Identity-Defining Beliefs on Social Media

Preprint · September 2022

DOI: 10.13140/RG.2.2.12438.22086/1

Daniel Williams
University of Sussex

Identity-Defining Beliefs on Social Media

Daniel Williams

Lecturer in Philosophy, University of Sussex, dgw23@sussex.ac.uk

ABSTRACT. When membership of a community depends on commitment to shared beliefs, the community is a belief-based coalition, and the beliefs are identity-defining beliefs. Belief-based coalitions are pervasive features of human social life and routinely drive motivated cognition and epistemically dysfunctional group dynamics. Despite this, they remain surprisingly undertheorized in social epistemology. This article (i) clarifies the properties of belief-based coalitions and identity-defining beliefs, (ii) explains why they often incentivize and coordinate epistemically dysfunctional forms of communication and cognitive labor, and (iii) argues that they provide a better explanation of many epistemic problems on social media than the concepts of epistemic bubbles, echo chambers, and gamification.

1.INTRODUCTION

"Have you considered not vaccinating your child?" The midwife's question took Maranda by surprise.[1] She had not considered this. The possibility had never even crossed her mind. When the midwife explained that one of her own children developed autism soon after being vaccinated, however, Maranda was interested. The midwife seemed kind and reasonable. She simply invited Maranda to do her own research.

After the midwife left, Maranda typed into Google, "Why not vaccinate?" In addition to streams of information answering her question, she came across a Facebook group, *Great Mothers Questioning Vaccines*. Now concerned, she expressed her new vaccine hesitancy to the group. She was flooded with responses: alarming anecdotes, links to credible-looking websites, evidence on the medical establishment's untrustworthiness, and more. The group struck Maranda as warm, intelligent, and even courageous. Soon she was convinced. Then she became more than convinced; she became fanatical. She began to evangelize on behalf of her new beliefs, bringing up the topic whenever possible in her offline

[1] See Storr (2021, 277–88).

interactions. When she reported such evangelism back to the group, her new co-believers showered her with praise, which energized her more. For Maranda, the online community became a real community, and her new beliefs became a central part of who she was.

Maranda's story seems to illustrate many phenomena often observed in the age of the internet and social media: an online subculture united by unfounded beliefs, the subculture's status as a source of genuine meaning and belonging for its members, the ease with which such members acquire supporting evidence for their convictions, the social praise and encouragement for those who affirm their commitment to such convictions, and more. In the past decade, similar phenomena seem to have emerged in many online subcultures, from the Manosphere to QAnon (Marwick and Partin 2022; Nagle 2017; Sunstein 2017). Moreover, some have worried that less extreme versions of such dynamics increasingly characterize and exacerbate more mainstream forms of political and cultural conflict (Bail 2021; Haidt 2022; Settle 2018; Van Bavel et al. 2021).

In trying to come to terms with some of the epistemic problems associated with social media, theorists have developed various constructs, including epistemic bubbles (Sunstein 2017), echo chambers (Nguyen 2020), and gamification (Nguyen 2021). This article argues that we will gain a better explanatory purchase on many of these problems with two very different concepts: *belief-based coalitions* and the *identity-defining beliefs* that form part of their membership criteria.[2] In posing a question on Great Mothers Questioning Vaccines, Maranda Dynda was quickly recruited into such a coalition. Because this community provided various benefits for its members, many within it—including Maranda herself—acquired a practical stake in holding and affirming its defining beliefs. This drove much of the individually biased reasoning in the community, but it also drove much of its epistemically dysfunctional group dynamics as well: the praise showered on those who displayed their commitment to the beliefs, the energy poured into constructing and sharing justifications of them, and—as Maranda would discover

[2] The term 'identity-defining belief' comes from (Rauch 2021).

when she later abandoned her beliefs and left the group—the collective punishment of heretics and apostates (Storr 2021, 277–88).

Once we understand the role of identity-defining beliefs in human social life, many features of the case, and many epistemic problems that we observe on social media more broadly, fall into place—or so I will argue in this article. After introducing belief-based coalitions and identity-defining beliefs (S2), I will explain why they often incentivize distinctive forms of group-based epistemic dysfunction (S3) and then apply these ideas to manifestations of such dysfunction on social media (S4).

2.BELIEF-BASED COALITIONS AND IDENTITY-DEFINING BELIEFS

Human beings are groupish (Boyer 2018; Pietraszewski et al. 2014; Tooby and Cosmides 2010). People develop strong attachments to groups, draw ingroup/outgroup boundaries, internalize costs and benefits to their communities, sacrifice to promote collective interests, and exhibit a suite of ingroup biases. When groups compete, such tendencies often extend to prejudice toward outgroup members. Although extensive research on "minimal group paradigm" experiments demonstrates that this groupishness can form surprisingly easily (Tajfel 1974), it typically takes its most powerful form in more stable communities, where similar tendencies often arise in groups of otherwise very different kinds.

Groupish tendencies are best understood in terms of motivations and capacities specialized for navigating a world of coalitions (Boyer 2018; Simler and Hanson 2016; Tooby and Cosmides 2010). In this section I describe the nature of coalitions and coalitional cooperation (2.1); I introduce the constructs of belief-based coalitions and identity-defining beliefs (S2.2); and I distinguish these ideas from the hypothesis that beliefs sometimes function as signals of group identity (S2.3).

2.1 COALITIONS AND COOPERATION

People typically join and stay within groups because they benefit from doing so. When these benefits result from the cooperation and coordination of group members, the group is a coalition, a bounded group that functions to promote the interests of its members. Because much of human cooperation occurs within such coalitions, our social world largely consists of—and has long consisted of—complex and often nested relationships between coalitions at multiple scales, from small-scale, fleeting alliances to larger and more stable cooperative units such as bands, tribes, sects, gangs, religions, ethnic groups, unions, subcultures, political parties, nation-states, and more (Boyer 2018; Storr 2021). These more enduring communities often exploit within-group cooperation not to achieve narrow, predefined goals but to promote the interests of group members in open-ended ways through social support, mutual aid, and opportunities for collective action (Tooby and Cosmides 2010; Williams 2021a).

Coalitions are pervasive because they are so powerful. Their reliance on cooperation makes them inherently fragile, however (Olson 1965; Tooby and Cosmides 2010). The source of this fragility is free riding. Coalitional cooperation provides benefits. Free riders attempt to reap such benefits without paying the costs required to sustain them for others. The problem is that this strategy is often instrumentally rational: even if all members of the coalition benefit from cooperation, individuals often benefit from letting other members pay the costs of cooperation for them.

Coalitions solve this problem in two basic ways: through *identification* and *incentivization*.[3] In the former case, coalitions only accept members that they identify as being committed to the group and its cooperative goals, overcoming the difficulty that individuals often benefit from exaggerating such cooperativeness.[4] Thus, groups of otherwise very different kinds often require initiation rites, behavioral restrictions, and rituals that function as costly and so credible signals of group commitment (Xygalatas 2022). Similarly, just as invading armies can burn the bridges behind them to commit—and so credibly signal a commitment—to fight on, coalitions often require members to act in ways that metaphorically burn their bridges to competing coalitions by selectively harming their reputation in the eyes of outsiders (Boyer 2018; Mercier 2020). Importantly, even behaviors that seem purely assortative—for example,

[3] These correspond to *partner choice* and *partner control* solutions in the cooperation literature (Sterelny 2012).

[4] Importantly, some signals of group identity are merely *assortative* and so do not confront this problem (Smaldino 2022).

wearing group-distinctive clothing or taking a public stand in favor of a group—can perform the same function if they are viewed unfavorably by those outside of the coalition (Iannaccone 1992).

In addition, coalitions also implement incentive structures—systems of social reward and punishment—that motivate cooperation (Boyd 2017). In this case, free riding is sanctioned either directly or reputationally, up to and including expulsion from the group, and those who pay their cooperative share are rewarded with greater reputations and status (Henrich and Muthukrishna 2021; Simpson and Willer 2015). One important feature of such reward systems involves prestige, an extremely motivating form of status rooted in respect, admiration, and deference (Ridgeway 2019; Storr 2021). Although there are different theories about why prestige hierarchies evolved in our species, one common function that they serve is motivating individuals to pay costs to advance the interests of a group, either through the carrot of enhanced status or the stick of its decline (Simpson and Willer 2015; Willer 2009). This process is often implemented explicitly in formal contexts, such as in nations that give medals to military heroes, but the underlying logic of competitive altruism—that is, competing to win status by sacrificing for others—is pervasive and often emerges spontaneously (Ridgeway 2019; Storr 2021; Willer 2009).

2.2 BELIEF-BASED COALITIONS

Coalitions have membership criteria, although such criteria are often implicit. As noted, commitment to the group and its cooperative goals are likely universal preconditions of coalitional membership. Beyond such requirements, however, membership criteria vary enormously. *Belief-based coalitions* can be understood as coalitions in which commitment to a group-specific (henceforth *identity-defining*) set of beliefs is a condition of entry.

Identity-defining beliefs can be loosely divided into two categories. First, because coalitional membership depends on group commitment, believing that one’s group has sufficiently positive qualities to warrant such commitment is likely necessary for most coalitions. When groups compete, this often means believing that one’s own group is more deserving of status and power than alternative groups.

Such group-serving and group-enhancing beliefs constitute *generic identity-defining beliefs* that recur in a similar form across many coalitions (Tajfel 1974; Williams 2022).

Second, some coalitions have *specific identity-defining beliefs* that concern the nature of reality beyond the group itself. Most religions are belief-based coalitions in this sense. They are coalitions in that religions typically leverage within-group networks of trust, mutual aid, and social support to promote the interests of their members (Norenzayan 2015). They are belief-based coalitions because commitment to a set of core convictions is normally a condition of membership. Political parties and activist groups are belief-based coalitions for similar reasons, as are communities organized around political ideologies or conspiracy theories. Of course, there is often some imprecision in which beliefs are identity defining, and identity-defining beliefs often evolve over time in response to external pressures and internal group dynamics, but these observations do not undermine the obvious fact that acceptance of specific beliefs at specific times is a condition of membership for many communities.

2.3 IDENTITY-DEFINING BELIEFS AND SIGNALING

Funkhouser (2017, 2022) argues that beliefs sometimes function as signals of group identity and commitment and draws on this framework to explain certain irrational group beliefs, such as some religious beliefs, scientific misperceptions, and unfounded political convictions (see also Bergamaschi Ganapini 2021; Simler and Hanson 2016; Williams 2021a). Although this hypothesis is consistent and complementary with the concept of identity-defining beliefs, it is important to note several differences.

First, because identity-defining beliefs often drive a form of motivated cognition in which people unconsciously subordinate the pursuit of truth to the goal of maintaining their group identity, they do frequently produce epistemic irrationality (Kahan et al. 2017; Williams 2021c). Nevertheless, we need not understand this process of identity-protective cognition in terms of signaling. Most of the appeal to signaling below (S3.2), for example, explores how people signal commitment to identity-defining beliefs, not how the beliefs themselves function as signals of group identity.

Second, although identity-defining beliefs often drive motivated cognition, they need not do so. For example, an individual might join a belief-based coalition because they formed its beliefs through purely epistemic means, and they might stay in that coalition only insofar as they can maintain such beliefs on epistemic grounds alone. Of course, as the emotional, social, or material benefits that this community provides to the individual increase, so will tendencies toward identity-protective cognition. It is difficult to get a person to understand something when their position and status within a valued community depends on them not understanding it. There are several forces that can counteract these tendencies, however. For example, people also typically care about accuracy and appearing reasonable; a group's success frequently depends on being responsive to reality; and many people belong to numerous cross-cutting communities, which constrains the degree to which any one group's beliefs can dominate cognition (Sen 2006). Further, some communities uphold norms that actively encourage within-group disagreement and reward epistemic virtue.

Finally, although Funkhouser's work focuses primarily on how identity-defining beliefs drive individual irrationality, they also often drive epistemically dysfunctional group dynamics, which is the main focus of this article.

3.THE SOCIAL EPISTEMOLOGY OF IDENTITY-DEFINING BELIEFS

Whenever beliefs are identity defining within communities that people are motivated to belong to, there are strong tendencies for their members to communicate in ways consistent with such beliefs (S3.1), to propagate and enforce them (S3.2), and to perform cognitive labor dedicated to rationalizing them (S3.3).

3.1 COMMUNICATIVE ADJUSTMENT

On an idealistic model of testimony, people communicate primarily to transmit accurate and relevant information of a sort that benefits others in shared projects of understanding and truth discovery (Nguyen 2021). This model is often simplistic and distortive. Even setting aside deliberate deception, in communicating we are frequently just as concerned with *advertising* or *hiding* traits of ourselves than with transmitting in-demand information, and such aims routinely come into conflict with the goal of simply informing others of what we know. That is, human communication is tailored in many ways and at multiple scales to make the communicator look good and to achieve strategic objectives (Boyer 2018; Dessalles and Grieve 2009; Joshi 2021; Kuran 1997; Loury 1994; Mercier 2020; Noelle-Neumann 1974; Simler and Hanson 2016). Thus, although we are an epistemically interdependent species reliant on social information, which information we receive often depends on the payoffs associated with communication, which are shaped in profound ways by incentives irrelevant to the truth of expressed viewpoints.

This dynamic characterizes many belief-based coalitions. Once communities have identity-defining beliefs, motivations to signal group commitment often incentivize those who care about their acceptance and status within the community to communicate in ways that are consistent with such beliefs (Loury 1994; Wolff 2021).

One might object that individuals who belong to belief-based coalitions must hold the relevant beliefs anyway and so will have no reason to adjust their communications. This intuition is misguided, however. The public acceptance of a group's identity-defining beliefs often masks significant heterogeneity in its members' private thoughts and attitudes (Joshi 2021; Noelle-Neumann 1974). People might secretly have contrary ideas, difficult questions, doubts, worries, potential counterevidence, and so on. If communicating these thoughts threatens their perceived commitment to the community's defining beliefs, however, they will be discouraged from expressing them (Loury 1994). In this way the evidence that group members are exposed to often becomes highly skewed toward confirming such beliefs. Of course, this process is dynamic and can easily spiral: as fewer group members dissent, the appearance of uniformity and so the costs of questioning or challenging identity-defining beliefs increases as well, leading more people to remain silent (Noelle-Neumann 1974).

One might also object that people will not join a belief-based coalition if they hold doubts about its defining beliefs, and that existing members of such coalitions can simply leave the group if they come to judge that its beliefs are unfounded. This objection is naïve, however: belief-based coalitions of many kinds—including sects, religions, social movements, activist groups, political parties, and more—are often a source of significant social and material benefits for people, and it is very easy to accrue significant non-fungible social capital within them. Given this, people routinely join and stay within belief-based coalitions for reasons independent of the truth of their beliefs, and the strength of these motivations will often easily outweigh the subjective costs of self-censorship.

Importantly, these social pressures can and do exist in the absence of identity-protective cognition. Even if I am in no way attached to a group's defining convictions, I will still often be motivated to adjust my communication if I am sufficiently attached to—or dependent on—the group itself. As noted above (S2.3), however, in many belief-based coalitions identity-protective cognition is pervasive: motivations to belong to such coalitions and win approval within them often produce corresponding motivations to form and maintain their defining beliefs (Williams 2021b). When this happens, the incentives deterring communications at odds with identity-defining beliefs are much stronger. Dissent is no longer merely a cue of disloyalty: by communicating evidence against group beliefs, it also frustrates the practical interests of others in believing what they want to believe.

3.2 ACTIVE BELIEF

Much behavior within belief-based coalitions goes beyond this communicative adjustment. In an insightful analysis, Storr (2021) points out that belief-based communities also often incentivize what he calls "*active belief*," a form of belief associated with a kind of zealous dogmatism. As the story of Maranda Dynda illustrates, this sometimes manifests as evangelism in which true believers seek to propagate group beliefs through persuasion, proselytizing, and arguing with—and sometimes attacking—outsiders who disagree. It is also often turned inward via the punishment of heretics and apostates, however, which can include everything from verbal attacks and attempts at reputational destruction all the way to imprisonment or murder (Loury 1994; Noelle-Neumann 1974).

From a self-interested perspective, active belief can seem puzzling. Even if one strongly believes something, why would one pay costs to propagate this belief or impose it on others? Of course, in some cases one also believes that the relevant beliefs *should* be propagated. Even here, however, there is a free rider problem: when beliefs are shared by one's community, why not let one's co-believers do the hard work required to transmit and enforce them?

In the framework developed here, active belief can sometimes be understood as a way of signaling commitment to identity-defining beliefs. Such signals are incentivized because—as Durkheim ([1912] 2008) observed in discussing ritual sacrifice in religions—signaling commitment to a group's ideas and values is a way of signaling commitment to the group itself (Storr 2021). This behavior can therefore arise even if nobody in the group independently desires to spread its beliefs. When propagating group beliefs constitutes a coalition's explicit goal, however, as it often does in political, cultural, and religious groups, active belief can be underpinned by reward systems that underlie any other form of within-group cooperation. In these cases, active believers are epistemic enforcers and warriors, motivated—as with those who enforce group norms and sacrifice to promote group interests more broadly (Boyd 2017; Ridgeway 2019)—by reputational and status rewards.

3.3 COGNITIVE LABOR

Like almost all sharing among non-kin, information sharing is regulated by social rewards and punishments (Boyer 2018; Dessalles and Grieve 2009; Mercier 2020; Simler and Hanson 2016). Although such rewards and punishments can be direct—the anger directed at a liar, for example, or the gratitude one feels for someone who provides helpful advice—they are often indirect, regulated by norms, reputations, and gossip (Boyer 2018; Dessalles and Grieve 2009; Mercier 2020). Importantly, such incentive structures—I will call them *epistemic reward systems*—do not merely motivate the transmission of preexisting information; they sometimes motivate individuals to expend time, energy, and resources producing or acquiring information for the purpose of sharing. Such *cognitive labor* can take many different forms—thinking hard, collecting data, infiltrating gossip networks, and so on—but it is unified in its function of producing or extracting socially valued information.

This epistemic reward system is clear and in part explicitly institutionalized in modern science, within which intellectual contributions are rewarded with—and so motivated by—recognition and status (Merton 1979), but the basic incentive structure is foundational to almost all communication and cognitive labor and underlies the unique forms of epistemic cooperation that human beings achieve.

Because people typically strive to form true beliefs relevant to their interests, epistemic reward systems tend to reward those who share information conducive to forming such beliefs and punish those who share irrelevant, inaccurate, or deceptive information (Mercier 2020). People do not always seek to build an accurate model of reality, however. In some cases, people seek information to rationalize predetermined conclusions. When this happens, the result is *rationalization markets*, information economies in which agents compete for social and sometimes financial rewards by producing and disseminating information tailored to justifying favored beliefs (Mercier 2020; Williams 2022). Such rationalization markets arise in belief-based coalitions whenever their members engage in identity-protective cognition and so seek out evidence and arguments that rationalize identity-defining convictions.

Just as competitive markets generally reward those who produce high-quality, low-cost goods and services, rationalization markets delegate the task of producing group-favored rationalizations to the minority of group members most effective at this activity. As with skillful defense lawyers who spin the truth to justify predetermined conclusions, such rationalization producers do not simply spread misinformation but frame and interpret reality in ways that rationalize group-favored beliefs. These rationalizations can therefore take many forms—empirical evidence, structured arguments, connections of favored narratives to preexisting knowledge, and so on—and they are often specialized not just for directly justifying favored conclusions but for neutralizing epistemic threats to them (see S4.2 below).

3.4 SUMMARY

In summary, belief-based coalitions are cooperative communities in which public commitment to group-specific beliefs is a condition of membership. I have argued that there are strong tendencies for such coalitions to interact with distinctive characteristics of human sociality in ways that incentivize and coordinate epistemically dysfunctional forms of communication, cooperation, and cognitive labor.

Of course, this explanatory framework is speculative in many ways, and like any framework its basic structure is highly idealized and neglects much psychological and social complexity. People have diverse and competing motives; they vary in important ways in temperament and cognitive ability; and people often belong to many communities, which can counteract the processes described here (S2.3). Moreover, groups obviously differ enormously in their goals and structure, and some groups have practices and norms designed to prevent the emergence of orthodoxies and encourage within-group disagreement. Nevertheless, I will now argue that—even bearing these important qualifications in mind—the schematic framework in this section provides a helpful explanatory purchase on some of the epistemic problems found on social media.

4. BUBBLES, CHAMBERS, AND GAMES

Belief-based coalitions and identity-defining beliefs are much older than social media. They feature in everything from the shared ancestor myths of tribal groups to the convictions of religious communities, self-serving ideologies of elite classes, and bizarre conspiracy theories affirmed by certain subcultures. Nevertheless, they also seem highly relevant to understanding social media. In recent years, many have worried about the epistemic behavior of online groups (Nagle 2017; Sunstein 2017). This includes online subcultures that exist in explicit opposition to mainstream beliefs, including groups organized around conspiratorial narratives (e.g., QAnon), misogyny (e.g., the Manosphere), racism (e.g., the alt-right), extremist religious views, and more. According to some, however, it also includes society's more mainstream political and cultural coalitions. On this view, the informational and incentive architecture of social media platforms encourages forms of communication and interaction that increase polarization and exacerbate intergroup conflict in society at large (Bail 2021; Haidt 2022; Settle 2018; Sunstein 2017).

The characteristics of belief-based coalitions seem relevant to many of the phenomena identified in support of such worries, including the existence of online groups and subcultures united by unfounded worldviews, the role of communication in signaling group identity and allegiance, the manifestations of active belief in evangelism and punishment of within-group dissent, the energy poured into justifying and protecting group-favored narratives, and more (Haidt 2022; Nagle 2017; Storr 2021). My aim in this final section is to develop this suggestion by favorably contrasting the framework outlined in this article with three alternative frameworks that focus on epistemic bubbles (S4.1), echo chambers (S4.2), and gamification (S4.3). Although the reflections here are tentative and speculative, I will present an initial case that belief-based coalitions provide a deeper and more flexible analysis of group-based epistemic dysfunction online than these more popular explanations.

Before this, however, three clarifications are important. First, my focus is on epistemically dysfunctional group dynamics on social media, not on all epistemic problems associated with social media. Second, many popular narratives about the epistemic costs of social media appear to be empirically unsupported. It is difficult to identify causal relations between social media and many of the trends that such narratives point to (e.g., political polarization), and in some cases the trends themselves—such as an allegedly unprecedent explosion of misinformation and conspiracy theorizing—are illusory (Altay 2022; Mercier 2020). Thus, my claim is not that belief-based coalitions illuminate a mythical epistemic catastrophe ushered in by social media, but that they provide a helpful lens for exploring how social media platforms interact with specific belief-based coalitions operating in specific contexts. Finally, although my focus is on certain forms of epistemic dysfunction, I do not mean to imply that social media only bring epistemic costs or even that such costs outweigh their benefits, of which there are many.

4.1 BUBBLES

One influential framework for understanding group-based epistemic dysfunction on social media appeals to epistemic bubbles, networks of like-minded people disproportionately exposed to evidence and arguments that confirm what they already believe. Although such bubbles are often called "echo chambers," I will follow Nguyen (2020) in avoiding that terminology (see S4.2).

Whatever term we use, the basic worry is that the high-choice information environment of the digital age encourages us to enclose ourselves in networks in which we rarely encounter disagreement with our beliefs (Pariser 2017; Sunstein 2017). Such opportunities are exacerbated by both natural psychological tendencies—for example, tendencies to congregate with like-minded others and seek out confirmatory evidence—and the architecture of online platforms themselves, which often seem to algorithmically filter our information diet to match our beliefs and values. According to proponents of this perspective, once we are enclosed in epistemic bubbles of this kind, we typically succumb to processes such as conformity, groupthink, and group polarization, and it is these bubble-driven dynamics that contribute to the epistemic fragmentation, polarization, and extremism allegedly observed today (Pariser 2017; Sunstein 2017).

There are two important things that this explanatory framework gets right. First, it draws attention to the role of social media and the digital age more broadly in facilitating belief-based coalition formation. To take just one example, the Incel ("Involuntarily Celibate") community, an online subculture that disproportionately attracts males with certain characteristics and attitudes (e.g., self-perceived low social status and misogyny), could probably not have emerged without the ways in which social media connects geographically distant individuals with similar traits and interests (Nagle 2017).

Second, online epistemic bubbles do exist at least to some degree in some contexts, especially when it comes to certain fringe subcultures and that segment of the population high in political engagement, as do certain characteristics such as conformity and group polarization often thought to be associated with them (Halberstam and Knight 2016; Sunstein 2017; Wojcieszak et al. 2022). Moreover, we should expect some degree of bubble formation whenever groups hold identity-defining beliefs, not just because people preferentially interact with ingroup members but also because identity-protective cognition drives individuals to insulate beliefs from disconfirmation and seek out rationalizations of them. From this perspective, then, extreme forms of online bubble construction can simply be viewed as one way in which certain belief-based coalitions protect identity-defining worldviews with offline parallels in how cults and sects sometimes live in isolated communes.

Nevertheless, understood as a general framework for understanding group-based epistemic dysfunction on social media, the concept of epistemic bubbles is inadequate for two reasons. First, although people do disproportionately consume information congenial to their political and cultural identities, for most people social media use seems to broaden the variety of viewpoints that they are exposed to (Acerbi 2019; Törnberg 2022). Thus, scientific research and commonsense observation align in noting that social media often features not isolated epistemic communities but constant intergroup contact and conflict. Second, contrary to the assumption that exposure to alternative viewpoints online will have beneficial epistemic consequences, such exposure often exacerbates dogmatism and division (Bail 2021; Garrett et al. 2014). For example, in an influential field experiment in which Democrats and Republicans were exposed to the tweets of politicians and opinion leaders from the other side, Bail et al. (2018) found that Republican participants expressed substantially more conservative viewpoints after such exposure and Democrat participants expressed slightly more liberal ones (see also Garrett et al. 2014).[5] This aligns with a significant body of research demonstrating that exposure to alternative beliefs online is often highly aversive for people and interpreted as "an attack upon their identity" (Bail 2021, 31).

The framework of belief-based coalitions helps to illuminate this situation. Society's mainstream political and cultural coalitions are in active competition for power and status. For such groups, epistemic isolation is therefore neither possible nor desirable, and the primary effect of social media is the exact opposite of epistemic bubbles: the provision of a public forum on which instant and spatially

[5] The effect was not statistically significant for liberals, however.

unconstrained communication facilitates a rate of exposure to different belief-based communities with no precedent in the offline world (Törnberg 2022). Moreover, it should not be surprising that exposure to contrary viewpoints under such conditions is often subjectively aversive and socially divisive. Belief-based coalitions often produce incentive structures—incentives rendered explicit and hyper-salient on social media (S4.3)—that motivate forms of communication and interaction antithetical to persuading outsiders. There are at least three primary manifestations of this, all of which are neglected by research on epistemic bubbles.

First, the appearance of uniform opinion within belief-based communities online is rarely simply a consequence of constricted information exposure and connection with like-minded others; it is often actively sustained by motivations to signal group identity and punish heretics.[6] Thus, Bail (2021) reviews evidence demonstrating that political moderates in the USA often self-censor and withdraw from online discourse primarily to avoid the wrath of activists and extremists from *within their own political parties* (see Haidt 2022; Hawkins et al. 2019). When combined with the fact that those highest in political engagement are the most partisan and by far the most likely to communicate about politics on social media, online exposure to the viewpoints of belief-based communities typically involves exposure to a degree of group consensus that does not reflect the real distribution of private beliefs and attitudes among the relevant group members (Törnberg 2022). This creates both a false *perception* of the degree of polarization and a heightened sense of outgroup homogeneity, thus increasing intergroup animosity (Bail 2021; Settle 2018). Of course, this process can easily spiral: as moderates self-censor and extremists become emboldened, the costs of internal dissent increase further (Wojcieszak et al. 2022).

Second, belief-based coalitions encourage signals of group commitment, which—under conditions of sharp intergroup conflict—often means selectively harming the communicator's reputation among

[6] Research on epistemic bubbles recognizes *reputational* influences on communication in addition to constricted information exposure (Sunstein 2017), but the generic and typically weak reputational pressures on conformity described in most research on belief polarization are different from the incentives for active punishment of heretics and apostates encouraged by belief-based coalitions. I am grateful to Elizabeth Edenberg for raising this objection.

outgroup members by demonizing them or expressing viewpoints that they will regard as outrageous (Funkhouser 2022; Mercier 2020; Williams 2021a; see S2.1 above). Such status-seeking antagonistic ingroup signaling appears to be ubiquitous on social media (Bail 2021; Bergamaschi Ganapini 2021; Brady et al. 2021; Grubbs et al. 2019; Osmundsen et al. 2021; Rathje et al. 2021; Settle 2018). Given this, it is not surprising that exposure to contrary viewpoints online is frequently aversive: it often involves encountering communication aimed at winning social approval within communities that are in active conflict with one's own (Suhay et al. 2018). In this way social media platforms often function not as "arenas for rational deliberation and political debate but as spaces for social identity formation and for symbolic displays of solidarity with allies and difference from outgroups" (Törnberg 2022, 10). As Tufekci (2018) puts it,

> [W]hen we encounter opposing views in the age and context of social media, it's … like hearing them from the opposing team while sitting with our fellow fans in a football stadium. Online, we're connected with our communities, and we seek approval from our like-minded peers. We bond with our team by yelling at the fans of the other one.

Finally, the concept of epistemic bubbles neglects how belief-based coalitions often produce rationalization markets that incentivize forms of cognitive labor dedicated to justifying identity-defining beliefs. Such rationalization markets are widespread and conspicuous on social media, however. From Twitter micro-influencers who cleverly affirm and rationalize group-favored narratives to popular Youtube videos showcasing a group's ideological critics being "destroyed," rationalization production is pervasive and crucial for understanding how many online communities protect unfounded worldviews and neutralize epistemic threats to them (Williams 2022). Once again, it should not be surprising that encountering communication of this kind does not have salutary epistemic consequences on those outside of the community. For such outsiders, the communication often appears—accurately—in the form of highly selective arguments in favor of conclusions that they do not agree with, which likely increases the perception that competing groups are highly biased.

In summary, then, the framework of epistemic bubbles rests on two mistakes: by neglecting the fact that most belief-based coalitions do not seek epistemic isolation, it ignores the ways in which social media often increases exposure to contrary viewpoints; and by neglecting the forms of communication incentivized by belief-based coalitions, it ignores how viewpoints on social media are often filtered and packaged in ways that are aversive and unpersuasive to those who do not already agree.

Nevertheless, it is important to qualify this analysis. We should only expect these dynamics in societies in which people come to social media platforms already attached to belief-based coalitions involved in acrimonious intergroup conflict. This is true in countries such as the USA, within which strong forms of partisan identification and affective polarization long predate the emergence of social media (Bail 2021; Finkel et al. 2020). In societies without sharp intergroup conflict, however, the incentives for ingroup signaling, active belief, and rationalization markets will be much weaker and so easily trumped by other motivations, such as appearing reasonable to audiences with heterogeneous views. This reflects a broader point: because people interpret the reward systems on social media through the lens of what they independently care about (S4.3), the effects of social media platforms depend on the nature of and relations between existing belief-based coalitions in the societies within which they exist (Benkler et al. 2018).

4.2 CHAMBERS

In an influential discussion, Nguyen (2020) argues that the social-epistemic structure most responsible for group-based epistemic dysfunction today—both on social media but also more broadly—is what he calls (in opposition to standard uses of this term) "echo chambers" (see also Jamieson and Cappella 2010). Unlike epistemic bubbles, in which relevant outside voices are simply omitted, Nguyen defines echo chambers as epistemic communities in which members share beliefs which include reasons to distrust those outside of the community, who are viewed as either dishonest or unreliable. Such communities can thus withstand exposure to contrary viewpoints precisely because they distrust their source. He illustrates this framework by appeal to conservative groups in the USA who share and consume the worldview of right-wing figures and media such as Rush Limbaugh, Fox News, and Breitbart, although he argues that echo chambers in general also thrive on social media (Nguyen 2021).

This analysis is correct to draw attention to the role of trust disparities between ingroup members and outsiders in sustaining many unfounded group-based worldviews. Although such disparities can be rational and arise in epistemically healthy groups (Levy 2021)—think of an epistemic community of scientific experts—Nguyen is right that such disparities are strongly antithetical to knowledge acquisition in many communities, either because of the degree of the disparity or because of the targets of trust and distrust. To take an extreme example, proponents of the conspiracy theory QAnon, which holds that the United States is secretly run by a ring of Satan-worshipping cannibalistic pedophiles, appear to place greater trust in the posts of an anonymous user "Q" onto the unregulated 8chan message board than in all mainstream sources of information.

Moreover, in some respects Nguyen's analysis superficially resembles the framework developed in this article, especially in the idea that holding certain beliefs constitutes part of the membership criteria for access to certain epistemic communities. Nevertheless, there are several important differences between the frameworks, and belief-based coalitions and identity-defining beliefs provide a superior explanation of group-based epistemic dysfunction.

First, as I have tried to show in this article, communities of co-believers often sustain internal reward systems that motivate and coordinate epistemically destructive forms of communication, cooperation, and cognitive labor. By focusing exclusively on the allocation of trust, the concept of echo chambers ignores this rich social fabric and its participatory dynamics, including motivations to signal group identity and commitment to identity-defining beliefs, to enforce and propagate such beliefs, and to perform cognitive labor dedicated to rationalizing them. Such phenomena are all crucial for understanding most forms of group-based epistemic dysfunction in general and on social media specifically, however.

We see this in the case of Maranda Dynda with which I began this article. Although the distrust of scientific authorities among members of anti-vaccination communities such as Great Mothers

Questioning Vaccines is clearly relevant for understanding their beliefs, the group also featured numerous other phenomena that went far beyond this. For example, Maranda describes how "warm and cosy" it felt to join a community of "strong, confident women," how you were "socially rewarded for going with the group," how the more you reported offline evangelism back to the group "the higher you moved up socially" within it, and how those who dissented or left the group experienced its collective wrath, insults, and hatred (Storr 2021, 280–86).

One sees similar dynamics in QAnon (Marwick and Partin 2022). For those who participate in the group, the motto of which is—tellingly—"Where we go one, we go all," it functions not just as an abstract *epistemic community* united by shared beliefs but as a real community that functions as a source of friendship and status for its members. Although the systematic distrust of mainstream authorities within this mostly online community is clearly relevant for understanding its beliefs, this distrust is combined with an incentive structure of social rewards and punishments that encourages various forms of active community participation, including extensive ingroup signaling, active belief, and rationalization markets (Berkowitz 2021). Drawing on qualitative fieldwork on the 8chan imageboard onto which "Q" posts, Marwick and Partin (2022) thus describe QAnon's community dynamics as a "collective, knowledge-making activity … designed to construct specific facts and theories that maintain QAnon's cohesion over time." And such participatory practices are not unique to QAnon. In a rich study of numerous online subcultures, including the misogynistic communities of the Manosphere, "alt-right" movements that initially emerged to prominence online, and extremely moralistic, hyper-progressive millennial subcultures on Tumblr, Nagle (2017) documents how all such communities manifest the dynamics of belief-based coalitions, including costly displays of allegiance to the group and its orthodoxies, the punishment of heretics, the encouragement of active belief, and the prestige conferred on those who affirm and cleverly rationalize shared convictions.

Thus, one problem with the concept of echo chambers is that its exclusive focus on trust ignores much of the epistemically relevant social fabric of many belief-based communities. In addition, however, it also misses how this fabric *interacts with* and *supports* the selective allocation of trust. For one thing, it is precisely because people tend to exhibit greater trust in ingroup members than outsiders that the ways in which belief-based coalitions often punish ingroup dissent can be so epistemically toxic. More importantly, however, by treating distrust of outsiders as merely a property of shared belief systems, the concept of echo chambers obscures how this distrust is often scaffolded by the incentives and dynamics within epistemic communities. Specifically, groups often actively seek out evidence and arguments that discredit those who challenge or threaten their identity-defining beliefs, and it is this collective demand for selective discrediting—and the associated willingness to attend to and admire those who satisfy this demand—that often fuels the cognitive labor required to satisfy it (Williams 2022). Thus, the prestige economies of online subcultures organized around identity-defining beliefs often select for the widespread dissemination of social media posts designed to derogate and so discredit outgroup members (Osmundsen et al. 2021; Rathje et al. 2021), but they also select for a class of cognitive laborers who achieve community esteem by neutralizing epistemic threats to group beliefs, often by undermining their source (Williams 2022).

This perspective diverges radically from Nguyen's. Specifically, although he concedes that echo chambers sometimes emerge because people value the community that they provide, he argues that "the most plausible explanation for the particular features of echo chambers is … [that they] are excellent tools to maintain, reinforce, and expand power through epistemic control" (Nguyen 2020, 149). Thus, he depicts people primarily as victims "trapped" within echo chambers manufactured by manipulative elites, which "prey" on them and that operate through techniques like those allegedly involved in cult "indoctrination" (Nguyen 2020).

Although this passive analysis of members of epistemically dysfunctional communities might be applicable in some cases, in most cases it is misleading. At the most general level, Nguyen's explanation of echo chambers seems to be underpinned by an image of human beings as credulous and easily duped into holding false beliefs by elite manipulation, which is at odds with a large body of empirical research demonstrating that people are highly vigilant social learners and extremely resistant to attempts at epistemic manipulation of this kind (Mercier 2020).[7]

More importantly, when we turn to real-world cases of group-based epistemic dysfunction, the analysis of group members as passive victims is often demonstrably mistaken. As already noted, online subcultures organized around unfounded worldviews typically involve active participation in valued communities within which shared beliefs are promoted and protected through internal reward systems and group participation. Thus, against a popular narrative in which any unsuspecting individual in society can be sucked into and "trapped" within such communities, a significant body of empirical research shows that they cater to populations independently receptive to their worldviews and to the community benefits that they provide (Marwick and Partin 2022; Phadke et al. 2021). Summarizing this research, Altay (2022, 10) therefore observes that people do not fall into online "rabbit holes" but "*jump in and dig*" (my emphasis).

However, the same point also applies to Nguyen's main example of an echo chamber in the epistemic community that consumes media such as Rush Limbaugh, Fox News, and primarily online sources such as Breitbart and Infowars. In this case, the empirical literature strongly suggests that such media has not arisen in a process by which elites have trapped and indoctrinated millions of people in an externally imposed belief system. Instead, the extraordinary market success of such media is primarily rooted in the ways in which it affirms and rationalizes the preexisting prejudices and attitudes of people with distinctive characteristics, including sympathy with white identity politics, reactionary social attitudes, and hostility to liberalizing social trends (Benkler et al. 2018; Iyengar and Hahn 2009; Williams 2022). Thus, although active community participation is much lower in these larger belief-based

[7] In more recent work, Nguyen (2021) suggests that this manipulation occurs not primarily via epistemic means but through the design of belief systems that seduce people with the pleasures of certainty and clarity. This is also implausible, however. The number of possible worldviews with such characteristics is infinite, so the characteristics themselves offer no explanation of why people would opt for some such worldviews over others; further, for those who hold the kinds of worldviews that Nguyen has in mind, including believers in QAnon and right-wing populist narratives, their belief in pervasive conspiracy, decline, and threat is likely highly distressing.

communities than in more tight-knit and fringe subcultures, the successful pundits and opinion producers are still much better thought of as *rationalization producers* than as elite manipulators. Rather than trapping an unsuspecting audience in an externally imposed worldview through manipulating their trust, the extensive arguments such producers give to distrust mainstream and liberal media constitute in-demand intellectual ammunition designed to insulate the worldview of their audience from contact with reality (Williams 2022).

Importantly, this analysis does not imply that the epistemically relevant characteristics of belief-based coalitions are *epiphenomena*. As I have tried to show, such characteristics are often *functional* in promoting, protecting, and rationalizing identity-defining beliefs. Moreover, once an informational architecture optimized for protecting and rationalizing favored beliefs and narratives exists, it often has independent effects. In the case of Fox News, for example, although it has achieved significant success by catering to the independently existing prejudices and attitudes of white social conservatives, there is now strong evidence that its profound conservative bias has independent effects on those who consume it (DellaVigna and Kaplan 2007). Finally, I am also not claiming—absurdly—that elite manipulation of public belief never occurs or is never successful. Nevertheless, if we understand the inhabitants of dysfunctional belief-based communities as credulous victims of externally imposed belief systems, we will lack the theoretical resources to identify the various causal pathways by which such manipulation operates.

4.3 GAMES

I will conclude by contrasting the framework developed here with one final explanation of epistemic dysfunction on social media that focuses on *gamification*. Specifically, I will focus on Nguyen's (2021) argument that the architecture of Twitter "gamifies" communication and public discourse in ways that have negative epistemic consequences, although Nguyen suggests that a similar argument applies to other social media platforms as well. Moreover, the general idea of understanding epistemic dysfunction within certain online communities such as QAnon in terms of their game-like properties is increasingly influential (Berkowitz 2021), and the arguments here generalize to all such analyses.

Although Nguyen's analysis of gamification is rich and subtle, the core idea can be understood by contrasting his depiction of non-gamified "natural" communication with Twitter communication. According to Nguyen (2021, 426), the "natural aim" of earnest discourse "is the collective pursuit of truth": "We aim to express what we think of as true, and to question and challenge each other's expressions, as part of our quest to understand the world." The informational and incentive architecture of Twitter corrupts these communicative aims by "inviting its users to *change the goals of their participation in discourse*—to simplify those goals in exchange for pleasure" (Nguyen 2021, 416). The source of these distorting pleasures, argues Nguyen (2021, 411), is the game-like character of Twitter, which offers "immediate, vivid, and quantified evaluations of one's conversational success" in the form of game-like "points" such as Likes, Retweets, and Follower Counts. Because maximizing such points is "simpler, clearer, and easier to apply" than achieving our natural communicative goals (Nguyen 2021, 412), we replace such goals with the more pleasurable ones of gamified communication:

> Pre-gamification, the aims of discourse are complex and many. Some of us want to transmit information or to persuade; some of us want friendship. Some of us want to join together in the pursuit of truth and understanding. Twitter gamifies discourse and, in so doing, offers us re-engineered goals for our communicative acts. Twitter invites us to shift our values along its pre-fabricated lines. We start to chase higher Likes and Retweets and Follower counts—*and those are very different targets*. (Nguyen 2021, 412; my emphasis)

This analysis is correct to draw attention to the effects of social media's reward systems on communication, which are undoubtedly psychologically powerful and epistemically consequential (Bail 2021; Storr 2021). Nevertheless, the concept of gamification does not provide the right framework for understanding these effects.

First, offline communication is just as regulated by social rewards and punishments as communication on Twitter. From an introspective point of view, it no doubt seems like the "natural aim" of discourse and conversation is the pursuit of truth and understanding, not least because human self-understanding is strongly biased toward presenting motives as noble in this way (Simler and Hanson 2016). As described above (S3.3), however, human communication is a form of social interaction scaffolded and regulated in almost all cases by social incentives rooted in phenomena such as reputations, norms, alliance formation, signaling, social pressures, and prestige (Boyer 2018; Dessalles and Grieve 2009; Mercier 2020; Simler and Hanson 2016; Storr 2021). Thus, outside of social media platforms, people are also highly sensitive to the distribution and strength of other people's attitudes and regulate their communication in many ways and at multiple scales in response to social feedback (Kuran 1997; Loury 1994; Noelle-Neumann 1974).

Of course, because people typically seek to form true beliefs relevant to their interests, epistemic reward systems do often favor and so facilitate the dissemination of accurate information that promotes knowledge and understanding (see S3.3). This is true on social media platforms as well, however, including Twitter (Altay 2022; Mercier 2020). Nguyen is correct that it is not *always* true on Twitter, *but it is not always true in offline life either* (Boyer 2018; Loury 1994; Mercier 2020; Noelle-Neumann 1974). For example, people frequently exaggerate, spread unfounded rumors and gossip, and self-censor in response to social incentives offline, and—as I have argued in this article—belief-based coalitions often motivate forms of communication antithetical to knowledge and understanding. Thus, it is misleading when Nguyen (2021, 416) contrasts the gamified nature of Twitter communication with "natural" communication by observing that "we have evidence aplenty that what makes something go viral [on Twitter] is not its truth, or the degree to which it promotes understanding." This is often just as true of offline communication (Boyer 2018; Mercier 2020; Mills 1994; Noelle-Neumann 1974). Among orthodoxies that went viral prior to the emergence of Twitter partly as a consequence of social goals and social pressures, for example, one might include the divine right of kings to exercise arbitrary power, the existence of evil Jewish conspiracies, the cognitive and moral superiority of dominant social groups, and more.

Given this, one way in which the concept of gamification leads us astray in understanding social media is in its assumption that the influence on—and corruption of—communication by social feedback and incentives is unique to social media. Nevertheless, one might concede this but argue that the specific way in which such feedback shapes communication on social media is still usefully understood in terms of gamification. This view is also mistaken, however.

For one thing, the analysis provided by gamification is incomplete. Even if communication on Twitter were aimed solely at maximizing quantities such as Likes, Retweets, and Followers, this fact is not informative without a supplementary account of what causes audiences to confer these rewards. Although such causes are clearly diverse, ranging from liking tweets featuring amusing dog videos to following accounts that produce good football commentary, the framework outlined in this article provides a deeper explanation of what motivates the allocation of social rewards within many belief-based communities. More importantly, however, communication on Twitter does *not* aim solely at the alleged pleasures associated with maximizing Twitter quantities. Instead, such communication is influenced by what the currency of Likes, Retweets, and Followers *represent*: namely, independently sought outcomes such as social approval, coalitional support, and prestige, and—in the case of other common social media events such as "pile-ons" and getting "ratioed"—independently feared outcomes such as reputational destruction, ostracism, and collective punishment.

This simple observation undermines the gamification analysis in two ways. First, it suggests that people do not fundamentally seek *pleasure* on Twitter. That is, although hedonic systems always play a proximate causal role in regulating behavior toward the achievement of goals, what people seek to achieve on Twitter is not pleasure itself but concrete social and material outcomes. As with goal pursuit more broadly, the process of achieving such outcomes is often actively unpleasant. Although *being* high status can be a source of pleasure, for example, *achieving* such status, especially in contexts of intense, heated political and cultural disputes, is often anything *but* pleasurable.

Second, because people respond to what Twitter metrics represent*,* people's interaction with and engagement on the platform is much more sophisticated than the concept of gamification implies. People do not exclusively seek to maximize Likes, Retweets, and Followers in the way that one might aim monomaniacally at winning points in games. As with ordinary social life, people interpret and contextualize social rewards and punishments on social media in sophisticated ways according to their nature and source. For example, we care much more about social approval and disapproval from those high in status and from ingroup members (Boyer 2018; Finkel et al. 2020; Ridgeway 2019), and we regulate our behavior—both offline and online—accordingly. Thus, a thousand likes from strangers online can be easily psychologically outweighed by a single passive-aggressive comment from a respected member of one's social network (trust me), and the likes that a tweet elicits from outgroup members can be barely registered relative to the deafening silence from one's co-believers.

In summary, then, the problem with the gamification analysis of Twitter is twofold: the social regulation—and frequent corruption—of communication is not unique to social media, and the ways in which this regulation occurs on social media is far more continuous with offline discourse than the concept of gamification implies. Moreover, these points generalize to attempts at understanding epistemic dysfunction on social media in terms of gamification more broadly: to the extent that online communities feature psychologically consequential game-like elements, these elements are manifestations of the incentive structure that underlies all human sociality and communication.

Crucially, this is not to deny that the specific incentive architecture of social media platforms such as Twitter has epistemically consequential—and often destructive—effects. Rather, it is a call to understand these consequences in ways that accommodate the distinctive character of all human sociality and communication. To that end, I will conclude with three speculations on this topic consistent with the framework and ideas outlined in this article.

First, although humans always evaluate other people's cooperativeness and group commitments, such evaluations likely work very differently on social media than in most offline contexts. Most importantly, in communities with significant offline contact and interaction, people can acquire an enormous amount of information about each other not mediated by deliberate impression management.[8] In fact, the benefits of impression management decrease in proportion to the degree of such scrutiny precisely because it becomes more difficult to influence people's assessment of you as their knowledge of you increases. On social media, however, *all* the information that you can acquire about others is mediated by deliberate attempts to communicate that information. This dramatically increases the incentives for impression management and might explain a common observation that social media has greatly amplified the performativity of much communication, including—in contexts of belief-based coalitions—competitive displays of ingroup allegiance (Tosi and Warmke 2020).

Second, although we are always sensitive to social rewards and punishments in communication, the fact that such incentives are so explicit, immediate, and hyper-salient on social media is no doubt consequential. In general, human motivations are plural, responsive to environmental contingencies, and must be traded off in complex ways. By adjusting the salience structure of the environment within which communication occurs, social media platforms likely have dramatic effects on such trade-offs. For example, the instant rewards you receive for a snarky Twitter post that cleverly affirms your ingroup's favored narrative might be outweighed by the reputational damage you incur among thoughtful individuals outside of your community. If the former rewards are instant and highly salient whereas the reputational damage is not explicitly signaled to you, however, you might opt for the former nevertheless, a situation that is exacerbated by the fact that most social media platforms lack any quick, low-cost way to express disapproval analogous to Likes.

Finally, the social feedback on communication in much offline discourse often arises in the context of two-person or small-group discussion. Thus, even though communication is heavily influenced by social feedback in such contexts, such feedback is often highly variable and idiosyncratic. When communicating on most social media platforms, in contrast, the potential audience is often much larger. Under such conditions, the benefits of tailoring communication in ways responsive to people's

[8] In technical terms, the information is mediated by cues, not signals.

idiosyncratic assessments are therefore smaller relative to the benefits of tapping into sources of positive feedback shared by larger audiences. For this reason, many social media platforms are plausibly specialized for emphasizing the rewards associated with affirming identities shared by large groups, which might explain a common intuition that communication on social media is much more groupish than most offline discourse.

REFERENCES

Acerbi, A. 2019. *Cultural Evolution in the Digital Age*. Oxford: Oxford University Press. [AU: Which city of publication?]

Altay, S. 2022. *How Effective Are Interventions against Misinformation?* PsyArXiv. https://doi.org/10.31234/osf.io/sm3vk.

Bail, C. 2021. *Breaking the Social Media Prism*. Princeton, NJ: Princeton University Press.

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115(37): 9216–21. https://doi.org/10.1073/pnas.1804840115.

Benkler, Y., R. Faris, and H. Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York: Oxford University Press. [AU: Which city of publication?]

Bergamaschi Ganapini, M. 2021. "The Signaling Function of Sharing Fake Stories." *Mind & Language*, n/a(n/a). https://doi.org/10.1111/mila.12373.

Berkowitz, R. 2021, January 11. *A Game Designer's Analysis of QAnon*. https://medium.com/curiouserinstitute/a-game-designers-analysis-of-qanon-580972548be5.

Boyd, R. 2017. *A Different Kind of Animal*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/hardcover/9780691177731/a-different-kind-of-animal.

Boyer, P. 2018. *Minds Make Societies: How Cognition Explains the World Humans Create*. New Haven, CT: Yale University Press.

Brady, W. J., K. McLoughlin, T. N. Doan, and M. J. Crockett. 2021. “How Social Learning Amplifies Moral Outrage Expression in Online Social Networks.” *Science Advances* 7(33): eabe5641. https://doi.org/10.1126/sciadv.abe5641.

DellaVigna, S., and E. Kaplan. 2007. “The Fox News Effect: Media Bias and Voting*.” *Quarterly Journal of Economics* 122(3): 1187–1234. https://doi.org/10.1162/qjec.122.3.1187.

Dessalles, J.-L., and J. Grieve. 2009. *Why We Talk: The Evolutionary Origins of Language*. Oxford: Oxford University Press. [AU: Which city of publication?]

Durkheim, É. 2008. *The Elementary Forms of Religious Life*. Oxford: Oxford University Press. [AU: Which city of publication?]

Finkel, E. J., C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. Van Bavel, C. S. Wang, and J. N. Druckman. 2020. “Political Sectarianism in America.” *Science* 370(6516): 533–36. https://doi.org/10.1126/science.abe1715.

Funkhouser, E. 2017. “Beliefs as Signals: A New Function for Belief.” *Philosophical Psychology* 30(6): 809–31. https://doi.org/10.1080/09515089.2017.1291929.

Funkhouser, E. 2022. “A Tribal Mind: Beliefs That Signal Group Identity or Commitment.” *Mind & Language* 37(3): 444–64. https://doi.org/10.1111/mila.12326.

Garrett, R. K., S. D. Gvirsman, B. K. Johnson, Y. Tsfati, R. Neo, and A. Dal. 2014. “Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization.” *Human Communication Research* 40(3): 309–32. https://doi.org/10.1111/hcre.12028.

Grubbs, J. B., B. Warmke, J. Tosi, A. S. James, and W. K. Campbell. 2019. “Moral Grandstanding in Public Discourse: Status-Seeking Motives as a Potential Explanatory Mechanism in Predicting Conflict.” *PLOS ONE* 14(10): e0223749. https://doi.org/10.1371/journal.pone.0223749.

Haidt, J. 2022, April 11. "Why the Past 10 Years of American Life Have Been Uniquely Stupid." *The Atlantic*. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/.

Halberstam, Y., and B. Knight. 2016. "Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter." *Journal of Public Economics* 143: 73–88. https://doi.org/10.1016/j.jpubeco.2016.08.011.

Hawkins, S., D. Yudkin, M. Juan-Torres, and T. Dixon. 2019. *Hidden Tribes: A Study of America's Polarized Landscape* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xz25v.

Henrich, J., and M. Muthukrishna. 2021. "The Origins and Psychology of Human Cooperation." *Annual Review of Psychology* 72: 207–40. https://doi.org/10.1146/annurev-psych-081920-042106.

Iannaccone, L. R. 1992. "Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives." *Journal of Political Economy* 100(2): 271–91. https://doi.org/10.1086/261818.

Iyengar, S., and K. S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59(1): 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x.

Jamieson, K. H., and J. N. Cappella. 2010. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. New York: Oxford University Press. [AU: City of publication?]

Joshi, H. 2021. *Why It's OK to Speak Your Mind*. New York: Routledge. [AU: City of publication?] https://www.routledge.com/Why-Its-OK-to-Speak-Your-Mind/Joshi/p/book/9780367141721.

Kahan, D. M., E. Peters, E. C. Dawson, and P. Slovic. 2017. Motivated Numeracy and Enlightened Self-government. *Behavioural Public Policy* 1(1): 54–86. https://doi.org/10.1017/bpp.2016.2.

Kuran, T. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

Levy, N. 2021. "Echoes of Covid Misinformation." *Philosophical Psychology* 0(0): 1–18. https://doi.org/10.1080/09515089.2021.2009452. [AU: Do you have the volume and issue number for this?]

Loury, G. C. 1994. “Self-Censorship in Public Discourse: A Theory of ‘Political Correctness’ and Related Phenomena.” *Rationality and Society* 6(4): 428–61. https://doi.org/10.1177/1043463194006004002.

Marwick, A. E., and W. C. Partin. 2022. “Constructing Alternative Facts: Populist Expertise and the QAnon Conspiracy.” *New Media & Society* 14614448221090200. https://doi.org/10.1177/14614448221090201.

Mercier, H. 2020. *Not Born Yesterday*. Princeton, NJ: Princeton University Press.

Merton, R. K. 1979. *The Sociology of Science: Theoretical and Empirical Investigations*, edited by N. W. Storer. Chicago: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo28451565.html.

Mills, C. W. 1994. “The Racial Contract.” In *The Racial Contract*. Ithaca, NY: Cornell University Press. https://doi.org/10.7591/9780801471353.

Nagle, A. 2017. *Kill All Normies*. Alresford: John Hunt Publishing. [AU: City of publication?]https://www.johnhuntpublishing.com/zer0-books/our-books/kill-all-normies.

Nguyen, C. T. 2020. “Echo Chambers and Epistemic Bubbles.” *Episteme* 17(2): 141–61. https://doi.org/10.1017/epi.2018.32.

Nguyen, C. T. 2021. “How Twitter Gamifies Communication.” In *Applied Epistemology*, edited by J. Lackey, 410–36. Oxford: Oxford University Press. [AU: City of publication?]

Noelle-Neumann, E. 1974. “The Spiral of Silence: A Theory of Public Opinion.” *Journal of Communication* 24(2): 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x.

Norenzayan, A. 2015. *Big Gods*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/paperback/9780691169743/big-gods.

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Osmundsen, M., A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen. 2021. “Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter.”

American Political Science Review 115(3): 999–1015. https://doi.org/10.1017/S0003055421000290.

Pariser, E. 2017. *The Filter Bubble*. New York: Penguin. [AU: City of publication?] https://www.penguinrandomhouse.com/books/309214/the-filter-bubble-by-eli-pariser/.

Phadke, S., M. Samory, and T. Mitra. 2021. "What Makes People Join Conspiracy Communities? Role of Social Factors in Conspiracy Engagement." *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW3), 223:1–30. https://doi.org/10.1145/3432922.

Pietraszewski, D., L. Cosmides, and J. Tooby. 2014. "The Content of Our Cooperation, Not the Color of Our Skin: An Alliance Detection System Regulates Categorization by Coalition and Race, but Not Sex." *PLOS ONE* 9(2); e88534. https://doi.org/10.1371/journal.pone.0088534.

Rathje, S., J. J. Van Bavel, and S. van der Linden. 2021. "Out-Group Animosity Drives Engagement on Social Media." *Proceedings of the National Academy of Sciences* 118(26): e2024292118. https://doi.org/10.1073/pnas.2024292118.

Rauch, J. 2021. *The Constitution of Knowledge: A Defense of Truth*. Washington, DC: Brookings Institution Press. [AU: City of publication?]

Ridgeway, C. 2019. *Status*. New York: Russel Sage Foundation. [AU: City of publication?] https://www.russellsage.org/publications/status.

Sen, A. 2006. *Identity and Violence*. New York: Penguin. [AU: City of publication?] https://www.penguin.co.uk/books/55882/identity-and-violence-by-amartya-sen/9780141027807.

Settle, J. E. 2018. *Frenemies: How Social Media Polarizes America*. Cambridge: Cambridge University Press. [AU: City of publication?] https://doi.org/10.1017/9781108560573.

Simler, K., and R. Hanson. 2016. *The Elephant in the Brain: Hidden Motives in Everyday Life*. New York: Oxford University Press. [AU: City of publication?]

Simpson, B., and R. Willer. 2015. "Beyond Altruism: Sociological Foundations of Cooperation and Prosocial Behavior." *Annual Review of Sociology* 41(1): 43–63. https://doi.org/10.1146/annurev-soc-073014-112242.

Smaldino, P. E. 2022. “Models of Identity Signaling.” *Current Directions in Psychological Science* 31(3): 231–37. https://doi.org/10.1177/09637214221075609.

Sterelny, K. 2012. *The Evolved Apprentice*. Cambridge, MA: MIT Press.

Storr, W. 2021. *The Status Game*. London: HarperCollins Publishers. [AU: City of publication?]

Suhay, E., E. Bello-Pardo, and B. Maurer. 2018. “The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments.” *International Journal of Press/Politics* 23(1): 95–115. https://doi.org/10.1177/1940161217740697.

Sunstein, C. 2017. *# Republic*. Princeton, NJ: Princeton University Press.

Tajfel, H. 1974. “Social Identity and Intergroup Behaviour.” *Social Science Information* 13(2): 65–93. https://doi.org/10.1177/053901847401300204.

Tooby, J., and L. Cosmides. 2010. “Groups in Mind: The Coalitional Roots of War and Morality.” In *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, edited by H. Høgh-Olesen, 91–234. Basingstoke: Palgrave-Macmillan. [AU: City of publication?]

Törnberg, P. 2022. “How Digital Media Drive Affective Polarization through Partisan Sorting.” *Proceedings of the National Academy of Sciences* 119(42): e2207159119. https://doi.org/10.1073/pnas.2207159119.

Tosi, J., and B. Warmke. 2020. *Grandstanding: The Use and Abuse of Moral Talk*. New York: Oxford University Press. [AU: City of publication?]

Tufekci, Z. 2018. “How Social Media Took Us from Tahrir Square to Donald Trump.” *MIT Technology Review*. https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump/.

Van Bavel, J. J., S. Rathje, E. Harris, C. Robertson, and A. Sternisko. 2021. How social media shapes polarization. *Trends in Cognitive Sciences* 25(11): 913–16. https://doi.org/10.1016/j.tics.2021.07.013.

Willer, R. 2009. “Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem.” *American Sociological Review* 74(1): 23–43. https://doi.org/10.1177/000312240907400102.

Williams, D. 2021a. “Signalling, Commitment, and Strategic Absurdities.” *Mind & Language* n/a(n/a). https://doi.org/10.1111/mila.12392.

Williams, D. 2021b. “Socially Adaptive Belief.” *Mind & Language* 36(3): 333–54. https://doi.org/10.1111/mila.12294.

Williams, D. 2021c. “Motivated Ignorance, Rationality, and Democratic Politics.” *Synthese* 198(8): 7807–27. https://doi.org/10.1007/s11229-020-02549-8.

Williams, D. 2022. “The Marketplace of Rationalizations.” *Economics & Philosophy*, 1–25. https://doi.org/10.1017/S0266267121000389.

Wojcieszak, M., A. Casas, X. Yu, J. Nagler, and J. A. Tucker. 2022. “Most Users Do not Follow Political Elites on Twitter; Those Who Do Show Overwhelming Preferences for Ideological Congruity.” *Science Advances* 8(39): eabn9418. https://doi.org/10.1126/sciadv.abn9418.

Arnaud Wolff, 2022. "The Signaling Value of Social Identity," Working Papers of BETA 2022-15, Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg.

Xygalatas, D. 2022. *Ritual*. New York: Profile Books. [AU: City of publication?] https://www.hachettebookgroup.com/titles/dimitris-xygalatas/ritual/9780316462402/.