Who Checks the Fact-Checkers? AI, Misinformation, and Knowledge Intermediaries

Abstract

The use of artificial intelligence (AI) to transmit information in democratic societies presents a new problem for scientific objectivity. This paper extends objectivity from knowledge production to knowledge transmission. It analyses how an overlooked type of agent—knowledge intermediaries—produces an overlooked type of trust—external trust. In practice, citizens rely on intermediaries to transmit trustworthy information. This paper argues that centralised techno-legal approaches, privileging AIs as epistemically superior intermediaries, risk epistemic harm to human intermediaries. In contrast, a decentralised intermediaries approach that values contestation among human intermediaries is a better way to transmit trustworthy information and resist misinformation.

1. Introduction

Artificial Intelligence (AI) is changing how information travels around democratic societies. In response, this paper extends considerations about scientific objectivity from *knowledge production* to *knowledge transmission*. Most significantly, this paper shows that scientific objectivity in knowledge transmission relies on a different type of trust—*external trust*. Internal trust is when *experts* trust the information fellow experts produce; external trust is when *nonexperts* trust the information experts produce. This overlooked type of trust relies on an overlooked type of agent—*knowledge intermediaries*. In the philosophy of science, Marcel Boumans, Maya Goldenberg and Sabina Leonelli pioneer the concept of knowledge intermediaries to analyse how information travels (Boumans et al. forthcoming). Experts specialise in knowledge *production*; knowledge intermediaries specialise in knowledge *transmission* from experts to citizens. This paper advances their analysis of knowledge intermediaries to show how citizens rely on a web of human intermediaries to access *externally trustworthy* information. So, this paper advances the misinformation debate in the philosophy of science from a largely *inward* failure in knowledge production and a lack of misinformation-proof research among experts to a *relational* type of failure in knowledge *transmission* and a lack of *external trust* among experts and lay publics under the condition of epistemic inequality. This foregrounds human intermediaries as an overlooked type of agent in the knowledge economy, promoting an overlooked type of trust that is critical for scientific objectivity. As a result, the use of AI in knowledge transmission presents a new test case for the central concept of scientific objectivity and how AIs should participate in the division of epistemic labour as knowledge intermediaries.

This paper constructs two competing frameworks for considering how AIs, as emerging technologies, should contribute towards information transmission. A centralised 'techno-legal' approach privileges AI as an epistemically enhanced intermediary to regulate how epistemically inferior human intermediaries handle information. In contrast, a decentralised 'intermediaries' approach dethrones AI as only one more intermediary among many human intermediaries that regulate each other. This paper shows that a decentralised ecosystem with human intermediaries contesting each other cultivates a robust information ecosystem that better supports the largely overlooked *external* trustworthiness of scientific knowledge. So, a centralised techno-legal approach does not merely face technical difficulties. It fundamentally misconceives the epistemology of scientific information transmission. In practice, computer scientists are presented with a vicious justificatory circle when they design AI to resist misinformation—broadly conceived as untrustworthy information. In order to know that alleged misinformation is truly misinformation, computer scientists must know that the process that selected the alleged misinformation was reliable. However, to know that the process that selected the alleged misinformation was reliable, computer scientists must know that the alleged misinformation is truly misinformation. Without independent access to either the (un)trustworthiness of the alleged misinformation or the reliability of the selection process, computer scientists cannot easily know whether the process was reliable or whether the alleged misinformation was truly misinformation. So, a centralised techno-legal approach risks an undercontested confidence in the competence of AI, which can cause significant epistemic harm to the information ecosystem.

Outside the philosophy of science, the democracy/epistocracy debate in political philosophy typically sees the democratic value of public deliberation and the antidemocratic risks of expert domination. However, the debate largely overlooks the democratic value of expertise. Using the philosophy of science and science and technology studies, Alfred Moore's critical elitism provides one of the most rigorous frameworks that defends the democratic value of expertise as a vital epistemic resource to inform collective judgements, guide collective actions and check political power. However, Moore largely leaves out how scientific information, once produced, travels around the information ecosystem. As a critical next step, this paper will extend a broadly critical elitist framework to *knowledge intermediaries*. An intermediaries approach provides a richer model of the social world, populated with more than citizens with a legal right to freedom of speech but largely ignorant of basic scientific facts, and a state informed by experts with the legal authority to outlaw specific types of speech. The social world is populated with a complex web of human intermediaries that shapes how information travels throughout the information ecosystem. This richer model of the social world uncovers the possibility that a complex web of human intermediaries may regulate information better

than techno-legal uses of AI and that techno-legal uses of AI may unintentionally but foreseeably cause epistemic harm to that web. Whatever AIs might do, computer scientists should consider how AI may support and not derail the social regulation of information through contestation among human intermediaries during public deliberations.

It is useful to outline the paper. In section 2, a centralised techno-legal approach towards AI is analysed. Section 2.1. explores the default techno-legal approach that primarily relies on AI to resist misinformation. Section 2.2. explores how computer scientists may use expert consensus to guide AIs and enhance techno-legal approaches. These sections show how these approaches risk overconfidence in fallible authorities, whether AIs or experts. In section 3, decentralised Millian approaches towards AIs are advanced. Section 3.1. explores the default Millian approach that primarily relies on free speech to resist misinformation. This section shows how the default Millian approach risks overconfidence in the fallible process of free speech. Section 3.2. explores how misinformation education may support free speech dynamics and enhance a Millian approach. However, misinformation education may overburden citizens. Section 3.3 explores how knowledge intermediaries may better support free speech dynamics. This preferred approach highlights how knowledge intermediaries are an overlooked epistemic agent needed for an overlooked type of trust—external trust—to restore public trust in science and resist the allure of misinformation.

2. Centralised Approaches

In the following sections, I will construct two approaches—centralised techno-legal approaches and decentralised Millian approaches—that put the social responsibility to resist misinformation onto a particular process. This type of approach empowers computer scientists to trust that the process that selected the alleged misinformation was reliable after all. However, this type of approach risks an unchecked (or undercontested) confidence in fallible procedures that may become unreliable. Who checks the fact-checking process?

2.1. A Techno-Legal Approach

In this section, I will explore techno-legal approaches that use AI to resist misinformation. From a Millian standpoint, I will argue that techo-legal approaches risk an unchecked confidence in a fallible agent—AI, or computer scientists—to know what is true independently of public deliberation.

Most policy-relevant topics contain significant misinformation. Broadly speaking, misinformation is unintentionally misleading information. Of course, false information can mislead. However, true information

can mislead, too. So, I will use ‘killer vaccine’ as a paradigmatic case of misinformation.[1] The European Medicines Agency reports about 0.001 reported fatal outcomes for every 100 administered COVID-19 vaccine doses.[2] Of course, deaths after vaccination are not the same as deaths from vaccination. Although vaccines might have causally contributed towards some of the deaths after vaccination, ‘killer vaccine’ is still significantly misleading because it assumes an extremely high standard for safety that is not typically used elsewhere. So, ‘killer vaccine’ is an untrustworthy piece of information. If citizens rely on it to review and revise their political judgements, they are likely to develop incompetent political judgements that are unreasonably risk-averse towards safe vaccines. Whatever the arguments for and against vaccine mandates might be, they did not mandate a ‘killer vaccine’.

It is tempting to think that the most effective way to reduce misinformation is through legislation. In return, public and private organisations are often tempted to think that AI is the most effective way to implement legislation. So, techno-legal approaches deliberately design legislation for emerging technologies—largely AI—to implement (Cordella and Gualdi 2025). In practice, AI mostly tracks the generic qualities of misinformation.[3] So, AI may permit misinformation without generic qualities and prohibit trustworthy information with the generic qualities of misinformation. As a result, techno-legal approaches risk low-quality assessments of hard cases. Nevertheless, techno-legal approaches allow AI to process a high quantity of information very quickly (Nakov et al. 2021). Consequently, techno-legal approaches might be justified on largely aggregative grounds: fast fact-checking, downranking, and censorship allow the expected benefits, when aggregated, to exceed the expected costs, even if the expected costs are high.

It is tempting to defer to AI for largely epistemic reasons (Hauswald 2025). It is plausible to presume that AI decision-making is generally more informed and rational than any human, and it is difficult to know enough about the data and algorithms used to make any particular AI decision to prove otherwise. Nevertheless, a human deference to AI risks a false sense of security when fallible AI regulates information (Romeo and Conti 2025). AI speech regulation includes a bundle of various measures that limit speech that is considered bad. For example, social media companies may use AI to delete and downrank specific types of speech on their platforms and suspend or ban specific users for their speech. While milder measures might not violate speech rights, they often reduce the efficacy of speech rights with ripple effects for the information base that

[1] https://www.bbc.co.uk/news/technology-54001894

[2] https://www.ema.europa.eu/en/human-regulatory-overview/public-health-threats/coronavirus-disease-covid-19/covid-19-medicines/safety-covid-19-vaccines

[3] One extensive analysis of the trust/trustworthiness distinction is O’Neill, Onora. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26 (2):293–300.

citizens can reasonably access during public deliberations and for the competence of their resulting political judgements.

Who checks the fact-checkers?[4] In a Millian spirit, to put the social responsibility on AI to resist misinformation is to risk an unchecked (or undercontested) confidence in fallible algorithms that may become unreliable.[5] As a countermeasure, computer scientists often test algorithmic decisions to check how competent particular algorithmic decisions are. For instance, Australia's AI Ethics Framework is the first framework to use 'contestability' as a core principle to protect individuals (Lyons et al. 2021). However, how contestability can and should be implemented is a highly context-specific question. In practice, computer scientists can only feasibly test a very small sample of the algorithmic decisions made. So, AI testing might make algorithmic decisions more competent within a very small range. Nevertheless, expert contestation may not make algorithmic decisions more competent everywhere or elsewhere. In practice, public deliberation remains critical to continuously contest a wide range of algorithmic decisions to check how competent algorithmic decisions are more generally.

This is a significant problem for techno-legal approaches. In order to know the reliability of fallible algorithms, computer scientists may rely on public deliberation to find the mistakes that fallible algorithms make. For instance, techno-legal approaches may blur the distinction between good legislation and legislation good for technological implementation. Rather than law governing technology, techno-legal approaches may empower technology to govern law. Most significantly, techno-legal approaches may neglect good but technology-unfriendly legislation and prioritise technology-friendly but defective legislation. For instance, public deliberation may show that techno-legal approaches use heavy-handed algorithms that are too prohibitive towards trustworthy information about vaccines, which heavy-handed algorithms misidentify as misinformation. In practice, censored information often has several sincere and smart defenders. Alternatively, public deliberation might show that techno-legal approaches use light-handed algorithms that are too permissive towards misinformation about vaccines that lack the generic qualities of obvious misinformation. So, AI does not always epistemically enhance public deliberation. In the opposite direction, public deliberation is often needed to epistemically enhance AI; public deliberation that a techno-legal use of AI may epistemically harm (as explored later).

---

[4] Eric Winsberg gives an extensive list of deeply disputable fact-checking Winsberg, Eric. 2024. "“Falsehoods Fly: Why Misinformation Spreads and How to Stop It” by Paul Thagard. Columbia University Press." *The Journal of Value Inquiry* https://doi.org/10.1007/s10790-024-09996-3.

[5] For a history of how considerations about fallibility guide political thought, see Schwartzberg, Melissa. 2007. "Jeremy Bentham on Fallibility and Infallibility." *Journal of the History of Ideas* 68 (4):563–85.

2.2. An Expert Approach

In this section, I will construct an expert approach as a way to improve techno-legal approaches. When computer scientists review fallible AIs, they can put the social responsibility on experts to guide them towards what is true. For instance, they may defer to expert organisations that develop consensuses about what is true about vaccines beyond a reasonable doubt. With expert consensus as an arbiter of truth in politics, computer scientists can know the (un)trustworthiness of alleged misinformation after all (Anderson 2011). Once computer scientists know that AI follows expert consensus, they can reasonably infer that AI tracks the truth more generally. The following of expert consensus is seen as good evidence of tracking the truth. However, I will argue that an expert approach risks an unchecked confidence in fallible consensuses that are not always true. So, who should choose the preferred bundle of expert consensuses?

As explored next, a central problem with an expert approach is that what qualifies as the right bundle of expert consensuses is not significantly easier to know than what qualifies as the truth. In practice, expert consensus is not a simple epistemic resource. More realistically, expert consensus is a complex bundle of many different types of agreements. Without independent access to the truth, computer scientists cannot easily infer if and when expert consensus converges on the truth.

The most demanding conception of consensus is a unanimous agreement for *epistemic* reasons: everyone agrees because everyone aimed to find the truth (Landemore 2012). The agreement is a byproduct of seeking the truth. However, this is often an infeasible standard. In practice, the fact of human fallibility is a good reason to expect some expert disagreement based on honest mistakes. So, an *allegedly unanimous* agreement might not be *actually unanimous*. It may merely overlook or exclude opposing experts. In practice, both the consensus experts and overlooked opposing experts may reasonably disagree because they use different high-quality data sets and research methods to produce conflicting information.

Alternatively, a unanimous agreement might include opposing experts because consensus experts convince them to accept the consensus for nonepistemic reasons despite their epistemic objections. The opposing experts agree with the consensus experts because they aimed to find a pragmatic agreement despite their ongoing principled disagreements. As Moore argues, deliberative acceptance allows opposing experts to temporarily suspend their scepticism for various pragmatic purposes if they judge that their objections were given a fair hearing (Moore 2017). For example, it helps experts to inform collective judgements, guide collective actions and check political power. So, the agreement is a direct product of seeking agreement. As a result, expert consensus is not a sign of truth. It is a sign of pragmatically justified agreement.

Secondly, expert consensus changes over time. As one of the most forceful defenders of scientific consensus, Peter Vickers has pioneered the concept of 'future-proof' science—scientific consensuses that the public can become reasonably confident will not change in the future (Vickers 2023, 18). His 30 paradigmatic cases of lasting scientific facts include 'the Sun is a star', 'DNA has a double helix structure' and 'the Earth is a slightly tilted, spinning, oblate spheroid' (Vickers 2023, 13–14). Whatever the merits of future-proof science might be, a future-proof approach does not work very well in politics. The *quality* of future-proof science might be high, but the *quantity* of future-proof science is low. It is plausible to presume that experts are most likely to discover future-proof answers to questions that are the easiest to answer and remain divided on questions that are harder to answer.[6] So, future-proof science is not enough to check the reliability of the fallible algorithms regulating public deliberations. In practice, public deliberation extends far beyond the small bundle of largely apolitical future-proof facts that Vickers identifies. As a result, using a small bundle of expert consensuses—say, future-proof science—risks mistaking unreliable AI for reliable AI. The following of future-proof science is not good evidence of tracking the truth more generally.

A less demanding definition of consensus is a less-than-unanimous agreement. However, this definition may become unattractively exclusionary. There might be strength in numbers, but there is not always truth in numbers. As explored next, a majority or super-majority by itself often only shows the quantity of agreement rather than the quality of agreement.

For instance, techno-legal approaches might be tempted to use metascience—the use of scientific research methods to research the quality of scientific research—to find trustworthy information. In return, computer scientists can use metascientific research to train AI. Whatever information gains the distrust of a broad metascientific consensus is seen as a good proxy for untrustworthy information.[7] Although metascience may help to find trustworthy information, I will argue that metascience is not always a good way to discover untrustworthy information. As explored next, metascience often relies on a general conception of science that seeks general qualities of good science, say, replicability, which threaten to become insensitive to the specific qualities of situated scientific practices.

Of course, some metascientific research does use various qualitative approaches such as case studies, historical analysis and ethnographies of labs. Nevertheless, metascientific research often relies on a highly quantitative conception of science. For example, meta-analysis, scientometrics, and replication studies rely

---

[6] For instance, Bryan Caplan argues that experts—particularly in economics—often agree on the basics and disagree on harder questions Caplan, Bryan. 2008. "Reply to My Critics." *Critical Review: A Journal of Politics and Society* 20 (3):377–413.

[7] For instance, Ioannidis, J. P. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8):e124. https://doi.org/10.1371/journal.pmed.0020124.

on bibliometrics, statistical tools, and computational methods to evaluate scientific outputs, bias, and reproducibility.[8] When situated scientific practices conflict with general metascientific principles, say, replicability requirements, it is not always the situated scientific practices that are defective because they are insensitive to general principles. As explored next, global assessment criteria are often defective because they are insensitive to the specific scientific goals and social resources of situated scientific practices (Leonelli 2017). In practice, the technology-intensive nature of much scientific research empowers technology producers to set research standards when they market their latest technologies as of the highest quality. Of course, technologies intend to facilitate high-quality research. However, technologies can unintentionally act as a filter on scientific research. In practice, technologies often shape the research standards of the scientific communities that use them and exclude or devalue the research of scientific communities that do not need to use them, given their specific purposes, or cannot afford to use them, given their limited means. So, the technological needs to meet global replicability requirements may wrongfully devalue or exclude situated research practices and unintentionally undercut the implicit trust among peers that typically characterises scientific research.

This shows that a central problem with a metascientific approach is that it may neglect the value of situated research that has adapted to work well for specific scientific goals in specific social circumstances. So, metascience cannot easily enhance the trustworthiness of AI because metascience itself is not always a reliable way to check the trustworthiness of information. As a result, using a big bundle of expert consensuses—say, a broad metascientific consensus—is that it risks mistaking reliable AI for unreliable AI. The fact that situated scientific practices may go against general metascientific principles—say, replicability requirements—is not always good evidence of untrustworthy information. Of course, consensus has highly valuable epistemic and pragmatic uses. However, consensuses are often too contested to justify algorithmic fact-checking, downranking or censorship of (mis)information at scale.

3. Decentralised Approaches

In the following sections, I will construct a decentralised Millian approach that puts the social responsibility to resist misinformation onto *public deliberation*. However, this risks an unchecked (or undercontested) confidence in fallible deliberations that may become unreliable. So, I will construct an education approach that puts a social responsibility onto citizens to educate themselves in ways that support Millian tendencies to

[8] For instance, Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop *et al.* 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1):0021. https://doi.org/10.1038/s41562-016-0021.

separate the truth from misinformation. However, this risks an ineffective—and, subsequently, unfair—burden on citizens. In contrast, I will construct an intermediaries approach that puts the social responsibility to resist misinformation onto knowledge intermediaries as a better way to enhance a Millian approach.

3.1. A Millian Approach

In this section, I will construct a Millian approach that uses free speech to resist misinformation. The novel contribution of this section is to apply the Millian argument against regulated speech to the Millian argument for free speech to show that a Millian approach largely overlooks the similarly realistic risks of an unchecked confidence in a general deliberative process to find what is true. It is not significantly easier to know if and when free speech has found the truth after deliberation than to know if and when particular people know the truth before deliberation.

J.S. Mill famously defends the epistemic virtues of free speech as a discovery process that finds the truth and better justifications for the truth (Turner 2021). So, I will construct a Millian approach that allows computer scientists to retreat to the general tendencies of deliberation: free speech tends to find the truth, whatever the truth might be. As a result, access to the reliability of public deliberation need not rely on evaluating the truth of the resulting knowledge claims. The general tendencies of free speech may give computer scientists independent access to the reliability of public deliberation instead.

In the first Millian case, an unpopular minority is right. In this case, strong speech rights—AI permitting most types of speech, except, say, direct incitements to violence—give the right minority the broadest freedom feasible to explain effectively why, say, vaccines are safe, to a wrong majority. In return, strong speech rights empower the conciliatory and newly converted majority to find out that they were wrong. As a result, the public accepts that free speech worked and exchanged error for truth (Mill 2011, 33). In the second Millian case, the unpopular minority is wrong. In this case, strong speech rights give a steadfast majority the broadest freedom feasible to find better justifications for why, say, vaccines are safe, and why the anti-vax minority is wrong. In return, the public accepts that free speech worked and found better justifications for the truth, produced by its collision with error (Shah 2021). In either case, strong speech rights in principle produce epistemic benefits that algorithmic censorship would threaten.

In practice, various context-specific factors can derail the general Millian tendencies of free speech. A Millian standpoint successfully sees that trust in fallible people—including computer scientists—to know the truth before public deliberation risks an unchecked (or undercontested) confidence in unreliable people. However, as explored next, a Millian approach largely overlooks that trust in a fallible deliberative process to

find the truth *also* risks an unchecked (or undercontested) confidence in an unreliable deliberative process. In the first anti-Millian case, an effective minority is wrong. Contrary to Mill, strong speech rights can give, say, an anti-vax minority the broadest freedom feasible to explain effectively why vaccines are unsafe to the majority. In return, a conciliatory and newly converted majority wrongly believes that free speech worked and exchanged error for truth. Nevertheless, free speech failed and exchanged truth for error. For instance, Ben Saunders argues that the temporary censorship of vaccine misinformation during pandemics is legitimate for Millian reasons (Saunders 2023). In emergencies, the more favourable conditions for Millian tendencies are absent. So, bad speech may become particularly harmful in emergencies, and free speech may become particularly ineffective. As a result, an epistemic risk with strong speech rights is that a wrong minority is empowered to spread harmful falsehoods when Millian conditions are absent.

In the second anti-Millian case, an ineffective minority is right. Contrary to Mill, strong speech rights can give a steadfast majority the broadest freedom feasible to find better rationalisations for why they believe that, say, pro-vax experts are wrong and that they are right. For instance, Daniel Williams argues that market-like dynamics can produce attractive rationalisations for false beliefs in return for social rewards such as attention and status (Williams 2023). So, the public may wrongly believe that free speech worked and found better justifications for the truth. Nevertheless, free speech failed and found better rationalisations for falsehoods. As a result, an epistemic risk with strong speech rights is that a wrong majority is empowered to find better rationalisations. In either case, strong speech rights in practice risk epistemic harms that AI in principle could reduce through fast fact-checking, downranking and censorship.

As explored next, it is difficult for computer scientists to know which way the scales tip. It is not too difficult for computer scientists to see the significant epistemic harm of *particular pieces of misinformation*. However, it is reasonably easy to overlook the significant harm of *suspending general rules*.[9] Even if the uncomfortable enforcement of strong speech rights protects particularly harmful pieces of misinformation about vaccines, the easy suspension of strong speech rights may also risk various harms when done more generally. Whatever harms strong speech rights potentially cause, they need to be balanced against the harms that easily suspending them also potentially cause.

Contrary to Saunders, the general rule of strong speech rights need not overlook exceptional circumstances. The general rule of strong speech rights may see how difficult it is to see exceptions to the exceptions. Independently of bad governments that would strategically abuse emergency powers, good governments may sincerely struggle to know when the exceptions do apply and if the exceptions to the exceptions do not apply.

[9] General rules are not universal or absolute: they may not extend to every situation.

Even if specific cases of censorship are legitimate in principle, they might remain infeasible or ineffective in practice. As Saunders himself observes, misinformation about vaccines could spread too quickly, and legitimate censorship could diminish trust in governments, with sincere governments often spreading confusing, misleading and false messages themselves (Lepoutre 2019). Ironically, Maxime Lepoutre argues that legislation that seeks to silence misinformation may accidentally amplify it. Firstly, the expressive power of the law shines a public spotlight on any speech the law prohibits. Secondly, the enforcement of the law can result in very public trials, with the perpetrators of misinformation able to publicise their prosecutions. Similarly, algorithmic prohibitions against misinformation may shine a public spotlight on misinformation and allow perpetrators to publicise their punishments.

In contrast, Lepoutre argues that more speech—counterspeech—may counter misinformation more effectively than more legislation. Firstly, Lepoutre argues that negative counterspeech that directly counters misinformation with informed corrections may unintentionally but foreseeably enhance the salience of misinformation since it must mention the misinformation to correct it. As an alternative, Lepoutre shows that the public may use positive counterspeech that indirectly counters misinformation with informed alternatives, which implicitly contradict misinformation without explicitly mentioning it. Secondly, Lepoutre argues that misinformation is often 'sticky' and that the epistemically harmful effects are often difficult to undo. So, Lepoutre shows that the public may use preemptive counterspeech that prevents misinformation from gaining popularity to begin with. The judicious use of positive preemptive counterspeech can populate the information ecosystem with informed views that implicitly contradict misinformation without explicitly mentioning it and prevent misinformation from gaining popularity to begin with. As a result, computer scientists should not overlook the possibility that verbal counterspeech may often remain more effective than algorithmic censorship.

3.2. An Education Approach

In this section, I will construct an education approach that puts a social responsibility on citizens to educate themselves about how to support the Millian tendencies of free speech to find the truth rather than spread misinformation. As a supportive measure, misinformation education does not aim to change the information ecosystem, but to change the citizens that populate it. Most significantly, misinformation education can help citizens develop digital skills (Goldstein 2021). For instance, Paul Thagard pioneers an AIMS framework that presents the acquisition, inference, memory and spread of trustworthy information and misinformation as significantly different (Thagard 2024). Thagard argues that informed citizens acquire trustworthy

information from reliable sources, use critical thinking to make valid inferences and find trustworthy information harder to remember, which hinders and slows the spread of trustworthy information. In contrast, misinformed citizens acquire misinformation from biased sources, use motivated reasoning to make invalid inferences, and find misinformation easier to remember, which helps to quicken the spread of misinformation. So, misinformation education may help citizens develop the digital skills to know when information has the qualities of misinformation. It is plausible to presume that misinformation spreads largely because citizens are inattentive rather than partisan (Benson forthcoming). In such cases, misinformation education may empower citizens to effectively identify misinformation at critical moments when they are willing to become attentive.

However, it is difficult for citizens to know how much misinformation education could tip the scales. Unless citizens already know what information qualifies as misinformation, they cannot easily know the efficacy of misinformation education. Despite their best efforts, citizens may largely remain or quickly become too cognitively biased. In practice, assessments of misinformation education often fix what information qualifies as misinformation, using different definitions for various substantive and methodological reasons (Rau and Premo 2025). Of course, most citizens across partisan divides accept that *undisputed misinformation* is misinformation. However, most citizens disagree that *disputed misinformation* is misinformation—particularly across partisan divides. In specific cases, citizens may know that disputed misinformation qualifies as misinformation. Not all *partisan* disagreement is *peer* disagreement. The most obvious cases of misinformation may fully follow Thagardian patterns. So, citizens can become confident that *obvious misinformation* is misinformation despite partisan disagreement. However, citizens should remain cautious that *unobvious misinformation* is misinformation because of partisan disagreement. They are hard cases with good reasons to judge either way. If citizens become confident at scale about what *disputed information* qualifies as *obvious misinformation* to assess how well misinformation education works, they risk an arrogant type of confidence in what information qualifies as misinformation.

As explored next, arrogance is best avoided. Firstly, a noninstrumental reason is that arrogance epistemically disrespects misinformed citizens as epistemic peers. It paternalistically presumes that some citizens have *generally* superior epistemic capabilities to know at scale what information qualifies as misinformation. For instance, a prevalent ‘post-truth’ narrative in politics demeaningly frames inaccurate or misinformed citizens as insincere and without regard for the truth, and weaponises the epistemic difficulties in finding accurate information on complex topics as motivational defects in political opponents to delegitimise and dismiss them (Hannon 2023). However, everybody is as human as everybody else. As epistemic peers, nobody is

generally more likely to be right in politics than anybody else (Joshi 2020). As a result, citizens should remain cautious about what disputed information qualifies as misinformation at scale, even if they are confident about specific cases of obvious misinformation.

Secondly, an instrumental reason to avoid an epistemically arrogant confidence in what information qualifies as misinformation is that arrogance raises the risk of error. As fallible agents, citizens may mistake trustworthy information for obvious misinformation. Over time, many mistakes would undercut the reliability of misinformation education itself. While misinformation education may contribute significantly towards resisting misinformation, it need not and should not do all the work. As explored next, knowledge intermediaries already do significant work to transmit trustworthy information and resist misinformation on behalf of the public. So, an intermediaries approach unburdens citizens and gives already epistemically enhanced intermediaries the social responsibility to resist misinformation on behalf of the public.

3.3. An Intermediaries Approach

In this section, I will construct an intermediaries approach as a better way to improve a Millian approach. In practice, citizens are often epistemically dependent on others to gain knowledge (Hardwig 1985). A knowledge intermediary is an epistemic agent that performs the distinctive social function of *transmitting* (or helping to transmit) information from the research community that produces it to a lay public.[10] In practice, citizens are not just epistemically dependent on experts (either as individuals or as communities). More realistically, citizens are epistemically dependent on a complex web of knowledge intermediaries that transmits trustworthy information from experts to them. For example, mainstream news media—including news stations and newspapers—hire journalists to investigate, report and critique information about significant topics in accessible formats. More recently, new online media—including podcasts and blogs—give a platform to researchers from universities and think tanks and opinion leaders from social organisations and political parties in friendly interviews and adversarial debates. More widely, 'issue publics'—individuals and civil society organisations especially interested in particular issues—gather otherwise dispersed information and epistemically enhance issue-specific preferences with a bigger and better information base (Elliott 2020). This complex web of intermediaries shapes how information travels from the research communities that produce information to the lay publics that use it (Herzog 2024, 15).

[10] Transmission is a fuzzy concept that includes whatever epistemic labour significantly aids the travel of information from research communities to lay publics. This includes science communication, science advocacy, knowledge brokering, knowledge mediation and data curation

A realistic philosophy of science has turned away from a purely epistemic or 'value-free' conception of objectivity and towards a more social or 'value-laden' conception of objectivity as a structured process of diverse contestation (Kitcher 2001; Longino 2002). As Kristina Rolin argues, scientific objectivity gives the public permission to trust scientific knowledge claims (Rolin 2021). With a social conception of objectivity, a structured process of diverse contestation makes scientific knowledge claims as robust as possible. Rolin argues that scientific objectivity is a *hybrid* concept that has an epistemic dimension of reliability and a neglected moral-political or ethical dimension of social responsibility. To achieve reliability and social responsibility, the division of epistemic labour in the knowledge economy produces different epistemically enhanced agents with different specialised skills. For instance, Michel Croce distinguishes between expert-oriented abilities and novice-oriented abilities (Croce 2018). Expert-oriented abilities are the virtues that empower experts to use their expertise to find and face new problems within their field. So, expert-oriented abilities may largely promote *reliable* research. In contrast, novice-oriented abilities are the virtues that empower experts to address a layperson's epistemic dependency on them properly. Novice-oriented experts may be sensitive to lay needs, intellectually generous, intellectually empathetic, and sensitive to lay epistemic resources. As explored next, novice-oriented abilities may better promote *socially responsible* research that follows sound moral values during scientific inquiry.

Although particular experts may develop very sophisticated novice-oriented abilities, it may overburden experts in general to strongly develop them. Most significantly, experts specialise in the production of information. So, it is plausible to presume that experts primarily develop *expert-oriented* abilities to *produce* information well. In particular, experts may produce information about vaccines that is trustworthy among fellow experts in specific contexts. I will call this 'internal trustworthiness'. The internal trustworthiness of information may depend on various epistemic qualities. For instance, experts may justifiably trust information they know is produced with high-quality data sets and research methods.[11] The expert-oriented abilities of experts help to make information internally trustworthy. However, experts may not produce information about vaccines that is trustworthy to citizens in different contexts. I will call this 'external trustworthiness'. The expert-oriented abilities of experts may not make information externally trustworthy. The external trustworthiness of information may be independent of various epistemic qualities. For instance, citizens may not trust the information that experts trust because citizens lack the expertise to assess the data and methods used. So, citizens may justifiably not trust the information about vaccines that experts

[11] Of course, *trustworthy* information is not always *true* because contradictory pieces of information may be equally trustworthy but cannot be equally true.

justifiably trust because the epistemic inequalities among them result in them using different signs of trustworthiness.

One way to fill this gap in the knowledge economy is with human intermediaries that specialise in the transmission of information. In contrast to the mechanistic calculations of AI, human intermediaries primarily develop *novice-oriented* abilities to *transmit* information well. The division of epistemic labour tends to produce epistemically enhanced human intermediaries with specialised skills that facilitate the lay uses of trustworthy information. In practice, a fallible but generally good intermediary may occasionally transmit particular pieces of misinformation despite their best efforts.[12] Nevertheless, it is plausible to presume that the specialisation of epistemic labour tends to produce generally good human intermediaries that, in aggregation, give low-information citizens reasonable access to trustworthy information. For example, the novice-oriented abilities of human intermediaries can make information about vaccines externally trustworthy to citizens. As Torsten Wilholt argues, the trustworthiness of experts as information producers partially relies on their attitudes towards the possible consequences of the information they produce (Wilholt 2013). So, intermediaries can help citizens assess the attitudes of experts towards possible consequences in order judge whether the experts produce what they consider socially responsible information. As a result, both experts and intermediaries are often needed for citizens to gain reasonable access to trustworthy information. The division of epistemic labour may privilege experts for *internally* trustworthy information, but it privileges intermediaries for *externally* trustworthy information.

Looking at the circumstances of scientific knowledge consumption, different publics in different situations use information for different purposes. In particular, narratives often frame the meaning of new information (Morgan 2022). When the narratives in and across publics change, the meaning and value of new information can change. So, human intermediaries can help to align information about vaccines with popular narratives among particular publics. Similarly, intermediaries can help to adapt popular narratives among particular publics to better accept trustworthy information about vaccines and better resist misinformation. For instance, Lepoutre argues that a distinctive type of counterspeech—narrative counterspeech—may counter misinformation more effectively than censorship since misinformation often uses popular narratives to spread (Lepoutre 2024). While no particular factual claim within a conspiracy theory may be seen as particularly compelling, some publics see the general narrative as very compelling. So, the general narrative is what causes the conspiracy theory to become popular, and the particular factual beliefs that come with it are largely a byproduct. As a countermeasure, skilled storytelling that understands how structure and style

[12] https://www.bbc.com/news/articles/c874nw4g2zzo

can contribute towards coherent, emotionally engaging and salient counternarratives may more effectively counter conspiratorial narratives than AI fact-checking, downranking and censorship.

In practice, experts must often make deeply disputed value judgements throughout the various stages of knowledge production (Reiss 2017). However, experts themselves cannot feasibly give every public a fair hearing. So, different intermediaries across the information ecosystem can represent different publics. It is plausible to presume that a representative intermediary that gives particular publics a fair hearing is more likely to align trustworthy information about vaccines with their narratives and adapt their narratives to produce effective counternarratives against misinformation. Most significantly, human intermediaries may tell stories that are sensitive to various situational factors—in particular, the specific information, interests and ideologies—among particular publics to produce effective counternarratives.

Firstly, different publics have different information. In particular, social inequities may result in different background information. For instance, as social standpoint theorists argue, the social standpoints of socially disadvantaged citizens may empower them to know that social institutions work particularly badly for them (Wylie 2003). In contrast, the social privileges of socially advantaged citizens may incline them to presume that social institutions tend to work well in general. As a result, trustworthiness might vary between socially advantaged and disadvantaged publics because the different levels of information among them change what levels of confidence in new information about vaccines may be justified. Secondly, different publics have different interests. In particular, social inequities may result in different political interests. When the basic structure of social institutions works particularly badly for socially disadvantaged citizens, they are often unprotected from facing unreasonably high fallouts from their personal decisions. In contrast, if the basic structure works well for socially advantaged citizens, they are largely protected from facing unreasonably high fallouts. As a result, trustworthiness may vary between socially advantaged and disadvantaged publics because they are presented with different levels of risk.

Thirdly, different publics have different ideologies. In particular, political minorities often pursue significantly different basic aims from political majorities. So, if information about vaccines is made to align with the values of a political majority, the preferred majority gains unfair epistemic—and, subsequently, social—advantages over political minorities (Thoma 2024). Most obviously, the preferred majority finds it easier to develop informed value judgments with reasonable access to information already aligned with their values. Conversely, political minorities find it harder to develop informed value judgements and easier for others to see their value judgements as uninformed, making competent and effective political participation comparatively harder for them. For instance, Rolin argues that scientific or intellectual movements (SIMs)—

collective efforts to pursue research programs in the face of resistance—are often apt allies for advocacy scholarship that takes social responsibility for giving voice to unheard groups. Most significantly, SIMs extend the pool of alternative value perspectives as a critical countermeasure to untrusted research (Rolin 2021). Conversely, political minorities may not trust information about vaccines that does not give their concerns a fair hearing. As a result, trustworthiness may vary between a political majority and political minorities because different ideologies may change what considerations should gain a fair hearing.

In practice, citizens can seek a reflective equilibrium between the information that peer deliberation judges as trustworthy and human intermediaries judge as trustworthy, without AI as a fixed authority adjudicating between them. In general terms, the first step is for citizens to use peer deliberation to revise what information they trust. The second step is for citizens to use human intermediaries to revise what information they trust. The third step is to repeat the process until an equilibrium is reached. So, an intermediaries approach does not put an unchecked (or undercontested) confidence into either peer deliberation or into human intermediaries. Any citizen confidence in peer deliberation is contested with human intermediaries, and any citizen confidence in human intermediaries is contested with peer deliberation. Without AI as an arbiter of truth in politics, this decentralised approach promotes the external trustworthiness of the information that survives continual contestation.

What regulates deliberation? A plausible answer is: more deliberation! In particular, 'meta-deliberation'—or, deliberation about deliberation—facilitates a reflective type of deliberation that deliberates about the advantages and defects of deliberation itself, and the participants are aware that such defects may also be present in meta-deliberation (Dryzek 2010). A central epistemic benefit of meta-deliberation is self-correction (Christiano 2012, 28). In particular, meta-deliberation can review the human intermediaries that inform public deliberation. Similar to experts, intermediaries have epistemic authority because they know more, but they are not above criticism (Moore 2017). Whatever the particular intermediary might be, the pieces of information that it transmits and the processes that it uses may become the subject of public deliberation. An intermediaries approach aims to contest and resist misinformation within the deliberative process with meta-deliberation. This promotes a system-level robustness. In other words, lay publics should primarily trust the process of deliberation rather than the participants in deliberation. An intermediary might not be trustworthy *in isolation*. However, an intermediary *embedded in a process of deliberation* may become trustworthy because it survives the strongest contestation available.

Unlike techno-legal approaches, an intermediaries approach does not aim to censor and prohibit misinformation about vaccines from entering public deliberation from the start with AI. An intermediaries

approach permits representative intermediaries—say, newspapers or podcasts—to aim to publicly justify the information about vaccines they transmit in the presence of wider contestation. As a countermeasure, rival intermediaries can aim to publicly contest misinformation. As explored next, this internal decentralised contestation of misinformation among intermediaries promotes the external trustworthiness of information better than the external centralised algorithmic censorship of misinformation. So, computer scientists should foreground contestation among fallible intermediaries rather than consensus among fallible experts to resist misinformation. However well human intermediaries may perform in absolute terms, they are not obviously worse than AI intermediaries, and human intermediaries may perform better.

Firstly, a noninstrumental reason to permit rather than prohibit misinformation about vaccines is to respect the moral integrity of political minorities with unpopular convictions. A centralised techno-legal approach is too insensitive to how specific situational factors change the external trustworthiness of information for particular publics. In practice, AI treats information as 'bare' facts (Boumans et al. 2025). In other words, the trustworthiness of information is treated as fixed independently of social context. For instance, Thagard wishes bots to police misinformation online (Thagard 2024, 3). However, as Winsberg argues, bots are distinctly unfit to interpret what information qualifies as misleading based on the psychological and social mechanisms that produce misinformation (Winsberg 2024, 3). Of course, misinformation may compromise the epistemic integrity of misinformed citizens. For instance, it may weaken their commitment to various epistemic norms. Nevertheless, algorithmic prohibitions against misinformation that unpopular minorities use to express their sincere moral judgments threaten to compromise their moral integrity and their ability to express their moral convictions as they judge best. In politics, literal truths are not the only—nor always the best—way to express moral judgements (Hannon 2021). As a result, an unchecked (or undercontested) confidence in, say, fact-checking bots, to resist misinformation threatens to harm the ability of human intermediaries to give unpopular minorities a fair hearing, compromising the moral integrity of the information ecosystem.

In contrast, a decentralised intermediaries approach facilitates a robust information ecosystem that both permits representative intermediaries to give unpopular minorities a fair hearing and permits rival intermediaries to resist misinformation through public contestation. As a result, human contestation is often better than AI censorship because it respects rather than compromises the strong speech rights of unpopular minorities. A respect for strong speech rights is critical to protect the moral integrity of political minorities committed to unpopular aims and the moral integrity of an information ecosystem committed to allowing unpopular minorities representation and allowing them a fair hearing.

Secondly, an instrumental reason to permit rather than prohibit misinformation about vaccines is to prefer verbal expressions of moral convictions to violent expressions. However bad misinformation might be, alternative ways to express the same moral convictions might be worse. An overly individualistic epistemology overlooks the epistemic significance of the collective information ecosystem. On the one hand, if prohibitive algorithms aim to censor misinformation, rival intermediaries are significantly less able to use public contestation to find out why the censored misinformation is attractive to particular publics. Censorship may *express that* misinformation is false.[13] However, censorship does not *explain why* misinformation is false. At best, fact-checking may only give generic explanations that miss what makes the misinformation particularly attractive to specific publics. In contrast, contestation among intermediaries often aims to explain why the misinformation is false to the specific publics in question. When AI censorship dominates human contestation, an unchecked confidence in prohibitive algorithms may epistemically harm the ability of human intermediaries to uncover enough about why the censored misinformation is attractive to particular publics and to produce effective counterspeech specifically for them.

On the other hand, if preachy algorithms amplify reliable information about vaccines, misinformed publics may see the information ecosystem as unrepresentative and rigged against them. Without contestation among intermediaries to help explain to particular publics why the disputed information is trustworthy, AI-amplified information may lack external trustworthiness. So, an unchecked confidence in either prohibitive or preachy algorithms may become socially irresponsible because the algorithms threaten to epistemically harm the ability of public deliberation to manage political disputes verbally. Politics is often seen as war by peaceful means. However, the less able human intermediaries can express the moral convictions of misinformed publics verbally, the more willing particular publics might become to express their moral convictions uncivilly or violently instead, especially if they feel significantly repressed.[14] So, human contestation is often better than AI censorship because it permits intermediaries to continue to represent misinformed publics verbally, which helps to reduce the risk that particular publics may feel the need to express their moral convictions uncivilly or violently instead.

Although an intermediaries approach has distinct advantages, this approach does have what I will call the symmetry problem. Whatever good intermediaries might do to gain the trust of particular publics, bad intermediaries can mirror to gain trust in misinformation about vaccines. The obvious solution to the

---

[13] For instance, Waldron defends the censorship of hate speech on largely expressive grounds Waldron, Jeremy. 2014. *The Harm in Hate Speech*. Harvard University Press.

[14] For instance, see Moore, Will H. 1998. "Repression and Dissent: Substitution, Context, and Timing." *American Journal of Political Science* 42 (3):851–73. https://doi.org/10.2307/2991732.

symmetry problem is to find a symmetry-breaker. With symmetry-breakers, AI can distinguish between good and bad intermediaries with seemingly symmetrical behaviour. For instance, expert consensus may empower AI to identify very bad intermediaries as the intermediaries that directly contradict the preferred bundle of expert consensuses about vaccines. Thagard argues that the abolition of very bad intermediaries is critical because efforts to re-inform misinformed people are unlikely to work (Thagard 2024, 3). So, if very bad intermediaries are not censored in advance of public deliberation, they may derail public deliberation and reduce the competence of the resulting political judgements.

Another solution to the symmetry problem is to concede that there are typically no symmetry-breakers. The absence of symmetry-breakers might be bad, but the pretence of a symmetry-breaker is worse. In practice, it is epistemically arrogant for computer scientists to confidently believe that their preferred bundle of expert consensuses gives them independent access to the truth despite peer disagreement. The typical absence of symmetry-breakers is a good reason to privilege public contestation to manage the persistent risk of misinformation, and to remain open to the live possibility that the preferred side is misinformed (Christiano 2012, 28). In practice, human intermediaries can use expert contestation to structure disagreement about what is not known. Similarly, they may use expert consensus to filter out false information for the publics they represent (Christensen et al. 2022, 93–94). However, human intermediaries do not become an arbiter of truth throughout the deliberative process. They remain embedded within ongoing contestation among rival intermediaries where any source of information remains open to revision and rejection (Shah 2021, 82–88). When algorithmic decisions about disputed information are seen as outside public deliberation and beyond revision and rejection, it protects fallible sources of information from public contestation and largely closes the possibility of self-correction as a result.

A knowledge intermediaries approach promises to restore trust in science. AI is dethroned as the arbiter of truth that regulates human intermediaries, as the privileging of any one fallible intermediary risks doing significant epistemic harm to the general contestation that the information ecosystem facilitates. More realistically, computer scientists should see AI as only one more fallible intermediary among many that is as open to contestation as human intermediaries. This approach is not anti-AI by default. It foregrounds the specific fragilities of AI that human intermediaries can counterbalance, with effective counterspeech that is sensitive to the specific situations of particular publics. Contrary to techno-legal approaches. AI need not and should not do all the work in knowledge transmission since AI does not evade the general dangers of centralising a fallible authority. In practice, contestation among human intermediaries does significant work to resist misinformation that the misuse of AI may epistemically harm.

5. Conclusion

The problem is not that computer scientists can easily know that techno-legal approaches do not work. The problem is that computer scientists cannot easily know that techno-legal approaches do work. So, computer scientists should not put the social responsibility to resist misinformation onto centralised techno-legal approaches. The social responsibility to resist misinformation is better put on human intermediaries. In practice, human intermediaries can more effectively promote the external trustworthiness of information and subsequently resist misinformation. In particular, human intermediaries are more able to align internally trustworthy information to the narratives of particular publics, and more able to adapt their narratives to accept internally trustworthy information and resist misinformation. As a result, computer scientists should become cautious of techno-legal approaches that may misuse fallible AIs to prohibit what information can enter public deliberation from the start, and potentially derail the effective counterspeech that contestation among human intermediaries otherwise facilitates. In its place, computer scientists should aspire to support and not derail contestation among human intermediaries as a social mechanism to resist misinformation within the deliberative process instead. A comparison between the centralised techno-legal regulation of information and no regulation assumes a false dichotomy. The decentralised social regulation of information through contestation among human intermediaries is an attractive alternative that techno-legal uses of AI may epistemically harm.

References

Anderson, Elizabeth. 2011. "Democracy, Public Policy, and Lay Assessments of Scientific Testimony." *Episteme* 8 (2):144–64.

Benson, Jonathan. forthcoming. "Is Fake News a Threat to Deliberative Democracy? Partisanship, Inattentiveness, and Deliberative Capacities." *Social Theory and Practice*

Boumans, Marcel, Joras Ferwerda, Maya J. Goldenberg *et al.* 2025. "Fostering Trustworthy Information: Countering Disinformation When There Are No Bare Facts." *Royal Society Open Science* 12 (6):250654. https://doi.org/https://doi.org/10.1098/rsos.250654.

Boumans, Marcel, Maya Goldenberg, and Sabina Leonelli. forthcoming. *Understanding Misinformation*. Cambridge University Press.

Caplan, Bryan. 2008. "Reply to My Critics." Critical Review: A Journal of Politics and Society 20 (3):377–413.

Christensen, Johan, Cathrine Holst, and Anders Molander. 2022. *Expertise, Policy-Making and Democracy*.

Christiano, Thomas, ed. 2012. Rational Deliberation among Experts and Citizens. Cambridge University Press.

Cordella, Antonio, and Francesco Gualdi. 2025. "Policymaking in the Digital Era: Exploring Techno-Legal Assemblages and Their Impact on Policy Formulation." *Government Information Quarterly* 42 (2):102023. https://doi.org/https://doi.org/10.1016/j.giq.2025.102023.

Croce, Michel. 2018. "Expert-Oriented Abilities Vs. Novice-Oriented Abilities: An Alternative Account of Epistemic Authority." *Episteme* 15 (4):476–98. https://doi.org/10.1017/epi.2017.16.

Dryzek, John S. 2010. Foundations and Frontiers of Deliberative Governance. Oxford University Press.

Elliott, Kevin J. 2020. "Democracy's Pin Factory: Issue Specialization, the Division of Cognitive Labor, and Epistemic Performance." *American Journal of Political Science* 64 (2):385–97. https://doi.org/https://doi.org/10.1111/ajps.12486.

Goldstein, Rena Beatrice, ed. 2021. Reconceiving Civic Competence for the Digital Age. Routledge.

Hannon, Michael, ed. 2021. Disagreement or Badmouthing? The Role of Expressive Discourse in Politics. Oxford University Press.

Hannon, Michael. 2023. "The Politics of Post-Truth." *Critical Review* 35 (1-2):40–62. https://doi.org/10.1080/08913811.2023.2194109.

Hardwig, John. 1985. "Epistemic Dependence." *The Journal of Philosophy* 82 (7):335–49. https://doi.org/10.2307/2026523.

Hauswald, Rico. 2025. "Artificial Epistemic Authorities." *Social Epistemology*:1–10. https://doi.org/10.1080/02691728.2025.2449602.

Herzog, Lisa. 2024. Citizen Knowledge: Markets, Experts, and the Infrastructure of Democracy. Oxford University Press.

Ioannidis, J. P. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8):e124. https://doi.org/10.1371/journal.pmed.0020124.

Joshi, Hrishikesh. 2020. "What Are the Chances You're Right About Everything? An Epistemic Challenge for Modern Partisanship." *Politics, Philosophy & Economics* 19 (1):36–61. https://doi.org/10.1177/1470594x20901346.

Kitcher, Philip. 2001. Science, Truth, and Democracy. Oxford University Press.

Landemore, Hélène. 2012. Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many. Princeton University Press.

Leonelli, Sabina. 2017. "Global Data Quality Assessment and the Situated Nature of “Best” Research Practices in Biology." *Data Science Journal* 16 (0):32.

Lepoutre, Maxime. 2019. "Can 'More Speech' Counter Ignorant Speech?" *Journal of Ethics and Social Philosophy* 16 (3)

Lepoutre, Maxime. 2024. "Narrative Counterspeech." *Political Studies* 72 (2):570–89. https://doi.org/10.1177/00323217221129253.

Longino, Helen. 2002. *The Fate of Knowledge*. Princeton University Press.

Lyons, Henrietta, Eduardo Velloso, and Tim Miller. 2021. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions." *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW1):Article 106. https://doi.org/10.1145/3449180.

Mill, John Stuart. (1859) 2011. *On Liberty*. Cambridge University Press.

Moore, Alfred. 2017. Critical Elitism: Deliberation, Democracy, and the Problem of Expertise. Cambridge University Press.

Moore, Will H. 1998. "Repression and Dissent: Substitution, Context, and Timing." *American Journal of Political Science* 42 (3):851–73. https://doi.org/10.2307/2991732.

Morgan, Mary S., ed. 2022. Narrative: A General-Purpose Technology for Science. Cambridge University Press.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop *et al.* 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1):0021. https://doi.org/10.1038/s41562-016-0021.

Nakov, Preslav, David P. A. Corney, Maram Hasanain *et al.* 2021. "Automated Fact-Checking for Assisting Human Fact-Checkers," paper presented at Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Survey Track,

O’Neill, Onora. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26 (2):293–300.

Rau, Martina A., and Anna E. Premo. 2025. "Systematic Review of Educational Approaches to Misinformation." *Educational Psychology Review* 37:43. https://doi.org/10.1007/s10648-025-10012-8.

Reiss, Julian. 2017. "Fact-Value Entanglement in Positive Economics." *Journal of Economic Methodology* 24 (2):134–49.

Rolin, Kristina H. 2021. "Objectivity, Trust and Social Responsibility." *Synthese* 199 (1):513–33. https://doi.org/10.1007/s11229-020-02669-1.

Romeo, Giuseppe, and Daniela Conti. 2025. "Exploring Automation Bias in Human–Ai Collaboration: A Review and Implications for Explainable Ai." *AI & SOCIETY* https://doi.org/10.1007/s00146-025-02422-7.

Saunders, Ben. 2023. "A Millian Case for Censoring Vaccine Misinformation." *Journal of Bioethical Inquiry* 20 (1):115–24.

Schwartzberg, Melissa. 2007. "Jeremy Bentham on Fallibility and Infallibility." *Journal of the History of Ideas* 68 (4):563–85.

Shah, Nishi. 2021. "Why Censorship Is Self-Undermining: John Stuart Mill's Neglected Argument for Free Speech." *Aristotelian Society Supplementary Volume* 95 (1):71–96. https://doi.org/10.1093/arisup/akab010.

Thagard, Paul. 2024. Falsehoods Fly: Why Misinformation Spreads and How to Stop It. Columbia University Press.

Thoma, Johanna. 2024. "Social Science, Policy and Democracy." *Philosophy & Public Affairs* 52 (1):5–41.

Turner, Piers Norris. 2021. "Introduction: Updating Mill on Free Speech." *Utilitas* 33 (2):125–32.

Vickers, Peter. 2023. Identifying Future-Proof Science. Oxford University Press.

Waldron, Jeremy. 2014. *The Harm in Hate Speech*. Harvard University Press.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Journal for the Philosophy of Science* 64 (2):233–53.

Williams, Daniel. 2023. "The Marketplace of Rationalizations." *Economics and Philosophy* 39 (1):99–123. https://doi.org/10.1017/S0266267121000389.

Winsberg, Eric. 2024. "“Falsehoods Fly: Why Misinformation Spreads and How to Stop It” by Paul Thagard. Columbia University Press." *The Journal of Value Inquiry* https://doi.org/10.1007/s10790-024-09996-3.

Wylie, Alison, ed. 2003. *Why Standpoint Matters*. Routledge.