R^2/Conditional R^2 = 0.000/0.174

which the user did and did not provide reasoning. Such a difference should not exist in condition Unassisted where users could not see the AI's predictions. The analyses in Table 9 show that this is indeed the case.

The negative correlation between providing reasoning and the user-AI agreement does not necessarily suggest a causation relationship, as this correlation can have 2 possible interpretations, one or both of which may be true:

(1) Users' providing reasoning reduces their agreement with the AI because the act of following up on their assessment with reasoning imposes pre-decisional accountability on them [70], resulting in users not necessarily accepting the AI's prediction.
(2) In condition Assisted where AI predictions were displayed, users explained their reasons on the tweets they were more confident about and less likely to be influenced by the predictions of the AI in the first place, resulting in the user-AI *disagreement* rate to be higher in these assessments compared to those without reasoning in the same condition.

Examining Interpretation 1. To determine whether interpretation 1 is true, we should look for evidences of users changing their mind after their initial assessment and examine whether these changes of mind are more frequent on tweets with reasons than those without. Such an effect would be seen in both conditions, although there could be an interaction effect involving the conditions as well.

Therefore, we used the dataset from Section 4.2.1 to see whether users' providing reasoning for their assessments correlates with a change in the value of their assessments (e.g., a user changing their assessment from accurate to inaccurate, or initially assessing a tweet as accurate, then changing their accuracy rating to inaccurate, until finally reverting to accurate again). We did not take into account when the reasoning had been submitted (before or after the change in accuracy rating), because although a reasoning may have been written after the change in accuracy rating, the user may have thought of it before, leading to the change in their rating.

The analysis outlined in Table 10 revealed that providing reasoning does in fact correlate with users changing their initial assessment rating and that the effect is statistically significant. This observation lends credence to interpretation 1, suggesting that asking users to provide reasoning does mitigate the negative effect of seeing their AI's predictions.

While it would be interesting to see how this effect varies by condition, in the entire dataset consisting of assessments from conditions Assisted and Unassisted, there were only 100 instances of users having changed the value of their accuracy ratings. Therefore, the marginal counts across segments of conditions × whether reasoning was provided would be too small to partition the data on condition (or include it as a factor).

Examining Interpretation 2. If interpretation 2 is true, then there should be evidence that users are more likely to submit reasons for assessments about which they are more confident. To determine whether this is the case, we performed the regression in Table 11 on the dataset from 4.2.1.

We found that confidence in an assessment does indeed have an effect on providing reasoning for the assessment and that effect is statistically significant. This observation suggests that in addition to interpretation 1, interpretation 2 may also be correct. Another explanation for this correlation is that providing reason about an assessment is likely to make the user more confident about their assessment. However, because of our data collection method, we do not have the means to test this hypothesis.

5 DISCUSSION AND FUTURE WORK

In this work, we begin to explore the potentials and the challenges of incorporating a personalized AI for determining content accuracy on social media. We evaluated how participants perceived such an approach and found that while many participants saw potential in adopting a personalized AI as a first-pass inspection that would reduce the volume of content in their feed down to items that are more likely to be trustworthy, others had reservations about it. Some of these reservations were related to a specific implementation or deployment of the tool that they assumed would be the one deployed on social media. For instance, one such concern was the lack of transparency about the reasoning behind the AI's decisions, which can be addressed by drawing from research in the field of Explainable AI [97] or by leveraging research in the field of text generation to generate reasons by using, as training data, the reasons that a user provides for their assessments [51]. Other users were concerned with the broader implications of enabling personalized content curation. For instance, some users were worried that they may not have the necessary knowledge to discern credibility of content and that they would rather receive assessments from experts. Other users went further, asserting that everyone should receive the one objective truth—a view that is at odds with the stance of some other participants who were antagonistic toward centralized moderation, and distrustful of platforms. Prior work has also reported on the two sides of this debate [54, 82].

Research has reported how users ask for assessments from certain sources they deem trustworthy and provide assessments to their social circle on social media and has advocated for streamlining this process by capturing accuracy assessments in structured form [54]. While the personalized AI in our study only took a user's assessments as training data, it could be configured to also use the

Table 9: Are users less likely to agree with the AI when they provide reasoning for their assessments? The model in the Table fit to the data from condition Assisted revealed that this is indeed the case. However, as expected, providing reasoning in condition Unassisted, where users do not see the AI's predictions, does not affect the user-AI agreement.

Independent Variable	Condition	$exp(\beta)$	CI	t test
whether the user provided reasoning for their assessment of a tweet*	Assisted	0.54	[0.38,0.75]	$z = -3.65$, **p<0.001*****
		Marginal R^2/Conditional R^2 = 0.016/0.221		
	Unassisted	1.16	[0.84, 1.59]	$z = 0.91$, $p = 0.36$
		MarginalR^2/ConditionalR^2 = 0.001/0.166		

*Dependent variable: whether the user's assessment of a tweet agrees with the AI's

Table 10: Is providing reasoning for one's assessment correlated with a change in the value (i.e., accurate or inaccurate) of the assessment? We fit the regression in the Table to our dataset and found that this is indeed the case.

Independent Variable	$exp(\beta)$	CI	z test
whether the user provided reasoning for their assessment of a tweet*	1.95	[1.20, 3.15]	$z = 2.71$, **p=0.007****

*Dependent variable: Whether the user's assessment of a tweet changed at any point Marginal R^2/Conditional R^2 = 0.016/0.348

Table 11: Are users more likely to submit reasons for assessments about which they are more confident? The regression in the Table fit to our dataset revealed that this is indeed the case.

Independent Variable	$exp(\beta)$	CI	z test
confidence of the user in their assessment of a tweet*	1.31	[1.17, 1.45]	$z = 4.91$, **p<0.001*****

*Dependent variable: whether the user provided reasoning for their assessment of a tweet Marginal R^2/Conditional R^2 = 0.021/0.325

assessments from the user's trusted sources in addition to, or even instead of, the user's own assessments. The configuration of whose assessments to take as training data can address the concerns of those users who want to rely on experts or their trusted sources for fact-checking information, and it has the added benefit that it can widen the reach of the assessments by those trusted sources to content that they have not explicitly assessed. Future work is needed to understand how participants would perceive such an approach.

Our work also sheds light on the issues that need to be resolved before a personalized AI can be safely deployed in this domain. In the sections that follow, we discuss these issues, offer directions for ways they could be dealt with, and call on future work to investigate these directions.

5.1 The Influence of AI's Predictions on Users' Judgments

In our user study, participants knew that AI was trying to learn from their assessments and that it could make mistakes. In fact, the participants were sent on a mission to teach the AI how to think like them and were on the lookout for instances where they disagreed with the AI. Nevertheless, they ended up being swayed by the predictions of the AI. Moreover, the influence of seeing AI's prediction on their assessments grew larger over time. This finding has important implications for content moderation through the use of not only a personalized AI, but also a centralized AI, e.g., one run by the platforms. In the case of a centralized AI that is introduced as a definitive oracle of the truth, the AI's influence can potentially be exacerbated, which can be consequential if the AI is ever wrong. Furthermore, our results offer insights to the body of work on adaptive recommender systems and content curation algorithms that, similar to the context of our study, are personalized AI models attempting to predict users' preferences [18, 19, 89]. In such scenarios, the set of recommended results could serve as a self-fulfilling prophecy, causing users to shift the ground truth (i.e., their actual preferences) to match what is offered to them. Prior work has also reported concerns and evidence about the potential for this effect [27].

5.1.1 On the Role of Providing Reasoning. Prior work reports that many cognitive biases and errors resulting from such biases can be mitigated by imposing pre-decisional accountability where decision makers can expect to be called upon to justify their choices [52, 70]. This approach has also been found to mitigate automation bias—over-reliance on a decision support system that has a mind of its own. In the context of our study, we found that users' providing reasoning for their assessment choices did in fact prevent their inflated agreement with the AI (compared to a baseline where they did not see the AI's predictions), and was correlated with a change in their accuracy rating.

If such a tool were to be deployed in the wild, users' interaction with the tool would look different from the context of our study. For the purpose of our study, we asked that users assess every tweet rather than leaving it to their discretion to decide which ones they wanted to assess. This constraint was because we wanted to control for model Hidden and Visible's performances across conditions Unassisted and Assisted and also because we needed enough data to get the AI to a point where its performance would be acceptable. In the real world, users cannot be expected to give feedback to the AI on every piece of content that they encounter. Conceivably, users may only intervene when they see mispredictions that they perceive

as egregiously incorrect. In such cases, users already disagree with the AI and therefore asking that they provide justifications for their choice would not help mitigate their possible over-reliance on the AI.

Therefore, an issue that needs to be addressed is how to reconcile the desire for efficiency, which is one of the motivations for exploring the use of AI in this domain in the first place, and encouraging users to treat all the AI's predictions with some amount of scrutiny by priming them to think critically and justify their agreement with the AI. For instance, future work can explore a scenario where users of such an AI are prompted every once in a while to indicate whether they agree or disagree with their AI's predictions and provide reasoning for their (dis)agreement, and whether the expectation of being called on for justification causes them to be more alert overall. The success of such an intervention, if it proves to be promising, may also depend on the frequency of the prompt and on what content the prompt is asked—for instance, those that the AI is more or less confident about, or those that the AI has labeled as accurate or conversely, as inaccurate.

5.2 Ethical Considerations

A question that may arise is whether enabling users to use a personalized AI such as ours on social media would lead to stronger filter bubbles, causing users to only receive content that supports their views and be shielded from divergent views [59, 91, 93]. The AI that we propose does not *filter* content out of the user's view and in fact, leaves users' feeds intact. It simply adds structured assessments on top of their feeds. A concern about capturing and displaying structured assessments could be that filtering posts based on structured metadata can be easier than sorting through unlabeled posts. For instance, by having accuracy predictions on every post, users may find it easier to discount items predicted as inaccurate. As the training data for these assessments is the user's assessments (or the assessments of other users that the user explicitly chooses, as discussed earlier in the Discussion), the predictions could be biased toward the user's views, thus making it easier for users to disregard opposing views.

This issue connects to an ongoing debate on whether users should only receive the content that is decided for them or whether they should be given the power to filter through content. As explained, amplifying structured assessments through the use of AI can result in selective exposure to attitude-reinforcing content [83] by putting more power in the hands of the users. However, we also recognize that centralized moderation has problems as well [41, 54, 60, 82] and the purpose of our work is to investigate alternative approaches and their potentials and challenges.

In fact, the choice between centralized vs democratized moderation is a choice between how much autonomy we want to give to individuals to make their own decisions as well as mistakes vs giving that power to the platforms or certain institutions. In the current information ecosystem, platforms hold the power of content moderation, feeding users content that they curate, and down-ranking, flagging, or removing other content that they have decided users should not see. Upholding this status-quo is a stance in this debate by itself, indicating that we have decided against giving individuals autonomy and that we believe platforms are appropriate arbiters of the truth. Our work, as an approach that empowers users to assess content, calls this status-quo into question. We call on researchers to reflect on the following questions:

(1) Do individuals have the right to see content that they want to see as well as not see content that they do not want to see? [26]
(2) Should we cede the power of content moderation to the platforms knowing that they are for-profit entities running on user engagement [43], they may not be politically neutral [38], and that centralized moderation does not address the needs of everyone [54, 64, 80]?

Meanwhile, in today's online social spaces, there are many opportunities for adding structured metadata contributed by users to content. A few examples are tags, ratings, and upvotes or likes. These spaces allow users to filter content based on such metadata. A question that needs serious thought is why and where we draw the line of allowing for filtering based on a certain set of metadata, and disallowing filtering based on a different set of metadata.

The fight against misinformation should be a collaborative effort between users and the platforms. In such a collaborative ecosystem, platforms could still provide fact-checking labels and articles but leave it to the users to decide whether and how they want to use the platform recommendations. The platforms can aspire to be a trusted source, without forcing their decisions on users.

A related concern in the context of a personalized AI is that users may become more confident or extreme in their views if they see that their AI agrees with their assessments. We tested this hypothesis in our study. We found that in the context of our study, users' confidence in their assessments in the cases of agreement with the AI was not different in a statistically significant way when they viewed the AI's predictions compared to when they did not. However, future work is needed to understand whether these findings generalize to a long-term adoption of the AI.

5.3 Considerations of a Real-World Deployment

To make this study tractable, we made several reductions to the problem that need to be addressed in a real-world deployment. One was that we curated our set of tweets manually according to a set of criteria. One such criterion was that the tweets contain verifiable claims. A question that future work should investigate is how to detect such content. This can be a nontrivial problem because the notion of what is verifiable can vary by individual. For instance, although we treated tweets that talked about life events especially with references to unknown individuals or events such as *"A friend of mine got COVID."* as unverifiable statements, these in fact, could be verifiable in the eyes of those users who are close to the tweet author. Another related problem is that although labels of accuracy may not be appropriate for opinions (rather than factual pieces) [54], opinions still have the potential to misinform. Another complication is how to assess those tweets that contain a number of claims, some of which may or may not be true. Future work should study how to capture and signal the credibility of posts in these cases.

Another criterion we used in filtering tweets was the presence of statistics and numbers as they are prone to change in a rather short time. For instance, a tweet claiming "COVID death toll is

nearing X" may be true today, but not the a month from now. In fact, across a longer period of time, the veracity of other claims can also be subject to change. For instance, the claim "Vaccines have minimized the risk of COVID down to the level of a seasonal flu" is not true today, but hopefully, may be at some point in the future. Therefore, an issue that needs to be studied is how to deal with the temporal aspect of content veracity, for AI training and testing as well as for demanding assessments from individuals. For instance, if we equip users with an AI (whether personalized or centralized) that signals the accuracy of content to them, should the AI's assessment consider the time the content refers to (topic time [58], in linguistic terms) relative to the time that it was posted? Two downsides of such an approach are that first, the content's accuracy status in the past may not be as consequential to the user as its status in the present, and second, the topic time may not always be clearly specified. Therefore, whether or not the AI signals the content's accuracy status as it would have been in the past, to prevent misinforming the user, it may be important to also signal whether the content is accurate in the present. Future work can study how this can be achieved, for instance, by having the AI gradually "forget" the ground truth that was captured in the past.

Yet as another simplification of the problem, we narrowed our scope to tweets whose main topic was COVID-19. A personalized AI for content moderation would be more helpful if its use were not restricted to a particular topic. A direction for future work would be to study the feasibility of such an AI, for instance, whether a user could realistically provide the number of datapoints needed from them for the AI to achieve a certain performance.

A two-item assessment of accuracy may not be sufficient for capturing all the complications that can arise in a context similar to our study. However, we decided to use a binary categorization in our task as an attempt to make the problem tractable for the purpose of the study, knowing that such categorization is common in misinformation studies [52, 54, 76]. We call on future work to investigate how best to capture assessments on social media posts.

6 LIMITATIONS

Due to the special considerations of the experiment design which we described before, in our study, we asked participants to assess every tweet. Having to provide assessments on each item may have primed our users to think more critically about the accuracy of content compared to if they had not have to provide feedback on every item. Therefore, it is possible that the effect of the influence of the AI on user judgment may be different (and potentially larger in magnitude) than what we observed if such a tool were to be deployed in the wild and that users did not have to pause and think whether they agree with the AI on every decision.

In our study, we curated a feed of tweets for our participants that had a rather balanced representation of the two sides of the debates in each subtopic related to COVID-19 (outlined in Table 1). In the real world however, the frequency of counter-attitudinal content that users encounter in their feed is more occasional. Future work should study whether our findings, e.g., those related to the influence of the AI on users' judgments, generalize to a scenario where users use such a tool in the feeds curated for them by the platforms.

For the purpose of the user study, we made certain decisions on the type of AI model and the data used to train the model. For example, we selected a SVM model to allow for quick retraining and updating of model predictions, and we limited the overall number of tweets to ensure that the user study would not be tedious. Choosing a different NLP model and/or asking users to annotate a larger set of tweets may affect the performance of the AI, in turn affecting users' perceived usefulness of a prospective AI. Therefore, future work is needed to understand how our results about users' perceptions of such an AI generalize to other settings with different NLP models and tweet datasets.

It is possible that a user's stance on the argument of a subtopic may change by assessing more tweets on the subtopic. This potential complication may cause the user's assessments on the subtopic to be inconsistent over time. If this happens across conditions Unassisted and Assisted, then the AI in condition Assisted may receive more (internally) consistent training data since the user has seen more tweets with which to curate their stance. Therefore, the AI in condition Assisted might perform better than the one in condition Unassisted, which could be confused about the potentially inconsistent datapoints on a certain subtopic provided by the user. A way to counter this potential effect would have been to randomly select the order the two conditions for participants. However, we determined that requiring more assessments without showing the AI's predictions (condition Unassisted) after the participants were already exposed to the AI's predictions (condition Assisted) would be odd and decided against such a setup. Nevertheless, we hypothesize that if users do become more consistent with themselves as they reflect on a certain topic, the effect is likely minimized by the end of the Seeding Step, when the user has already assessed multiple examples of each subtopic. We leave it to future work to determine the existence of such an effect.

There is a possibility of heterogeneity among our participants with respect to their attitudes toward AI or AI-assisted identification of misinformation. We did not ask about participants' attitudes. Similarly, we did not ask about whether participants had a Computer Science background nor did we restrict recruitment based on familiarity with AI. Future work should investigate how participants' perceptions of the AI or our other findings differ across various segments of the user population, for instance those who are antagonistic toward AI-assisted identification of misinformation.

Although we restricted our task on Mechanical Turk to US-based participants, there is a possibility that users not located in the US may have used a VPN to hide their real location [14, 46]. Our institution's data privacy policy prevented us from capturing IPs and therefore, we are not able to ascertain whether the IPs are likely to be VPN IPs.

7 CONCLUSION

Concerns about freedom of speech, autonomy of individuals in deciding what content to consume, and the misalignment in incentives between users and platforms can render centralized content moderation by the platforms non-ideal. Researchers have identified democratized approaches to misinformation, where instead of deciding what users should or should not consume, users are empowered to make more informed decisions about what content

to believe and share. One such approach is to enable users to assess content and to capture their assessments as structured data, as assessing can help users have accuracy on top of their mind—reducing the likelihood that they share misinformation—and has the potential to warn the user's social circle about inaccuracies. However, assessments by a user or the user's social circle cannot match in scale to the amount of information to which the user is exposed. In this work, we attempt to deal with the scale problem by exploring the potentials and the challenges of incorporating a personalized AI for determining content accuracy on social media, which takes as training data a user's assessment of content and predicts how the user is likely to assess other content. Such an AI could act as a first-pass inspection of the accuracy of content that a user encounters—directing their attention to items that they likely will find trustworthy, or conversely, items that are probably inaccurate and would benefit from the user's explicit assessment. Through a user study we investigated how users perceive such an AI for content moderation. The user study involved users interacting with a personalized AI that would learn a user's assessments of a feed of tweets, show its predictions of whether the user would find other tweets (in)accurate, and evolve according to the user feedback about whether it is correct or incorrect in its predictions. Through a controlled experiment, we also studied whether users would be swayed by seeing the predictions of the AI in their decision on how to assess content. We found that users were in fact influenced by seeing the predictions of the AI and in fact, over time became more reliant on the predictions of their AI. However, this influence was mitigated when they provide reasoning for their assessment. We draw from our empirical observations to identify design implications and directions for future work.

REFERENCES

[1] [n.d.]. *About Verified Accounts.* Retrieved August 25, 2022 from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

[2] [n.d.]. *Label Sleuth.* Retrieved August 25, 2022 from https://www.label-sleuth.org

[3] 2018. *Facebook apologises for blocking Prager University's videos.* Retrieved August 25, 2022 from https://www.bbc.com/news/technology-45247302

[4] 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

[5] Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting fake news using machine learning: A systematic literature review. *arXiv preprint arXiv:2102.04458* (2021).

[6] Jennifer Nancy Lee Allen, Antonio Alonso Arechar, Gordon Pennycook, and David Rand. 2020. Scaling up fact-checking using the wisdom of crowds. (2020).

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

[8] Supanya Aphiwongsophon and Prabhas Chongstitvatana. 2018. Detecting fake news with machine learning method. In *2018 15th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON).* IEEE, 528–531.

[9] Marc-André Argentino. [n.d.]. *QAnon and the storm of the U.S. Capitol: The offline effect of online conspiracy theories.* https://theconversation.com/qanon-and-the-storm-of-the-u-s-capitol-the-offline-effect-of-online-conspiracy-theories-152815

[10] Cory L Armstrong and Melinda J McAdams. 2009. Blogs of information: How gender cues and individual motivations influence perceptions of credibility. *Journal of Computer-Mediated Communication* 14, 3 (2009), 435–456.

[11] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[12] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–16.

[13] 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).

[14] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis* 20, 3 (2012), 351–368.

[15] 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.

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

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

[18] Hugo L Borges and Ana C Lorena. 2010. A survey on recommender systems for news data. In *Smart Information and Knowledge Management.* Springer, 129–151.

[19] Lukas Brozovsky and Vaclav Petricek. 2007. Recommender system for online dating service. *arXiv preprint cs/0703042* (2007).

[20] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces.* 454–464.

[21] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[22] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[23] 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, 4 (2020), 1073–1095.

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

[25] 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

[26] Caroline Mala Corbin. 2009. The First Amendment right against compelled listening. *BUL Rev.* 89 (2009), 939.

[27] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 585–592.

[28] Pranav Dixit and Ryan Mac. 2018. How WhatsApp Destroyed A Village. *Buzzfeed News* (2018).

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

[30] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 183–193.

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

[32] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems.* 2371–2382.

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

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

[35] Elia Gabarron, Sunday Oluwafemi Oyeyemi, and Rolf Wynn. 2021. COVID-19-related misinformation on social media: a systematic review. *Bulletin of the World Health Organization* 99, 6 (2021), 455.

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

[38] 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

[39] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics* 83, 5 (2014), 368–375.

[40] Kirsten Gollatz, Felix Beer, and Christian Katzenbach. 2018. The turn to artificial intelligence in governing communication online. (2018).

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

[42] Ulrike Gretzel and Daniel R Fesenmaier. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce* 11, 2 (2006), 81–100.

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

[44] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems* 117 (2021), 47–58.

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

[46] Peter D Harms and Justin A DeSimone. 2015. Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology* 8, 2 (2015), 183–190.

[47] Kurtis Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, Dillanie Sumanthiran, Ji-ze Jang, Md Ali, Ehsan Hoque, et al. 2021. Could you become more credible by being White? Assessing impact of race on credibility with deepfakes. *arXiv preprint arXiv:2102.08054* (2021).

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

[49] 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.

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

[51] Touseef Iqbal and Shaima Qureshi. 2020. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences* (2020).

[52] 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.

[53] 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.

[54] 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.

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

[56] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 324–332.

[57] Jan Kirchner and Christian Reuter. 2020. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–27.

[58] Wolfgang Klein. 1994. Learning how to express temporality in a second language. In *Società di linguistica Italiana, SLI 34: Italiano-lingua seconda/lingua straniera: Atti del XXVI Congresso*. Bulzoni, 227–248.

[59] Silvia Knobloch-Westerwick and Jingbo Meng. 2011. Reinforcement of the political self through selective exposure to political messages. *Journal of Communication* 61, 2 (2011), 349–368.

[60] András Koltay. 2022. The Protection of Freedom of Expression from Social Media Platforms. *Mercer Law Review* 73, 2 (2022), 6.

[61] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[62] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[63] Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, et al. 2019. Detecting fake news using machine learning and deep learning algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. IEEE, 1–5.

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

[65] 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.

[66] Jaume Masip, Eugenio Garrido, and Carmen Herrero. 2004. Facial appearance and impressions of 'credibility': The effects of facial babyishness and age on person perception. *International journal of psychology* 39, 4 (2004), 276–289.

[67] Maria D Molina and S Shyam Sundar. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society* (2022), 14614448221103534.

[68] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 441–450.

[69] Kathleen L Mosier and Linda J Skitka. 1999. Automation use and automation bias. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 344–348.

[70] Kathleen L Mosier, Linda J Skitka, Mark D Burdick, and Susan T Heers. 1996. Automation bias, accountability, and verification behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 40. SAGE Publications Sage CA: Los Angeles, CA, 204–208.

[71] Kathleen L Mosier, Linda J Skitka, Kristina J Korte, M Mouloua, and R Parasuraman. 1994. Cognitive and social psychological issues in flight crew/automation interaction. *Human performance in automated systems: Current research and trends* (1994), 191–197.

[72] Mohsen Mosleh, Gordon Pennycook, Antonio A Arechar, and David G Rand. 2021. Cognitive reflection correlates with behavior on Twitter. *Nature Communications* 12, 1 (2021), 1–10.

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

[74] 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.

[75] 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.

[76] 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.

[77] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[78] 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.

[79] 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/

[80] 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.

[81] John Reed. 2018. Hate speech, atrocities and fake news: The crisis of democracy in Myanmar. *Financial Times. Retrieved from https://www.ft.com/content/2003d54e-169a-11e8-9376-4a6390addb44* (2018).

[82] 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.
[83] David O Sears and Jonathan L Freedman. 1967. Selective exposure to information: A critical review. *Public Opinion Quarterly* 31, 2 (1967), 194–213.
[84] 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.
[85] Emily V Shaw, Mona Lynch, Sofia Laguna, and Steven J Frenda. 2021. Race, witness credibility, and jury deliberation in a simulated drug trafficking trial. *Law and human behavior* 45, 3 (2021), 215.
[86] Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, Lucy Yip, Liat Ein-Dor, Lena Dankin, Ilya Shnayderman, Ranit Aharonov, Yunyao Li, Naftali Liberman, Philip Levin Slesarev, Gwilym Newton, Shila Ofek-Koifman, Noam Slonim, and Yoav Katz. 2022. Label Sleuth: From Unlabeled Text to a Classifier in a Few Hours. https://arxiv.org/abs/2208.01483
[87] Robert Shrimsley. 2016. *Facebook photos: snap judgments.* Retrieved August 25, 2022 from https://www.ft.com/content/dbcdf744-7ac6-11e6-b837-eb4b4333ee43
[88] Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Accountability and automation bias. *International Journal of Human-Computer Studies* 52, 4 (2000), 701–717.
[89] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon. com. *Ieee internet computing* 21, 3 (2017), 12–18.
[90] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[91] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160.
[92] Sara Spray. 2016. *Facebook Is Embroiled In A Row With Activists Over "Censorship".* Retrieved August 25, 2022 from https://www.buzzfeed.com/saraspary/facebook-in-dispute-with-pro-kurdish-activists-over-deleted
[93] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576.
[94] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. 1–9.
[95] Alan R Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the robotic age. *Commun. ACM* 61, 9 (2018), 22–24.
[96] Jinping Wang, Maria D Molina, and S Shyam Sundar. 2020. When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior* 107 (2020), 106278.
[97] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*. Springer, 563–574.
[98] 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.
[99] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.
[100] 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.
[101] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

A THE PERSONALIZED AI SYSTEM

For training a personalized model for each user that evolved as the user provided more assessments, we needed an AI that could be retrained quickly enough so that the user could see the AI reacting to their input and adjusting its assessments within the time-frame of the user study. Therefore, we deployed an instance of an open-source system for text annotation and building text classifiers called Label Sleuth [2, 86]. Label Sleuth is an interactive visual system consisting of both a frontend and backend, which allows users to interactively label text elements while the system automatically trains in the background, resulting in a text classification model which is iteratively updated as more labels are provided by the user. For the purposes of this study, we bypassed Label Sleuth's frontend (as we built our own custom UI tailored to this user study) and had our experimental system communicate with the API of Label Sleuth's backend to continuously submit a user's assessments, check for updates to the user's model, and retrieve the most recent model's predictions.

Label Sleuth trains a binary classifier that predicts which items from a document (in our case tweet texts) likely belong to the positive category (i.e., the category of interest). The positive category in our case was the category "inaccurate". The model training happens in iterations, with Label Sleuth starting a new model training iteration as more labels are provided. To allow for fine-grained control over when a new iteration is invoked, Label Sleuth offers two customizable settings: The first model positive threshold and changed element threshold. The first model positive threshold denotes after how many initial positive (i.e., inaccurate) labels a model should start to be trained for the user. The changed element threshold indicates the number of changes in user labels (either the user assessing new tweets or changing their previous assessments) relative to the last trained model that are required to trigger the training of a new model. For the purposes of our work, we needed to set these thresholds small enough to allow for a user's model to update at least a few times in reaction to user inputs. This reactivity was important especially in the condition Assisted to convey to the users that their model was indeed personalized and learning from their assessments over time. However, the threshold should not be too small to result in updated models whose predictions keep flip flopping compared to the previous models or do not differ from the previous models at all (and therefore cause wasted computation). Based on empirical testing, we set both first model positive threshold and changed element threshold to 4.

Behind the scenes, Label Sleuth trains a Support Vector Machine (SVM) model for the user, and therefore is lightweight enough to allow for fast retraining of the model when new labels arrive. While retraining of a model is fast, it still takes time approximately in the order of a minute. After the training of a new model for a user by Label Sleuth, our experimental platform needed to determine which, if any, predictions by the new model were different from before, record them, and signal them to the user in condition Assisted. This processing would incur an extra delay. The delay of model training and processing model predictions, along with the fact that a new model would be triggered to start training after a new set of 4 tweets are assessed meant that each feed had to be long enough (i.e., consist of enough tweets) so that the predictions for a user and for a feed could change at least a few times before the user finished assessing the tweets of that feed. Making the feeds too long however, would make the task tedious and possibly result in a loss of participants' attention after some point.

Considering these constraints, we empirically tested feeds of different lengths, and decided that each feed would consist of 26

tweets. With 26 tweets in each feed, each of the models Hidden and Visible would have seen 52 examples by the end of their life cycle. This is because model Hidden is trained on and used to predict the assessments of the tweets belonging to the Seeding Step and condition Unassisted; and model Visible is trained on and used to predict the assessments of the tweets belonging to the Seeding Step and condition Assisted. In fact, another purpose that providing free-text reasons for 3 tweets in each feed served was delaying users a little longer.

B DATA CLEANING

We cleaned the text of the tweets both for feeding to the AI as well as presenting to our participants. We removed strings of the form @username at the beginning of tweets, because these were mostly present on tweets that were replies to other tweets (but that nevertheless were self-contained to be included in our experimental task). We replaced other occurrences of @username for the AI with a special token. These mentions in the middle of the text were used for instance to cite a quote from another account. The reason we replaced these mentions with a generic token for the AI was because the occurrence of each one was too rare in our small dataset for the AI to learn any meaningful information from it. For the same reason, we replaced the links for the AI with another special token. We removed hash signs (used in hashtags) for the AI but kept the word that followed the sign. For both the AI as well as our study participants, we removed indicators of thread at the end of a tweet's text (e.g., 2/5). Only the tweet texts were used for training users' personalized AI models. Other features such as tweet author or date were not considered because the occurrence of each unique value of those features would be too rare in our small dataset to help with training.

C USERS' REASONING ABOUT WHY THE AI WORKS OR FAILS

In reasoning about the AI's performance, many users had come up with their own mental models of how the AI made its predictions, and in light of their theory, had decided to either excuse the AI for its mistakes, hold it up to higher standards, or simply rationalize its correct predictions and mispredictions. This phenomenon is similar to how users hypothesize how the black box Facebook curation algorithm chooses content for their news feed [32, 33]. One such model was that the AI was using keywords in deciding the accuracy of a tweet and therefore did not understand semantics (N=3), a concern that has been reported about centralized AIs for content moderation before [40]. A class of theories speculated that the AI's mispredictions were due to (or more frequent for) tweets discussing certain subtopics (N=17), without (or with complex) supporting citations and links (N=11), with mixed or yet undecided factuality or which mixed opinions and facts (N=6), of a certain veracity (e.g., the AI made more mistakes on the tweets that were in fact accurate) (N=6), with complicated wording or technical jargon (N=4), sarcasm (N=4), nuances in wording (N=2), or when tweets had the appearance of being legitimate but in fact were not, because for example, they included dog whistles or twisted facts (N=4). Another interesting theory was that the AI was attempting

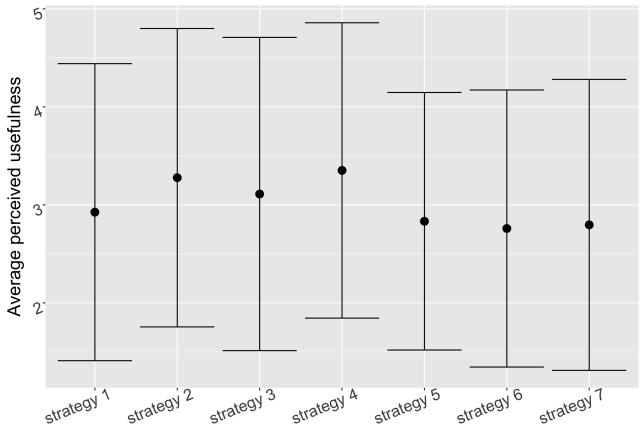


Figure 10: Users ratings of how useful they find each of the content moderation strategies described in Section D on a 5 point likert scale. The means are rather close to each other and the standard deviations are very wide, indicating a high variance among users.

to categorize users into one of two sides and predict the labels accordingly, but that some users may not fit in either side (N=1).

> *"The ones where people were using extreme language, but in a sarcastic way that I think the bot misread as literal."*

D A COMPARISON OF DIFFERENT CONTENT MODERATION STRATEGIES

To capture our participants' perception of a personalized AI for determining the accuracy of content in comparison with other content moderation strategies, we asked them to rate the usefulness of a set of such strategies in the post-study survey. These strategies differed along the three dimensions described in Section 2: some were centralized, others democratized; some relied on humans as moderators, others incorporated automation; and some took action against misinformation by removing the misinforming content, others by flagging it instead. The approaches are described in Table 12:

Participants rated the usefulness of each strategy on a 5-point likert scale. Figure 10 shows the mean and the standard deviations around the mean for each of these strategies. Figure 11 additionally shows the distributions of participants' responses about to what extent they consider each of the content moderation strategies useful. As these figures demonstrate, no one strategy appeals to all users, and users' opinions of each strategy vary widely. For each strategy, there are avid proponents and those who are set against it.

We examined participants' free-text responses to understand their perceived pros and cons of each strategy. Some participants were in general proponents of content moderation regardless of the strategy used. They believed the universal advantages of content moderation are that they create a curated environment for user, the user would not have to fact-check every piece on their own, and that because misinformation is a rampant problem, dealing with

Table 12: The content moderation strategies about which we asked our participants' perceptions and where they fall along the 3 dimensions discussed in Section 2.

Strategy	Centralized vs Democratized	Human moderation vs Automation	Flagging vs Removal of content
1. Social media platforms running their own AI to identify and remove misinformation from a user's newsfeed	centralized	automation	removal
2. Social media platforms running their own AI to identify and flag misinformation from a user's newsfeed.	centralized	automation	flagging
3. Social media platforms employing human moderators to identify and remove misinformation from a user's newsfeed.	centralized	human	removal
4. Social media platforms employing human moderators to identify and flag misinformation from a user's newsfeed.	centralized	human	flagging
5. Social media platforms enabling all users to assess posts for accuracy and allowing each user to specify whose assessment they want to see.	democratized	human	flagging
6. Social media platforms enabling all users to not only assess posts, but also run their personalized AI for predicting content accuracy based on the assessments they have provided (similar to this study).	democratized	human & automation	flagging
7. Social media platforms enabling all users to assess posts, run their personalized AI for predicting content accuracy (similar to this study), and see the predictions of the personalized AI of other users.	democratized	human & automation	flagging

it needs complementary resources. There were other participants who were against any content moderation strategy, because they wanted to see all the information there is and decide for themselves, they believed all these strategies would ultimately be deployed by the platforms who cannot be trusted, or they visit social media to look for opinions and not facts.

An advantage that was cited in support of democratized approaches was that they would be less biased than social media platforms who are currently the authority in centralized moderation. Some participants also believed that it would be helpful to learn about the judgment of other users they find trustworthy on content they encounter on social media. The concerns cited against democratized approaches were that users may not assess in good faith or may be biased, personalized assessments may result in the spread of misinformation or confusion about what is true, they may result in the development of echo chambers, some people do not have the ability to determine what is misinformation, or that because users already know which of their sources are trustworthy, do not need a tool to signal the trustworthiness of content to them.

> *"I think giving too much power to the platforms users could be an issue. One bad apple could lead to many very quickly and they they're flagging everything that doesn't fit their agenda, truth or not."*

> *"Social media is filled with different opinions, finding out what is right and what is wrong is hard. Personalized AI may be good, its [sic] a way to help enhance your views and possibly weed out the false news. However, I want accuracy. I support my views, but I support the truth. If my views are wrong, I want to know that."*

The advantages cited in support of using AI in content moderation described the AI as being more efficient, more trustworthy than unknown or anonymous human moderators, and not having bias or morals. On the other hand, some believed that because humans understand the nuances of wording better, they are better equipped to detect misinformation.

Any participant responses that discussed removal vs flagging of misinforming were against removal. This was because they believed content removal would equate censorship and would be a danger to free thinking, what is considered true changes over time, and sometimes users want to read certain posts regardless of whether they are accurate.

> *"It greatly depends on the user themselves. I prefer that everyone has a choice to choose what they find most useful. I mind much less if an AI or human flags something as "potentially" misleading, but I have a huge issue with information being fully removed. What is considered "true" changes over time, a case in point being the origin of the covid virus itself, so it is dangerous to remove information ever in my opinion. Flag is ok. Remove is not OK."*

The Objectivity of the Truth. In response to this question, three users had explicitly asserted that the truth is objective and therefore were against personalized content moderation approaches because they believed people should not be allowed to curate personalized truth for themselves. We examined the assessments from these users to understand how often the agreed with each other about what the truth is. The agreement in assessments among these 3 users was 85%. They disagreed on the accuracy of 12 tweets (out of 78). On 7 of those 12 tweets, at least 2 accounts of disagreements among the 3 users had a confidence rating of 3 (somewhat confident) or higher. This is an interesting result, showing that the three users that believed on the existence of a single truth, in some cases disagreed between themselves on what that truth is. Nonetheless, although these users do not always agree on what the objective truth is, it is possible that they are willing to change their assessment if an authority notified them about it.

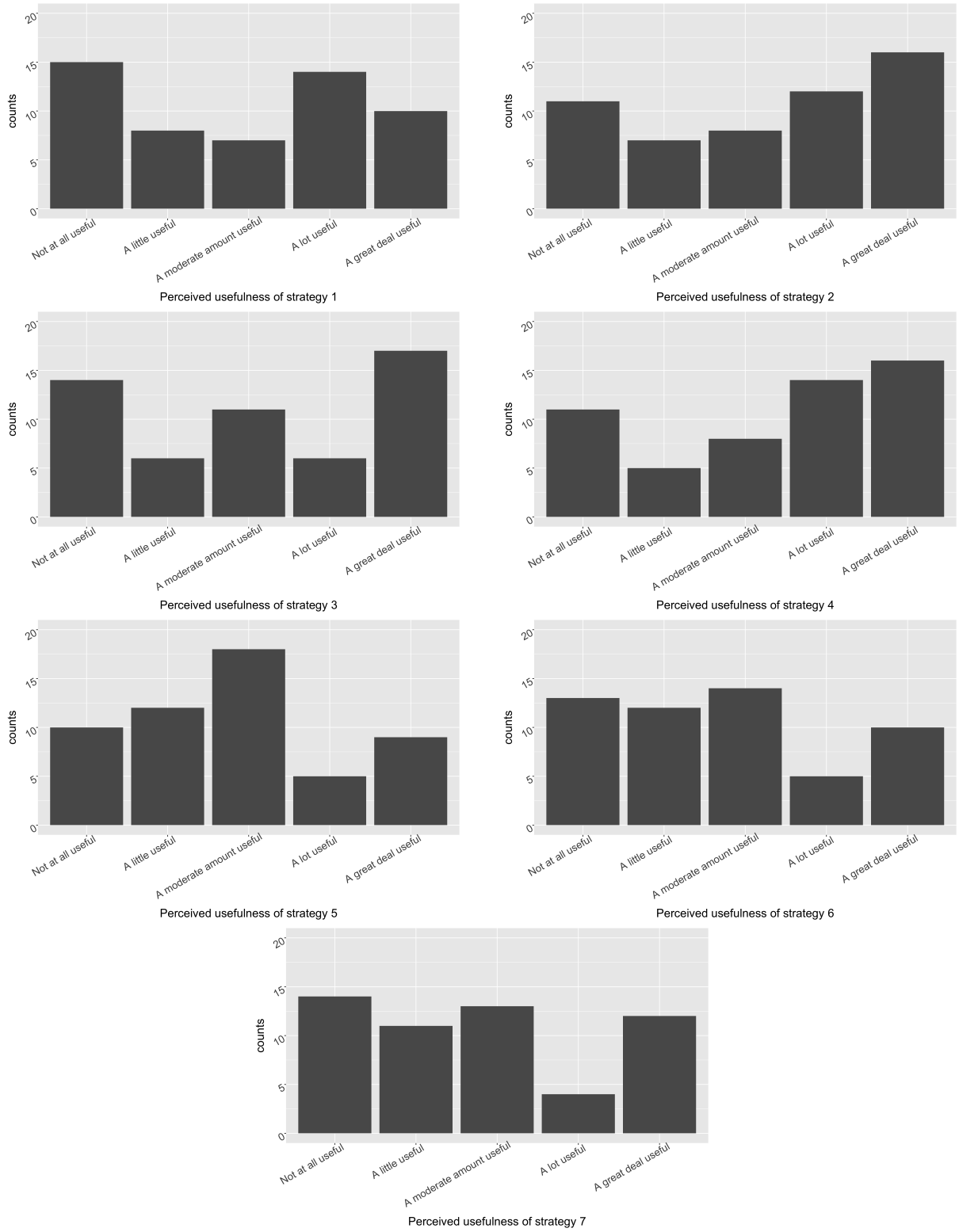


Figure 11: The participants' perceived usefulness of the content moderation strategies described in Section D.

Table 13: Does users' agreement with their personalized AI change over time even in the Seeding step where users are not exposed to the AI's predictions? The analysis in the Table suggests that it does, and the difference in user-AI agreement across the two conditions Assisted and Unassisted grows larger over time even in the Seeding step.

Independent Variable*	$exp(\beta)$	CI	z test
model (whether the model that assessed a tweet for a user was Hidden or Visible)	0.98	[0.85, 1.12]	$z = -0.33$, $p = 0.74$
iteration of the model	1.07	[1.03, 1.12]	$z = 3.24$, **p=0.001****
model × iteration	1.07	[1.00, 1.13]	$z = 2.05$, **p=0.04***

*Dependent variable: Whether a user's assessment of a tweet agrees with the AI's Marginal R^2/Conditional R^2 = 0.006/0.175

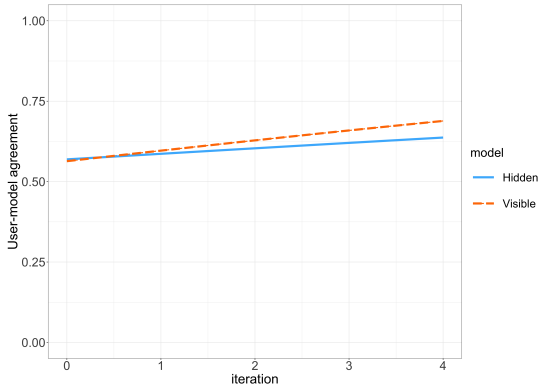


Figure 12: The effect of the interaction between model and iteration on the predicted user-model agreement at iterations when the user was still in the Seeding step. Over time, the difference is user-model agreement across the two models increases. The difference in these initial iterations however, is smaller than the difference observed in Figure 9 that also includes the iterations after the Seeding step.

E EXAMINING THE DIFFERENCE IN USER-MODEL AGREEMENT WHEN USERS WERE STILL IN THE SEEDING STEP

As discussed in the Appendix Section A, the initial version of the model for a user would not be trained until after the first model positive threshold was met, meaning that the user had assessed at least 4 tweets as inaccurate. While for all users the first model positive threshold was met in the Seeding step, for some it was met much later than the others within the feed, because it was dependent on the accuracy ratings that they gave to the tweets in the feed. Therefore, the iteration of the model at which the users transitioned from the Seeding step to the next step was also different across users. In 4.2.1 we established that the user-model agreement across the 2 models Hidden and Visible grows larger over time. We wanted to further examine the difference when the user was still in the Seeding step and not yet exposed to the predictions of the AI, and observe whether it would be different from the overall trend that we observed before. Therefore we performed the regression in 13 on that portion of the dataset when the users had still been in the Seeding step. The results were similar to ones we observed for the entire dataset.

Figure 12 shows the effect of the interaction effect between iteration and model for when users were still in the Seeding step. We observe both from the statistical tests as well as the figure that there is a difference in user-model agreement across the two models from the beginning. If the data is from when the users were still in the Seeding Step (and not yet exposed to the predictions of their model), why is there a difference in user-model model agreement across the two models?

Imagine that the model Visible for a user p at iteration 1 has predicted that the tweet t is inaccurate. The future iterations of model Visible for the user are also likely to retain the same prediction for tweet t, since after each retraining of the model, the predictions for only a few tweets would change as the model does not flip flop widely on its predictions from one iteration to the next. Therefore, when the prediction of the model Visible about tweet t is eventually shown to the user in condition Assisted at some future iteration, the user agrees with the AI, and therefore it is as if the user has also agreed with the predictions of an earlier iteration of the model. In summary, the user-model agreement is higher for model Visible in the Seeding step because 1) in the future iterations when the user sees the predictions of a later model, they shift their "ground truth" (i.e., the accuracy assessment of a tweet) to match what the model predicts, and 2) the predictions of that later model largely match the predictions of the earlier versions of the model, including those in the Seeding step.

F TASK INSTRUCTIONS

Before interacting with each of the 3 feeds, the platform presented users with instructions on what they needed to do next. Figures 13, 14, and 15 show screenshots of these instructions.

Task Instructions

You will see a feed of tweets related to COVID as shown in the picture.

For all the tweets on the feed, use your best judgment to assess whether they are accurate or inaccurate and how confident you are in your belief. There is also a textbox where you can provide your reasoning for why you believe the tweet is or is not accurate.

Upon marking the the tweet as accurate/inaccurate, two other questions appear in order to capture your confidence in your assessment and your reasoning.

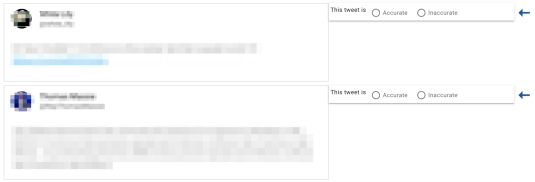
We ask that you **provide your reasoning for at least 3 of your assessments**. You are encouraged to provide your reasoning for as many more tweets as you would like, as it would greatly help our research.

Your task is to use the assessment pane next to each one to label their accuracy. We will use the labels that you provide to train a personalized Artificial Intelligence (AI) model for you which intends to learn and predict how you would label other tweets.

Once you have assessed and marked your level of confidence for all the tweets as well as provided your reasoning for at least 3 of your assessments, you will be able to proceed to the next step.

Figure 13: Instructions before the Seeding step.

We need a few more accuracy labels from you. The next page will show you another feed of tweets. We ask that you also assess the tweets on that feed.
You are only able to change the assessments of **4** tweets at a time. The assessments that you can change at a certain time are marked with a blue arrow, similar to the picture below.

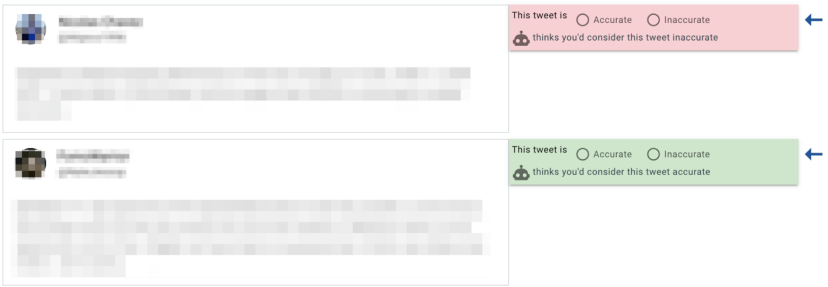


Similar to the last step, **once you have assessed and marked your level of confidence for all the tweets as well as provided your reasoning for at least 3 of your assessments**, you will be able to proceed to the next step.

Figure 14: Instructions before condition Unassisted.

On the next page, you will see the last feed of tweets. For each tweet, we will show the AI's prediction of how you would assess the tweet. Your task is to guide the AI to become better at learning your assessments by indicating whether you agree or disagree with the AI's predictions. You will do this by explicitly assessing tweets as accurate or inaccurate.

Although you will be able to see the AI's predictions of your assessments on all the tweets of the feed, similar to the previous step, you are only able to change the assessments of 4 tweets at a time. The assessments that you can change at a certain time are marked with a blue arrow, similar to the picture below.



As you are guiding the AI, a list of updated predictions will appear on the side. These are the tweets that the AI has changed its prediction on, based on the feedback that you have given to it. We encourage you to explore this pane. In this pane, you can click on any of the tweet previews (orange arrow shown on the picture below) to bring the actual tweet into view.

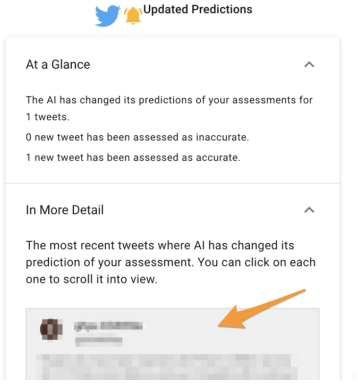


Figure 15: Instructions before condition Assisted