advances.in/psychology

July 16, 2024
Doi: 10.56296/aip00018
ISSN: 2976-937X

REVIEW

Scaling crowdsourcing interventions to combat partisan misinformation

Clara Pretus[1,2,*], Helena Gil-Buitrago[2], Irene Cisma[3], Rosamunde C. Hendricks[2], & Daniela Lizarazo-Villarreal[1,2]

Received: March 21, 2024 | **Accepted:** July 3, 2024 | **Published:** July 16, 2024 | **Edited by:** Jonas R. Kunst

[1] Department of Psychobiology and Methodology of Health Sciences, Universitat Autònoma de Barcelona. [2] Hospital del Mar Research Institute, Barcelona, Spain. [3] Institut de Neurociències, Universitat Autònoma de Barcelona. *Please address correspondence to Clara Pretus, clara.pretus@gmail.com, Department of Psychobiology and Methodology of Health Sciences, Universitat Autònoma de Barcelona, Carrer Fortuna, Campus de la UAB, 08193 Bellaterra, Spain.

Partisan misinformation undermines people's ability to make decisions based on accurate information, posing a threat to democracy and liberal values. Current interventions to counter misinformation are less effective when it comes to politically polarizing content, especially among extreme partisans who share the most misinformation. A new line of research suggests that crowdsourcing interventions, or using laypeople's judgments to help people spot misinformation, provide an additional layer of content moderation that can help overcome these limitations. We present a model that explains when crowdsourcing interventions will be successful based on three factors: trust in fact-checking sources, dissonance with previous beliefs, and crowd size. These three factors are often at odds in politically polarized social media environments, where more trusted sources may be less willing to provide dissonant opinions, resulting in smaller fact-checking crowds. Based on this model, we discuss how crowdsourcing interventions could be scaled in a way that is ethical and leverages network analysis methods to connect people with neighboring communities outside their ideological echo chambers. Finally, we propose venues for future research in the field of crowdsourcing interventions that lie at the intersection between individual-level and system-level solutions to partisan misinformation.

Keywords: crowdsourcing, fact-checking, misinformation, trust, network analysis

1. INTRODUCTION

Misinformation has been listed as the “most severe short-term risk the world faces”, as reported by members of the World Economic Forum (World Economic Forum, 2024). One reason for this is that it threatens democracy and can destabilize society through ideological polarization (Au et al., 2021; Spohr, 2017). Polarization is fueled by people's tendency to amplify moralized content that favors “my side”, regardless of its accuracy (Marie et al., 2023; Pretus et al., 2023; Rathje et al., 2021). This process can arise from rational belief updating given a set of priors (Cook & Lewandowsky, 2016). Efforts to correct misinformation through debunking are effective among the general population (Chan et al., 2017) but have reduced effectiveness among far-right partisans (DeVerna et al., 2022; Martel et al., 2024; Pretus et al., 2023; Rathje et al., 2022), who spread the largest share of online misinformation (Guess et al., 2019). These limitations could partly be overcome by a multi-layered approach to content moderation (Bak-Coleman et al., 2022), including crowdsourcing strategies where social media users share responsibility for fact-checking content (Pretus et al., 2024). In the current work, we explore how crowdsourcing interventions could be scaled to provide a fast first-line response to content moderation on social media.

Crowdsourcing, or the act of obtaining needed services, ideas, or content by asking for contributions from a group, especially from an online community, rather than from traditional suppliers (Allen et al., 2021), has yielded significant achievements for humanity. One example is Wikipedia, the largest ever-evolving encyclopedia fully managed by human volunteers, who continuously improve over 6.7 million articles in English at a rate of 21 edits per second (Wikimedia Foundation, 2024, June 12; Wikimedia Statistics, 2024). Crowdsourced knowledge by non-experts can also help people discern the veracity of political statements (Espina Mairal et al., 2024), which could have wide-ranging effects on how we moderate content on social media. This is partly because, when aggregated, laypeople's judgments can be used to accurately discern reliable news sources (Pennycook & Rand, 2019) and true news headlines (Arechar et al., 2023). Critically, collective accuracy judgments are well received even by extreme partisans, who reduce partisan misinformation sharing in alignment with the number of people who think a given post is misleading (Pretus et al., 2024). Finally, crowdsourcing could also reinforce fact-checking behaviors by presenting them as socially desirable (Gimpel et al., 2021) and creating a sense of shared responsibility over content moderation. Thus, crowdsourcing interventions could meet current demands for more immediate fact-checking strategies that are effective among extreme partisans.

The question of how to implement crowdsourcing interventions in real-world informational ecosystems is not a trivial one. “Community Notes”, a community-based fact-checking system by Twitter/X, currently uses laypeople's knowledge for content moderation. However, the reliance on single users to add contextual information to misleading posts followed by its validation by a diverse set of users causes delays in fact-checking, which often occurs only after posts have circulated widely and accumulated thousands of impressions (Renault et al., 2024). This limitation could be overcome by allowing everyone to tag posts as misleading and presenting fact-checks in an aggregate manner ('number of people who think a post is misleading'). The aggregate approach to crowdsourcing elicits questions such as who can fact-check which content, and should everyone be able to see everyone else's fact-checks?

If everyone could see everyone else's fact-checks, then people could debunk content they simply disagree with, and coordinated bot attacks could be used to bring down posts by political opponents. If people were only exposed to fact-checks by accounts they follow,

they would be more likely to fact-check content that is incoherent with their ideological bubble, exacerbating echo chambers. Conversely, if people were solely exposed to fact-checks generated outside of their community, then people may mistrust those fact-checks. Importantly, the way an algorithm filters who can see whose ratings should be transparent to users, who have the right to know how information is presented to them. Hence, the issue of how to implement crowdsourcing interventions is far from resolved.

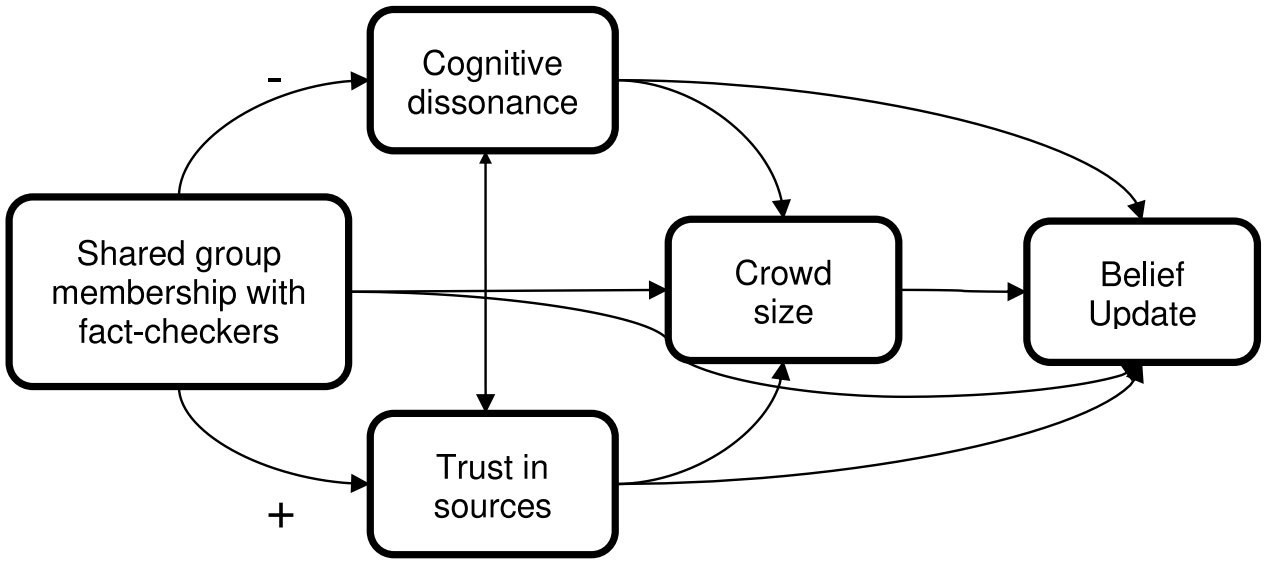
We present a theoretical model to help address these questions (see Figure 1). The model includes three key psychological factors that should be considered to implement crowdsourcing interventions in a way that promotes belief updates. These factors include dissonance with previous beliefs, trust in fact-checking sources, and crowd size. We review the literature supporting why each of these elements is important for effective fact-checking. Cognitive dissonance, trust, and crowd size are often at odds in polarized social media environments due to strong ingroup/outgroup tensions. In these contexts, ingroup sources will be more trustworthy but will provide fewer dissonant fact-checks, while outgroup sources will generate more dissonant fact-checks but will inspire less trust. We propose ways to balance out these factors when implementing crowdsourcing interventions in the real world by leveraging network analysis methods. Finally, we suggest directions for future research that will help elucidate how to optimize crowdsourcing interventions to combat misinformation at scale.

2. HOW DO CROWDSOURCING INTERVENTIONS WORK?

Crowds across the globe can generate high-accuracy judgments (Arechar et al., 2023). This happens because of a phenomenon known as the *wisdom of the crowds*, which was first discussed more than a century ago (Galton, 1907), and in modern days has been compared to the *"Ask the audience"* resource from *Who Wants to be a Millionaire?* (Surowiecki, 2004). The concept is rooted in the idea that collective judgment from a diverse and large group of individuals can be more accurate than that of isolated experts. For instance, aggregate judgments from politically balanced groups of laypeople have been found to closely align with professional fact-checkers (Allen et al., 2021). In the

Figure 1

Elements of Crowdsourcing Interventions that Influence Belief Update



context of social media and misinformation, crowdsourcing from a broad base of internet users can be leveraged to evaluate the veracity of news articles, in other words, crowdsourcing could help appropriately identify and fact-check misinformation.

The wisdom of crowds arises from adding diverse viewpoints and information (Larrick et al., 2012). Accuracy then originates from the principle that, while individual estimates may vary widely, their average is likely to be closer to the truth due to the dissolution of individual biases and errors. For instance, using a diverse pool of evaluators has proven to be effective in assessing the accuracy of news articles on civic topics, health-related information, and other content (Allen et al., 2021). In another study, crowd workers evaluated the truthfulness of statements related to the COVID-19 pandemic using a customized search engine (Roitero et al., 2020). Their task involved judging the accuracy of eight statements while researchers analyzed factors such as worker background, biases, and cognitive abilities. Despite the variability in internal agreement between crowd workers, their combined judgments generally aligned with expert labels. A later longitudinal study was conducted by re-launching the task multiple times with both novice and experienced workers (Roitero et al., 2023). The longitudinal study replicated the results, showing that workers were able to detect and objectively categorize online (mis)information related to COVID-19. Although these results are promising, current evidence points to specific conditions in which the accuracy of crowdsourced judgments can be optimized.

2.1 When do crowds generate accurate judgments?

Recent studies suggest crowds can generate more accurate judgments under certain conditions. When it comes to the composition of the crowd, researchers have found that evaluations from heterogeneous crowds have a greater alignment with expert assessments. For instance, Espina Mairal et al. (2024) found that political heterogeneity boosted the accuracy of collective estimates when evaluating the trustworthiness of political messages. This is because diverse and politically balanced groups ensure the inclusion of a range of perspectives to bear on the veracity of news content. The study found that politically heterogeneous social influence distinctively improved the ability of individuals to discriminate between true and false statements relative to a control benchmark with no social feedback. Conversely, homogeneous groups significantly increased their political bias after social influence. This is connected to the fact that people's political alignments strongly influence their willingness to classify news as misinformation. For instance, Coscia and Rossi (2020) found that individual biases significantly impact the assessment of content veracity by simulating the behavior of users in a crowdsourced content policing system using agent-based modeling. This computational modeling approach simulates the actions and interactions of autonomous agents to assess their effects on the system as a whole. Simulated users' judgments were influenced by factors such as their degree of polarization, tolerance levels, and propagation attitudes. These results underscore the need for nuanced strategies to increase crowd heterogeneity.

In terms of crowd size, Espina Mairal et al. (2024) find that more than 3 or 4 individuals significantly boost the accuracy of trustworthiness estimates. Thus, larger crowds are more likely to offer accurate collective estimates. As a reference, several studies report high levels of external agreement with expert assessments based on the evaluations of 10 crowd workers (La Barbera et al., 2020; Roitero et al., 2018; Roitero et al., 2023).

How crowd worker responses are collected and aggregated also affects the accuracy of crowdsourced judgments. Importantly, individual responses should be collected

independently from one another to preclude social influence effects between fact-checkers which can lead to lower accuracy performance (Hong et al., 2019). For instance, Condorcet's jury theorem elucidates the need for individual judgments to be independent for accuracy to improve with larger group sizes (Hahn et al., 2019). In other words, increasing the number of individual judgments will not improve accuracy if they are not independent. Moreover, the scales used to rate the truthfulness of a given news vary widely across fact-checking websites and research studies. Some agencies use four to seven-level scales, such as *True, Partly True, Misleading,* and *False* (Lee et al., 2023), while some studies have experimented with 0 to 100 and 0 to infinite scales (Roitero et al., 2018). Recent work suggests that greater agreement with expert judgments can be achieved by merging these ratings into 2 or 3 categories, such as *True, In between*, and *False* (Roitero et al., 2023). Of note, merging percent ratings into fewer categories could artefactually increase external agreement with experts. The same authors underlie the need to use average crowd worker ratings to obtain more accurate aggregate ratings instead of relying on the internal agreement among workers, which is often low. Thus, how collective estimates are collected and processed is also an important issue to consider when designing effective fact-checking strategies.

Finally, researchers and social media platforms can use quality checks to optimize the accuracy of crowdsourced fact-checks. Quality checks can include test questions that help discern high-quality from low-quality fact-checkers, for example by asking them to rate known accurate and inaccurate statements. Another quality check involves asking fact-checkers to justify their assessments. For instance, Roitero et al. (2023) asked crowd workers to support their accuracy judgments with a written justification of over 15 words as well as the URL of the website they had used as a source of verification. They found that crowd workers who used text selected from the source website to justify their responses, as well as more free text reworking that information, provided more accurate evaluations. The same study found that assessments from crowd workers who participated in more than one set of evaluations (experienced workers) were of higher quality than those of novice workers. Thus, the accuracy of crowdsourced fact-checks can be improved not only by fine-tuning the composition of the crowd and the data collection process but also by including additional checks that discern high-quality from low-quality fact-checkers.

Of note, most of the reviewed studies were conducted in Western contexts and focused on specific social media platforms (e.g., Allen et al., 2021; Coscia & Rossi, 2020; Espina Mairal et al., 2024; Roitero et al., 2018, 2023). Beyond the Western context, Arechar et al. (2023) found that aggregated crowd judgments on the veracity of news were highly accurate across all 16 examined countries on 6 continents. Nevertheless, additional cross-cultural replications are needed to understand how crowds perform in different cultural contexts and social media platforms.

The effectiveness of crowdsourcing significantly depends on the context, specific implementation strategies, and participants' political biases. Despite that, the reviewed body of evidence suggests that when properly structured, crowdsourcing can be used as an accurate mechanism for evaluating information at a scale and speed that traditional fact-checking cannot match.

3. HOW DO CROWDSOURCING INTERVENTIONS INFLUENCE PEOPLE'S BELIEFS?

For crowdsourced fact-checks to be effective, they need to be both accurate and believed by other social media users. One of the main strengths of using crowdsourcing approaches is that crowds provide normative information. Collective decisions reflect a shared consensus or societal norm that guides individuals'

behavior and decision-making in a way that aligns with the values and knowledge of the group as a whole (Larrick et al., 2012). This collective decision-making process can legitimize outcomes and establish standards that individuals are more likely to accept and follow. Social norms significantly influence people's behavior across a variety of outcomes such as prejudice reduction, energy conservation, economic decisions, health choices, and many other domains (McDonald & Crandall, 2015). While descriptive norms impact behavior by reflecting common actions within a community, injunctive norms drive behavior through societal expectations. For example, people may save energy because they conform to the energy-saving behaviors of their neighbors (descriptive norm) or because they perceive saving energy as a socially desirable behavior (injunctive norm). When it comes to fake news, injunctive norms about the desirability of reporting misinformation increased the reporting of fake news, while descriptive norms, or letting participants know that other people report fake news, did not. In contrast, the combined application of both injunctive and descriptive social norm messages was found to be more effective in improving reporting behavior regarding fake news (Gimpel et al., 2021).

Normative information can also impact virality. In a study examining the impact of crowd wisdom on fact-checking social media content, researchers analyzed the spread of posts that underwent crowd fact-checking on Twitter's Birdwatch platform (Drolsbach & Pröllochs, 2023). Unlike previous studies that focused on misinformation fact-checked by third-party organizations, this study found that posts identified as misleading by the crowd were less likely to go viral compared to non-misleading posts. These findings suggest that people rely on social cues to determine what content they are willing to share and emphasize crowdsourcing as an effective method to mitigate misinformation.

Individuals rely on social influence to address their need for accurate reality perceptions, strong relationships, and a favorable self-concept (Cialdini & Goldstein, 2004). Specifically, normative information can help people generate more accurate evaluations that are also aligned with group norms, affirming their identity as group members. Misinformation challenges this logic; even if fake news is often widely shared and appears normative, it is still inaccurate. Thus, as proposed by the Identity-based Model of Belief (Van Bavel & Pereira, 2018; Van Bavel et al., 2024), identity concerns often override accuracy goals, leading individuals to align their beliefs with group norms over objective facts. For instance, individuals are more likely to share misinformation aligned with group-relevant values, especially when they highly identify with their reference groups (Pretus et al., 2023). Importantly, identity concerns not only influence people's informational context and who they trust but also who they accept fact-checks from. Thus, addressing misinformation effectively requires interventions that provide accurate information while also resonating with people's social identity.

In sum, crowdsourcing interventions offer a way to present accurate and normative information, thus aligning precise descriptions of reality with social identity motives. Engaging a broad audience in fact-checking allows for collective wisdom and offers a space for a shared truth that can resist misinformation distortions. We next discuss which conditions need to be met so that crowdsourced fact-checks are not only accurate but also influential in changing people’s beliefs.

3.1 When do crowds influence people’s beliefs?

Crowdsourcing interventions stand out as a promising strategy to counter misinformation. However, they rely on people's willingness to accept corrections of information they may agree with and believe new information they may disagree with. We explore three factors that can influence people's willingness to

update their beliefs in response to crowdsourced accuracy judgments. First, receiving fact-checks that are incongruent with one's beliefs can lead to cognitive dissonance, which seems to either facilitate or hinder belief update. Secondly, who is part of the crowd and how the crowd is labeled can influence how much trust people place in crowdsourced accuracy judgments. Finally, the size of the crowd affects how much a given accuracy judgment is perceived as normative, impacting people's propensity for belief update. We discuss how each of these factors influences the design of effective crowdsourcing interventions to counter misinformation and promote belief update.

3.1.1 Cognitive Dissonance

Cognitive dissonance has been defined as the underlying tension arising from inconsistencies between an individual's behavior and their thoughts and beliefs (Festinger, 1957), or from simultaneously holding two conflicting beliefs (Cognitive Dissonance: A History in Tweets, 2011; Cooper, 2012; Fontanari et al., 2012). Cognitive dissonance can occur in a variety of situations such as forced compliance, when individuals are forced to perform actions that are inconsistent with their values, beliefs, or behaviors; or when gaining new information that is not aligned with previous beliefs (Annu & Dhanda, 2020). In the context of misinformation, cognitive dissonance might occur when individuals are exposed to fact-checks that are incongruent with their previous beliefs or disrupt their tendency to actively seek, engage with, and share ideologically congruent information. Individuals are motivated to reduce the psychological discomfort elicited by cognitive dissonance by either changing their previous beliefs to match the newly acquired knowledge or by reinforcing their previous beliefs and avoiding any conflicting information (Kaaronen, 2018). For instance, studies have found that the psychological discomfort triggered by retractions of misinformation can translate into a continued belief in misinformation (Susmann & Wegener, 2022). Hence, understanding how and when cognitive dissonance influences the efficacy of fact-checks in promoting attitudinal and behavioral change is critical to the development of crowdsourcing interventions against misinformation.

Cognitive dissonance can hinder individuals' willingness to reconsider their values and perspectives, leading them to reject or scrutinize information that is incongruent with their preexisting convictions (Lewandowsky et al., 2012). For instance, exposure to opposing views on social media has been found to exacerbate political polarization (Bail et al., 2018). The underlying cognitive process behind this phenomenon is known as *cognitive immunization*, which describes strategies through which individuals devalue information that contradicts their prior beliefs (Kube, 2023), thus contributing to belief maintenance. This is related to confirmation bias or people's tendency to seek out evidence that aligns with their current outlook (Voinson et al., 2015) and disregard contrary viewpoints by modifying their attention levels and information recall (Al Marrar & Allevato, 2022). As a result, people may either devalue or reframe information that conflicts with their values. For instance, climate change skeptics might reject the evidence of its occurrence by either undermining the credibility of the source (devaluating the information) or by stating that changes in climate are part of a natural process (reframing the information). In this sense, cognitive dissonance represents a burden to fact-checking attempts as engagement in cognitive immunization can prevent belief update despite disconfirming information.

One relevant strategy individuals use to reduce the psychological discomfort elicited by cognitive dissonance is altering elements of their social context. Particularly, in online social environments, people may prefer to connect with like-minded others to avoid exposure to conflicting information from their social media feeds. For instance, individuals who are usually very hostile online might surround themselves

with other people who also provoke hostility (Kaaronen, 2018). Online platforms like Facebook and Twitter/X facilitate interactions among individuals who share similar perspectives (Cinelli et al., 2020), which is exacerbated by people's tendency to interact with like-minded others, a phenomenon known as homophily (Fu et al., 2012). Homophily significantly shapes personal informational environments, influencing the information individuals are exposed to, the attitudes they adopt, and their interactions (McPherson et al., 2001). Homophily also promotes the formation of echo chambers, which contribute to the reinforcement of beliefs, polarization, and the spread of misinformation (Acemoglu et al., 2021; Jiang et al., 2021; Törnberg, 2018). In this sense, cognitive dissonance and homophily might be tightly intertwined, since homophily reduces the chance of experiencing cognitive dissonance.

Cognitive immunization may also be overcome given a large enough discrepancy between expected and actual outcomes, also known as prediction error (Corlett, 2018). Studies suggest that the size of prediction errors directly predicts belief update (Vlasceanu et al., 2021a): the higher the predicted error, the greater the probability of updating beliefs, and therefore, the lower the chances of cognitive immunization. For instance, participants who were asked to guess how much global temperatures would increase and estimated temperatures much higher or lower than actual projected temperatures were more likely to update their beliefs (Kube, 2023). Interestingly, belief update still happened among those engaging in cognitive immunization given large prediction errors. This is in line with Bayesian models of belief update, which describe rational agents who update their "prior" beliefs based on new evidence, leading to a revised set of "posterior" beliefs. In the context of political news, studies have found that individuals with high analytical thinking skills are better at incorporating new information and accurately updating their beliefs (Tappin et al., 2021). Thus, when taken as ground truth, fact-checks can function as new evidence that helps individuals update their beliefs in the face of misinformation.

Belief polarization poses an enigma in light of the large body of evidence supporting people's adherence to the rules of Bayesian inference. While Bayesian reasoning involves updating prior beliefs in line with new evidence, belief polarization occurs when people who receive the same information update their beliefs in opposing directions (Cook & Lewandowsky, 2016). This phenomenon is considered "irrational" because it involves contrary updating, which appears to violate optimal responding as dictated by Bayes' theorem. Attempts at explaining belief polarization from a rational perspective propose more elaborate Bayesian models known as Bayesian networks. These models incorporate the notion that people's interpretation of new evidence is moderated by additional belief variables that are part of the general prior belief (Cook & Lewandowsky, 2016). For instance, two individuals with different prior beliefs about climate change may interpret the same scientific data in ways that strengthen their respective views because their interpretation of the evidence is moderated by additional belief variables, such as their trust in science and their political ideology. Thus, networks that incorporate a variety of beliefs could help disrupt belief polarization by offering a range of new ways of interpreting evidence.

Can this situation be extrapolated to political beliefs and misinformation behaviors? Studies suggest that political tolerance increases by exposing people to a range of diverse ideas, and to reasons supporting opposing ideas (Mutz, 2002). On social media, users who are part of politically and socially heterogeneous networks have been found to engage more in political expression, including following or liking more politicians on social media, and sharing more political messages (Barnidge et al., 2018). A greater acceptance of contrary viewpoints is possible when people belong to a diverse enough online

community. Heterogeneous compared to homogenous settings have been found to enhance truth discernment among social media users, contributing to belief revision against their partisan ideology (Espina Mairal et al., 2024). In this sense, heterogeneous settings would create an opportunity for individuals to update their beliefs due to greater political tolerance.

Crowdsourcing interventions can disrupt echo chambers by increasing network heterogeneity and nudging users toward alternative viewpoints. This could be achieved by gathering accuracy judgments from heterogeneous sources, which are most effective in bypassing partisan biases and fact-checking political speech as previously mentioned (Espina Mairal et al., 2024). However, because cognitive dissonance seems to facilitate and hinder belief update depending on multiple factors, attention should be paid to the amount of dissonance generated by crowdsourcing interventions. Some of these factors include the degree of belief polarization, the magnitude of prediction errors, the risk for cognitive immunization, and individual traits such as analytic thinking ability. We suggest that, at least in some cases, excessively dissonant opinions can put a person at risk for cognitive immunization, while not enough dissonance can foster the reinforcement of false beliefs shared by ideological communities. Future research should consider different levels of cognitive dissonance to understand if they contribute differently to belief updating across diverse political, cultural, and social media contexts. This research could be used to fine-tune the composition of the fact-checking crowd in a way that promotes an optimal level of cognitive dissonance.

To conclude, cognitive dissonance is a key element to consider when scaling crowdsourcing interventions to combat misinformation. Different implementation strategies may trigger different levels of cognitive dissonance among different online communities, which may impact their ability to promote belief update. For that, crowdsourcing interventions that aim to reduce echo chamber effects need to consider the political and ideological preferences of social media users and the online communities they are embedded in.

3.1.2 Trust in Fact-Checking Sources

Similarly to cognitive dissonance, trusting the source of accuracy judgments is key for crowdsourcing interventions to be successful, especially in polarized contexts (Pretus et al., 2024). According to most persuasion theories, it is easier to trust information given by highly credible sources than by less credible ones (Pornpitakpan, 2004). Source credibility dimensions include *expertise* and *trustworthiness*. *Expertise* is the perceived capability of a communicator to make correct assertions, while *trustworthiness* is the degree to which an audience perceives the assertions made by a communicator as the most valid one possible (Hovland et al., 1953). Source trustworthiness is more effective than expertise in diminishing people's reliance on misinformation, as evidenced by Pluviano et al. (2020). In this study, COVID-19 vaccine skeptics increased their intentions to take the vaccine to a greater extent after receiving information from a trusted source shared by an ordinary person than after receiving information from a questionable source shared by a healthcare professional. Although this study relied on a sample of undergraduate students and may not generalize across other contexts, these results suggest that trust in fact-checking sources is a critical element to consider when scaling crowdsourcing interventions.

People seem to be highly sensitive to *who* provides corrections to misinformation. One of the most relevant aspects of a fact-checker people consider is their group identity (Reinero et al., 2023). Fact-checks are more effective in reducing misinformation sharing when they are generated by people in the same identity group (Reinero et al., 2023), especially if they are perceived as highly credible (Liu et al., 2023). For

instance, Pretus et al. (2024) found a 25% reduction in participants' likelihood of sharing inaccurate information when a *Misleading* count was added next to the *Like* count. This strategy was more effective when the *Misleading* count reflected in-group norms, that is when collective accuracy judgments came from people in the same identity group. This may be due to people being more likely to believe information that comes from the ingroup, regardless of its veracity (Pereira et al., 2023; Swire et al., 2017), while out-group information is most often followed by negative comments (Wojcieszak et al., 2021). Thus, ingroup sources may be more effective in eliciting trust in crowdsourcing interventions, especially when the identity of these sources is made available to social media users.

Knowing that corrections to in-group misinformation have been predominantly generated by outgroup sources could undermine fact-checking efforts. Studies suggest that correction strategies that come from outgroup members can have a backfire effect (Nyhan & Reifler, 2010; Reinero et al., 2023). For instance, individuals believe misinformation even more after witnessing outgroup corrections (Reinero et al., 2023). Yet, backfire effects are infrequent in practice (Nyhan, 2021) and other studies have found no evidence of them in response to corrective information (Ecker et al., 2023; Wood & Porter, 2019). Even in the absence of backfire effects, people are less likely to believe information that originates from groups outside their own (Pereira et al., 2023). This may be due to partisans perceiving outgroup information as a threat to their identity (Pereira et al., 2023) and experiencing high levels of dissonance about values and essential norms they do not share. This dissonance can lead to distrust and negative evaluations of outgroup members (Chambers & Melnyk, 2006). As a consequence, individuals may ignore or attribute less credibility to fact-checks from sources outside their group. Crowdsourcing interventions could help increase trust in fact-checks by offering corrections from a broad and heterogeneous group of fact-checkers who do not share a single identity.

The relationship between group membership and trust in sources is exacerbated in polarized contexts. The effectiveness of most debunking misinformation strategies has been found to decrease when the topic is politically polarized (Chan & Albarracín, 2023). This may be caused by the fact that polarized contexts offer greater incentives to conform to ingroup norms and distance oneself from outgroup norms (Pretus et al., 2024). This situation creates a prone environment for ingroup political elites to influence the public's opinion and makes in-party elite's misinformation even more compelling to the public (Orchinik & Rand, 2023), regardless of its content (Traberg et al., 2024). Thus, how to label fact-checking sources in crowdsourcing interventions becomes an especially relevant question in politically polarized contexts such as the U.S.

The issue of trust in fact-checking sources affects all types of fact-checking efforts, from fact-checking agencies to social media platforms that issue their own fact-checks. Fostering trust has become increasingly challenging due to a global “erosion of trust” in governmental institutions and social media companies over the last decade (Prange-Gstöhl, 2016). This trend has been intensified by the development of social media itself, which has facilitated the spread of online misinformation and conspiracy theories, thereby increasing polarization and partisan hostility (Cinelli et al., 2020; Hassan & McCrickard, 2019; Pham et al., 2020). This scenario has emerged from a social media incentive structure that rewards likability and virality over accuracy and trustworthiness (Globig et al., 2023). Crowdsourcing interventions can shift this incentive structure by promoting fact-checking as a desirable behavior, which everyone is partly responsible for. Studies suggest that people are willing to accept this level of responsibility, especially when they are aware

that their contributions may influence public discourse (Allen et al., 2021). Crowdsourcing interventions may thus help foster trust in fact-checks by encouraging people to participate in its generation.

3.1.3 Crowd Size

Finally, the number of people who participate in fact-checking a given piece of information can influence how much people believe crowdsourced accuracy judgments. Fact-checks from larger crowds may be interpreted as more accurate and trustworthy than fact-checks from smaller crowds. For instance, people are less willing to share misinformation with every additional person participating in collective accuracy judgments (Pretus et al., 2024), highlighting the impact of the absolute size of the fact-checking crowd. The effect of additional fact-checkers was independent of the proportion of negative compared to positive feedback and the virality of the post (posts receiving thousands compared to dozens of *Likes*). Despite this, the relative amount of people who fact-check a post compared to the number of people who endorse a social media post is also important. The same study revealed that individuals were less likely to share posts that were flagged as misleading by a crowd that was 80% (compared to 20%) of the size of the crowd who had *liked* them (Pretus et al., 2024). Positive social media engagement metrics such as *Likes* can counteract the effect of collective accuracy judgments. Therefore, both the absolute and the relative crowd size of fact-checkers are important for successful crowdsourcing interventions.

Aside from being more influential, larger crowds can generate more accurate collective estimates, as explained above (see *When do crowds generate accurate judgments?*). Specifically, crowds of at least 3 or 4 individuals have been found to significantly boost the accuracy of collective judgments (Espina Mairal et al., 2024), and commonly used crowdsourcing approaches involve at least 10 fact-checkers (La Barbera et al., 2020; Roitero et al., 2018; Roitero et al., 2023). Thus, crowd size is not only important to generate persuasive fact-checks but also to produce truly reliable ones.

Crowd size can be heavily impacted by partisan dynamics and social media network structures. In-group members may be less willing to fact-check in-group sources and more willing to fact-check out-group sources, regardless of how deserving of fact-checks they are. This can lead to reduced fact-checking within ideological communities and increased fact-checking between ideological communities. As a result, more aggressive communities with a greater ability to mobilize large crowds could generate more convincing fact-checks - backed by a larger number of fact-checkers - against political opponents. This would allow larger and more active communities to dominate online discourse by deliberately canceling factual information they perceive as threatening. This possibility should be taken into account when designing crowdsourcing interventions, which should guarantee that fact-checks do not come from within a single homogenous community.

Overall, cognitive dissonance, trust in fact-checking sources, and crowd size are critical factors that influence belief update in the context of misinformation. Effective interventions against misinformation should provide corrections that are incongruent enough with prior beliefs to facilitate belief update, without inducing cognitive immunization. Moreover, to foster acceptance of this new information, collective accuracy judgments should come from trusted sources in different communities. Cognitive dissonance and trust are often at odds in polarizing contexts. While ingroup members may inspire trust but offer few dissonant judgments, outgroup members may offer dissonant judgments but inspire little trust. This results in smaller and less effective crowds willing to fact-check in-group content within communities. Crowdsourcing interventions should rely on crowds that are large and diverse enough to

produce a dissonance level that motivates individuals to revise their convictions while remaining credible and trustworthy. In the next section, we discuss different ways to balance out these elements when implementing crowdsourcing strategies.

4. IMPLEMENTATION

Throughout this review, we have presented three factors that are important for successful crowdsourcing interventions against misinformation: cognitive dissonance with previous beliefs, trust in fact-checking sources, and crowd size. In this section, we discuss how these factors could be optimized when implementing crowdsourcing interventions to effectively reduce belief in misinformation. We propose that a key aspect to consider when addressing this question is the placement of individuals in conversational networks and how information spreads inside network structures.

In the present work, we define conversational networks as dynamic networks where nodes are social media users and connections between nodes represent followership, that is, a scenario where at least one of the users follows the other. The strength of these connections is influenced by the amount of interaction between users via likes, comments, and reposts. This definition is based on the fact that, despite considerable differences between social media platforms in terms of their algorithm, design, and users, most platforms create interactions based on followership (e.g., TikTok, Twitter/X, Instagram, Snapchat, Youtube). While this definition is useful for the current theoretical approach, it should be regarded with caution as differences between platforms might influence how such a network may be built. Additionally, while real world conversational networks may continuously change as social media users dynamically choose to “follow” and “unfollow” other accounts, the same networks may be studied as static entities in experimental research, especially in cross-sectional experimental designs that disregard temporal dynamics. Our proposed implementations should thus be considered in light of the dynamic nature of conversational networks in real world social media environments.

The formation of collective memories, and by extension beliefs, is influenced by the conversational networks people are embedded in as well as the individual-level sociocognitive processes that are sparked during conversational recall (Coman et al., 2016; Vlasceanu et al., 2021b, 2023). Conversations within communities cause people to synchronize their beliefs with people close to them in the conversational network. The more closely connected an individual is to another person inside the network, the more similar their beliefs become (Vlasceanu et al., 2021b; Wilkes-Gibbs & Clark, 1992). Connection between any two individuals is defined by direct interaction (conversation) with each other. Therefore, people close in the network converse directly with each other, whereas people further away from each other are only indirectly connected through conversations with a third, fourth, or fifth (...) person. As a result of these interactions, people's beliefs become more synchronized and coordinated (Vlasceanu et al., 2021b; Wilkes-Gibbs & Clark, 1992). Echo chambers could consequently be considered an example of a conversational network whose users form a tightly-knit community with highly synchronized beliefs.

To break out of the echo chambers and belief synchronization across networks, people need to come into contact with individuals who fulfill the features mentioned in this review; individuals who create reasonable dissonance with prior beliefs, who can be trusted, and who represent a sufficient crowd size to be convincing. Furthermore, in the interest of transparency, social media users should be able to know who participated in fact-checking. Thus, fact-checkers should be similar enough to the receivers of those fact-checks that they can be trusted, yet sufficiently different to create dissonance and influence people’s beliefs. We propose two

types of solutions to achieve these results.

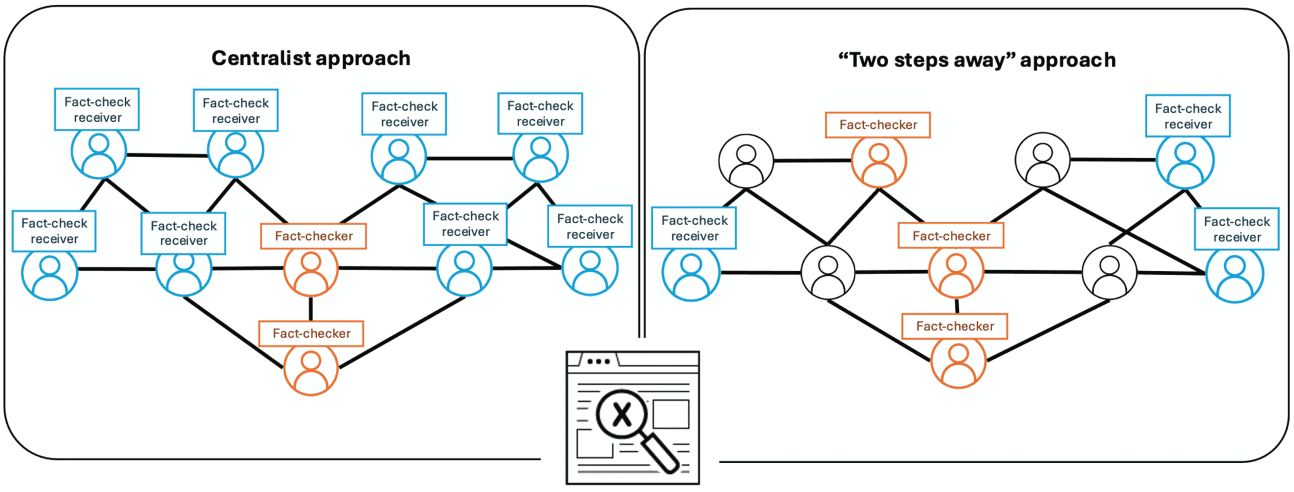
4.1 The Centralist Approach

One approach entails selecting fact-checkers based on their absolute location within a network. The idea would be to allow everyone to fact-check but only display fact-checks from individuals who are connected to more than one online community and occupy more central locations in their networks (see Table 1 and Figure 2, left panel). These individuals are termed "Centrals" and can leverage beliefs from several groups due to their various group memberships. Because of how social norms evolve through social networks, central individuals can be more influential in shifting norms and correcting misconceptions (Gelfand et al., 2024). Thus, highly interconnected individuals may significantly affect the community as a whole. Central individuals may be acknowledged and trusted by several different ingroup members, yet because they also belong to other groups, they could also bring additional or opposing beliefs to the table. Thus, receiving fact-checks from Centrals could make less interconnected ingroup members more reflective and create the needed dissonance to break through echo chambers.

The centralist approach comes with several aspects to consider. First, from an ethical perspective, allowing only select individuals to influence others in the network could undermine the principle of fairness, since others in the network would not have the same opportunities to fact-check. In this sense, being able to fact-check could serve as an incentive for people to connect with several distinct communities, fostering network heterogeneity. Related to this, incentives to become a Central may be stronger for actors that aim to use fact-checking illegitimately as a tool to achieve greater influence and undermine specific messages. How these actors may be distinguished from natural centrals remains to be elucidated. Secondly, when targeting the most central people in a network, their held beliefs will spread to their various connections for better or for worse. Thus, if a given Central is highly partisan or prone to misinformation, this too will spread

Figure 2

Implementation Strategies for Crowdsourcing Interventions



Note. In the centralist approach (left panel), centrally located individuals within a network who hold multiple group memberships function as fact-checkers for the rest of the users in a network (absolute location). In the "Two steps away" approach (right panel), individuals receive fact-checks from users outside of their immediate ideological

across the network and reach numerous other people. Conversely, if a given central is politically moderate, they could be uninterested in politics altogether and be less motivated to fact-check partisan posts in the first place, reducing the overall amount of fact-checking taking place in the network. The question of whether individuals who are central in social media networks are more or less partisan, especially in the light of their multiple group memberships, remains to be addressed.

4.2 The "Two Steps Away" Approach

A second solution involves selecting fact-checkers based on their relative location to the individuals receiving their corrections. That is, allowing everyone to fact-check misinformation but displaying corrections from "fact-checkers who are two steps away from you" (see Table 1 and Figure 2, right panel). Of note, the label "two steps away" has been chosen to represent fact-checkers who are not part of the

Table 1

Hypotheses to be tested and limitations of each of the proposed implementations

Strategy	Hypotheses to be tested	Limitations
Centralist Approach	Centrals with diverse group memberships are less partisan Centrals are more willing to provide dissonant fact-checks Centrals are trusted by members of different groups A crowd of Centrals is heterogenous, independent, and large enough to generate accurate judgments	Fewer users are able to fact-check Interested parties could become Centrals and misuse their ability to fact-check
Two Steps Away Approach	People outside of one's immediate ideological bubble are more willing to provide dissonant fact-checks People outside of one's immediate ideological bubble are perceived as more trustworthy than people far out of one's bubble A crowd of people outside of one's immediate ideological bubble is heterogenous, independent, and large enough to generate accurate judgments	Less central users receive fewer (but more trusted) fact-checks Need to establish the optimal distance in steps between fact-checkers and fact-check receivers

user's immediate informational environment but also not too far from it. The optimal distance in steps between fact-checkers and fact-check receivers, either two or more steps, remains to be elucidated. This solution rests on the assumption that extreme partisans, who are more prone to misinformation (Guess et al., 2019), are located in more isolated and tight-knit online communities and connect to other more moderate groups only indirectly through distant connections. The idea would be to rely on distant connections (at least two steps away) with more moderate groups of center-right and center-left individuals, respectively, to provide fact-checks. In the US center-left and center-right individuals have been found to be less susceptible to misinformation than their more extreme counterparts (Guess et al., 2019). Additionally, a study across 45 countries on 6 continents found center-left and center-right voters to be more knowledgeable about politics than people in the middle and at the extremes of the political spectrum (De keersmaecker et al., 2024). Thus, center-left and center-right individuals could help provide more accurate fact-checks to isolated parts of the network that may be hard to reach through centrals. At the same time, extreme partisans may be more receptive to fact-checks from individuals who display more moderate forms of their own ideology (center-left and center-right, respectively) as compared to individuals with opposing ideologies. Thus, using collective accuracy ratings from people who are "two steps away" may help create enough dissonance with previous beliefs from a large enough crowd, while preserving a certain level of trust.

Using either the centralist or the "two steps away" approach comes with several limitations. Firstly, fact-checkers in either approach could misuse their ability to fact-check content to promote messages they agree with, for instance, by giving them a high credibility score. This would allow partisans to enhance the persuasive impact of ingroup-consistent information. One way to overcome this would be to use fact-checking exclusively as a dissuasive tool rather than a persuasive tool. This could be achieved by allowing fact-checkers to "downvote" content by marking it as "misleading" but not "upvote" content by marking it as "credible". This strategy may come with its own limitations since crowds could similarly organize to "downvote" accurate claims. To counteract both possibilities, our proposed strategies attempt to diversify the composition of fact-checking crowds for any given post, making it more difficult for people to organize and collectively downvote posts they disagree with.

Another open question is whether these solutions would apply to different cultural contexts, where political identities may not be organized along left-right political divides. Because we focus on network structure rather than political content, our framework is flexible enough to accommodate diverse contexts as long as isolated individuals with absolutist stances receive fact-checks from more central individuals who we assume would have less ideologically congruent attitudes. For instance, it would be useful in contexts where extreme tight-knit communities are formed around alternative political cleavages such as separatism *versus* anti-separatism, or support for *versus* resistance to autocratic regimes and specific political figures. These assumptions are generally met in polarized political contexts, where attitudes are sorted based on group membership (Mason, 2016). However, in non-polarized contexts as well as in the case of non-partisan misinformation, attitudes are not necessarily distributed in line with group membership. These scenarios pose fewer challenges when it comes to mistrust against defined outgroups and reluctance to fact-check ingroup-coherent misinformation. For instance, individuals in less polarized contexts such as the UK are equally responsive to crowdsourced accuracy judgments from the ingroup compared to anonymous social media users (Pretus et al., 2024). Thus, fact-checking in these scenarios could be done by

individuals both inside or outside ideological bubbles.

Of note, we have intentionally avoided labeling certain users as those who "hold accurate beliefs" and using them as fact-checkers. While some strategies are available to identify high-quality fact-checkers (see *When do crowds generate accurate judgments?*), their judgments are ultimately validated by comparing them to expert assessments issued by professional fact-checking agencies. While expert judgments can be a reliable source of knowledge, this would entail that social media companies act as arbiters of truth based on the assessments of select agencies. In this work, we propose that accurate beliefs may arise from exposure to different points of view, and thus can be approximated by selecting a diverse pool of fact-checkers or by choosing fact-checkers who are exposed to multiple perspectives. We have proposed two implementations that we believe would generate accurate corrections based on the judgments of heterogeneus, independent, and large enough crowds, three necessary criteria to achieve accuracy (see *When do crowds generate accurate judgments?*). However, the fact that these implementations would successfully result in crowds that fulfill these three criteria needs to be validated in the light of empirical evidence (see Table 1).

In this section, we have discussed two potential solutions to optimize the efficacy of crowdsourcing interventions that involve trade-offs between cognitive dissonance, trust in fact-checking sources, and crowd size. Both approaches rely on several assumptions that need to be tested before any implementation takes place. We next discuss future avenues of research that could help clarify whether these assumptions hold in real-world online environments.

5. FUTURE DIRECTIONS

To maximize the effectiveness of crowdsourcing interventions and minimize the risk of coordinated attacks against given political messages, we propose limiting whose fact-checks are visible to each user based on the relative or absolute location of the fact-checker in social networks. Ensuring a diverse set of fact-checkers necessarily requires leaving some of them out, especially those who could have a greater conflict of interest. This work is an attempt to define how to do this filtering in a way that is systematic and transparent. We believe our proposed implementations can generate sufficient cognitive dissonance based on a large enough crowd, all the while preserving some level of trust in sources and algorithm transparency. This new approach comes with several open questions that should be addressed.

Firstly, while we provide evidence that trust in sources, cognitive dissonance, and crowd size can impact belief updating, several core assumptions in our model remain to be tested. One of these assumptions is the potential non-linear effect of dissonance on belief updating: the idea that increasing levels of dissonance have a positive effect on belief updating, while at least in some cases too much dissonance will prompt cognitive immunization. Future research should assess whether this pattern emerges in at least certain conditions. Another assumption is that choosing fact-checkers in central locations of the network or those who are "two steps away" from any given user would generate a diverse enough group of fact-checkers able to provide accurate judgments on any given issue. Testing these two assumptions is a critical first step to validate our proposed model and implementations.

Secondly, both proposed solutions rest on the assumption that there is an association between absolute network location and political orientation or at least political extremity. Studies suggest partisan echo chambers foster polarization (Hobolt et al., 2023). Thus, one might expect that extreme partisans are embedded in more isolated and tight-knit communities

while moderates occupy more central locations and are connected to a greater number of different communities. Political orientation has proven to be a valuable predictor of many outcomes related to misinformation, such as exposure and engagement with fake news (Guess et al., 2019) and resistance to fact-checking (Rathje et al., 2022). However, other work suggests that the traditional left-right spectrum falls short when it comes to capturing important aspects of political behavior (Costello et al., 2023). Future research should assess where in the network different groups with varying degrees of political extremity are located, and how connected they are to neighboring communities. This may vary in different social media platforms (e.g., Facebook *versus* Twitter/X) and across cultural groups. Thus, an in-depth analysis of the ideological composition of any given network should precede any attempt at implementing the proposed solutions.

When it comes to crowd size, one question that remains is what percentage of individuals across different ideological communities are willing to fact-check misinformation that is ideologically congruent with their own and neighboring communities. Extreme partisans may be less willing to fact-check ingroup sources, while moderates may be less interested in politics and thus less likely to fact-check partisan misinformation out of indifference. This would cause center-right and center-left individuals to receive fewer fact-checks from neighboring communities of more center-leaning moderates on one side and, respectively, far-right and far-left partisans on the other side. This may not be particularly detrimental, because center-left and center-right voters are less susceptible to misinformation (Guess et al., 2019) and more politically knowledgeable than people at the extremes and center of the political spectrum (De keersmaecker et al., 2024). Thus, these groups may benefit less from fact-checking than more extreme groups. Knowing how much each group is willing to fact-check different political sources and how responsive it is to different crowd sizes would help fine-tune whose ratings should users have access to.

Additionally, because social media users should be able to know whose fact-checks they are exposed to, how to label fact-checkers becomes a relevant issue. One question is whether individuals at the far right of the political spectrum, who could benefit the most from fact-checking (Guess et al., 2019), trust individuals who are ideologically close, but in different ideological communities, such as center-right voters. Studies suggest that, while individuals are attracted by similarly minded others (Byrne & Nelson, 1965), they prefer people with ideologically consistent opinions and high levels of certainty over these opinions (Zimmerman et al., 2022), two traits that characterize political extremists (Zmigrod et al., 2019). These results leave little room for moderates with ambiguous or ambivalent political attitudes to be liked, even among other moderates (Zimmerman et al., 2022). Future studies should thus assess whether labeling fact-checkers as “center-right” or “center-left” has detrimental effects on people's tendency to update their beliefs, especially among neighboring communities of extreme partisans. In this case, alternative labels that refer to network structure rather than political orientation, such as “fact-checkers who are 2 steps away from you”, would be recommended.

Finally, another venue for future research when it comes to increasing the heterogeneity of fact-checking crowds would be to use machine learning algorithms to assess the political leanings of different individuals in the crowd. These assessments could be used to adjust the weight of each contributor's input to crowdsourced judgments, resulting in collective estimates balanced for political affiliation. Moreover, in the interest of transparency, the extent of the contributions from fact-checkers of different political leanings could be communicated to users consuming crowd-checked content. While this approach might be

interesting to pursue in the future, it also has its own set of concerns, especially data privacy considerations and the reliability of political affiliation estimates using machine learning.

Elucidating these questions will be key to confirming the suitability of network-based approaches to define whose ratings people are exposed to. This would allow using systematic criteria to regulate fact-checking that apply to everyone in an ethical manner, without discrimination, and in a way that is transparent and accessible to social media users. This approach could be especially interesting to open-source social media platforms such as Mastodon or Bluesky Social, which are already aligned with these principles.

6. CONCLUSION

We have argued that crowdsourcing interventions will be most effective if they convey dissonant feedback from a large enough crowd that people trust. This approach will be most beneficial for extreme partisans who share most misinformation but also mistrust traditional fact-checking sources. We believe this can be achieved by allowing everyone to fact-check, but making these ratings visible to users in a way that precludes individuals and organizations from abusing this capability to overturn political opponents. We have proposed two ways to find a balance between these three elements based on the user's relative and absolute locations in social media networks while preserving transparency and proposed future research to address remaining questions. Current attempts to employ crowdsourcing approaches to fact-checking, such as “Community Notes” on Twitter/X, often arrive too late. By allowing users to fact-check content in an aggregate manner, social media companies could provide an extra layer of content moderation by allowing people to share this responsibility while offering a first-line response to misinformation.

7. CONFLICTS OF INTEREST

The authors declare no competing interests.

8. ACKNOWLEDGEMENTS

H.G. and R.H. were supported by a European Union grant (grant no. 101070930). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the granting authority can be held responsible for them. D.L. was supported by a BIAL Foundation grant (Ref. 133/2022).

REFERENCES

- Acemoglu, D., Ozdaglar, A., & Siderius, J. (2021). *A model of online misinformation*. https://doi.org/10.3386/w28884
- Al Marrar, M., & Allevato, E. (2022). Cognitive dissonance: Affecting party orientation and selective recall of political information. *Athens Journal of Social Sciences*, *9*(2). https://doi.org/10.30958/ajss.9-2-2
- Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36). https://doi.org/10.1126/SCIADV.ABF4393
- Annu & Dhanda, B. (2020). Cognitive dissonance, attitude change and ways to reduce cognitive dissonance: A review study. *Journal of Education, Society and Behavioural Science*, *1*, 48–54. https://doi.org/10.9734/JESBS/2020/V33I630236
- Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513. https://doi.org/10.1038/S41562-023-01641-6

- Au, C. H., Ho, K. K. W., & Chiu, D. K. W. (2021). The role of online misinformation and fake news in ideological polarization: Barriers, catalysts, and implications. *Information Systems Frontiers 2021*, 1–24. https://doi.org/10.1007/S10796-021-10133-9
- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9216–9221. https://doi.org/10.1073/PNAS.1804840115
- Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, *6*(10), 1372–1380. https://doi.org/10.1038/s41562-022-01388-6
- Barnidge, M., Huber, B., de Zúñiga, H. G., & Liu, J. H. (2018). Social media as a sphere for "risky" political expression: A twenty-country multilevel comparative analysis. *The International Journal of Press/Politics 23*(2), 161–182. https://doi.org/10.1177/1940161218773838
- Byrne, D., & Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology*, *1*(6), 659-663. https://doi.org/10.1037/h0022073
- Chambers, J. R., & Melnyk, D. (2006). Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Personality and Social Psychology Bulletin*, *32*(10), 1295-1311. https://doi.org/10.1177/0146167206289979
- Chan, M. P. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour, 7*(9), 1514-1525. https://doi.org/10.1038/S41562-023-01623-8
- Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science, 28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55* (Volume 55, 2004), 591–621. https://doi.org/10.1146/ANNUREV.PSYCH.55.090902.142015
- Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5
- Cognitive Dissonance: A History in Tweets. (2011). *Perspectives on Psychological Science*, *6*(1), 98-101. https://doi.org/10.1177/1745691610393526
- Coman, A., Momennejad, I., Drach, R. D., & Geana, A. (2016). Mnemonic convergence in social networks: The emergent properties of cognition at a collective level. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8171–8176. https://doi.org/10.1073/PNAS.1525569113
- Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science, 8*(1), 160–179. https://doi.org/10.1111/TOPS.12186
- Cooper, J. (2012). Cognitive dissonance theory. In *Handbook of Theories of Social Psychology: Volume 1* (Vol. 1, pp. 377-397). SAGE Publications Ltd, https://doi.org/10.4135/9781446249215
- Corlett, P. (2018). Delusions and prediction error. *Delusions in Context*, 35–66. https://doi.org/10.1007/978-3-319-97202-2_2
- Coscia, M., & Rossi, L. (2020). Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, *17*(167). https://doi.org/10.1098/RSIF.2020.0020
- Costello, T. H., Zmigrod, L., & Tasimi, A. (2023). Thinking outside the ballot box. *Trends in*

Cognitive Sciences, *27*(7), 605–615. https://doi.org/10.1016/J.TICS.2023.03.012

- De keersmaecker, J., Schmid, K., Sibley, C. G., & Osborne, D. (2024). The association between political orientation and political knowledge in 45 nations. *Scientific Reports 2024*, *14*(1), 1–10. https://doi.org/10.1038/s41598-024-53114-z
- DeVerna, M. R., Guess, A. M., Berinsky, A. J., Tucker, J. A., & Jost, J. T. (2022). Rumors in retweet: Ideological asymmetry in the failure to correct misinformation. *Personality and Social Psychology Bulletin*, *50*(1), 3–17. https://doi.org/10.1177/01461672221114222
- Drolsbach, C. P., & Pröllochs, N. (2023). Diffusion of community fact-checked misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–22. https://doi.org/10.1145/3610058
- Ecker, U. K. H., Sharkey, C. X. M., & Swire-Thompson, B. (2023). Correcting vaccine misinformation: A failure to replicate familiarity or fear-driven backfire effects. *PLOS ONE*, *18*(4), e0281140. https://doi.org/10.1371/JOURNAL.PONE.0281140
- Espina Mairal, S., Bustos, F., Solovey, G., & Navajas, J. (2024). Interactive crowdsourcing to fact-check politicians. *Journal of Experimental Psychology: Applied, 30*(1), 3–15. https://doi.org/10.1037/xap0000492
- Festinger, L. (1957). *A Theory of Cognitive Dissonance* [Book]. Stanford University Press.
- Fontanari, J. F., Bonniot-Cabanac, M. C., Cabanac, M., & Perlovsky, L. I. (2012). A structural model of emotions of cognitive dissonances. *Neural Networks*, *32*, 57–64. https://doi.org/10.1016/J.NEUNET.2012.04.007
- Fu, F., Nowak, M. A., Christakis, N. A., & Fowler, J. H. (2012). The evolution of homophily. *Scientific Reports 2012, 2*(1), 1–6. https://doi.org/10.1038/srep00845
- Galton, F. (1907). Vox populi. *Nature, 75,* 450–451.
- Gelfand, M. J., Gavrilets, S., & Nunn, N. (2024). Norm dynamics: interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, *75*, 341–378. https://doi.org/10.1146/ANNUREV-PSYCH-033020-013319
- Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, *38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389
- Globig, L. K., Holtz, N., & Sharot, T. (2023). Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife*, *12*, e85767. https://doi.org/10.7554/ELIFE.85767
- Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). https://doi.org/10.1126/SCIADV.AAU4586
- Hahn, U., Von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science, 11,* 194–206. https://doi.org/10.1111/tops.12401
- Hassan, T., & McCrickard S. (2019). Trust and trustworthiness in social recommender systems. In *Companion proceedings of the 2019 world wide web conference*. 529-532. https://doi.org/10.1145/3308560.3317596
- Hobolt, S. B., Lawall, K., & Tilley, J. (2023). The polarizing effect of partisan echo chambers. *American Political Science Review*, 1–16. https://doi.org/10.1017/S0003055423001211
- Hong, H., Ye, Q., Du, Q., Wang, G. A., & Fan, W. (2019). Crowd characteristics and crowd wisdom: Evidence from an online investment community. *Journal of the Association for Information Science and Technology*, *71*(4), 423–435. https://doi.org/10.1002/ASI.24255
- Hovland, C., Janis, I., & Kelley, H. (1953). *Communication and persuasion.* New Haven, GT: Yale University Press.
- Jiang, B., Karami, M., Cheng, L., Black, T., & Liu,

H. (2021). *Mechanisms and Attributes of Echo Chambers in Social Media*. https://doi.org/10.48550/arXiv.2106.05401

- Kaaronen, R. O. (2018). A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance. *Frontiers in Psychology*, *9*, 408127. https://doi.org/10.3389/FPSYG.2018.02218
- Kube, T. (2023). Factors influencing the update of beliefs regarding controversial political issues. *The Journal of Social Psychology*. https://doi.org/10.1080/00224545.2023.2253981
- La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., & Spina, D. (2020). Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. *Advances in Information Retrieval*, *12036*, 207. https://doi.org/10.1007/978-3-030-45442-5_26
- Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 227–242). Psychology Press.
- Lee, S., Xiong, A., Seo, H., & Lee, D. (2023). "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review, 4*(5). https://doi.org/10.37016/MR-2020-126
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, Supplement*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018
- Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the fact-checkers: The role of source type, perceived credibility, and Individual differences in fact-checking effectiveness. *Communication Research*. https://doi.org/10.1177/00936502231206419
- Marie, A., Altay, S., & Strickland, B. (2023). Moralization and extremism robustly amplify myside sharing. *PNAS Nexus*, *2*(4). https://doi.org/10.1093/PNASNEXUS/PGAD078
- Martel, C., Rathje, S., Clark, C. J., Pennycook, G., Van Bavel, J. J., Rand, D. G., & van der Linden, S. (2024). On the efficacy of accuracy prompts across partisan lines: An adversarial collaboration. *Psychological Science*, *35*(4), 435–450. https://doi.org/10.1177/09567976241232905
- Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, *80*(S1), 351–377. https://doi.org/10.1093/POQ/NFW001
- McDonald, R. M., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147–151. https://doi.org/10.1016/J.COBEHA.2015.04.006
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(Volume 27, 2001), 415–444. https://doi.org/10.1146/ANNUREV.SOC.27.1.415
- Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, *96*(1), 111–126. https://doi.org/10.1017/S0003055402004264
- Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, *118*(15), e1912440117. https://doi.org/10.1073/PNAS.1912440117
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/S11109-010-9112-2
- Orchinik, R., & Rand, D. G. (2023, October 29). Pro-climate statements from Elon Musk can persuade republicans on climate change. https://doi.org/10.31234/osf.io/v9mzk
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(7), 2521–2526. https://doi.org/10.1073/PNAS.1806781116
- Pereira, A., Harris, E., & Van Bavel, J. J. (2023).

Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes and Intergroup Relations*, *26*(1), 24–47. https://doi.org/10.1177/13684302211030004

- Pham, D. V., Nguyen, G. L., Nguyen, T. N., Pham, C. V., & Nguyen, A. V. (2020). Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access*, *8*, 78879–78889. https://doi.org/10.1109/ACCESS.2020.2989140
- Pluviano, S., Della Sala, S., & Watt, C. (2020). The effects of source expertise and trustworthiness on recollection: the case of vaccine misinformation. *Cognitive Processing*, *21*(3), 321–330. https://doi.org/10.1007/S10339-020-00974-8
- Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281. https://doi.org/10.1111/J.1559-1816.2004.TB02547.X
- Prange-Gstöhl, H. (2016). Eroding societal trust: a game-changer for EU policies and institutions? *Innovation: The European Journal of Social Science Research*, *29*(4), 373–390. https://doi.org/10.1080/13511610.2016.1166038
- Pretus, C., Javeed, A. M., Hughes, D., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2024). The Misleading count: an identity-based intervention to counter partisan misinformation sharing. *Philosophical Transactions of the Royal Society B*, *379*(1897). https://doi.org/10.1098/RSTB.2023.0040
- Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation and resistance to fact-checking. *Journal of Experimental Psychology: General*, *152*(11), 3116–3134. https://doi.org/10.1037/xge0001436
- Rathje, S., Roozenbeek, J., Steenbuch Traberg, C., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the Editors of Psychological Science: Meta-Analysis Reveals that Accuracy Nudges Have Little to No Effect for U.S. Conservatives: Regarding Pennycook et al. (2020). *Psychological Science*. https://doi.org/10.25384/SAGE.12594110.V2
- Rathje, S., van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(26), e2024292118. https://doi.org/10.1073/PNAS.2024292118
- Reinero, D. A., Harris, E. A., Rathje, S., Duke, A., & Van Bavel, J. J. (2023, May 11). Partisans are more likely to entrench their beliefs in misinformation when political outgroup members correct claims. https://doi.org/10.31234/osf.io/z4df3
- Renault, T., Restrepo-Amariles, D., & Troussel, A. (2024). Collaboratively adding context to social media posts reduces the sharing of false news. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4800565
- Roitero, K., Demartini, G., Mizzaro, S., & Spina, D. (2018). How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. *CIKM Workshops on Rumours and Deception in Social Media.*
- Roitero, K., Soprano, M., Portelli, B., De Luise, M., Spina, D., Mea, V., Della, Serra, G., Mizzaro, S., & Demartini, G. (2023). Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*, *27*(1), 59–89. https://doi.org/10.1007/S00779-021-01604-6
- Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., Mizzaro, S., & Demartini, G. (2020). The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? *International Conference on Information and Knowledge Management, Proceedings*, 1305–1314. https://doi.org/10.1145/3340531.3412048

- Spohr, D. (2017). Fake news and ideological polarization. *Business Information Review*, *34*(3), 150–160. https://doi.org/10.1177/0266382117722446
- Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday & Co.
- Susmann, M. W., & Wegener, D. T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory and Cognition*, *50*(2), 435–448. https://doi.org/10.3758/S13421-021-01232-8
- Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(12), 1948–1961. https://doi.org/10.1037/XLM0000422
- Tappin, B. M., Pennycook, G., & Rand, D. G. (2021). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General*, *150*(6), 1095–1114. https://doi.org/10.1037/xge0000974
- Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0203958
- Traberg, C. S., Harjani, T., Roozenbeek, J., & van der Linden, S. (2024). The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports 2024*, *14*(1), 1–18. https://doi.org/10.1038/s41598-024-54030-y
- Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/J.TICS.2018.01.004
- Van Bavel, J. J., Rathje, S., Vlasceanu, M., & Pretus, C. (2024). Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, *56*, 101787. https://doi.org/10.1016/J.COPSYC.2023.101787
- Vlasceanu, M., Dyckovsky, A. M., & Coman, A. (2023). A network approach to investigate the dynamics of individual and collective beliefs: Advances and applications of the bending model. *Perspectives on Psychological Science*, *19*(2), 444–453. https://doi.org/10.1177/17456916231185776
- Vlasceanu, M., Morais, M. J., & Coman, A. (2021a). Network structure impacts the synchronization of collective beliefs. *Journal of Cognition and Culture*, *21*(5), 431–448. https://doi.org/10.1163/15685373-12340120
- Vlasceanu, M., Morais, M. J., & Coman, A. (2021b). The effect of prediction error on belief update across the political spectrum. *Psychological Science*, *32*(6), 916–933. https://doi.org/10.1177/0956797621995208
- Voinson, M., Billiard, S., & Alvergne, A. (2015). Beyond rational decision-making: Modelling the influence of cognitive biases on the dynamics of vaccination coverage. *PLOS ONE*, *10*(11). https://doi.org/10.1371/journal.pone.0142990
- Wikimedia Foundation. (2024, June 12). Wikipedia. https://es.wikipedia.org/wiki/Wikipedia
- Wikimedia Statistics(2024). Retrieved March 21, 2024, from https://stats.wikimedia.org/#/all-projects
- Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194. https://doi.org/10.1016/0749-596X(92)90010-U
- Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2021, February 4). Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter. https://doi.org/10.31219/osf.io/xwc79
- Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

- World Economic Forum (2024). *The world is changing and so are the challenges it faces*. Retrieved March 21, 2024, from https://www.weforum.org/agenda/2024/01/ai-disinformation-global-risks/
- Zimmerman, F., Garbulsky, G., Ariely, D., Sigman, M., & Navajas, J. (2022). Political coherence and certainty as drivers of interpersonal liking over and above similarity. *Science Advances*, *8*(6), 1909. https://doi.org/10.1126/SCIADV.ABK1909
- Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2019). Cognitive inflexibility predicts extremist attitudes. *Frontiers in Psychology*, *10*, 424519. https://doi.org/10.3389/FPSYG.2019.00989