Measuring Epistemic Trust: Towards a New Lens for Democratic Legitimacy, Misinformation, and Echo Chambers

DOMINIC ZAUN EU JONES, University of Illinois Urbana-Champaign, USA
ESHWAR CHANDRASEKHARAN, University of Illinois Urbana-Champaign, USA

Trust is crucial for the functioning of complex societies, and an important concern for CSCW. Our purpose is to use research from philosophy, social science, and CSCW to provide a novel account of trust in the 'post-truth' era. Testimony, from one speaker to another, underlies many social systems. Epistemic trust, or testimonial credibility, is the likelihood to accept a speaker's claim due to beliefs about their competence or sincerity. Epistemic trust is closely related to several 'pathological epistemic phenomena': democratic (il)legitimacy, the spread of misinformation, and echo chambers. To the best of our knowledge, this theoretical contribution is novel in the field of social computing. We further argue that epistemic trust is no philosophical novelty: it is measurable. Weakly supervised text classification approaches achieve F_1 scores of around 80 to 85 per cent on detecting epistemic distrust. This is also, to the best of our knowledge, a novel task in natural language processing. We measure expressions of epistemic distrust across 954 political communities on Reddit. We find that expressions of epistemic distrust are relatively rare, although there are substantial differences between communities. Conspiratorial communities and those focused on controversial political topics tend to express more distrust. Communities with strong epistemic norms enforced by moderation are likely to express low levels. While we find users to be an important potential source of contagion of epistemic distrust, community norms appear to dominate. It is likely that epistemic trust is more useful as an aggregated risk factor. Finally, we argue that policymakers should be aware of epistemic trust considering their reliance on legitimacy underwritten by testimony.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; *Empirical studies in collaborative and social computing*; • **Applied computing** → **Law, social and behavioral sciences**.

Additional Key Words and Phrases: epistemic trust, epistemology, social epistemology, political epistemology, misinformation, echo chambers, democratic legitimacy, institutional legitimacy

ACM Reference Format:
Dominic Zaun Eu Jones and Eshwar Chandrasekharan. 2024. Measuring Epistemic Trust: Towards a New Lens for Democratic Legitimacy, Misinformation, and Echo Chambers. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 462 (November 2024), 33 pages. https://doi.org/10.1145/3687001

1 Introduction

How sociotechnical systems (e.g. social media) affect the distribution and processing of information has been a concern of CSCW research for many years. The effects of misinformation are serious and CSCW perspectives are often critical [1, 90]. Closely related to the 'post-truth' phenomena are echo chambers and institutional legitimacy. In this paper, we synthesise research in social epistemology, social science, and CSCW in order to provide an account of these phenomena revolving around the notion of *trust*. Our purpose is to bring perspectives of trust from social epistemology in to CSCW

Authors' Contact Information: Dominic Zaun Eu Jones, dzjones2@illinois.edu, University of Illinois Urbana-Champaign, Urbana, Illinois, USA; Eshwar Chandrasekharan, eshwar@illinois.edu, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.

ACM 2573-0142/2024/11-ART462
https://doi.org/10.1145/3687001

and provide evidence that these perspectives lead to valid and useful measurement of trust as it underlies these phenomena.

Trust is crucial for the functioning of complex societies. There are at least two complementary accounts of this. First, and perhaps most commonly, is the game theoretic, cooperative, economic, and more generally behavioural account [9, 48, 65, 100]. For example, trust lowers transaction costs in an economy and is related with higher economic growth and equality [48]. Game theoretic experiments are often used to measure trust [3, 9]. The second is the epistemic account. For example, Reglitz [79], who we will find useful in our synthesis, relates fake news and legitimacy with people's perceptions of their fellow citizens factual and moral judgements. These accounts focus more on trust in information and testimony, rather than directly measuring cooperative behaviour or reciprocity. It is the *source* that is the main subject of trust. The source may be a person, as in Judge-Advisor systems [88], or perhaps a news site [47] or institution. This paper is primarily concerned with the epistemic account.

Knowledge, or at least (justified and unjustified) beliefs, are existentially important. For example, I know that if I sign a legal contract I will be held to its provisions. If I have some justified belief about the effects of an economic policy then this will affect my level of support or opprobrium directed towards it. Most of my knowledge and beliefs will be ultimately underwritten by testimony. It is, as a general rule, impossible to travel to every corner of the world and perceive first-hand all events. We are dependent on each other [47, 88]. Testimony is transmitted in many forms, but those most interesting to us presently are news and social media.

Testimony is not, as another general rule, blindly trusted. Nor is it rejected out-of-hand. Most of us inhabit some epistemic space between total gullibility and skepticism. It would be unfortunate if farmers did not believe seasonal climate forecasts in the planting of their crops. It would be similarly unfortunate if we believed that each and every representative of the government were secretly reptiles. Hence the importance of *epistemic trust* (or *testimonial credibility*). Epistemic (dis)trust tells us if, and why, testimony from some specific entity is accepted or rejected. Epistemic trust depends on one's view of a speaker's competence and sincerity. We argue that epistemic trust is deeply linked with misinformation, echo chambers, and institutional legitimacy. The problem is: can we measure it? We provide evidence that we can. Further, we apply a measurement model to social media data for ecological validation. Among other things, we find community-level epistemic norms are important, which aligns with prior work [16, 28].

Social computing has a growing interest in misinformation and credibility [23, 42, 53], and civic engagement [52, 75, 82]. It has a long history of interest in trust, especially highlighting that it is *social*, not *technological* [25, 94]. This work approaches the issue from a different angle. We do not attempt to fact-check, detect misinformation, or pop echo chambers. Instead, we argue that epistemic trust is both theoretically useful, practically measurable, and complementary to the current approaches in the field.

While trust is not new to CSCW, to the best of our knowledge this is the first work to emphasise it in a social epistemic sense. Theory allows us to view unifying threads between misinformation, echo chambers, and institutional legitimacy. Furthermore, we argue these threads are crucial complements for how we currently view these phenomena.

Summary of Contributions. Our contributions to the CSCW community are as follows. First, we provide a novel synthesis of theory from the fields of philosophy, social science, and CSCW as it relates to trust and 'pathological epistemic phenomena.' It relates the notion of 'epistemic distrust' with misinformation, echo chambers, and democratic (il)legitimacy. We contrast it with common approaches to misinformation research in CSCW and argue it complements these approaches. Second, we demonstrate the applicability of this synthesis by operationalising the notion of epistemic

trust and provide evidence that it can be measured in real communities, connecting it with other CSCW research. We do so by describing a novel, to our knowledge, NLP task. Third, we discuss the implications and identify further research directions.

Road-map. Section 2 and 3 detail theoretical motivations and contributions. The key aim is to argue that epistemic trust is theoretically linked with 'pathological epistemic phenomena': problems such as democratic (il)legitimacy, misinformation, and echo chambers. To do so, we need to discuss the field of social epistemology and the concept of epistemic trust. We then need to discuss each of the pathological epistemic phenomena above in turn and how they relate to epistemic trust. Section 4 presents a method to measure epistemic trust and provides evidence for validation. Section 5 investigates the results from the measurement model applied to social media data. It also serves as further validation. Section 6 discusses findings and policy implications. Section 7 concludes with limitations and future work.

2 Background and Theoretical Motivations

2.1 Social and Political Epistemology

Briefly, epistemology (from Greek 'episteme,' or knowledge) is the philosophical field concerning knowledge, justification, belief, and related prerequisites like truth or evidence. While the term itself is not more than two hundred years old, the field 'is at least as old as any in philosophy' [91].

Epistemology historically and typically focuses on the individual [29]. A relatively new field, social epistemology introduces social concerns into the field. Political epistemology may be understood as a subfield of social epistemology, though the intersection of political philosophy and epistemology has been a concern since at least Plato. Edenberg and Hannon [22] argue that it has 'only recently been recognised as a distinctive subfield' due to the recency of social epistemology itself, and as recent political developments (e.g. Trump) provide significant and popular impetus.

Prima facie, epistemology is relevant to government because good decisions require good information. Of course, intersections between epistemology and governance are not new. Plato argues in the *Republic* that states should be governed by philosopher kings due to the virtue of their knowledge and wisdom and its application to statecraft. The more modern liberal (as in Liberalism) assumption around free discourse was notably described by J.S. Mill in *On Liberty*. According to Mill, censorship is wrong not just for instrumental reasons (i.e. free speech and debate encourages truth and discourages falsehood) but that it prevents fulfilment of our truth-seeking goals as a 'progressive being.' More modern liberal scholars generally follow this line of reasoning though varying on their stance on regulation and connection to higher moral principles (for something canonical, see Rawls [77] p. 197).

Neither are we limited to philosophy in the analytic tradition. Habermas' 'Discourse Principle' [10] states that some choice or action 'is justified only if those affected by [it] could accept it in some reasonable discourse.'

Anderson [4] understands democracy as an epistemic system, where 'the problems we need to solve demand the utilization of information that is highly dispersed across society.' This way of framing democracy is not dissimilar from the way economics frames markets as information aggregators. The price signal in our case is the vote, or perhaps the conversation. Anderson thus models democracy in three ways: Condorcet's Jury Theorem, the Diversity Trumps Ability theorem [37], and Dewey's experimentalist model. While the argument is interesting, the core for us is that '[to] realize the epistemic powers of democracy, citizens must follow norms [including those that] institute deliberation and reason-giving.' Indeed, the legitimacy of law itself depends on some form of justification. Brennan [12] criticises this justification on the grounds that it relies upon a strong assumption that voters genuinely hold (empirical and normative) beliefs and are rational. Brennan

argues that many voters instead vote (or view something as legitimate) for identity reasons. de Ridder [18] argues that *deep disagreements* undermine the the assumption of reasonable debate crucial to many epistemic accounts of democracy. These disagreements can involve fundamental and long-lasting disagreement on epistemic and moral norms. For example, what counts as evidence (typically, whose testimony), or positions on abortion. Clearly public justification depends on shared epistemic norms, but de Ridder argues deep disagreements cause parties to 'easily ... think the other less cognitively virtuous,[1] fundamentally misguided, or badly informed.' de Ridder argues this leads to 'cognitive polarization', and leads us to discuss epistemic dependence and trust.

Epistemic trust, and the epistemology of testimony generally, are the key concepts from social epistemology this thesis is built around. Briefly, epistemic trust is the likelihood that a listener will accept (or deny) a speaker's testimony, based on perceptions of their *competence* or *sincerity*. Before detailing it further, we make the case that it is important for understanding several important epistemic phenomena.

2.2 Pathological Epistemic Phenomena and Why They are Generally Considered to be Bad

Pathological epistemic phenomena is a term used here to unite a few social structures or properties understood to be related to a 'post-truth' society. The discussion below will enable us to skip further use of that label and ground ourselves more surely.

Three pathological epistemic phenomena are of concern. First is democratic or institutional illegitimacy. Second is misinformation or disinformation. Third are echo chambers or filter bubbles. Our core argument is that the notion of epistemic trust is crucial in all three phenomena. We focus here on their epistemic content.

2.2.1 Democratic and institutional (il)legitimacy. Our account of democratic legitimacy will largely follow Larmore [50]. Consider what Larmore calls the 'circumstances of politics': fundamental and endemic reasonable disagreement over the right and good. Each of us has reasonable differences in our positive and normative understandings of justice, distribution of resources, the role of government. We cannot simply wish these circumstances away or appeal to some a priori true moral framework on which to converge. We must be able to live together, and ideally, attain levels of cooperation that make complex economies and societies possible. The task of a state is to ensure this, via coercion, which requires *legitimation* via a *legitimation story*.

A state is legitimate when people generally *believe* it is entitled to enact laws to govern them. Indeed, legitimacy generates an obligation for citizens to abide (or at least, not undermine) those laws. *Perceptions* of legitimacy alone generate authority, but if the legitimation story is truly accepted then legitimacy is a stable. It provides a robust foundation on which to coordinate, cooperate, and build complex societies. Extreme costs of oppressive coercion need not be paid as everyone is, so to speak, along for the ride. Note that authority secured by coercion alone — authoritarian oppression — does not deliver this stability (and hence is not legitimate).

Under what conditions should this legitimation story be accepted? When those subject to fundamental principles can (but not necessarily do) see reason from their perspective to accept them. That is: fundamental principles are *justifiable* to a person with *their own perspective*.

The emphasis above is not random. Each phrase emphasised has important epistemic content. A person is unlikely to see reason from their perspective to accept, say, that an election was free and fair when their perspective involves massive government conspiracy. Justification relies on *testimony*. Perspectives do also and may also be subject to echo chambers or misinformation. This

[1]Epistemic virtue and vice is related with the acceptance or belief in misinformation: [59, 60]. Many of the recommendations in [21] amount to epistemic virtue.

is where the philosophical rubber hits the pragmatic road. Fundamental political principles with some sort of legitimation deficit are subject to legitimacy risks. Arguably, this characterises the political climate at the time of writing. It may be that reduced levels of cooperation, coordination, and political stability are because the epistemic foundations of legitimacy are eroding.

We need not constrain ourselves to fundamental political principles either. Specific institutions can experience legitimacy risks. Tucker [95, 96] is informative here. If they are to be stable, institutions require 'incentives-values compatibility.' For our purposes, values-compatibility means that an institution can *justify* itself in a manner similar to the above. Central banks are a useful example as they rely primarily on *credibility* (and are the subject of [95]). If they experience a negative legitimacy shock, or the fundamental principles do, then they are unlikely to work. Indeed, international institutions and cooperation can be conceived of in similar terms [96].

Nor do we need to constrain ourselves to fundamental political principles (roughly, a constitution). Estlund [24] asserts[2] that 'democracy is roughly a case of *symmetrical legal subjection*': that is, each citizen is subject to the coercive authority of all other citizens collectively. We are legally and coercively dependent on each other based on our moral equality. This moral equality implies equal input into law- and decision-making processes, implying a *symmetrical epistemic dependence*. Reglitz [79] makes this point. Since laws require justification, and evidence, we inevitably rely on the testimony of our fellow citizens. The ultimate legitimacy of law relies on our *epistemic trust* in other's testimony — that we consider it reliable and given in good faith. The relationship is complex. Fuerstein [26] argues that justification,[3] requiring certain epistemic standards, is a necessary condition for (warranted) epistemic trust. Fuerstein argues that political epistemology is akin to a prisoner's dilemma, where skepticism is warranted. In his account, justification serves as the cooperation technology. In any case, epistemic trust and legitimacy are closely connected theoretically.

Finally, it is important to note that trust is crucial not only for the legitimacy of the state, institutions, and law alone. Trust underlies cooperation (which legitimacy secures), enabling more complex social and economic arrangements. Declines in trust are likely to reduce living standards and increase the likelihood of political volatility and even violence.

2.2.2 Misinformation and disinformation. For our purposes, misinformation is any information that is false. The term *disinformation* is often used to denote some malicious intent behind its spreading, but it can usefully be grouped with misinformation. The use and abuse of misinformation has a long history but has, in the last decade, become one of the more important policy issues. The World Health Organisation declared the COVID-19 pandemic an 'infodemic' [17] due to the damaging role of misinformation. The average US adult was exposed to and remembered one or more pieces of misinformation during the 2016 Presidential election [2]. Misinformation has contributed to the January 6 Capitol attack and the Russia-Ukraine war. Foreign disinformation or influence operations are well documented [90]. Clearly false information can put lives and livelihoods at stake.

Misinformation is aided and arguably abetted by social media. Vosoughi et al. [98] found that misinformation spread faster and reached more people than the truth. It can be finely targeted [56]: see the Cambridge Analytica scandal. 70 per cent of Americans view the spread of misinformation online as a major threat as of 2022 — and only 4 per cent view it as no threat at all [87].

Most literature focuses on *belief* as the ultimate harm that misinformation causes. That is, they take the Millean view that systemic epistemic problems stem from a decrease in the ratio of truth to falsehood. Per Vosoughi et al. [98], '[foundational] theories of decision-making, cooperation,

[2]Following Viehoff [97].
[3]Rather, the 'Liberal Principle of Justification', see Rawls [78] and Larmore [50].

communication, and markets all view some conceptualization of truth or accuracy as central to the functioning of nearly every human endeavour.' Per Alcott and Gentzkow [2], fake news 'imposes private and social costs by making it more difficult for consumers to infer the true state of the world.'

Consider again the increase in the ratio of falsehood to truth. Excusing the problematic metaphor of a marketplace of ideas, we might consider this an example of asymmetric information. It is a market for lemons scenario: if falsehood is so common, who can be trusted and what is true? Starbird [90] notes how disinformation campaigns can operate this way by sowing doubt in others and ourselves.

Theory from social epistemology suggests that doubt alone — epistemic distrust or testimonial skepticism — is all that is needed for harm. Belief is not actually required: *perceptions* suffice. Per Reglitz [79], if citizens believe misinformation is effective and widespread — even if it is not — then this undermines epistemic trust between citizens. If we believe that a significant proportion of those we are legally and epistemically dependent upon are likely to fall for brazen falsehoods then we are less likely to view them as epistemic equals. To put it another way, *we stop viewing other citizens as legitimate*. Rini [81] provides a similar account. Rini suggests that due to the poor operational security of the Russian disinformation campaign[4] it may have been primarily acting through this epistemic trust channel. Russia's strategy, according to Rini, is to cause a sort of epistemic collapse based on inducing testimonial skepticism through 'testimonial sabotage.' That is, plant some obvious untruths (the more incoherent the better), expose the fakes, and further channel them along partisan divides (as trust in testimony is related to identity). Note that this is precisely the same as epistemic trust, although Rini is less concrete about the role of perceptions. Lynch [54] explicitly links perception of disagreement to 'cognitive polarisation': the mere perception of deep epistemic disagreement (potentially caused by identity-expressive speech sowing confusion or outright nihilistic trolling) can lead to belief polarisation.

Perceptions can be demonstrated. Knuutila et al. [49], via a survey of global risk perceptions, find that while people are worried about misinformation, it is unrelated with the actual prevalence of it. Almost all Americans view it as a threat [87]. Lima et al. [51] provide breakdowns of what actors users tend to blame for creating, disseminating, or failing to prevent misinformation.

Finally, we shouldn't immediately identify expression with belief: there is evidence of a disconnect. Hannon [34] argues that many perceived political disagreements (including factual ones) are identity-based cheerleading, or cheap talk. While partisanship affects factual *expressions*, Bullock et al. [13] conclude from experiment that a partisan gulf in actual *belief* 'may be more illusory than real.' A political expression, even of fact, may be less about a careful evaluation of direct or testimonial evidence and more about cheering on your side like a sports team. Pennycook and Rand [71] report a disconnect between belief and sharing behaviour is largely driven by inattention. Fundamental attribution error in recipients of sharing may allocate malice where inattention is more appropriate.

2.2.3 Echo chambers. Parisier [68] describes a 'filter bubble' as an algorithm that filters information based on our personal tastes. It isolates us in our own information world. Typically also known as 'echo chambers', they operate via recommender systems on the supply side and selective exposure on the demand side. The ultimate effect is that users of social media are informationally segregated into communities that agree with each other. If we no longer share facts, let alone values, how is democracy to work?

[4]For an overview, see Howard et al. [39]. Tellingly '[s]urprisingly, these campaigns did not stop once Russia's IRA was caught.'

Inhabiting an echo chamber can result in group polarisation. Sunstein [92] highlights two reasons: reputation-seeking behaviour and limited argument pools. It's important to note here that what Sunstein calls 'enclave deliberation' is not prima facie bad — take the example of low-status or alienated individuals who might otherwise be ignored. Echo chambers today are qualitatively different. Like misinformation, they have been supercharged by recommender systems and social media.

The literature generally takes selective exposure as the fundamental cause of echo chambers. Nguyen [63] provides useful nuance. He separates information bubbles based on selective exposure from those based on epistemic trust. In this account, 'epistemic bubbles' occur when different information is excluded (potentially by accident) and may be punctured merely by introducing them. These are distinguished from 'echo chambers', which function by systematically discrediting the testimony of those outside of the group. That is, they induce *epistemic distrust* in sources of information outside the bubble, and in this way they enforce beliefs via *evidential pre-emption* ('they would say that, wouldn't they?'). Indeed, it is possible for an agent inside one of these structures to act in an epistemically virtuous way (e.g. seek out evidence) but fail to update their beliefs away from those of the group. The 'QAnon' phenomenon is likely an example of a Nguyen-type echo chamber.

Evidence is mixed on echo chambers. Del Vicario et al. [19] find echo chambers in conspiracy and science news. Many studies of social media find these sorts of communities [93]. Neo-nazis openly congregate on platforms like Telegram [33, 99]. Small pieces of content like push notifications contribute to echo chamber effects [83]. Given that many Americans at least sometimes consume news on social media [73], echo chambers here are an issue. However, Dubois and Blank [20] find that those interested in politics or who consume diverse media avoid echo chambers. Nelson and Taneja [62] find that the fake news audience are also exposed to media the rest of the online population consumes.

Indeed, the community structures indicative of echo chambers exist on social media. But it does not follow, even theoretically, that inhabitants will not be exposed to information that might pop their bubble. It is likely that their persistence is better explained with in-group reputation effects or Nguyen-type epistemic trust effects.

2.2.4 Pathological epistemic phenomena are interrelated. It is hard to tease apart each of the above phenomena. Each is highly endogenous. Echo chambers, and polarisation generally, contribute to the spread of misinformation [19, 67]. Efstratiou and De Cristofaro [23] argue that polarisation and misinformation adherence are closely tied. Misinformation, via selective exposure, might be causal of echo chambers. Both can undermine democratic or institutional legitimacy by undermining the justification it requires, or making some perspectives unreachable. Clearly, inhabitants of the QAnon echo chamber do not view the government as legitimate.

The fundamental theoretical argument of this paper is that these phenomena are usefully understood through the lens of *epistemic trust*. It underlies legitimacy through *justification*. It can cause echo chambers to be resistant to intervention. A breakdown in trust is a direct target, and probable cause, of misinformation. Combined with epistemological theory, its measurement is a step towards greater understanding of these phenomena, and ideally better policy.

2.3 Most Research on Misinformation Focuses on Belief and Accuracy

Misinformation is by far the most studied of the pathological epistemic phenomena above. While the notion of epistemic trust (if not by that name) is not novel to social computing, we argue that a focus on it suggests novel interventions and experiment. A direct focus on democratic and institutional legitimacy is, to the best of our knowledge, novel.

As above, the primary harm underlying most misinformation research is the fact that it is both believed and untrue, or at least unverified. There is a focus on accuracy, reflected in the goal of misinformation detection and (potentially automated) fact checking.

Ecker et al. [21] provide a recent review of the psychological drivers of misinformation. It focuses on the *belief* of misinformation, rather than any knock-on effects it may have on epistemic trust. In another review, Pennycook and Rand [71] find that 'poor truth discernment is associated with lack of careful reasoning and relevant knowledge,' in addition to cognitive heuristics.

Juneja and Mitra [44] detail the human and technical infrastructure underlying fact checking. They also report that fact-checkers are distrustful of automated approaches. Micallef et al. [61] similarly focus on fact-checkers. He and He [35] do the same for a community of debunkers on Reddit. Kaufman et al., [46] find that crowdsourced detection of truthfulness is effective. Lu et al. [53] find that an AI labelling news as fake or not nudges people into agreement (even if it is wrong). Jahanbakhsh et al. [41] add accuracy assessments to a social media platform, but notably also include trust in specific users. There is a vast literature in NLP on detecting misinformation and fake news, using both content, social context, and external information [40, 45, 64, 86]. Pennycook et al. [70] find the 'illusory truth effect': prior exposure to fake news increases perceptions of accuracy.

Algorithmic auditing of recommender systems is a line of research directly related to echo chambers, although it is common to motivate it with misinformation. Srba et al. [89] find that it is possible to burst recommender-induced epistemic bubbles (per Nguyen, though the authors use the term 'filter bubble') on YouTube. This is done by watching content debunking misinformation.

2.4 Intervention: If We Tell Grandpa it is Untrue or to Stop Watching Internet Weirdos all Our Problems shall be Solved?

Literature around interventions also generally concerns perceived accuracy and correction or so-called 'backfire effects.' Nyhan [66] finds that corrections are somewhat effective, but importantly discusses the role of the epistemic environment and group identity, Ecker et al. [21] largely dismiss backfire effects.

In their review, Efstratiou and De Cristofaro [23] suggest several interventions. In particular, they are mostly focused on reasoning and accuracy: reducing cognitive load, priming for accuracy, avoiding cognitive biases in design. Some focus on the quality of the deliberative environment or the prevalence of views generally. Ecker et al. [21], in addition to debunking, suggest 'prebunking,' which acts to immunise against misinformation.

We argue that epistemic trust can complement a focus on accuracy-based intervention. Of course, accuracy is a critical part of any epistemic system. However, there is a disconnect between actual belief and expression (or sharing) [13, 14, 34, 71, 76]. Whether this stems from identity or inattention, it suggests there will be a non-trivial group who will not be reached by accuracy-based interventions. Petersen and Osmundsen [72] find that some individuals spread 'hostile rumours' because they have a nihilistic 'need for chaos' driven by social marginalisation. Explicitly, this '[implies] that the ultimate policy solution ... does not lie in fact-checking or small nudges.' Fact-checking also depends on epistemic trust in the fact-checkers. If they are viewed as compromised then fact-checking will not work. If it is true that some inhabit Nguyen-style echo chambers [63], then as their epistemic trust in external sources of information is compromised, fact-checking will not work. Disinformation operations can operate by inducing testimonial skepticism, or a general reduction in epistemic trust. This will reduce the effectiveness of fact-checking. Finally, per Reglitz [79], the *perception alone* that misinformation is a problem and that our fellow citizens may be fooled is enough to reduce epistemic trust generally.

This suggests that the current accuracy-first approach is best complemented. Indeed, there are limits to machine detection of misinformation in the presence of disagreement [31]. A focus on epistemic trust and its measurement will allow us to build tools that can observe, at a community or user level, a fundamental driver of pathological epistemic phenomena.

3 Epistemic (Dis)Trust (or Testimonial (In)Credibility)

Fuerstein [26] defines epistemic trust as the propensity to accept a speaker's claim due to belief in their *reliability or competence* (the likelihood that their claims are true), and their *sincerity* (the likelihood that they express their belief accurately and in good faith). Audi [5] describes it in the same way, though he uses the term *testimonial credibility*.

Consider the utterance 'Republicans are stupid.' This is an example of epistemic distrust based on competence. 'Democrats are liars' is an expression of epistemic distrust based on sincerity.

How does epistemic trust operate? Testimony is a crucial source of belief (and knowledge) [5]. But not every piece of testimony we hear forms a belief in our mind. If I hear that 'the French Revolution was a secret plot orchestrated by the lizard-people,' I am certain not to believe it (in part because shit happens [55]). The testimony has been effectively *filtered*. In the case where I have low epistemic trust in the speaker, I am liable to filter any testimony they make. This is the claim of *effective filtering*: that this reliably prevents beliefs based on false testimony, following Grodniewicz [32]. The process is complex — Grodniewicz builds on both epistemological theory and empirical results to argue that it is not effective in real-time but it is effective in the long-run. Of course, if our view is that the harm of misinformation comes from belief then the filtering hypothesis should be augmented by socio-affective drivers [21]. But, as we have argued above, actual belief may not be necessary for harms to result.

Epistemic trust has a target: it is directed at someone or something. Consider again the sentence 'Republicans are stupid.' The entity 'Republicans' is the target of distrust.

3.1 Is Epistemic (Dis)Trust Measurable?

We have argued that there is a disconnect between belief and expression. Since we are considering Reddit, where users cannot simply click 'share' to share content, the inattention explanation [71] is less concerning. We must contend with the assertion that the expressions may not be sincerely held and perhaps spread by nihilistic trolls or by deliberate influence operations [72, 81]. Although arguably both motivations do indicate a lack of trust in 'the system,' broadly defined.

First, people who do not filter testimony of epistemic (dis)trust may still well incorporate it into their beliefs. In this sense, the underlying motivation of spreading distrust may be less important than sheer volume of expression. This lack of filtering may be aided and abetted by echo chamber and group polarisation dynamics [92]. Reputation-seeking behaviour, a desire to fit in, or even to express the perceived in-group identity [34] may compromise filtering. Continued exposure to these beliefs via the illusory truth effect [70] may transform their sincerity. Per Audi [5], testimony does not require sincerity to be incorporated as a justified belief.

Second, even though these expressions may not be sincere (for whatever reason), it may be enough to motivate action.

Third, if we are not able to defeat the concerns, then we should view measurements of epistemic trust as being biased upwards. Under a strong assumption of uniform insincere participation across communities, we would still be able to rely on relative measurements if not the precise one. Dropping the assumption would require priors about nihilistic trolling and disinformation operations, which are possible to construct.

Of course, this assumes that everyone who holds their (dis)trust sincerely has some similar baseline in social media participation and the propensity to express their distrust. Measurements on

social media, whatever their bias, should not be confused as measurements of (dis)trust in society generally.

One further concern is that trust may not be as measurable as distrust. Expressions of (dis)trust are likely to be sparse and driven by salience. While I (presumably) have well-formed and justified views on the trustworthiness of many sources, I could not enumerate and express them all without prompt. Ultimately, the internet is not a place where one may find expressions like 'I find the government trustworthy' and be able to accept it without a hint of irony or suspicion of sarcasm.

Of course, the more general concept of trust has a long history of measurement. Per Bauer and Freitag, this is often achieved through surveys developed in the mid twentieth century [7]. This is despite innovations with game-theoretic experiments (e.g. [9]), implicit association tests, etc. Compared to these efforts, we use a more limited definition of trust and measure with social media comment data in the wild.

Despite the challenges, we argue that we can both operationalise and measure epistemic trust.

4 Measuring epistemic trust

4.1 Sentence and Community Data

We use data from Reddit over the period of January and February 2020. Reddit is a social media platform similar to a forum. It is organised into separate communities called 'subreddits' (denoted by an 'r/', then the name), sometimes with sharp behavioural and norm (including epistemic) discontinuities between them. For example, decorum and the type of evidence acceptable in r/AskHistorians differs strongly from r/Conspiracy. Activity from any given user account can be tracked over time including all communities they participate in. Data are available via the Pushshift API or historical archive [8]. In this section, we focus on *sentences*.

Reddit is a large text corpus. For one month alone, compressed, all comments and submissions to the platform total around 30 gigabytes. The platform itself is growing, so the size is increasing over time. Therefore, it is impractical to use the entire corpus. To efficiently use these data, we select a subsample.

Considering the subject, political subreddits are the most natural subsample. It is unlikely that any user on a subreddit dedicated to cats in amusing positions is trading government conspiracies. Hoffman et al. [36] construct a set of political subreddits they term the 'Reddit Politosphere.' However, extremist subreddits, which are generally part of the political subsample, often get banned. To control for this (and the fast-moving nature of Reddit generally) we expand the seed list by using user-subreddit overlap statistics: notably the probability that a user posts in subreddit A given they participate in subreddit B. We take the top 20 similar subreddits after a post frequency-inverse poster frequency reweighting. Jaccard similarity would also be appropriate here. Ultimately, 954 communities are sampled.

Finally, all comments made in the expanded set of political subreddits in any given month are split into sentences. Emojis, URLs, and non-alphanumeric characters are removed.

4.2 Approach

We build an epistemic (dis)trust classifier at the sentence level. No labelled dataset for epistemic trust exists. This makes training and validation difficult. We adopt a weakly supervised classification approach using techniques from information retrieval and natural language inference[5]. We then validate a small sample manually. One might consider this an approach to bootstrap, given appropriate human labelling, into a gold-standard dataset for epistemic trust. We leave this task for

[5]Code will be available at https://github.com/dzjones/epistemictrust

future work. Nonetheless, replicating other successes in weakly supervised text classification, the distrust models perform well.

Classification of sentences or documents usually requires a large dataset with human verified gold-standard labels. Weakly supervised text classification does away with this requirement. Taxonomies, relating concepts and examples in a tree structure, are helpful [57, 85].

Information retrieval. If we have a strong prior about the semantic structure of expressions of epistemic (dis)trust, information retrieval is useful. These priors are realised in the form of a taxonomy. We take sincerity and competence as taxonomic children of epistemic (dis)trust generally. Trust and distrust, both for sincerity and competence, are thus grandchildren nodes of epistemic trust. Underneath these categories lie specific keyphrases. These priors, given a form as a taxonomy, can then be rendered into *queries* (using the keyphrases) for an information retrieval system. The full taxonomy is detailed below.

We use SentenceBERT (SBERT) [80], in particular the `msmarco-distilbert-base-v4` model, to generate embeddings at the sentence level. This model is trained on the MSMARCO question-answering dataset [6]. This model is particularly useful as it is 'asymmetric': we are not looking for documents that exactly match a query given the many ways it can be expressed. The dataset contains real user search queries on Microsoft's *Bing* search engine. Embeddings are normalised to the unit sphere.

4.3 Taxonomy

4.3.1 Competence. This category tracks whether the speaker is expressing (dis(trust in the competence of something (i.e. it is (un)likely what they say is true). The keyphrases for competence distrust are: 'stupid', 'incompetent', 'ignorant', 'idiot', 'sheep', 'insane', 'moron', 'dumbass', 'clown', 'living in your bubble', 'incoherent', 'nonsense', 'accept reality', 'irrational', 'retarded'[6], 'no proof', 'intellectually dishonest', 'misleading', and 'indoctrinated'.

Some (e.g. 'stupid', 'ignorant') are straightforward. Some require additional explanation. 'Sheep', for example, is a term commonly used to imply someone follows the crowd, or worst, has been deceived (see also 'indoctrinated'). 'Living in your bubble', 'accept reality', and similar phrases are often used when a target is being accused of being in an echo chamber, expressing false testimony. Interestingly, while 'no proof' and 'intellectually dishonest' might seem a more natural fit in sincerity, sentences matching these queries tended to express competence-based distrust.

For trust, the phrases are: 'intelligent', 'reasoned', 'informed', 'evidence', 'accurate', and 'truthful'.

4.3.2 Sincerity. This category tracks whether the speaker is expressing (dis)trust in the sincerity of something (i.e. they are likely to be lying, or they do not truly hold the belief expressed). The keyphrases are: 'liar', 'dishonest', 'untrustworthy', 'corrupt', 'inhuman', 'immoral', 'disinformation', 'misinformation', 'propaganda', 'fake news', 'unreliable source', 'paranoid', 'shill', 'bias', 'discredited', 'manipulated', and 'scam'.

As above, phrases like 'liar' or 'dishonest' should be clear. Included are several phrases related to misinformation: literally 'misinformation', 'fake news', 'propaganda'. While information that is incorrect may be shared inadvertently, the use of these terms typically is to cast suspicion on the sincerity of the sharer. 'Corrupt', 'inhuman', and 'immoral' are not direct epistemic qualities, but typically imply some malign intention of the target. 'Unreliable source', and 'discredited' are more often used to describe a source of information, though this can often blow back on a speaker (whether by deliberate intention of the accuser, or by the fundamental attribution error).

For trust, the phrases are 'trustworthy', 'reputable', 'sincere', and 'genuine'.

[6]Unfortunately, this term is still common online.

4.4 Training

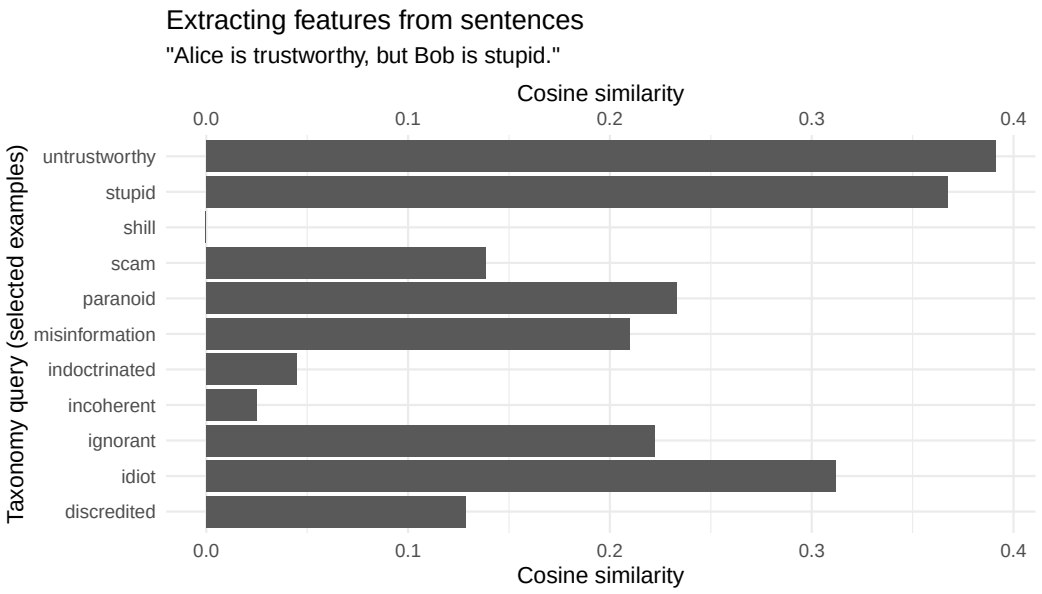

Fig. 1. Example (truncated) vector resulting from computing cosine similarities between query vectors computed by SentenceBERT. These cosine similarities serve as features for classification.

The SBERT model is applied to every sentence in the training set, resulting in a 784-dimensional embedding. Each keyphrase in the taxonomy (corresponding to a leaf) is also fed through the model to form a query embedding. This allows us to compute the cosine similarity between every sentence embedding and every query embedding (Figure 1).

The aim is then to construct a binary classifier for each of the four categories of epistemic (dis)trust: trust based on competence, distrust based on competence, trust based on sincerity, and distrust based on sincerity. To do so, we first construct a training dataset for each classifier based on the query similarity information.

A given classifier (e.g. distrust based on competence) corresponds to a set of query leaves in the taxonomy (e.g. 'stupid', 'incompetent', etc.). Positive examples for each classifier (e.g. expressions of distrust based on competence) are formed by sentences that have a cosine similarity of at least 0.5 for any one of its associated query leaves. This is a rough threshold and forms one hyperparameter of the model.

The set of positive sentences may be small. To expand the set — and to fill in semantic gaps that the taxonomy queries may have missed — we use k-nearest-neighbour data augmentation. That is, we find the k nearest sentence embeddings to every sentence in the positive example set. This process is slow, but significantly accelerated by FAISS [43], a library providing efficient and GPU-accelerated vector similarity search.

Finally, negative examples are sampled uniformly at random of the sentences not included in the positive set. Since the semantic space of expressions of not-epistemic-trust is so large, we oversample the negative set by a factor of 20 to 1. This does not harm the in-sample model performance results.

Once a dataset is prepared, it can be fed into any classifier. Logistic regressions and random forests [11] work well. The model performs well using the cosine similarities between the sentence

Table 1. Sentences in each training data category. Expansion positives refers to the additional sentences found via nearest-neighbour search. Negatives are (over)sampled uniformly at random from sentences not in the positive set.

	Competence		Sincerity	
	Distrust	Trust	Distrust	Trust
Positives	192,301	98,684	130,991	8,236
Expansion positives	256,398	110,224	185,421	13,139
Negatives	3,927,002	1,934,756	2,568,877	167,965

and the query embeddings as a form of feature engineering. Given resource constraints, and the large size of the data, this is effective.

The training data for the model was the expanded political subreddit set for January 2020. Size statistics can be found in Table 1.

4.5 Validation

Once trained, the model was applied to the expanded political subreddit set for February 2020. For validation purposes, the model has not seen any of the expressions.

Table 2. Example sentences. An 'x' indicates a positive classification by a random forest.

(Notion)	Competence		Sincerity	
(Polarity)	Trust	Distrust	Trust	Distrust
you are ignorant and a liar		x		x
democrats are idiots		x		
democrats are liars				x
republicans are dumb		x		
you're smart	x			
democrats are intelligent, trustworthy	x		x	

Table 2 reports classification results for the random forest models for constructed sentences. The sentences were chosen to be clear examples of epistemic (dis)trust. The model successfully classifies each into the correct category.

More comprehensive validation requires a large validation set. Since no labelled dataset yet exists for the task, we manually labelled a small validation set. We did so for each of the four classifiers. Our prior was that the frequency of expressions of distrust is quite low, in general, which was borne out in the data (see Section 5). Therefore, we selected an equal fraction of sentences classified as positive examples and negative examples, effectively oversampling the positive class. We manually labelled 200 sentences each in the positive and negative class, so each classifier has 400 examples, and distrust overall having 800.

While labelling, several sentences might be considered borderline cases. Without greater context (impractical at scale, but perhaps ultimately required) we separated the validation into two cases. First, the conservative case where borderline sentences are labelled as negative. Second, the ideal case, where borderline sentences are labelled as positive.

The distrust models performed adequately, with most models reporting F_1 scores of 75 to 85 per cent (Table 3). This suggests that epistemic trust is indeed measurable. Ideal cases, by definition, outperformed the conservative cases, but the results remain in either case.

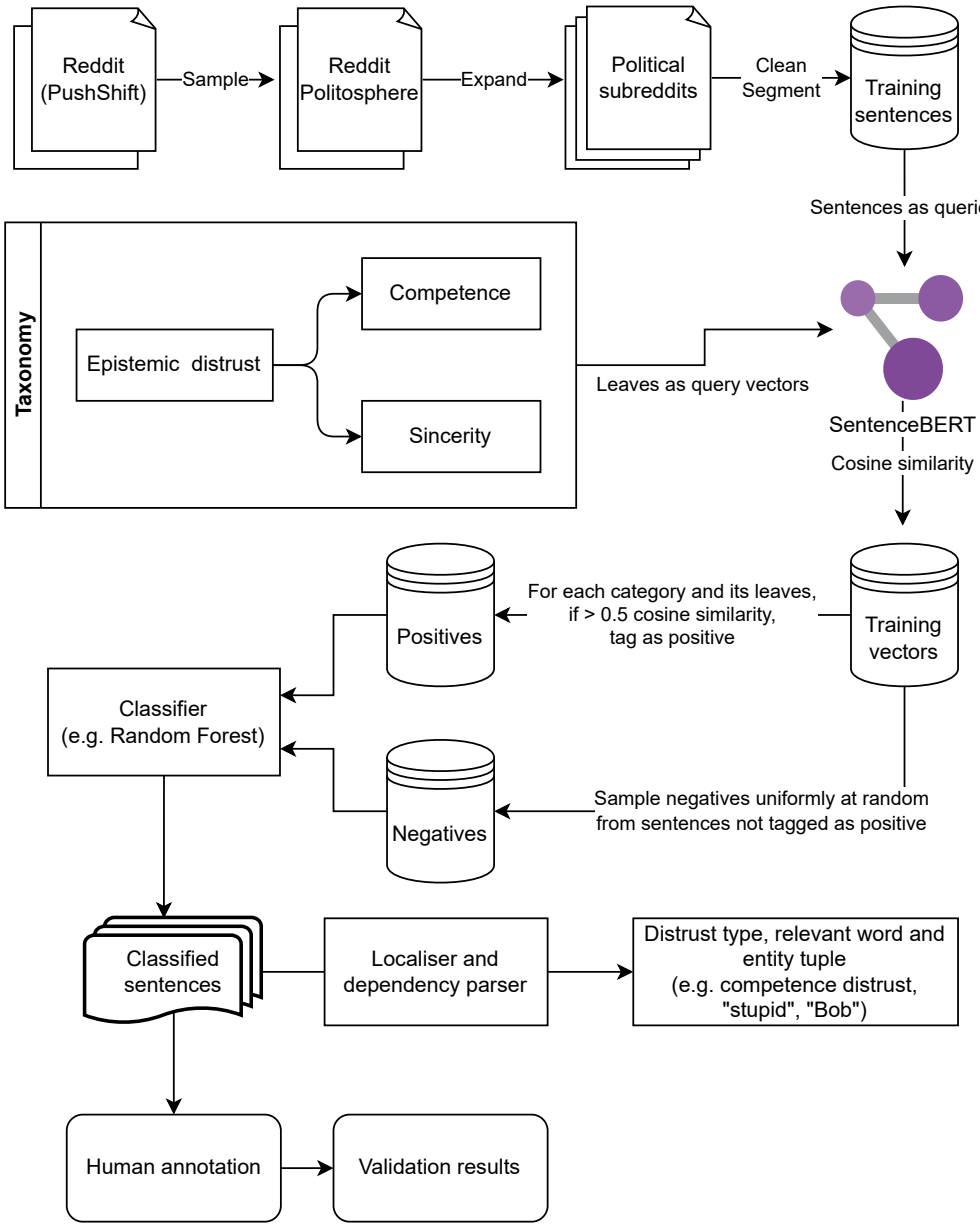

Fig. 2. Data and modelling flow diagram.

Notably, there was a large disparity between model classes. Random forests performed the best, outperforming multilayer perceptrons, XGBoost, and L_2 penalised logistic regression. This result holds if the multilayer perceptrons are trained on the raw embeddings — in addition, training a random forest is much faster. Indeed, even fine-tuning BERT[7] underperforms the models based on SentenceBERT cosine similarity features. This is not as surprising as it may seem: the other models

[7]More precisely, DistilBERT [84].

Table 3. F_1 scores for the competence and sincerity distrust models. Scores are reported as percentages, averaged over 10 separate estimations. Standard deviations are reported in brackets where available. Fine-tuned BERT results are from one estimation due to resource constraints.

	Competence distrust		Sincerity distrust	
	Ideal	Conservative	Ideal	Conservative
Random forests	84.1 (0.65)	78.4 (0.52)	80.6 (0.46)	74.9 (0.5)
MLP	75.4 (1.31)	71.1 (1.07)	64 (2.92)	60.8 (2.91)
BERT (Fine-tuned)	75	72	65	62
XGBoost	77.6 (1.19)	73.1 (1.67)	69.3 (1.86)	65.9 (2.18)
Logit (L_2 regularised)	80.4 (0)	75.1 (0)	76.2 (0)	72.5 (0)

act as a 'final layer' on top of SentenceBERT, whereas the fine-tuned BERT model is working with less information. It is likely that a task-specific deep architecture could outperform the random forest. That said, the simplicity and speed of the classical feature-engineering approach is desirable. Indeed, the adequate performance of a logistic regression indicates that this model should be included as a baseline for future improvements. The logistic regression also has the benefit of relative ease of interpretation.

We select random forests as the baseline model. One further validation step is to look at feature importances. The traditional way to do this for a random forest is to look at the mean decrease in impurity. Considering the weakly-supervised context this may not be appropriate. Instead, we opt for permutation-based feature importance. We compute importances using a permute-and-relearn method: first, permute any given feature, re-train the model, and compute the difference in objective function (for more discussion, see [38, 58]). The objective function is the F_1 score on the validation set, and so the importance measure is the difference in F_1 between the baseline and the feature-permuted model. Figure 3 shows the results. We do not use these results to refine the model as the validation set is small and this may introduce unwanted bias.

A few broad observations. First, query vectors from all categories seem important to both models, though precisely which ones differ across models. This increases our confidence that competence and sincerity are, while related, different measures. Second, models seem to use queries across the whole taxonomy. Third, there are a few some query vectors which don't seem to affect performance that much (e.g. 'intelligent'). Fourth, some query vectors improve one model but detract from another (e.g. 'ignorant').

4.6 Distrust is Measurable, Trust Less So

Above we report only distrust results. This is because the models trained on the trust taxonomy perform exceptionally poorly: an accuracy not meaningfully different from zero. This is likely because the culture of online posting is not conducive to these sorts of expressions — perhaps there is an atmosphere of cynicism, or perhaps there is a low level of general trust in politics [74]. It may be that when an entity is 'trustworthy' it is simply not salient enough to comment on.

This raises a question on how we are to operationalise trust. Given we can only effectively measure one direction (whether this is due to sparsity or some other reason), the use of relative or ordinal information (e.g. deviations from average, z-scores) are useful. This generates a hypothesis: subreddits with a strong epistemic culture that are not directly focused on some phenomena (e.g. fake news debunking) should rank low for epistemic distrust.

For future work, a larger human-labelled dataset constructed by several labellers would improve training and validation.

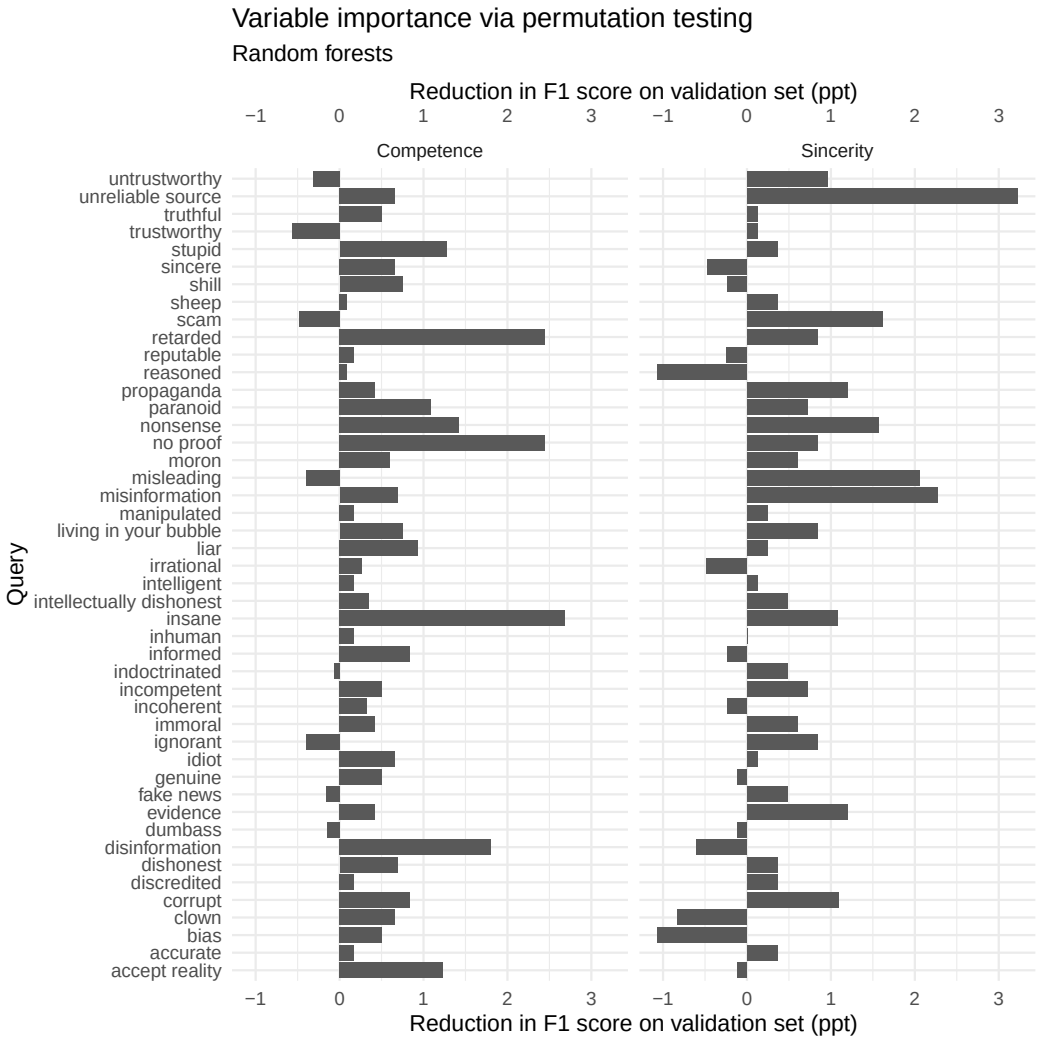

Fig. 3. Random forest permutation feature importances. Importances are the difference between the baseline F_1 and the feature-permuted and re-trained random forest. Queries were permuted in isolation. Query vectors from all categories seem important to both models, though some query vectors don't seem to affect performance that much. The precise effect differs between the models, increasing our confidence that we are measuring different sub-concepts.

4.7 Entity Resolution and Disambiguation

The final step in the pipeline is to extract the entity being targeted by the expression of (dis)trust. Co-reference resolution or named entity recognition are applicable here. However, there is one complication: One sentence may express two or more different forms of epistemic (dis)trust. For example: "Alice is trustworthy, but Bob is stupid." This sentence expresses trust in the sincerity of Alice, and distrust in the competence of Bob. To disambiguate, it is important to first localise key words driving the classification of epistemic (dis)trust.

To do this, we first localise the relevant key word of each classifier by replacing each word with '[MASK]' and highlighting whichever masked word caused the largest drop in classification probability. We then use the distance in a dependency parse tree between this word and nominal subjects to identify the target of distrust.

5 Ecological validation

Above we have argued that epistemic distrust is measurable and presented a reasonable measurement scheme. Here, we argue that this measure is useful when applied to actual communities. Recall that epistemic distrust is closely related to democratic (il)legitimacy, misinformation, and echo chambers. One would expect that higher levels of distrust are associated with more severe pathological epistemic phenomena. A key focus in this section will be the 'distrust density', or the fraction of sentences in any aggregation that are classified as distrust. There are three sensible ways to aggregate: by community, by user, and by the target of distrust. We consider each in turn.

5.1 Community (Dis)Trust Density

Figure 4 plots the measured distribution of epistemic trust on the political community sample. It plots the 'distrust density', or the share of positively-classified sentences of the total in a given community. As is common in work on social media, the distribution is heavily skewed: focusing on the horizontal axis, or the ranking of communities by distrust density, roughly one eighth of the communities have a distrust density above 3 per cent. In terms of the relative importance of competence and sincerity, shown in the colour breakdown of the vertical axis, it appears that expressions are roughly evenly split between the two. Sentences expressing both are in the minority.

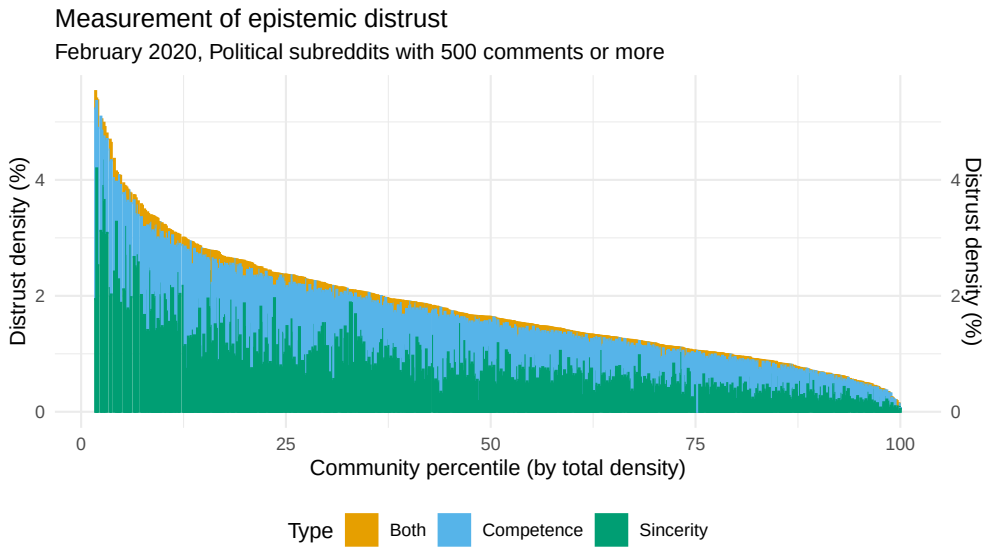

Fig. 4. The measured distribution of epistemic trust on subreddits with 500 or more posts in the month of February 2020. Like most social media data, it is heavily skewed.

A community must have at least 500 posts in the month of February 2020 in order to contribute to the results. Removing small or relatively inactive communities should increase our confidence that results are not due to noise. Figure 5 controls for this and shows that the distribution of distrust

density is broadly the same when we control for community size (as number of comments). That is, instead of the community density ranking, we scale each community by the share of comments in the corpus. A larger community will account for more of the horizontal axis than a small one. The median distrust density controlling for community size or not is a little below 2 per cent. Communities at the left end of the distribution — those with very high distrust densities — tend to be quite small.

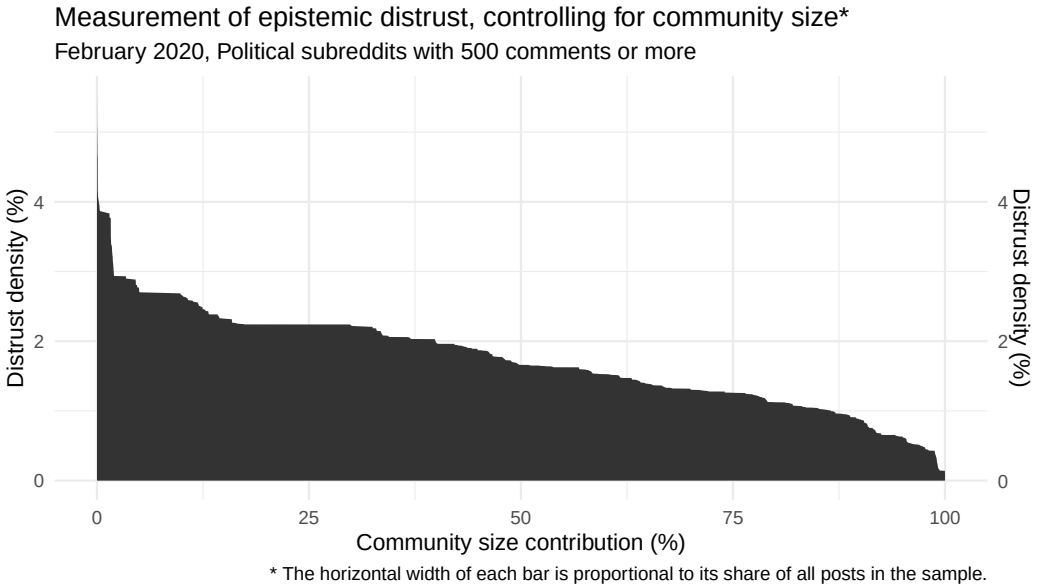

Fig. 5. The total distrust density distribution, controlling for the size of each subreddit as number of comments. This is qualitatively similar to Figure 4.

The question remains on how to measure trust as opposed to distrust (see paragraph in section 4.6). Community-based z-scores may be appropriate, or perhaps measures using quantiles. Figure 6 splits the communities into quintiles of their distrust density rank. It is worth summarising some stylised facts at this point informed by the above figures. First, expression of epistemic distrust is relatively rare (though not exposure to it), with densities of around 4.5 per cent in the most distrustful communities. Second, distrust based on sincerity is predominant, but not overly so, and there is significant variance across communities. Third, distrust based on sincerity seems to have a higher 'elasticity' than that based on competence. That is, we can see that communities in the highest quintile of distrust have average sincerity densities around 0.6 percentage points higher than average competence densities. However, the relationship flips for the lowest quintile: average sincerity densities are slightly below. As we move up the ranking to more distrustful communities, sincerity seems to account for more of the increase at any stage. Fourth, expressions of distrust tend to be of either competence or sincerity and only very rarely both, though not trivially rarely.

Why might more distrustful communities rely more on maligning sincerity rather than competence? To sketch a hypothesis: consider conspiratorial communities (e.g. r/conspiracy). Almost by definition these communities have low epistemic trust. Indeed, this aligns with the measurement: r/conspiracy is in the highest quintile with a distrust density of 2.9 per cent. Two thirds of distrustful expressions within r/conspiracy are due to distrust based on sincerity. The nature of a conspiracy

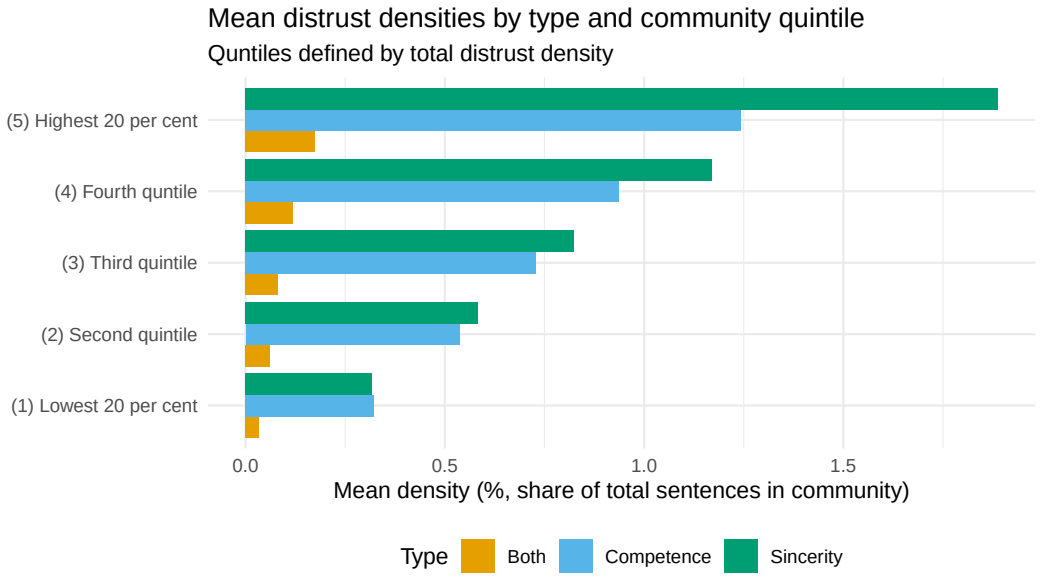

Fig. 6. Splitting the communities into quintile groups, we find a steady increase in expressions of distrust based on competence. Those based on sincerity, however, seem to respond more to rank.

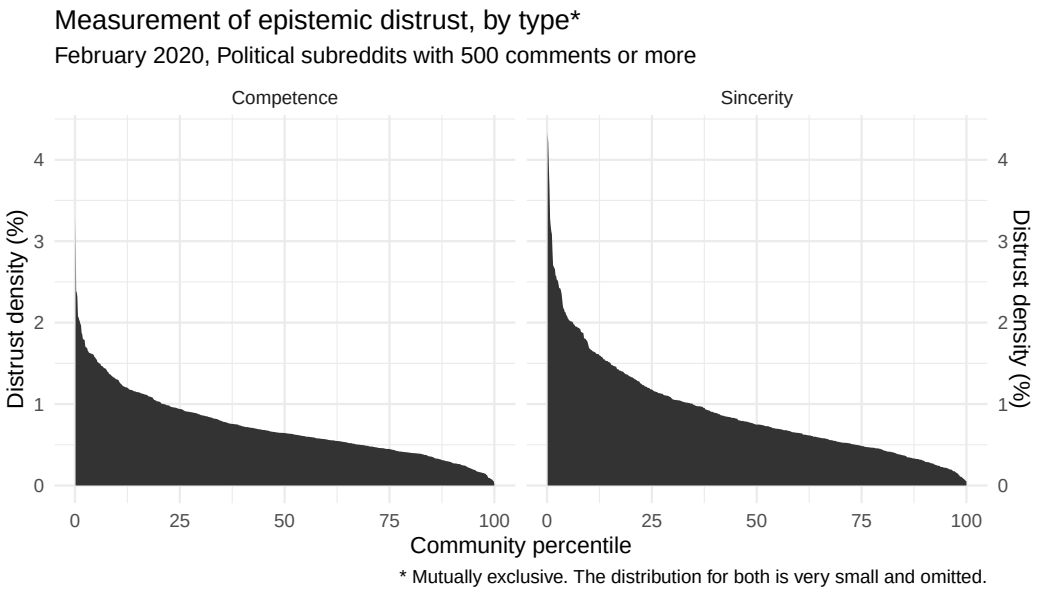

Fig. 7. The measured distribution of epistemic trust on subreddits with 500 or more posts in the month of February 2020, cut by type. The sincerity distribution has more mass and is more skewed than competence.

is less that those that control the world are stupid but more that they are malign, corrupt, and most importantly, lying. Distrust based on sincerity might be more responsive to overall distrust, and

account for the lion's share, due to conspiratorial traits becoming more likely. Of course, this is only a sketch, and correlational at this point, but worth considering for future research.

Figure 7 shows the distribution of distrust density again, with community ranking on the horizontal axis and disaggregated by type. The 'elasticity' of sincerity is noticeable in the distributions as well: note the sharper curve on the left of the sincerity distribution. The predominance of sincerity is also visible here: note the higher mass.

Characterising communities. Now we have established several stylised (statistical) facts of epistemic distrust, we should add qualitative colour. The following two tables characterise the communities at either end of the distributions in Figure 7.

Table 4. Top 10 highest and lowest subreddits for competence distrust density, or the fraction of all sentences classified as competence distrust

. Note the sample contains only subreddits with 500 or more posts in the month of Feburary 2020.

	Subreddit (Highest)	Competence (%)	Subreddit (Lowest)	Competence (%)
1	BernieSandersSucks	3.29	mmt_economics	0.00
2	NeverTrump	2.39	foreignpolicy	0.06
3	donaldtrump	2.37	AskHistorians	0.07
4	trumptweets	2.31	brealism	0.08
5	shitfascistssay	2.07	boltaction	0.08
6	daverubin	2.07	distributism	0.09
7	Full_news	2.03	europeanunion	0.09
8	LibertarianUncensored	2.02	GBPolitics	0.09
9	drumpfisfinished	1.98	transtimelines	0.11
10	Mueller	1.97	econmonitor	0.13

Table 4 shows the highest and lowest 10 subreddits by the fraction of sentences classified as competence distrust. High-distrust communities tend to be centred around Donald Trump (e.g. r/NeverTrump, r/trumptweets, r/Mueller), other 2020-election-related figures (r/BernieSandersSucks), or extreme political philosophies (e.g. r/shitfascistssay, r/LibertarianUncensored). It is, perhaps, not surprising that communities like this top the charts: Trump is a controversial figure, to put it lightly.

Communities that rank lowest for competence distrust tend to have a technocratic policy-wonk flavour (e.g. r/mmt_economics, r/foreignpolicy), have strong epistemic standards enforced by moderation (e.g. r/AskHistorians, r/brealism), or are centred around obscure political or economic philosophies (e.g. r/distributism, r/mmt_economics: this rough classification is not mutually exclusive). The common thread here seems to be strong epistemic communal norms or small, tightly-knit communities. That is: epistemic distrust is kept in check by external coercion or reputation. This is a connection to the evolution of trust in the prisoner's dilemma.

It is worth considering these low-distrust (or high-trust) communities in more detail. r/AskHistorians [28] is a community built around historical questions and their answers. Their rules wiki[8] contains extensive rules for how exactly these questions are asked and answered, enforced by moderation. For example, questions should not be 'loaded', encouraging debate and agenda-pushing more than disinterested information-seeking. Answers should be accurate and rely on good sources. These are all strong epistemic norms which likely contribute to the low distrust of r/AskHistorians. It also serves as additional validation of the theory and measurement. The r/brealism community is

[8]https://www.reddit.com/r/AskHistorians/wiki/rules/

similar. It is built around 'fact-based' discussion of Brexit, and this reliance on evidence is the first rule of the community.

Policy-wonkish communities also seem likely to have higher-quality epistemic norms than usual. Discussion on these communities (e.g. r/foreignpolicy and r/mmt_economics) might be about the pros and cons of any given policy (or projection) and are primarily evidence-based.

We might broadly classify the top 10 communities based on competence distrust as revolving around 'politics', whereas the bottom 10 revolve around 'policy'.

Table 5. Top 10 highest and lowest subreddits for sincerity distrust density. Note the sample contains only subreddits with 500 or more posts in the month of Feburary 2020.

	Subreddit (Highest)	Sincerity (%)	Subreddit (Lowest)	Sincerity (%)
1	ActiveMeasures	4.33	politicalcartoons	0.00
2	ModernPropaganda	4.20	1022	0.05
3	media_criticism	3.89	SwitchHacks	0.07
4	ImpeachmentWatch	3.65	longrange	0.08
5	WikiLeaks	3.28	mmt_economics	0.09
6	propaganda	3.19	transtimelines	0.10
7	Mueller	3.12	mutualism	0.11
8	redacted	3.09	1911	0.11
9	worldevents	2.86	boltaction	0.11
10	Donald_Trump	2.70	georgism	0.11

What about distrust based on sincerity? Table 5 reports the top and bottom 10 communities based on the share of sentences classified as sincerity distrust. The communities clearly differ from those drawn from competence, although there is some overlap. Several of the high-distrust communities revolve around misinformation (e.g. r/ActiveMeasures, r/ModernPropaganda), are conspiratorial (r/redacted), or are related to Donald Trump (r/Donald_Trump, r/Mueller). Among the low-distrust communities, we see again niche political and economic philosophies (r/mmt_economics, r/mutualism, r/georgism), but also several gun-related subreddits (r/1022, r/1911, r/longrange).

It is clear why communities revolving around misinformation score high ('misinformation' is a query vector, after all). r/ActiveMeasures, for example, discusses information operations and explicitly calls out misinformation. This highlights that the nature of the community is important, and that we are not exactly measuring epistemic norms.

We do not know why several gun-related communities score low on sincerity-based distrust. It may be that these communities are tight-knit (and where reputation, or at least its perception, is important). It may also be that they are primarily discussing technical details (not unlike the policy wonk communities): perhaps precise differences between different models of firearms.

Distrust contagion? Can distrust can spread across communities? Perhaps user similarity is predictive of distrust. We computed a subreddit-user overlap (similarity) graph in the process of expanding the political subreddit sample. We can use this graph to begin answering this question.

To refine the question, and motivate why the user overlaps might be useful, a rough sketch of a data-generating process may help. Users, perhaps varying in their intensity of media usage, choose to split their time across several communities (i.e. a repertoire). User preference would play a role here, and perhaps each user has some initial level of distrust which may be affected by their consumption. Users presumably bring their specific level of distrust with them when they split their time across communities. We might then ask whether the amount of user overlap between

communities is related to their distrust densities. In other words: can we predict distrust density based on user overlap?

Figure 8 sketches an answer. It uses a measure of user overlap: the conditional probability that, given a user participates in community A, they also do so in B, as weights. Then it computes a weighted average of each community's distrust density. 'Similar' communities will have higher weights. We might consider this as a very rough (and statistically cavalier) graph convolution. We compare this predicted density to the *actual* distrust density.

There does appear to be a positive relationship between a community's actual distrust density and the predicted density. The relationship looks nonlinear, and is robust when we remove outliers. Variance is high, indicating that user overlaps aren't perfect predictors: internal community epistemic norms presumably dominate.

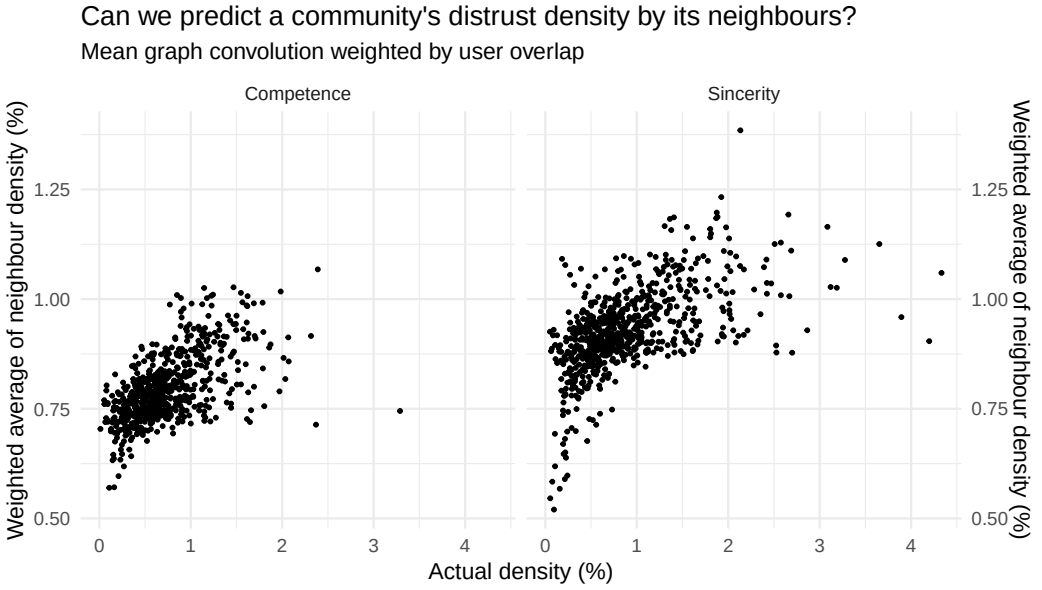

Fig. 8. We can define a graph of subreddits with (directed) edges between them weighted by user overlaps as conditional probabilities. A simple graph convolution — that is, taking the weighted average distrust density across a neighbourhood has a positive relationship with the actual distrust density. We can predict a community's level of distrust based on similar communities, although internal norms presumably account for most of the variance.

Expression and exposure. Distrust densities, the fraction of sentences in a community expressing distrust, lie in the low single digits. But what about exposure to distrust? Back-of-the-envelope calculations can help. That is: even if expression density looks quite low, should we still be worried about it from the reader's perspective? Assume a distrust density of 1 per cent, each thread averaging 20 comments, and that distrust probabilities are uniform across all comments. In this over-simplified case, a user would have to read through five threads in order to be exposed to one expression of distrust. This may not sound like much, but over the course of a week, month, or several months, the likelihood that any given user encounters several expressions of distrust is high.

Of course, without reliable media consumption data we can go little further than this. Expressions of distrust may also be indicative of deeper attitudes held by the community, rather than relevant

solely for potential exposure. Deliberate (or otherwise: [15]) invasions of high-distrust users might also increase likelihoods. Ultimately, the evolution of distrust and its effects on users, in non-well-mixed environments (echo chambers), is key to understanding pathological epistemic phenomena.

5.2 Specific Users

Following our sketch of a data-generating process above, it is clear that users are important. Here we focus on users in the data, instead of community. Figure 9 shows the cumulative distribution of expressions of distrust by user.

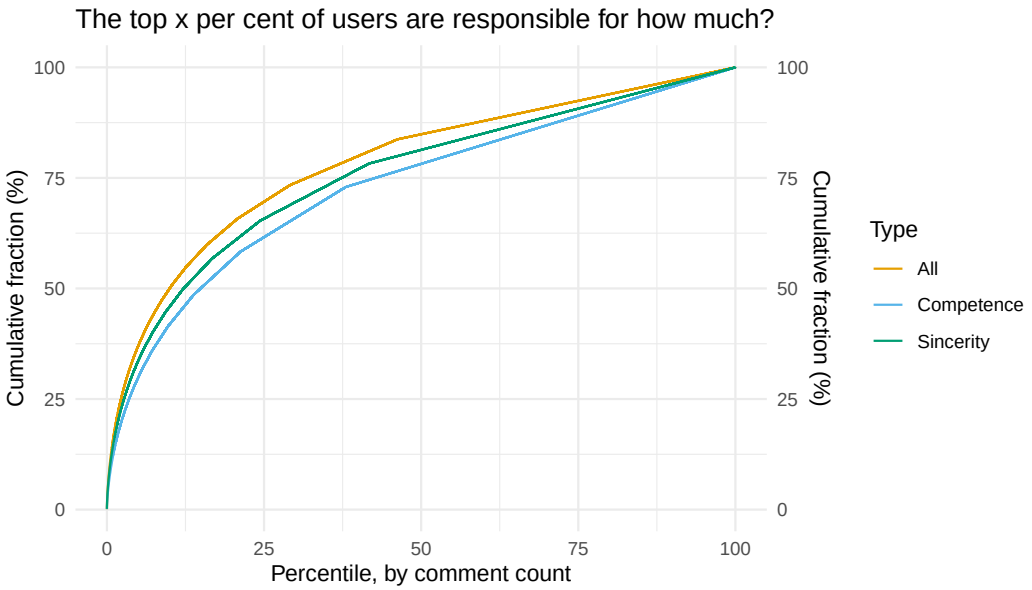

Fig. 9. User CDF. Kolmogorov-Smirnov tests indicate that the distributions are different statistically, although perhaps not qualitatively. The 'All' line refers to the entire sample.

There are 262,246 user accounts in the sample. As one might expect, the distribution is skewed: the top 25 per cent of users (roughly 65,500 accounts) are responsible for roughly 70 per cent of comments expressing distrust. The top 10 per cent of users (roughly 26,000 accounts) are responsible for half of all expressions. Arguably, 26,000 accounts is not a small core, indicating that distrust is a prevalent issue[9]. These users would be a useful sample to track over time and correlate with community-level measures of distrust. Thus, we spend the rest of this section sketching how these high-distrust users, defined as the top 10 per cent, differ from the sample broadly.

In the sketch of a data-generating process, users split their time between several communities. Following this, Figure 10 shows 'user fragmentation': how many different communities do users express distrust in? High-distrust users are shown alongside the entire sample. Note that high-distrust users are far more likely to participate in several communities. While three quarters of users in the entire sample express distrust in a single subreddit, only one quarter of high-distrust users do. That is: high-distrust users are more likely to spread distrust across different communities.

This may be because there are only a limited amount of things to comment on in any given community. High-distrust users are likely also more intense participators.

[9]Indeed, this aligns with several surveys of distrust in America [74].

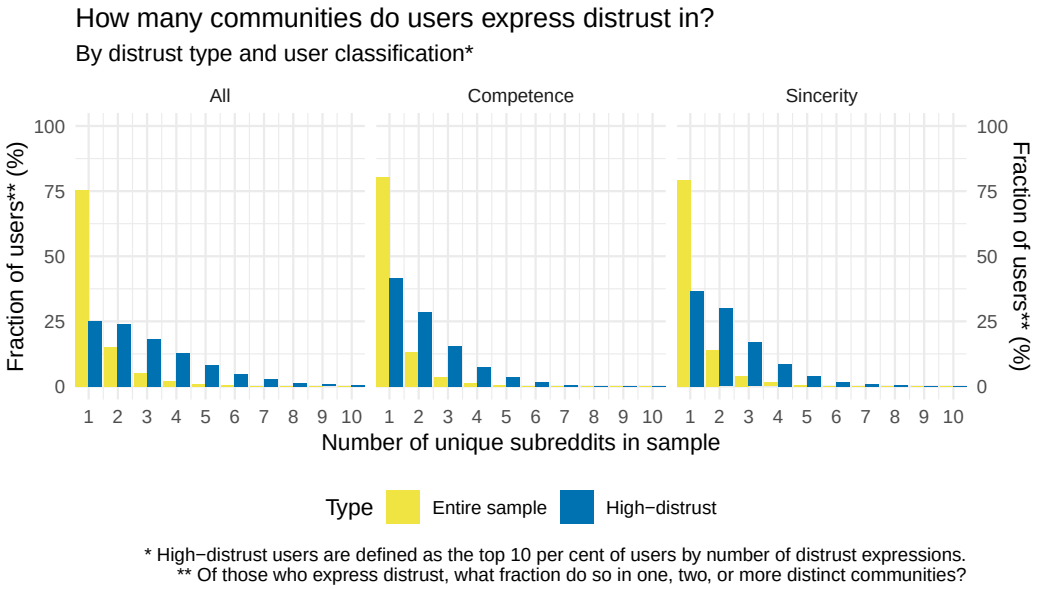

Fig. 10. User fragmentation. Each bar reads the fraction of users (in a given category) that express distrust in n different subreddits. High-distrust users are more likely to spread across several communities.

Table 6. Top and bottom 10 subreddits by increase in high-distrust user likelihood to participate. A difference of 1, for example, means a high-distrust user is 1 percentage point more likely to participate in a given subreddit compared to the background distribution.

	Subreddit (Top 10)	Difference (ppt)	Subreddit (Bottom 10)	Difference (ppt)
1	politics	1.62	worldnews	-0.61
2	worldpolitics	0.82	news	-0.53
3	conspiracy	0.70	Cringetopia	-0.45
4	Libertarian	0.54	BlackPeopleTwitter	-0.33
5	China_Flu	0.43	personalfinance	-0.27
6	WayOfTheBern	0.41	TwoXChromosomes	-0.25
7	AskTrumpSupporters	0.40	HistoryMemes	-0.24
8	ukpolitics	0.31	WhitePeopleTwitter	-0.18
9	Christianity	0.31	technology	-0.17
10	neoliberal	0.30	SandersForPresident	-0.17

We can further split the sample into all users who express distrust (the 'background' distribution) and high-distrust users. Are there specific communities that high-distrust users are more likely to express distrust in over the background distribution? Table 6 shows the top and bottom 10 subreddits by difference in these likelihoods.

Communities that high-distrust users tend to favour include conspiratorial communities (e.g. r/conspiracy) and general-interest political (e.g. r/politics, r/worldpolitics[10]). Communities which they disfavour include general-interest news subreddits (e.g. r/worldnews, r/news) and communities

[10]At the time, this community was political. By twist of fate, it is now a pornographic community.

where moderation is strong (e.g. r/BlackPeopleTwitter, r/TwoXChromosomes). It is not immediately clear why there is such a difference between general-interest politics and news communities.

There does appear to be a strong partisan bias to these communities, with the most favoured being tilted to the right.

5.3 Who is the Target?

The measurement method also reports the most relevant word (see Section 4.7) and the associated entity or subject. That is: in the sentence 'Bob is stupid,' 'stupid' is the relevant word, and 'Bob' is the subject. We explored two case studies (the subreddits r/BernieSandersSucks and r/shitfascistssay). In general, the localisation approach works well, though the dependency parse identification requires significant human validation. Therefore, we hesitate to draw sweeping conclusions. The approach requires fine-tuning (e.g. adding object of prepositions or modifiers of nominals).

One might expect r/BernieSandersSucks to largely target Bernie Sanders with distrust. However, in almost all cases the subject is either a 'you' (i.e. the commenter they are replying to) or a vague group (e.g. 'everyone', 'ignorant hicks'). It is likely that earnest participation in the community carries the presumption that Bernie Sanders is epistemically untrustworthy. Therefore the sentiment is not explicitly and frequently expressed. It seems that most of the expressions of distrust in this community arise from debate, potentially with invading interlocutors (i.e. those who believe Bernie Sanders does not suck). r/shitfascistssay, however, seems to direct distrust more towards the communal target (fascists).

6 Discussion

Social and political epistemology are useful for CSCW theory. Much of what whizzes around the internet is testimony. The notion of epistemic trust — the credibility of one's testimony — is theoretically linked to several pathological epistemic phenomena. While causality may be tricky, there are reasonable theoretical grounds uniting democratic (il)legitimacy, misinformation, and echo chambers around the concept of epistemic trust. Therefore, developing an effective and robust measurement apparatus has the potential to both simplify and deepen our understanding of these phenomena. Their social, economic, and political importance cannot be understated.

6.1 Theoretical Implications

Epistemic trust is theoretically related to pathological epistemic phenomena. Democratic (il)legitimacy, misinformation, and echo chambers all have important epistemic content. Being fundamentally social, *testimony* is core in each. Therefore, epistemic trust, underlying testimony, is a useful unifying concept. The measurement of epistemic trust should allow us to simplify and deepen our understanding.

Epistemic trust is measurable. Epistemic trust is not a philosophical curiosity. We can break epistemic trust down into that based on competence and sincerity. Further, we can operationalise these concepts into a taxonomy and measure it on real data. The baseline random forest model achieves an F_1 score of 84 per cent on competence-based distrust and 81 per cent on sincerity-based distrust on a human-annotated validation set. That said, trust, as opposed to distrust, does not appear to be as easily measurable.

Expression of epistemic distrust is relatively rare, though there is substantial heterogeneity. The fraction of sentences classified as epistemic distrust ranges between 0 and 5 per cent across approximately 1000 communities related to politics. That said, the distribution is skewed: only 10 per cent of communities have densities over 3 per cent. A user's participation and consumption

patterns will affect how much distrust they may be exposed to. But it seems likely a user will, over time, encounter some expressions of distrust.

6.2 Implications for Online Moderation

Conspiratorial communities and those focused on controversial political topics tend to express higher levels of distrust. Looking at the communities which score the highest on each type of epistemic distrust leads to two commonalities. Conspiratorial communities, almost by definition, have high levels of epistemic distrust. Communities focused on controversial and relevant political topics (for example, Donald Trump) also see high levels of distrust, although this may be due to increased debate. These sorts of communities are relevant for studies of misinformation and echo chambers, underscoring the utility of epistemic trust.

Communities with strong, moderated, epistemic norms or are tight-knit tend to express lower levels of distrust. Conversely, communities with low measurements of distrust tend to encourage good epistemic norms and enforce them by moderation. Evolutionary game-theoretic effects may operate here, too. Small, tight-knit communities focused on niche political topics also seem to have low levels of distrust.

Users are important carriers of epistemic distrust. Similarity between communities can be defined as the number of users that participate in both. Using these as weights, one can predict the level of distrust in one community given the level of distrust in similar communities. This is evidence to suggest that users are important carriers of epistemic distrust between communities, and may motivate studies of contagion.

The distribution of distrust expression across users is heavily skewed. The top 10 per cent of distrust-expressing users (roughly 26,000) are responsible for half of all expressions. Arguably, this is not a small core of users.

High-distrust users tend to participate in more communities than average. Compared to all users who express distrust, the top 10 per cent are far more likely to participate in more than one community. While this may reflect the fact that high-distrust users, by sheer volume, are intense participators, the point remains that the high-distrust core of users participates in many more communities than average. High-distrust users also seem more likely to participate more in politically right-leaning communities.

Community norms are important. Epistemic trust, while useful, is not the be-all-and-end-all. Previous results and two case studies indicate that the norms and presumptions of participation of communities are highly relevant for what is expressed. For example: we might expect a community built around the presumption that Bernie Sanders 'sucks' to express that frequently. However, it does not, perhaps because the presumption is in the community's name ('BernieSandersSucks') and does not bear repeating. Ultimately, the norms and nature of each community are of first-order importance, and we might consider epistemic distrust as something akin to a risk factor, more useful in aggregate.

6.3 Methodological Implications

Weakly supervised techniques can be effective when combined with social theory. Weakly supervised text classification techniques are effective for classification tasks like news topic or positive/negative sentiment. They can be less reliable when asked to detect more latent sentiments like epistemic trust. However, with an operationalisation informed by social theory, these methods can be effective.

Entity recognition is often implicit or about amorphous groups. The target of an expression of distrust is often explicit in topic-based communities. When there is a target, it is often an amorphous group ('them'). Community context should be included in any entity recognition system if we are to look at specific sentences.

6.4 Policy Implications

Institutions that rely on some form of legitimacy, or are vulnerable to misinformation, should take steps to understand and observe epistemic (dis)trust in relevant communities. This is a broad definition: almost all government institutions, and presumably many corporate ones, fit.

For example, health authorities have grasped the deleterious effects misinformation can have on vaccination campaigns. Direct expression of epistemic distrust is also a component. Monitoring of how (dis)trust, rather than just misinformation, spreads would improve their functioning.

Central banks are another example. Monetary policy relies on central banks remaining credible: that is, expectations are aligned appropriately. This is far less likely if the central bank is considered epistemically untrustworthy. Central banks should consider this risk.

Social media platforms like Reddit, Twitter, and Instagram should also consider how their users are expressing or consuming epistemic distrust. To the extent they develop and apply anti-misinformation or anti-echo-chamber policy, monitoring of epistemic trust will play a role. Likewise, CSCW researchers should consider measuring epistemic trust when studying pathological epistemic phenomena. Indeed, Thimbleby et al. [94] argue that designing systems around trust 'will allow more flexible systems.' Epistemic trust is also relevant for inclusion efforts in library and information science [69].

Of course, if institutions are to follow the policy advice given above, further work is required to more rigorously demonstrate the utility of measuring epistemic trust. It also needs to be easily measurable by people without specialised skills. A 'legitimacy observatory' that allows policymakers to do this across communities and over time, rigorously backed both theoretically and empirically, would be useful.

This 'legitimacy observatory' might take in a list of institutions that are important for democratic governance, or might be focused on a specific institution. For example, a health authority. Social media posts and comments, sorted by topic, would be a sensible high-level aggregation. For example, vaccination programmes, education and outreach, etc. Finally, the observatory would measure distrust densities for each of these topics, over time and disaggregated by source community. A health official, then, might be able to get a historical measure of distrust based on sincerity in comments made about their vaccination programme. This might be combined with other methods to identify users who are more likely to spread rumours, per Ghenai and Mejova [27]. A political scientist may be able to query several different institutions they believe are most important for democratic legitimacy.

We have focused on monitoring. Are interventions warranted? And if so, what kind? We found that communities with strong and enforced epistemic norms are less likely to be distrustful (e.g. r/AskHistorians). While enforcement may be effective, it is costly and unlikely to ever gain traction on communities like r/Conspiracy. Boosting more trusting or well-normed communities in recommender systems may be an effective intervention, but may risk of reinforcing distrust. More research is needed before we could confidently recommend an intervention. Unfortunately, it seems that the most effective interventions are on a societal level. For example, improving the openness and perceived fairness of government. At an institutional level, the nature of distrust might be related with the type of intervention: distrust based on competence likely demands something different to that based on sincerity.

7 Future work and limitations

Are we actually measuring epistemic distrust? While initial validation results seem promising, this approach is not without limitations, or the methods without constraints. Here, we suggest further avenues of research to address limitations or improve confidence.

Establish relationships between epistemic trust and pathological epistemic phenomena. We have good theoretical reasons to believe epistemic (dis)trust is related to democratic (il)legitimacy, the spread of misinformation, and the formation of echo chambers. However, rigorous empirical work is required. By putting links between epistemic trust and pathological epistemic phenomena on solid ground, any future application will be more confident in its results, and policy advice more sound. Several research questions jump to mind. What is the relationship between misinformation spreading and epistemic distrust at the community level? Do users with high levels of epistemic distrust help to form echo chambers, and how do we track this? Can we establish any causal effects?

Construct an expanded and improved human-annotated dataset for validation. Since the taxonomy was manually constructed, it may be missing important components. This would bias any measurement. An expanded taxonomy would be useful. Additionally, due to resource constraints, we only managed to annotate a small validation set. Ideally, a validation set should contain several thousand positive and negative examples across each of the epistemic (dis)trust categories. Several human annotators should contribute to this effort, meaning the construction of a codebook and annotation procedure is sensible. An expanded and improved validation dataset would increase confidence in any downstream results. This data could also be used in a jury learning context [30].

Consider zero-shot learning approaches. The modelling approach we used was a hybrid between large language models and classical feature-engineering. Zero-shot learning should be investigated, perhaps using models like ChatGPT.

Use data from communities on different platforms. We trained and evaluated only on Reddit. While we find community norms to be important, Reddit operates differently to other social networks. Using data from X (née Twitter), Facebook, or even news site comments would be useful.

Measure epistemic trust over time. Due to resource constraints, we only included two months of data. One for training and the other for validation and investigation. Given more time, epistemic trust should be tracked across time in both the community and user dimensions. It may be that epistemic distrust could serve as an early warning for increased propensities to share misinformation, or for the formation of an echo chamber. Using data across time would also increase confidence that results are robust to semantic drift.

Account for irony and sarcasm. We do not account for irony or sarcasm, which is frequent online.

8 Conclusion

Pathological epistemic phenomena like misinformation, echo chambers, and democratic legitimacy are crucially linked by the notion of *epistemic (dis)trust*: the degree to which we (dis)trust testimony based on its speakers perceived competence or sincerity. We synthesise this contribution to CSCW theory using theory from social and political epistemology. Further, we find evidence that epistemic distrust is measurable and not a philosophical curiosity. By applying the measurement model to social media, we further find a degree of ecological validation. To the best of our knowledge, a focus on epistemic trust is novel in CSCW and its measurement is novel in natural language processing. Further work in improving the measurement and validation of epistemic trust has the potential to

improve our understanding and efforts to combat misinformation, echo chambers, and perceptions of institutional illegitimacy.

References

[1] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–34. https://doi.org/10.1145/3579520

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236. https://doi.org/10.1257/jep.31.2.211

[3] Carlos Alós-Ferrer and Federica Farolfi. 2019. Trust Games and Beyond. *Frontiers in Neuroscience* 13 (Sept. 2019), 887. https://doi.org/10.3389/fnins.2019.00887

[4] Elizabeth Anderson. 2006. The Epistemology of Democracy. *Episteme: A Journal of Social Epistemology* 3, 1 (2006), 8–22. https://doi.org/10.1353/epi.0.0000

[5] Robert Audi. 2011. *Epistemology: A Contemporary Introduction to the Theory of Knowledge* (3rd ed ed.). Routledge, New York.

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 (Oct. 2018). https://doi.org/10.48550/arXiv.1611.09268 arXiv:1611.09268 [cs]

[7] Paul C. Bauer and Markus Freitag. 2017. *Measuring Trust*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.1

[8] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. arXiv:2001.08435 (Jan. 2020). https://doi.org/10.48550/arXiv.2001.08435 arXiv:2001.08435 [cs]

[9] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, 1 (July 1995), 122–142. https://doi.org/10.1006/game.1995.1027

[10] James Bohman and William Rehg. 2017. Jürgen Habermas. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/habermas/

[11] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Jason Brennan. 2021. Does Public Reason Liberalism Rest on a Mistake? Democracy's Doxastic and Epistemic Problems. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0009

[13] John G. Bullock, Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. Partisan Bias in Factual Beliefs about Politics. *Quarterly Journal of Political Science* 10, 4 (Dec. 2015), 519–578. https://doi.org/10.1561/100.00014074

[14] John G. Bullock and Gabriel Lenz. 2019. Partisan Bias in Surveys. *Annual Review of Political Science* 22, 1 (May 2019), 325–342. https://doi.org/10.1146/annurev-polisci-051117-050904

[15] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 31:1–31:22. https://doi.org/10.1145/3134666

[16] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–25. https://doi.org/10.1145/3274301

[17] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 Social Media Infodemic. *Scientific Reports* 10, 1 (Oct. 2020), 16598. https://doi.org/10.1038/s41598-020-73510-5

[18] Jeroen de Ridder. 2021. Deep Disagreements and Political Polarization. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press, 0. https://doi.org/10.1093/oso/9780192893338.003.0013

[19] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The Spreading of Misinformation Online. *Proceedings of the National Academy of Sciences* 113, 3 (Jan. 2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[20] Elizabeth Dubois and Grant Blank. 2018. The Echo Chamber Is Overstated: The Moderating Effect of Political Interest and Diverse Media. *Information, Communication & Society* 21, 5 (May 2018), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

[21] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The Psychological Drivers of Misinformation Belief and

Its Resistance to Correction. *Nature Reviews Psychology* 1, 1 (Jan. 2022), 13–29. https://doi.org/10.1038/s44159-021-00006-y

[22] Elizabeth Edenberg and Michael Hannon. 2021. Introduction. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0001

[23] Alexandros Efstratiou and Emiliano De Cristofaro. 2022. Adherence to Misinformation on Social Media Through Socio-Cognitive and Group-Based Processes. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 488:1–488:35. https://doi.org/10.1145/3555589

[24] David Estlund. 2021. Epistocratic Paternalism. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0007

[25] Batya Friedman, Peter H. Khan, and Daniel C. Howe. 2000. Trust Online. *Commun. ACM* 43, 12 (Dec. 2000), 34–40. https://doi.org/10.1145/355112.355120

[26] Michael Fuerstein. 2013. Epistemic Trust and Liberal Justification. *Journal of Political Philosophy* 21, 2 (2013), 179–199. https://doi.org/10.1111/j.1467-9760.2012.00415.x

[27] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–20. https://doi.org/10.1145/3274327

[28] Sarah A. Gilbert. 2020. "I Run the World's Largest Historical Outreach Project and It's on a Cesspool of a Website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 19:1–19:27. https://doi.org/10.1145/3392822

[29] Alvin Goldman and Cailin O'Connor. 2021. Social Epistemology. In *The Stanford Encyclopedia of Philosophy* (winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/epistemology-social/

[30] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3502004

[31] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445423

[32] J P Grodniewicz. 2022. Effective Filtering: Language Comprehension and Testimonial Entitlement. *The Philosophical Quarterly* (Oct. 2022), pqac064. https://doi.org/10.1093/pq/pqac064

[33] Jakob Guhl and Jacob Davey. 2020. A Safe Space to Hate: White Supremacist Mobilisation on Telegram. *Institute for Strategic Dialogue* 26 (2020).

[34] Michael Hannon. 2021. Disagreement or Badmouthing? The Role of Expressive Discourse in Politics. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0017

[35] Lu He and Changyang He. 2022. Help Me #DebunkThis: Unpacking Individual and Community's Collaborative Work in Information Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 413:1–413:31. https://doi.org/10.1145/3555138

[36] Valentin Hofmann, Hinrich Schütze, and Janet B. Pierrehumbert. 2022. The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 1259–1267. https://doi.org/10.1609/icwsm.v16i1.19377

[37] Lu Hong and Scott E. Page. 2004. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proceedings of the National Academy of Sciences* 101, 46 (Nov. 2004), 16385–16389. https://doi.org/10.1073/pnas.0403723101

[38] Giles Hooker, Lucas Mentch, and Siyu Zhou. 2021. Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or There Is No Free Variable Importance. arXiv:1905.03151 (Oct. 2021). https://doi.org/10.48550/arXiv.1905.03151 arXiv:1905.03151 [cs, stat]

[39] Philip N. Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2018. *The IRA, Social Media and Political Polarization in the United States, 2012-2018.* Project on Computational Propaganda.

[40] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 754–763. https://doi.org/10.18653/v1/2021.acl-long.62

[41] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 18:1–18:42. https://doi.org/10.1145/

3449092
[42] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 524:1–524:40. https://doi.org/10.1145/3555637
[43] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (July 2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572
[44] Prerna Juneja and Tanushree Mitra. 2022. Human and Technological Infrastructures of Fact-checking. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 418:1–418:36. https://doi.org/10.1145/3555143
[45] Hamid Karimi and Jiliang Tang. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. arXiv:1903.07389 (April 2019). https://doi.org/10.48550/arXiv.1903.07389 arXiv:1903.07389 [cs, stat]
[46] Robert A. Kaufman, Michael Robert Haupt, and Steven P. Dow. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 553:1–553:18. https://doi.org/10.1145/3555611
[47] Kari Kelton, Kenneth R. Fleischmann, and William A. Wallace. 2008. Trust in Digital Information. *Journal of the American Society for Information Science and Technology* 59, 3 (Feb. 2008), 363–374. https://doi.org/10.1002/asi.20722
[48] Stephen Knack and Philip Keefer. 1997. Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics* 112, 4 (1997), 1251–1288. jstor:2951271 https://www.jstor.org/stable/2951271
[49] Aleksi Knuutila, Lisa-Maria Neudert, and Philip N. Howard. 2022. Who Is Afraid of Fake News? Modeling Risk Perceptions of Misinformation in 142 Countries. *Harvard Kennedy School Misinformation Review* (April 2022). https://doi.org/10.37016/mr-2020-97
[50] Charles E. Larmore. 2020. *What Is Political Philosophy?* Princeton University Press, Princeton, New Jersey.
[51] Gabriel Lima, Jiyoung Han, and Meeyoung Cha. 2022. Others Are to Blame: Whom People Consider Responsible for Online Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (April 2022), 106:1–106:25. https://doi.org/10.1145/3512953
[52] Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 65:1–65:26. https://doi.org/10.1145/3449139
[53] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 461:1–461:27. https://doi.org/10.1145/3555562
[54] Michael P. Lynch. 2021. Political Disagreement, Arrogance, and the Pursuit of Truth. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0014
[55] Pete Mandik. 2007. Shit Happens. *Episteme* 4, 2 (June 2007), 205–218. https://doi.org/10.3366/epi.2007.4.2.205
[56] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological Targeting as an Effective Approach to Digital Mass Persuasion. *Proceedings of the National Academy of Sciences* 114, 48 (Nov. 2017), 12714–12719. https://doi.org/10.1073/pnas.1710966114
[57] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-Supervised Hierarchical Text Classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Honolulu, Hawaii, USA, 6826–6833. https://doi.org/10.1609/aaai.v33i01.33016826
[58] Lucas Mentch and Giles Hooker. 2016. Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research* 17, 26 (2016), 1–41. http://jmlr.org/papers/v17/14-168.html
[59] Marco Meyer and Mark Alfano. 2022. Fake News, Conspiracy Theorizing, and Intellectual Vice. In *Social Virtue Epistemology*, Mark Alfano, Colin Klein, and Jeroen de Ridder (Eds.). Routledge.
[60] Marco Meyer, Mark Alfano, and Boudewijn de Bruin. 2021. Epistemic Vice Predicts Acceptance of Covid-19 Misinformation. *Episteme* (July 2021), 1–22. https://doi.org/10.1017/epi.2021.18
[61] Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. 2021. Fakey: A Game Intervention to Improve News Literacy on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 6:1–6:27. https://doi.org/10.1145/3449080
[62] Jacob Nelson and Harsh Taneja. 2018. The Small, Disloyal Fake News Audience: The Role of Audience Availability in Fake News Consumption. *New Media & Society* 20 (Jan. 2018). https://doi.org/10.1177/1461444818758715
[63] C. Thi Nguyen. 2020. Echo Chambers and Epistemic Bubbles. *Episteme* 17, 2 (June 2020), 141–161. https://doi.org/10.1017/epi.2018.32
[64] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1165–1174. https://doi.org/10.1145/3340531.3412046 arXiv:2008.07939 [cs]

[65] Martin A. Nowak. 2006. Five Rules for the Evolution of Cooperation. *Science* 314, 5805 (Dec. 2006), 1560–1563. https://doi.org/10.1126/science.1133755
[66] Brendan Nyhan. 2021. Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions. *Proceedings of the National Academy of Sciences* 118, 15 (April 2021), e1912440117. https://doi.org/10.1073/pnas.1912440117
[67] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review* 115, 3 (Aug. 2021), 999–1015. https://doi.org/10.1017/S0003055421000290
[68] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Viking, London.
[69] Beth Patin, Tami Oliphant, Danielle Allard, LaVerne Gray, Rachel Ivy Clarke, Jasmina Tacheva, and Kayla Lar-Son. 2021. At the Margins of Epistemology: Amplifying Alternative Ways of Knowing in Library and Information Science. *Proceedings of the Association for Information Science and Technology* 58, 1 (Oct. 2021), 630–633. https://doi.org/10.1002/pra2.515
[70] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. 2018. Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology: General* 147, 12 (2018), 1865–1880. https://doi.org/10.1037/xge0000465
[71] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (May 2021), 388–402. https://doi.org/10.1016/j.tics.2021.02.007
[72] Michael Bang Petersen, Mathias Osmundsen, and Kevin Arceneaux. 2023. The "Need for Chaos" and Motivations to Share Hostile Political Rumors. *American Political Science Review* (Feb. 2023), 1–20. https://doi.org/10.1017/S0003055422001447
[73] Pew Research Center. [n. d.]. Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/
[74] Pew Research Center. 2022. Public Trust in Government: 1958-2022. https://www.pewresearch.org/politics/2022/06/06/public-trust-in-government-1958-2022/
[75] ashwin rajadesingan, Carolyn Duran, Paul Resnick, and Ceren Budak. 2021. 'Walking Into a Fire Hoping You Don't Catch': Strategies and Designs to Facilitate Cross-Partisan Online Discussions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 393:1–393:30. https://doi.org/10.1145/3479537
[76] Steve Rathje, Jon Roozenbeek, Jay J. Van Bavel, and Sander van der Linden. 2023. Accuracy and Social Motivations Shape Judgements of (Mis)Information. *Nature Human Behaviour* 7, 6 (June 2023), 892–903. https://doi.org/10.1038/s41562-023-01540-w
[77] John Rawls. 1999. *A Theory of Justice* (rev. ed ed.). Belknap Press of Harvard University Press, Cambridge, Mass.
[78] John Rawls. 2005. *Political Liberalism* (expanded ed ed.). Columbia University Press, New York.
[79] Merten Reglitz. 2022. Fake News and Democracy. *Journal of Ethics and Social Philosophy* 22, 2 (July 2022). https://doi.org/10.26556/jesp.v22i2.1258
[80] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. arXiv:1908.10084 (Aug. 2019). https://doi.org/10.48550/arXiv.1908.10084 arXiv:1908.10084 [cs]
[81] Regina Rini. 2021. Weaponized Skepticism: An Analysis of Social Media Deception as Applied Political Epistemology. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press. https://doi.org/10.1093/oso/9780192893338.003.0003
[82] Chiara Rossitto. 2021. Political Ecologies of Participation: Reflecting on the Long-term Impact of Civic Projects. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 187:1–187:27. https://doi.org/10.1145/3449286
[83] Madelyn Rose Sanfilippo and Yafit Lev-Aretz. 2019. Topic Polarization and Push Notifications. *First Monday* (Aug. 2019). https://doi.org/10.5210/fm.v24i9.9604
[84] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108 (Feb. 2020). https://doi.org/10.48550/arXiv.1910.01108 arXiv:1910.01108 [cs]
[85] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4239–4249. https://doi.org/10.18653/v1/2021.naacl-main.335
[86] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. arXiv:1809.01286 (March 2019). https://doi.org/10.48550/arXiv.1809.01286 arXiv:1809.01286 [cs]
[87] Laura Silver and Pew Research Center. [n. d.]. Americans See Different Global Threats Facing the Country Now than in March 2020. https://www.pewresearch.org/fact-tank/2022/06/06/americans-see-different-global-threats-facing-the-country-now-than-in-march-2020/

[88] Janet A. Sniezek and Lyn M. Van Swol. 2001. Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes* 84, 2 (March 2001), 288–307. https://doi.org/10.1006/obhd.2000.2926
[89] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, and Maria Bielikova. 2023. Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems* 1, 1 (Jan. 2023), 6:1–6:33. https://doi.org/10.1145/3568392
[90] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 127:1–127:26. https://doi.org/10.1145/3359229
[91] Matthias Steup and Ram Neta. 2020. Epistemology. In *The Stanford Encyclopedia of Philosophy* (fall 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/epistemology/
[92] Cass R Sunstein. 2000. Deliberative Trouble? Why Groups Go to Extremes. *The Yale Law Journal* 110 (2000), 49.
[93] Ludovic Terren and Rosa Borge-Bravo. 2021. Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research* 9 (March 2021), 99–118. https://rcommunicationr.org/index.php/rcr/article/view/94
[94] Harold Thimbleby, Steve Marsh, Steve Jones, and Andy Cockburn. 2018. Trust in CSCW. In *Computer-Supported Cooperative Work* (1 ed.), Stephen A.R. Scrivener (Ed.). Routledge, 253–271. https://doi.org/10.4324/9780429462276-16
[95] Paul Tucker. 2018. *Unelected Power: The Quest for Legitimacy in Central Banking and the Regulatory State.* Princeton University Press, Princeton.
[96] Paul Tucker. 2022. *Global Discord: Values and Power in a Fractured World Order* (1st ed.). Princeton University Press, Princeton.
[97] Daniel Viehoff. 2014. Democratic Equality and Political Authority. *Philosophy & Public Affairs* 42, 4 (Sept. 2014), 337–375. https://doi.org/10.1111/papa.12036
[98] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (March 2018), 1146–1151. https://doi.org/10.1126/science.aap9559
[99] Samantha Walther and Andrew McCoy. 2021. US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism. *Perspectives on Terrorism* 15, 2 (2021), 100–124. jstor:27007298 https://www.jstor.org/stable/27007298
[100] Paul J. Zak and Stephen Knack. 2001. Trust and Growth. *The Economic Journal* 111, 470 (2001), 295–321. https://doi.org/10.1111/1468-0297.00609

Received July 2023; revised January 2024; accepted March 2024