British Journal of Political Science (2026), **56**, e16, 1–16
doi:10.1017/S0007123425101294

British Journal of Political Science

ARTICLE

Muting the Liars: A Democratic Response to Disinformation

Yi-Hsuan Huang

Department of Political Science, Swarthmore College, Swarthmore, PA, USA
Email: yhuang6@swarthmore.edu

(Received 21 July 2024; revised 6 December 2025; accepted 16 December 2025)

Abstract

Disinformation poses a serious threat to democracy, yet regulating it risks infringing on freedom of speech. This article defends the democratic legitimacy of regulating disinformation by distinguishing it from two similar forms of speech: 'false opinion' and 'toxic persuasion'. I argue that disinformation, as deliberate falsehoods intended to manipulate citizens' political judgment, does not merit protection. Regulation, on this account, is normatively legitimate and desirable when it safeguards citizens' ability to function as meaningful decision makers in the democratic common world. I then propose a dual-track model to identify removable content. Paired with regular review, transparency obligations, and an appeal process, this framework offers principles that help democracies to balance between protecting expressive freedom and resisting disinformation.

Keywords: Disinformation regulation; content regulation; disinformation; free speech; social media

Disinformation poses a pervasive threat to democracy. Through altered footage, deepfakes, and fabricated 'news', disinformation campaigns can manipulate public opinion at unprecedented rates and scales (see Association for Progressive Communications 2021; Jamieson 2018; Marwick and Lewis 2017). In response to this informational paralysis, content regulation has become increasingly common in democracies. For example, in the United States, major social media companies, such as Facebook, had self-imposed rules to remove information that interferes with 'the functioning of political processes' (Meta 2024). In the UK, platforms have to comply with the regulatory directives from the Office of Communications (Ofcom) to eliminate false content. Similar protocols also operate in France, Germany, and other democracies (Coe 2024; Craufurd-Smith 2019; Loebbert 2022). Empirical studies show that these measures help to reduce the circulation of false messages and frustrate disinformation networks by increasing the costs of dissemination (Clifford and Powell 2019; Fielitz and Schwarz 2020). Yet the normative legitimacy of these interventions still awaits careful examination.

Is content regulation compatible with freedom of speech? This question has already attracted sustained legal attention in democracies. Especially in the United States, many commentators have warned that regulation can violate the constitutional protection of free speech (Bennett and Livingston 2020; Brison and Gelber 2019; Bromell 2022; Epstein 2020; Macfarlane 2021; Samples 2019; Susskind 2022). As a scholar of political theory, I approach this question from a different angle. Instead of focusing on constitutional compatibility, I look into whether it is normatively

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

defensible for democracies to censor false online speech. Even if regulation is deemed permissible in a constitutional framework, a normative investigation has important value for democratic politics in its own right – since online platforms have become the major site where public opinion is formed (for example, Forestal 2021), if what content can be subject to removal is not *consistently defined and normatively justified*, we end up relying on any regulating entities, be it a private platform or governmental agency, to impose arbitrary restrictions on everyday political speech (Bromell 2022; Susskind 2022). Although this inquiry was motivated by the debates in American politics (particularly, whether private companies can remove posts or deplatform speakers under the First Amendment), it responds to the broader dilemma confronting most democracies: to balance between protecting free expression and safeguarding political processes.

This article assesses the normative tension between free speech and falsehood regulation. Recent scholarship usually concurs that restrictions can be placed on commercial advertising, defamation, child pornography, incitement, and terrorist advocacy, but remains divided over whether hate speech should be regulated (see, for example, Heinze 2016; Waldron 2014). As a result, the normative discussion of speech regulation mainly focuses on instances of hateful remarks, leaving the legitimacy of falsehood restrictions rather underexplored (Brison 2021; Brown 2015; Gross and Kinder 1998; Heinze 2016; Leaker 2020; Waldron 2014; Whitten 2021). On intuitive grounds, falsity *per se* does not constitute sufficient reasons for sanction. Error is a common human experience. To the extent that online platforms have become the major site for political conversations, regulation can hinder public deliberation, endangering both personal and political freedoms. However, if not *all* falsehoods are morally problematic, any account seeking to defend content regulation must be able to answer two interrelated questions: in what respects does disinformation differ from other factually false political speech, and why should that distinction(s) warrant restriction?

In what follows, I contend that disinformation is conceptually different from two other categories of speech: false opinion and what I call 'toxic persuasion' – rhetoric that strategically represents the speaker's view to appeal to audiences. Drawing from Hannah Arendt, I argue that opinion, despite potential errors, conveys the speaker's unique perspectives and lived experiences. Disinformation, as deliberate falsehoods seeking to mislead audiences' political judgment, should not be regarded as opinion but as *lies*. A democratic society protects factually inaccurate opinions as it respects members' unique perspectives and their equal authority in collective governance. Lies that seek to distort collective decisions do not bear the same moral weight. Additionally, while persuaders sometimes employ problematic rhetoric to mobilize, disinformation speakers differ in the way that they make baseless claims without credible evidence, avoid questions, and discredit any competing sources.

By articulating these distinctions, this article seeks to provide both a theoretical and practical pathway for democratic citizens and policy makers to navigate two competing ends: ensuring the free exchange of opinions while protecting collective decisions from outright manipulation. Borrowing from the insight of Shiffrin (2014), I argue that the legitimacy of regulation lies in the need to protect citizens as effective thinkers, and, in turn, as meaningful decision makers in collective governance. Given the inherent challenge of identifying disinformation, I conclude that legitimate regulation can only aim to *reduce* rather than eliminate disinformation. Along this reduction framework, I propose a two-track identification system to further narrow down the content that can be subject to removal. Democracies should also seek to install procedural checks and balances, such as periodical reviews, transparency, and an independent appeal system, to prevent misuse.

The Dilemma of Disinformation Regulation: Two Competing Ends

In July 2023, Taiwan's Ministry of National Defense issued a public clarification in a press conference: no, its annual Han-Kuang drill *was not* a 'rehearsal' for the then-President Tsai to 'flee' to the United States (Wang 2023). Fact-checking organizations traced the viral claim and found

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

that the original post, which alleged Tsai of 'intending to flee with Taiwanese assets' after 'provoking the Chinese government', came from Weibo and resurfaced on Facebook every year during the drill (MyGoPen 2021; Taiwan FactCheck Center 2023). Similar fabrications have unsettled the democratic world. For example, AP News now has a dedicated page of 'Not Real News: A look at what didn't happen this week'. One recent entry confirmed that – no, the federal government *does not* give each migrant a $5,000 gift card at the border. Another one verified that – no, there *was not* a crowd chanting 'F–k Joe Biden' at a football game. The video has been digitally altered to insert the chant (Goldin 2023).

Disinformation can take various forms. Common examples include false connections (the headlines or visuals do not match the content), false context (genuine content is shared with false contextual information), imposter content (genuine content is impersonated to mislead), and fabricated content (the story is completely made up to deceive) (Wardle 2020). A 2021 European Parliament report catalogs a few common strategies of disinformation actors: spreading false content through automated bot networks, masking content sponsorship (to give the false impression that it comes from real grassroots activists), impersonating authoritative media and public figures, and digitally altering or fabricating images and videos (Colomina et al. 2021).

For definitional clarity, this article joins the growing body of media studies to identify disinformation as deliberate false messages intended to mislead and deceive, which differentiates it from unintentional falsehoods as misinformation (Freelon and Wells 2020; Jack 2017; Wardle 2020; Wardle and Derakhshan 2017). Specifically, I refer to disinformation as *false content deliberately created to mislead and manipulate the political judgments of its audience*. This means that a piece of disinformation contains at least three elements: (i) the speaker believes that a statement is factually erroneous, (ii) the speaker consciously presents as what they believe to be factually true, and (iii) the *purpose* of this misrepresentation is to misguide the audience's political reasoning.[1,2] Although the third element is not always included in the definition of disinformation studies, there is no doubt that most disinformation campaigns intend to impact political events (for example, Colomina 2021; Engler 2019). Disinformation raises questions in normative political theory, not because it misinforms listeners on any random matters. Rather, as Bennett and Livingston (2020, 3) point out, disinformation actors aim to 'advance political goals', and their strategies often involve 'discrediting opponents, disrupting policy debates, influencing voters, inflaming existing social conflicts, or creating a general backdrop of confusion and informational paralysis'. This article thus follows Bennet and Livingston's observation and limits its focus to misleading content that has a *political orientation*. The scope of analysis thereby excludes, for example, commercial marketing ploys or a friend's mischievous pranks.

Social media and internet platforms allow information to be transmitted at an unprecedented rate, making democratic discourse vulnerable to manipulation. Content regulation (the removal of false content or fake accounts, specifically) has since become a common defensive tool. In the United States, such measures are largely voluntary initiatives undertaken by private platforms in the absence of statutory mandates. By contrast, the United Kingdom's Online Safety Act 2023 obliges platforms to curb disinformation and misinformation following the directives of Ofcom (Coe 2024). France's 'bills against manipulation of information' empower electoral candidates to seek judicial removal of false content that could compromise electoral integrity, with judges required to issue rulings within forty-eight hours (Craufurd-Smith 2019). While these interventions can impede the spread of deceptive materials and enhance the costs of disinformation campaigns (Clifford and Powell 2019; Fielitz and Schwarz 2020), their legitimacy remains ambiguous.

[1]This combines Shiffrin's definition (2014, 12) of lies and Freelon and Wells' definition (2020, 145) of disinformation.

[2]It is worth noting that disinformation need not succeed in deceiving its audiences to mislead. False content can impair collective decisions by inundating audiences, increasing the cost for individuals to identify credible content, and disabling public discussion by depriving a sense of shared realities (Engler 2019).

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

In legal studies, opinions are divided over the constitutionality of falsehood regulation (Bromell 2022; Samples 2019; Susskind 2022). In the United States, critics argue that any governmental intervention infringes the First Amendment, whereas others hold that a careful reading of Section 230 of the Communications Act could justify a more expansive role of the government (Franks 2019). In Europe, legislative restrictions on speech appear more readily embedded in constitutional frameworks. Article 10 of the European Convention on Human Rights clearly states that exercising the freedom of speech 'carries with it duties and responsibilities, [and] may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society'. Nonetheless, legal disputes persist. For example, in 2022, Germany's Court of Cologne ruled that its Network Enforcement Act (NetzDG, which requires social media platforms to moderate false information) was in partial violation of the EU law – in part because the Federal Office of Justice was deemed not 'independent enough' to supervise media services (Loebbert 2022). Despite extensive discussions in legal literature, a focused normative investigation remains lacking.

On the normative grounds, two interrelated questions beg more clarity: how does disinformation differ from other false accounts, and does the difference justify its removal? Without consistent principles that distinguish disinformation from other false expressions, even well-intended regulation can end up censoring speech that ought to be protected. In fact, opponents of regulation often question whether a consistent distinction is even tenable. The CATO Institute argues that the definition of disinformation is too vague to translate into regulating guidelines (Samples 2019). Even parties that see regulation in a more positive light, such as the European Parliament, conceded that 'It is not yet clear where the dividing line is between legitimate political persuasion and illegitimate manipulation of thoughts' (Colomina et al. 2021). And yet, as our political life is increasingly mediated by digital platforms, false speech has a much higher chance of inflicting systemic harm. Outright rejections of interventions can render democracies defenseless against co-ordinated attacks. Addressing the threats of disinformation, therefore, requires a careful re-articulation of the boundaries of free speech in contemporary contexts.

If falsity alone is insufficient to justify removal, to answer the question of legitimacy, we need to first examine what is at stake when liberal democracy does not protect false speech. In the next section, I review three principal justifications for free speech and evaluate how each responds to the challenge of disinformation.

Free Speech, False Opinion, and Lies

Free speech is often defended for three primary values: the pursuit of truth, individual freedom, and the functioning of a democratic state (Barber 1984; Brison 2021; Bromell 2022; Moon 2021). In other words, speech is not protected simply to have 'more speech', but because its protection brings some essential goods for liberal democracy (Brison 2021). Hence, a defense of free speech must explain how it advances the targeted value in a given context. This section reviews three free speech defenses in the context of disinformation. I then argue that the democratic defense is the one that poses the strongest *objection* to falsehood regulation – and yet, even on this account, disinformation as deliberate falsehoods still falls *outside* the ambit of protection.

Truth-seeking is one of the most commonly referenced reasons to defend expressive freedom in liberal societies. John Stuart Mill (2015[1859]) famously argues that even factually erroneous opinions should be freely aired, since public debate and deliberation may expose neglected truths. In the same spirit, Justice Brandeis characterizes free speech as a 'means indispensable to the discovery and spread of political truth', and insists that the remedy for fallacies is 'more speech, not enforced silence' (Whitney v. California 1927). However, this truth-seeking optimism is subject to empirical challenges. Recent studies find that false beliefs have a strong stickiness in our

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

belief systems (Baekgaard et al. 2019; Gilbert 1991; Kunda 1990; Taber and Lodge 2016). Some experiments even show that counterspeech can entrench rather than dislodge errors (Nyhan and Reifler 2010; Lewandowsky et al. 2012).[3]

Taking account of today's media environment, Wu (2018) notes that the First Amendment was written in a time when speech was highly censored and scarce, and many of its assumptions no longer hold. Now the amount of speech often overwhelms individuals and impedes meaningful engagement with one another. Similarly, Napoli (2019) points out that online platforms do not provide the conditions for truth discovery as Millians envisioned, since the filter bubbles and algorithms rarely expose users to diverse opinions or motivate high-quality debates.[4] This means that if truth were the main concern, tolerating blatant falsehoods can be counterproductive. Restrictions on disinformation can potentially relieve information fatigue and create space for productive discussion. Therefore, the truth defense of free speech does not seem to give definitive reasons to refuse to limit disinformation.

Another common defense frames free speech as an essential means to self-realization. Indeed, the ability to verbalize internal thoughts is central to personal freedom. But the right to free speech is never absolute. Defamation, fraud, and false advertising are well-established domains where deception incurs legal sanction. On the other hand, as social media and new technologies enable new expressive activities, it can be argued that we never had the 'liberty' to broadcast our words to a global audience. Regulating online spaces is thus different from policing in-person conversations. Given the level of influence of online platforms, one can argue that the scope of expressive freedom in online spaces should be more restrictive.[5] No matter where the judgment falls, unless one endorses free speech absolutism, it is well recognized that expressive freedom can be subject to reasonable constraints. For those who seek to contain disinformation, the 'freedom' to disseminate deceptive information is simply not a meaningful one to protect. Especially when regulation targets online platforms and leaves expressions in other domains intact, the self-realization defense does not seem to give sufficient reasons to deny content regulation.[6]

Finally, free speech is defended as a component of democratic life (for example, Post 2011). The freedom to articulate one's views and convictions allows citizens to communicate with one another. It enables the key functions of collective self-government, such as agenda-setting, the defense of interests, negotiation, and persuasion (Barber 1984, 178–79). Post (2011) maintains that an opinion can be factually erroneous, but having the freedom to express it and impact public discourse is what affirms a person's equal standing in the collective decision-making process. Free speech is therefore *constitutive* of democracy. Sanctioning some viewpoints while permitting others jeopardizes the political equality of citizens (Heinze 2016). Through these lenses, the democratic defense gives rise to a compelling objection to falsehood censorship. If the purpose of free speech is to allow citizens to communicate freely as political equals, factual accuracy alone is not a legitimate reason to deny protection. Restricting false content online will deprive individuals of equal opportunities to participate in the formation of public opinion and to influence collective decisions in the way they see fit.

[3]The recent study of Wood and Porter (2019) points to some uncertainties in this claim.

[4]Some might argue that this can be solved by improving algorithms. But it is unclear if, without quality gatekeeping of the information pool, amending algorithms is sufficient to provide a truth-seeking environment.

[5]Another argument is that if we conceive individual freedom not only as the ability to express, but also the ability to self-realize as rational agents, a personal freedom defense may even give reasons to limit disinformation. For example, Bonotti and Seglow (2021) point out that fake news exploits listeners' cognitive biases and thwarts their autonomy.

[6]Drawing from Mill, one related argument against speech regulation posits that while some speech is indeed problematic, the state should not be entrusted with regulation decisions. Giving the government the right to do so will lead to the abuse of power and ultimately infringe on individual freedom. While the concern of misuse is valid, as I explained earlier, our expressive right is never absolute, and the state is *already* entrusted to regulate many domains of speech. In *Conclusion: Reclaiming a Democratic Collective Voice*, I respond more specifically to the worry of governmental abuse in the context of disinformation control.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

Yet condoning the democratic harm of disinformation can run into what Bonotti and Seglow (2022) call the problem of 'internal balancing'. That is, safeguarding the speaker's right to utter problematic information can preclude others from enjoying the same freedom effectively. In reply to Post, Brown (2015) points out that hate speech often silences victims and undermines their 'fair and real opportunities' to impact public opinion. If equal rulership is what is at stake, democracies might have a positive reason to *limit* hate speech rather than leaving it unchecked (Brown 2015, 194–96). In a similar vein, extensive deceptive information can counteract the democratic purpose of free speech – to allow *meaningful* exchanges between citizens in collective governance.

I argue that breaking this impasse requires that we draw a fine line between respecting diverse opinions and being indifferent to all kinds of non-facts. In this regard, Hannah Arendt's *Truth and Politics* offers a helpful roadmap. On the one hand, Arendt (2006 [1967], 232–36) insists that factual truth is unshakable, independent of interpretation, and distinct from opinion. The stubbornness of factual truth is manifest in a statement such as 'Germany invaded Belgium in August 1914', whose validity can be affirmed even by 'the most extreme and sophisticated believers in historicism', and a contrary claim is 'clearly an attempt to change the record' (Arendt 2006 [1967], 235–45). On the other hand, Arendt (2006 [1967], 236) warns that truth 'peremptorily claims to be acknowledged and precludes debate, and debate constitutes the very essence of political life'. This is not to say that truth stifles democracy. Quite the contrary; for Arendt, truth is what underpins the formation of sound opinion. At the same time, political communication requires more than stating facts, but also making such facts relevant to our community.

Imagine that one says, 'It rains'. On Arendt's account, the meteorological fact that 'water drops from the sky' will not have any public bearing unless it gives rise to a shared implication for the community (that 'the river wall needs repair' or 'the water-restriction policy should end'). Arendt acknowledges that in forming an opinion, 'our thinking is truly discursive, running . . . through all kinds of conflicting views', and some opinions will be mistaken. But an opinion is always an attempt to communicate to our fellow members: this is how 'I' see the world (Arendt 2006 [1967], 236–37). Re-reading Arendt, Zerilli (2016, 265) thus stresses that having diverse opinions is the foundation of a democratic life – a democratic common world emerges only 'when there is a plurality of worldviews', since '[o]ur sense of what is common, "the sameness of the object," can appear only when it is seen from multiple perspectives'.

Competing claims naturally arise. Two individuals can witness the same rain, and one attributes it to supernatural power, the other to meteorology. Arendt offers no conclusive formula for reconciling opinions (that might or might not be informed by truth), but she does suggest that ideal opinion formation involves envisioning oneself in the standpoints of others (Arendt 2006 [1967], 237). The more positions that one can inhabit, the better quality the opinion is. The implicit lesson here, I argue, is that effective political communication requires a willingness to move beyond material disagreements and recognize the lived experience that an interlocutor seeks to convey. By drawing on Arendt, I do not mean that citizens must be able to practice this ideal mode of communication at all times. Nor do I imply that every opinion deserves equal intellectual weight in political reasoning. What I mean to stress is that since opinions are the vehicles of personal beliefs and preferences, members of a democracy only get to create a shared world with others as political equals when their opinions can enter freely into the public realm and be subject to consideration or objection.

Recognizing this, it is only reasonable that Arendt draws a sharp distinction between opinion and lie. If the equal moral weight of an opinion stems from respecting a speaker's shared authority in collective governance, deliberate falsehoods lose their bearings. Regardless of its empirical accuracy, an opinion is an attempt to describe 'how things appear' to one's eye. A lie, by contrast, does not 'reflect upon personal truthfulness' and, instead, is a conscious misrepresentation of the speaker's view (Arendt 2006 [1967], 245). While self-deception is always possible, disinformation does not fall into that category. Rather, as deliberate falsehoods created to mislead, disinformation belongs to what Sissela Bok (1978, 17) classifies as 'clear-cut lies' – whose 'intention to mislead is

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

obvious, where the liar knows that what he is communicating is not what he believes, and where he has not deluded himself into believing his own deceits'.

Revisiting Arendt's account allows me to draw two implications for this inquiry. First, indeed, as Post contends, the freedom to participate in the formation of public opinion is what affirms one's equal authority in democratic governance. But the purpose of this freedom is to ensure that personal viewpoints and convictions can enter the public realm and shape collective decisions. A deceptive message that carries no authentic views does not give rise to the same moral claim. Second, if we take free speech's democratic function seriously, democratic government may even have a positive duty to protect collective decisions from outright manipulation. Bok (1978, 20) has a sharp description of how lies undermine our agency as decision makers: 'All our choices depend on our estimates of what is the case; these estimates must in turn often rely on information from others. Lies distort this information and therefore our situation as we perceive it, as well as our choices'. In the same way, disinformation deprives individuals of the chance to make meaningful choices among viable options.

Here, Shiffrin's critical analysis of the relationship between free speech and lies offers a valuable lens to analyze the threat that disinformation poses to citizens' political agency. Shiffrin (2014, 82–87) argues that dominant free speech theories tend to only emphasize our interests as listeners (who seek unrestricted access to information) or speakers (who wish to express without interference). What these frameworks neglect, she contends, is our interests as *thinkers* – who listen in order to process and apply information, and who speak, not merely to express, but to formulate and update our thoughts through the conversations with others. On this account, thinking, the ability to interpret interactions and develop judgment, is what unites the purpose of both listening and speaking. By feeding others with untruthful information, what lies erode is a person's moral agency as a reflective thinker (Shiffrin 2014, 23–24, 90–92).

Departing from Shiffrin, I do not claim that free speech is only meant to serve our role as thinkers. What I mean to stress is that individuals are listeners, speakers, and thinkers all at once. And these roles can come into tension. Protecting a speaker's right to disseminate falsehoods can compromise others' ability to receive credible information and exercise judgment. In the case of disinformation, deceptive speech denies others as co-founders of a common world by distorting the epistemic basis on which their judgments rest. In this light, free speech becomes counterproductive if it only ensures participation in public discourse without enabling meaningful collective governance. Regulating disinformation is not only compatible with free speech's democratic function, but also desirable – if we are to protect authentic exchanges in the creation of a shared world.

Toxic Persuasion and Norm-Disrupting Speech

I have argued that disinformation differs from factually false opinion since it does not convey authentic personal beliefs and, therefore, is not entitled to protection. However, if a speaker's authenticity is what determines the normative status of false speech, a further question arises: where would *persuasive rhetoric* fall within the very same framework? To mobilize an audience, political speakers often *knowingly* simplify complex events, exaggerate the benefits of a policy, or downplay anticipated costs. These tactics very much blur the line between an honest opinion and a lie. Nonetheless, persuasion plays an important role in political speech, and it seems troubling to conflate it with disinformation. In this section, I differentiate disinformation, as norm-disrupting speech seeking to disturb public deliberation, from ordinary persuasion. I argue that, although both disinformation and persuasion can misrepresent how the speaker perceives the objective world, only the former warrants restrictions.

Some persuasion strategies are indeed *toxic* for meaningful deliberation. In *Gorgias*, Plato famously warns that mobilizing rhetoric can conceal a speaker's convictions. It is a 'knack' rather

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

than a 'craft', 'one that perpetrates deception by means of shaping and coloring, smoothing out and dressing up' (Plato and Zeyl 2012 [1987], 24–25). Such strategic rhetoric seeks to mobilize persuadees without full disclosure of the speaker's underlying rationale. For similar reasons, Keith Dowding (2016, 13–14) argues that non-manipulative persuasion is only possible when the speaker and the listener both agree on some criteria on how to verify the evidence in a persuasive claim, even at some abstract level. Absent such an agreement, a prompted behavior change nonetheless involves manipulation since the speaker will have to appeal to reasons that they do not genuinely share to mobilize the listener. (For example, imagine a science believer convinces a magic believer to take a pill not because of its scientific benefits but because of its superpower.)

It is helpful to identify the moral defects in persuasion when striving for better communication. But one might question whether a 'purified' mode of communication is achievable, or even desirable, for our political life. As Markovits (2006, 258–59, 262) points out, strategic language is a natural human experience: 'In different contexts, the same statements can mean very different things; the same delivery affects different audiences in diverse ways [. . .] and the wrong voice was unduly limit the prospects of being heard'. Not to dismiss the potential harm of toxic persuasion, its moral murkiness, as I see it, is that by presenting something consistent with the listener's pre-existing view, the speaker bypasses a conscious (and often critical) evaluation process of the audience, leaving their genuine reasoning unexposed and unexamined. And yet, if democratic communication values two-way exchange rather than one-way preaching, strategic rhetoric can actually help to explore common grounds and enhance mutual understanding when it makes audiences feel more seen and heard. More importantly, conflating persuasion with disinformation overlooks important differences between them and the normative issues that ensue.

Indeed, common persuasion can involve conscious overstatement or understatement, such as 'we've created more jobs than any government in the last two years' or 'a *thousand* billionaires only pay 3 percent in taxes on average', when the data only covered 400 families, or oversimplification and ambiguity, such as 'voting for me is voting for gender equality'. But are these claims morally indifferent to a statement such as 'a suitcase of ballots was brought into the polling station' or 'the federal government is giving immigrants $5,000 gift cards'? The key distinctions, I maintain, lie in the way in which the speech *is used* to penetrate the political world and the way that the speaker responds to the shared norms of public discourse. When persuasion, rather than deception or manipulation, is the primary goal, a speaker usually avoids completely baseless claims and is willing to provide evidence or respond to questions when challenged. In contrast, a disinformation speaker tends to (i) make baseless assertions without credible evidence, (ii) refuse to justify or correct claims when pressed, and (iii) urge listeners to disengage from other sources and rely solely on the speaker's authority.

To illustrate these distinctions more concretely, consider first the content of a narrative. Toxic persuasion tends not to be completely baseless and can retain a measure of validity in certain contexts. Imagine a self-proclaimed feminist candidate making a case that 'voting for me is voting for gender equality'. While exaggerated, such a claim may still be reasonable if the candidate has paid sustained attention to related issues. The situation is very different when, for example, during the campaign period of the marriage equality referendum in Taiwan, opponents of same-sex marriage asserted that if the new legislation is passed, 'we are no longer allowed to call our mom "mom" or our dad "dad", and that these familial terms would be replaced with gender neutral "parent 1" and "parent 2"' (Liang 2016). This kind of statement has no grounds in any given context. The claim may be somewhat less troubling if there were precedents of policing this language, which was not the case.

The second distinction concerns the speaker's responsiveness to questions or challenges. A persuader might employ a selective or misleading framing. But we can generally expect them to provide further justification, or even correction, when pressed. For example, in defending a tax policy, a persuader might highlight only favorable statistics yet still be willing to explain how the numbers were calculated, reference credible sources, or acknowledge mistakes. Whereas a

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

disinformation speaker often remains unresponsive, simply reiterates the same claim, or supplies fabricated 'evidence'. This question-avoidant attitude also marks the third distinction: while a persuader might selectively cite evidence in favor of their view, they are generally willing to respond to counter-sources and explain the discrepancies. Disinformation speakers, on the other hand, tend to delegitimize all the rival claims, calling them fake or politically tainted, and urge listeners to disengage from other sources and rely exclusively on the speaker's authority.

Beerbohm and Davis (2023) describe this latter conduct as a form of epistemic abuse (what they call 'political gaslighting') – in which the speaker induces the audience to *suspend cognitive exercises* in their belief-forming process, and makes them reliant on the speaker's authority for political judgment. Speech in this manner is different from mobilization. While both mobilizers and gaslighters may present information selectively, it is the latter who attempts to limit listeners' thought process, discourage engagement from alternative sources, and tether 'the listener's loyalty' to the speaker and political leaders (Beerbohm and Davis 2023, 875). Beerbohm and Davis (2023, 874) also suggest that persuasion can be somewhat 'innocuous' when one can 'reasonably expect' that the speaker may 'spin on the facts' for a given purpose. In their example, a swimmer might anticipate that their coach will exaggerate the odds of winning as a motivational strategy; the exaggeration becomes less problematic when there is a shared understanding that speech may be used in this way.

Likewise, in our political life, misrepresentation is often anticipated. We thus expect public speakers to be ready to face questioning. Under this premise, misleading persuasion becomes less troubling, not because dishonesty is blithely welcomed, but because every public speech is subject to a process of scrutiny. Challenges, debates, and interrogations are the cornerstone of political communication, since it is only through these mechanisms that citizens *test a speaker's truthfulness*, with a shared understanding that some lies will inevitably sail under the radar. While what constitutes a credible justification is always subject to debate, this process becomes meaningless if the speaker simply evades the scrutiny altogether – repeating baseless claims, dismissing all counter-evidence, refusing to answer questions, and urging audiences to detach from competing information. In doing so, the speaker not only avoids justifying their claims but also denies the norms of public discussion. It is in these respects that disinformation, as norm-disrupting, manipulative speech, differs from ordinary persuasion.[7]

By making these arguments, I do not suggest that there is a perfect system to differentiate disinformation from toxic persuasion. Some cases will inevitably fall in the gray area. However, there are two implications that this analysis seeks to foreground. First, these criteria help to resist the relativist tendency that equates speech with very different orientations as morally indifferent. Second, disinformation, particularly as norm-disrupting, manipulative speech, is not entitled to the same protection as persuasion or mobilization. Taken together, my analysis is meant to show that the differences between persuasion and disinformation warrant not only conceptual separation but also differentiated normative treatment.

Indeed, this section suggests that distinguishing between toxic persuaders and disinformation disseminators requires observing the course of action of a speaker. As indecisive as this sounds,

[7]Rosenblum and Muirhead (2019) offer a related observation in their account of what they term 'new conspiracy theory'. They argue that classic conspiracy theories, however implausible, often adhered to certain 'scientific and journalistic' conventions. Whereas new conspiracism increasingly relies on sheer repetition and unfounded accusation – that is, 'if one cannot be certain that a belief is entirely false, with the emphasis on entirely, then it might be true – and that's true enough' (Rosenblum and Muirhead 2019, 20, 43). This shift in conspiracist rhetoric can indeed blur the line between disinformation and conspiracy theory. Yet I maintain that the two terms are meant to characterize the speech in question as distinct political acts. Although disinformation often draws on conspiratorial narratives, its defining feature is that the speaker deliberately misrepresents their views. In contrast, Rosenblum and Muirhead (2019, 20) define conspiracy theory as 'an attempt to explain a specific event' that typically seeks to 'reveal secret plots' behind it. In this framing, the conspiracist is presumed to be sincere – or at least their intent remains indetermined. The term thus describes an ideological stance, marked by its distrust towards authority, rather than a calculated effort to deceive.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

this is also how the court assesses problematic speech, such as libel or slander: by examining how the speech in question is deployed in a context. Democratic decision making is a multistep process. To detect a manipulator hence requires looking closely at how speech is employed to serve the speaker in a sequence of events. An obvious example is when the speaker self-contradicts, such as denying the safety of vaccines but receiving them secretly. Seeing speech as a political act, a disinformation speaker often reveals that their goal is not to persuade on a matter of genuine concern, but to *mislead* – by discrediting opponents, delegitimizing other sources of authority, their primary goal is often to obstruct public communication.

When disinformation is widely condoned, it generates democratic costs beyond distorted policy choices; it leads to the paralysis of the public sphere. In *Truth and Politics*, Arendt (2006 [1967], 246) offers a vivid foresight of what a society dominated by 'organized lying' looks like: 'The result of a consistent and total substitution of lies for factual truth is not that the lies will now be accepted as truth, and the truth be defamed as lies, but that the sense by which we take our bearings in the real world – and the category of truth vs. falsehood is among the mental means to this end – is being destroyed'. Pervasive disinformation breeds what McKay and Tenove (2021, 708) call 'epistemic cynicism', where citizens lose trust in any information and withdraw from public discussion. If enabling democratic communication is the primary reason we protect false speech, reducing the amount of disinformation is, in fact, crucial to preserving the very space where political discourse takes place.

Muting the Liars: A Democratic Model of Regulation

I have argued that disinformation differs from factually false opinion and toxic persuasion and thus falls outside of the scope of free speech. Nevertheless, to say that restricting a category of speech does not *inherently* infringe on expressive freedom is different from saying that a specific regulation is legitimate. Designing appropriate regulation raises a separate set of concerns. Given how fluid and ubiquitous speech is, the primary challenge for policy makers is to ensure that disinformation controls do not stray beyond legitimate targets and suppress other utterances. Since different constitutional frameworks allow different forms and degrees of interventions, the goal of this section is not to offer jurisdiction-specific prescriptions, but rather, to articulate the normative considerations behind institutional design and suggest potential responses. Following earlier theorization, I propose a 'reduction model' and a two-track identification system to identify content that can be subject to removal. While I leave the choice of regulating entity open, I outline factors that favor legislative interventions.

Disinformation poses a serious threat to democratic governance. However, if what distinguishes disinformation from false opinion and toxic persuasion is its *deceptive orientation*, accurate detection seems to require reading the speaker's mind. A speaker's genuine intent can be murky and disguised, and misjudgments can end up censoring legitimate speech and endangering political communication. For this reason, I follow Susskind's suggestion (2022) that regulatory models should aim to *reduce* the amount of disinformation rather than *eliminate* it.[8] More precisely, I argue, a reduction model should make two important concessions. First, it should recognize that detecting every piece of disinformation is impossible, and neither should this be the goal. Reduction rather than elimination is a necessary trade-off to protect expressive freedom. Second, removal decisions should rest on active indicators of manipulation. This means that the range of content that can be legitimately removed will be *smaller* than the actual scope of disinformation.

[8]Though I have borrowed the term 'reduction model' from Susskind (2022), my regulation proposal differs from his. Susskind proposes to sort platforms on their scales and the level of social risks they pose. Large-scale and high-risk platforms should be subject to more stringent regulation and reduce false content more rigorously. His proposal does not explain how to differentiate disinformation from other types of falsehood and whether they deserve different treatment.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

How exactly does this reduction model operate? My suggestion is that a regulatory framework should distinguish between two categories of information and assign different procedures. The first category includes materials where manipulation is self-evident: a bot account, an edited video, a fabricated news headline, or a doctored image. In addition, since disinformation is usually generated by malicious entities, there are often traces of co-ordinated attacks. Security analysts and fact-checking companies often use Coordinated Inauthentic Behavior (that is, multiple inauthentic accounts pushing identical content), IP tracing, or linguistic traits to identify a connection with disinformation campaigns (for example, Rastogi and Bansal 2022). In these cases, the evidence of distortion is explicit, and the content does not represent sincere belief. Here, media platforms or regulators can proceed with removing the content.

In contrast, some information can be completely inaccurate or baseless, but the signs of manipulation are not immediately discernible. This kind of content falls into the second category, which requires a more careful detection process. Identifying disinformation in these cases requires observing closely the context in which the speech is made and the speaker's course of action – whether they dismiss the need for evidence, refuse to answer questions, and discourage audiences from engaging alternative sources. This process demands significant time and resources, and can risk chilling public discourse if applied to every user. Limiting the targets is therefore necessary.[9] In practice, one can envision restricting such scrutiny to high-visibility accounts, the accounts of celebrities, posts and messages that garner wide attention, or content that covers significant events (such as a presidential debate). In France, for instance, the 'bills against manipulation of information' focus on false information spread within three months before elections that has the potential to undermine 'the reliability of the elections' (Craufurd-Smith 2019).[10]

In terms of regulating entities, when all else is equal, a legislative framework is likely to yield more effective outcomes. This is because, firstly, legislation can provide consistent standards across platforms and better counteract the issue of account migration, whereby users shift to less-regulated venues to evade restrictions (Innes and Innes 2023). For example, the case of Gettr, founded by the team of Donald Trump after his suspensions from Facebook and Twitter (McGraw et al. 2021), underscores the importance of consistent regulation across sites. Legislative interventions that apply to *all* platforms can more effectively close these regulatory gaps.

Secondly, authenticating contested content may require systemic co-operation across platforms. For example, an original photo might appear on Facebook and then be modified and reposted on Twitter. Decision makers might need access to the original image to identify the sign of manipulation. *When* and *how* information is circulated can also be of significance. The same message can have different implications posted at different times. Judging whether the content carries manipulative intent may require cross-analysis of data across sites. Policy interventions that facilitate such co-operation (such as the European Union's Code of Practice on Disinformation, which establishes networks across platforms to ensure swift fact-checking during election periods) are likely to achieve more effective outcomes than social media self-regulation (European Commission 2025). Conversely, platform self-regulation can have the advantage of being flexible and proactive, and responds faster to new disinformation strategies. Policy makers and regulation advocates have to weigh the relative merits within their legal and political systems.

In addition to the dual-track system, I propose three guidelines to ensure that regulation facilitates, rather than hinders, the democratic function of free speech. First, democracies should refrain from penalizing users who inadvertently share false information. Even if a post shows the

[9]In US. vs. Alverez (2012), the Supreme Court ruled that the Stolen Valor Act was unconstitutional, in part because it was too broad, and could apply to cases where the lie does not cause harm. By limiting this scrutiny to public speech and public figures, a well-designed, narrow regulation can avoid this problem.

[10]Note that the French model differs from my two-track proposal. It relies on candidates to initiate appeals, and the removal decision is not based on the speaker's authenticity but on whether the false information has the potential to undermine elections. Nonetheless, it serves as a good case of how targets can be narrowed.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

intent to deceive (say, a fabricated headline), the *act* of sharing should not incur sanctions, unless there is clear evidence that the user knowingly shares the content to mislead. This draws a complex line: while such content may be removed for self-evident manipulation, individual users should not face penalties for failing to fact-check, nor have their accounts suspended. This restraint is to ensure that citizens' rights to free speech are robustly protected. As argued earlier, the freedom to share personal perspectives, even when factually mistaken, is the foundation of collective governance. This freedom should not be unreasonably curtailed unless there are demonstrable signs of manipulation.

Second, decisions regarding content moderation should be subject to sufficient oversight, periodic review, and public transparency. Some mechanisms must be in place to ensure that removal decisions are not left to the whims of private companies or governmental organs. One potential strategy to establish checks and balances is to compose an independent advisory committee to oversee removal decisions and develop regulatory guidelines, with or without legislative mandates or penalties. More importantly, any content removal should undergo regular review and be transparent. For instance, social media platforms may be required to report all removed content to an independent committee or official department on a monthly basis, and these reports should be publicly accessible. Finally, an open and fair appeal channel should be in place. Appeals against removal decisions can be adjudicated by a court or an ad hoc committee, ideally by a body distinct from the original decision maker. These mechanisms are crucial to enable civic oversight and provide avenues for remedy in cases of misjudgment.

Before the conclusion, let me address a common concern raised by the allies of regulation proponents. Some might argue that the framework I propose is too permissive towards harmful information shared by those who genuinely believe it. For example, one might genuinely perceive vaccines as ineffective and the coronavirus as a hoax. Proponents of content regulation may argue that such falsehoods should be removed regardless of the speaker's intent to deceive – and to some extent, I agree. Some false information invokes significant public health concerns or threatens national security and may be legitimately removed to protect the community. For example, one can envision a restriction that applies to all COVID-related falsehoods, and the signs of manipulation are irrelevant. However, in these cases, we are not relying on a *democratic* argument for restricting false accounts, but a national safety or exigency one. The investigatory focus shifts from 'whether regulating some false speech is justifiable from the standpoint of democracy' to 'whether the emergency in question outweighs our commitment to democratic values'. For the latter, it is clear that regulation incurs democratic costs, and the purpose is not to protect a meaningful democratic space but to safeguard the safety or stability of a community. Restrictions of this kind of information fall in a different normative realm and demand their own justifications. Proponents will have to explain why the particular case warrants a violation of democratic freedom, which is beyond the scope of this article.

Conclusion: Reclaiming a Democratic Collective Voice

Speech is ubiquitous. There will always be cases that fall at the margins. Yet this does not mean that categories are meaningless. The primary goal of this article is to clarify the conceptual distinction between disinformation and other similar speech. Differentiating disinformation from false opinion and toxic persuasion, I explained why they deserve different moral treatment and proposed regulatory designs grounded in their theoretical distinctions. In this section, I review common objections to regulation, offer clarifications, and articulate my contributions to the existing debates.

First, a prominent objection focuses on the risk of governmental abuse. The central worry is that authorizing the state to regulate speech sets dangerous precedents – a concern articulated clearly by Mill's *On Liberty*. Even if the speech in question is problematic, on this account,

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

granting the state the regulatory power risks legitimizing infringements on expressive freedom. Brison (2021, 117) reviews this objection in the context of hate speech and points out that its underlying assumption is that it is *impossible* to draw reasonable boundaries between different categories of speech and restrict only the illegitimate ones. She contends that this assumption is nonetheless overstated; determining what constitutes hate speech is not *significantly more difficult* than judging what counts as other regulated offenses, such as defamation or plagiarism (Brison 2021, 117). On the one hand, I share Brison's view that sound judgment is attainable. While borderline cases will always arise, they should not preclude maintaining these categories. On the other hand, I take this line-drawing anxiety seriously. By articulating the differences between disinformation and similar utterances, this article aims to help scholars, policy makers, and citizens avoid the slippery slope of treating all false accounts as morally indifferent.

A second concern is that an endorsement of disinformation regulation might give authoritarian leaders the excuse to suppress political critics. Empirically, it is true that the language of 'fake news' or 'disinformation' has been employed to delegitimize opponents (Farkas and Schou 2018). But the same problem applies to the term 'free speech', which has increasingly been used to defend problematic expressions (Castelli Gattinara 2017; Leaker 2020). These instances precisely underscore the need for conceptual clarity – my analysis should provide us with the theoretical tools to respond to the false accusations of 'fake news' and 'disinformation', and explain why the speech in question *does not* meet the criteria for restriction.

To be sure, regulation always carries the risk of misuse. Although my account defends the justifiability of content regulation, the argument is not that every democracy *must* sanction disinformation. Well-designed regulation requires robust accountability systems and appropriate oversight. Where institutional capacity falls short in securing these features, content regulation can indeed undermine expressive freedom. Nevertheless, disinformation regulation is not the first speech restriction in democracies. Traditional news outlets, for example, need to be subject to the scrutiny of the Federal Communications Commission (FCC). Instead of a blanket rejection, what we can more usefully do is examine helpful precedents and identify effective counterbalances against misuse. In this regard, the suggested policy guidelines in this article should also help citizens to call out regulatory measures that are at odds with democratic principles.

Last but not least, I do not present content regulation as a panacea. Complementary measures, such as media literacy programs, pre-warning strategies, and restrictions on political advertising, also play vital roles in limiting the reach of manipulative content. Nonetheless, the unique value of regulation is that it underlines democracy's special obligations to protect opinion, rather than deliberate falsehood. By upholding free speech's normative value, regulation should help restore the norms of public discourse in democratic spaces.

To revisit my arguments: confronting disinformation requires differentiating deceptive falsehoods from other false accounts. This article contends that disinformation should not be conflated with factually false opinion, as it does not represent genuine personal beliefs. And while some degree of misrepresentation is common in political persuasion, disinformation is marked by the speaker's refusal to respond to challenges with credible justification and proof. Since free speech is meant to protect meaningful exchanges between citizens, I argue that it need not extend protection to deceptive content intended to mislead political judgment.

Through these analyses, my work calls for enhanced attention to the conundrum that disinformation poses to democratic worlds: protecting the freedom of speech while safeguarding the integrity of political processes. A polity without free speech cannot be a democracy. Yet a polity without meaningful public exchanges cannot be truly self-governing. Freedom of speech is shallow if it does not enable valuable conversations among fellow members. Citizens and policy makers need to be aware of the trade-offs in the choices they make. Properly designed, I argue, content regulation can be a legitimate step to protect citizens as meaningful decision makers in shaping a common world.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

Acknowledgements. I owe special thanks to Susan Bickford, Alexandra Oprea, Jennifer Forestal, and Jeff Spinner-Halev for their thoughtful feedback. I am also indebted to Nicolás de la Cerda, Kai Yui Samuel Chan, Jihyun Jeong, and Madeleine Austin for their critical engagement and careful readings along the formation of this project. Finally, I thank audiences at the Association of Political Theory 2024, members of the UNC theory workshop, three anonymous reviewers, and the editors of the *British Journal of Political Science* for their valuable comments, which greatly improved the manuscript.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None to disclose.

References

Arendt H (2006) Truth and politics. In Kohn J (ed.), *Between Past and Future*. London: Penguin Books, 223–259.

Association for Progressive Communications (2021) Disinformation and freedom of expression. Available at: https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/2-Civil-society-organisations/APC-Disinformation-Submission.pdf (accessed 5 April 2024).

Baekgaard M, Christensen J, Dahlmann CM, Mathiasen A and Petersen NBG (2019) The role of evidence in politics: Motivated reasoning and persuasion among politicians. *British Journal of Political Science* **49**(3), 1117–1140.

Barber BR (1984) *Strong Democracy: Participatory Politics for a New Age*. Berkeley: University of California Press.

Beerbohm E and Davis RW (2023) Gaslighting citizens. *American Journal of Political Science* **67**(4), 867–879.

Bennett L and Livingston S (2020) *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States The Disinformation Age*. New York, NY: Cambridge University Press.

Bok S (1978) *Lying: Moral Choice in Public and Private Life*. New York: Vintage Books.

Bonotti M and Seglow J (2021) Freedom of expression. *Philosophy Compass* **16**(7), 1–13.

Bonotti M and Seglow J (2022) Freedom of speech: A relational defence. *Philosophy & Social Criticism* **48**(4), 515–529.

Brison SJ (2021) Free speech skepticism. *Kennedy Institute of Ethics Journal* **31**(2), 101–132.

Brison SJ and Gelber K (2019) *Free Speech in the Digital Age*. Oxford: Oxford University Press.

Bromell D (2022) *Regulating Free Speech in a Digital Age: Hate, Harm and the Limits of Censorship*. Cham, Switzerland: Springer.

Brown A (2015) *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

Castelli Gattinara P (2017) Framing exclusion in the public sphere: Far-right mobilisation and the debate on Charlie Hebdo in Italy. *South European Society and Politics* **22**(3), 345–364.

Clifford B and Powell HC (2019, June 6) De-platforming and the Online Extremist's Dilemma. Available at: https://www.lawfaremedia.org/article/de-platforming-and-online-extremists-dilemma (accessed 31 March 2024).

Coe P (2024) Tackling online false information in the United Kingdom: The Online Safety Act 2023 and its disconnection from free speech law and theory*. *Journal of Media Law* **15**(2), 1–30.

Colomina C, Sánchez Margalef H and Youngs R (2021) The impact of disinformation on democratic processes and human rights in the world. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf (accessed 30 August 2025).

Craufurd-Smith R (2019) Fake news, French Law and democratic legitimacy: Lessons for the United Kingdom? *Journal of Media Law* **11**(1), 52–81.

Dowding K (2016) Power and persuasion. *Political Studies* **64**(1_suppl), 4–18.

Engler A (2019) Fighting deepfakes when detection fails. Available at: https://www.brookings.edu/articles/fighting-deepfakes-when-detection-fails/ (accessed 30 August 2025).

Epstein B (2020) Why it is so difficult to regulate disinformation online. In Bennett L and Livingston S (eds.), *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. New York, NY: Cambridge University Press, 190–210.

European Commission (2025) The 2022 code of practice on disinformation. Available at: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation (accessed 30 August 2025).

Farkas J and Schou J (2018) Fake news as a floating signifier: Hegemony, antagonism and the politics of falsehood. *Javnost - The Public* **25**(3), 298–314.

Fielitz M and Schwarz K (2020) Hate not found?! Deplatforming the far right and its consequences. Available at: https://www.idz-jena.de/fileadmin/user_upload/Hate_not_found/IDZ_Research_Report_Hate_not_Found.pdf (accessed 31 March 2024).

Forestal J (2021) Beyond gatekeeping: Propaganda, democracy, and the organization of digital publics. *Journal of Politics* **83**(1), 306–320.

Franks MA (2019) "Not where bodies live: " The abstraction of internet expression. In Brison SJ and Gelber K (eds.), *Free Speech in the Digital Age*. Oxford: Oxford University Press, 137–149.

Freelon D and Wells C (2020) Disinformation as political communication. *Political Communication* **37**(2), 145–156.

Gilbert DT (1991) How mental systems believe. *American Psychologist* **46**(2), 107–119.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

Goldin M (2023, December 15) Not real news: A look at what didn't happen this week. AP News. Available at: https://apnews.com/article/fact-check-misinformation-b17dd3acc23d199557af144625d0b404 (accessed 31 March 2024).
Gross KA and Kinder DR (1998) A collision of principles? Free expression, racial equality and the prohibition of racist speech. *British Journal of Political Science* **28**(3), 445–471.
Heinze E (2016) *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.
Innes H and Innes M (2023) De-Platforming disinformation: Conspiracy theories and their control. *Information, Communication & Society* **26**(6), 1262–1280.
Jack C (2017) Lexicon of lies: Terms for problematic information. Available at: https://datasociety.net/library/lexicon-of-lies/ (accessed 10 May 2024).
Jamieson KH (2018) *Cyberwar: How Russian Hackers and Trolls Helped Elect a President? What we don't, can't and do know*. Oxford: Oxford University Press.
Kunda Z (1990) The case for motivated reasoning. *Psychological Bulletin* **108**(3), 480–498.
Leaker A (2020) *Against Free Speech*. Lanham: Rowman & Littlefield Publishers.
Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N and Cook J (2012) Misinformation and its correction. *Psychological Science in the Public Interest* **13**(3), 106–131.
Liang WY (2016) Anti-Marriage-Equality Group Warns Parents Will "Disappear;" LGBTQ+ Advocates Denounce the Campaign as Fearmongering. Available at: https://www.storm.mg/article/196720 (accessed 17 July 2024).
Loebbert F (2022) Germany: Administrative court of cologne grants Google and Facebook interim relief; holds network enforcement act partially violates EU law. Available at: https://www.loc.gov/item/global-legal-monitor/2022-03-30/germany-administrative-court-of-cologne-grants-google-and-facebook-interim-relief-holds-network-enforcement-act-partially-violates-eu-law/ (accessed 31 March 2024).
Macfarlane E (2021) *Dilemmas of Free Expression*. Toronto: University of Toronto Press.
Markovits E (2006) The trouble with being earnest: Deliberative democracy and the sincerity norm*. *Journal of Political Philosophy* **14**(3), 249–269.
Marwick A and Lewis R (2017) Media manipulation and disinformation online. Available at: https://datasociety.net/wp-content/uploads/2017/05/DataAndSociety_CaseStudies-MediaManipulationAndDisinformationOnline.pdf (accessed 4 April 2024).
McGraw M, Nguyen T and Lima C (2021) Team Trump quietly launches new social media platform. Available at: https://www.politico.com/news/2021/07/01/gettr-trump-social-media-platform-497606 (accessed 30 August 2025).
McKay S and Tenove C (2021) Disinformation as a threat to deliberative democracy. *Political Research Quarterly* **74**(3), 703–717.
Meta (2024) Misinformation. Available at: https://transparency.meta.com/policies/community-standards/misinformation (accessed 20 July 2024).
Mill JS (2015) On liberty. In Philp M and Rosen F (eds.), *On Liberty, Utilitarianism and Other Essays*. Oxford: Oxford University Press, 5–112.
Moon R (2021) Does freedom of expression have a future?. In Macfarlane E (ed.), *Dilemmas of Free Expression*. Toronto: University of Toronto Press, 15–34.
MyGoPen (2021) Realtime FactCheck | Misleading Post Claims Han Kuang Drills Are "Exercises on How to Flee to the U.S". Available at: https://www.mygopen.com/2021/08/exercise.html (accessed 31 March 2024).
Napoli P (2019) *Social Media and the Public Interest: Media Regulation in the Disinformation Age Social Media and the Public Interest*. Columbia: Columbia University Press.
Nyhan B and Reifler J (2010) When corrections fail: The persistence of political misperceptions. *Political Behavior* **32**(2), 303–330.
Plato and **Zeyl DJ** (2012) *Gorgias*. Indianapolis: Hacket Publishing Company.
Post R (2011) Participatory democracy and free speech. *Virginia Law Review* **97**(3), 477–489.
Rastogi S and Bansal D (2022) Disinformation detection on social media: An integrated approach. *Multimedia Tools and Applications* **81**(28), 40675–40707.
Rosenblum NL and Muirhead R (2019) *A Lot of People Are Saying: The New Conspiracism and the Assault on Democracy*. New Jersey: Princeton, Princeton University Press.
Samples J (2019) Why the government should not regulate content moderation of social media. CATO Institute, **865**, 1–31. Available at: https://object.cato.org/sites/cato.org/files/pubs/pdf/pa_865.pdf (accessed 4 April 2024)
Shiffrin SV (2014) *Speech Matters: On Lying, Morality, and the Law*. Princeton, NJ: Princeton University Press.
Susskind J (2022) *The Digital Republic: On Freedom and Democracy in the 21st Century*. New York, NY: Pegasus Books.
Taber CS and Lodge M (2016) The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Political Psychology* **37**(S1), 61–85.
Taiwan FactCheck Center (2023) False: Online Rumor Claims Taiwan's Military Drills Are a "Rehearsal for the President's Escape"? Available at: https://tfc-taiwan.og.tw/articles/9370 (accessed 31 March 2024).
US. v. Alvarez (2012).
Waldron J (2014) *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press

Wang YL (2023) Defense Ministry Refutes Claim That Han Kuang Drills Are Tsai Ing-wen's "Escape Rehearsal," Says It's PRC Disinformation Aimed at Intimidating the Public. Available at: https://tw.news.yahoo.com/%E9%A7%81-%E6%BC%A2%E5%85%89%E6%BC%94%E7%BF%92%E6%98%AF%E8%94%A1%E8%8B%B1%E6%96%87%E9%80%83%E8%B7%91%E6%BC%94%E7%B7%B4-%E5%9C%8B%E9%98%B2%E9%83%A8-%E4%B8%AD%E5%85%B1%E5%81%87%E8%A8%8A%E6%81%AF%E6%83%B3%E5%A8%81%E6%87%BE%E6%B0%91%E5%BF%83-082801973.html (accessed 31 March 2024).

Wardle C (2020) Understanding information disorder. Available at: https://firstdraftnews.org/long-form-article/understanding-information-disorder/ (accessed 31 March 2024).

Wardle C and Derakhshan H (2017) Information disorder: Toward an interdisciplinary framework for research and policy making. Available at: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (accessed 31 March 2024).

Whitney v. California (1927) 274 U.S. 357.

Whitten S (2021) *A Republican Theory of Free Speech: Critical Civility A Republican Theory of Free Speech: Critical Civility*. Cham, Switzerland: Palgrave Macmillan.

Wood T and Porter E (2019) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* **41**(1), 135–163.

Wu T (2018) Is the first amendment obsolete? *Michigan Law Review* **117**(3), 547-581.

Zerilli L (2016) *A Democratic Theory of Judgment*. Chicago: University of Chicago Press.

Cite this article: Huang Y.-H (2026) Muting the Liars: A Democratic Response to Disinformation. *British Journal of Political Science* 1–16. https://doi.org/10.1017/S0007123425101294

https://doi.org/10.1017/S0007123425101294 Published online by Cambridge University Press