Centre for International
Governance Innovation

CIGI Paper No. 336 — October 2025

Freedom of Thought on Social Media: Supporting User Navigation of False Content

Richard Mackenzie-Gray Scott

Centre for International
Governance Innovation

CIGI Paper No. 336 — October 2025

Freedom of Thought on Social Media: Supporting User Navigation of False Content

Richard Mackenzie-Gray Scott

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Senior Program Manager Jenny Thiel
Publications Editor Christine Robertson
Publications Editor Susan Bubak
Graphic Designer Sepideh Shomali

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Table of Contents

About the Author

Richard Mackenzie-Gray Scott works across the areas of human rights, digital technologies, constitutional studies and international law and relations, comprising research, teaching, policy engagement and legal practice. He guest lectures for the Policy Lab on AI and Bias at the University of Pennsylvania, works with the Centre for International Governance Innovation to provide policy guidance and research on matters of digital governance, and is a strategic advisor for Research Integrity Chain, a new tech start-up providing software that utilizes blockchain to protect intellectual property and authenticate data provenance for the purposes of distinguishing between human- and computer-generated content.

His research has been published in leading peer-reviewed journals, reported on in the press — including newspapers such as *The Herald* and *The Times* — and has been referred to by the International Committee of the Red Cross, the North Atlantic Treaty Organization Cooperative Cyber Defence Centre of Excellence and the Public Administration and Constitutional Affairs Committee of the UK Parliament. He has also written for *Prospect Magazine*, Tech Policy Press and *Verfassungsblog*, among other outlets.

He has provided evidence to the Scottish government, UK government and UK Parliament, and has worked on cases before the United Nations High Commissioner for Refugees, the International Centre for Settlement of Investment Disputes, the London Court of International Arbitration, the UK Supreme Court, and the Court of Appeal of England and Wales. He also served on the International Bar Association's Human Rights Institute Task Force on Drones.

Richard completed a post-doctorate at the University of Oxford, where he was a fellow at the Bonavero Institute of Human Rights and St Antony's College, and funded by the British Academy, studying the challenges and opportunities presented by existing and emerging digital technologies to human rights and democratic governance. He holds a Ph.D. from the University of Nottingham, a law degree from the University of Glasgow, and studied under a Grotius Fellowship and a Reuben Lipman Scholarship at the University of Michigan. He read economics and law during his undergraduate studies at the University of Stirling, where he was a student-athlete.

He is the author of *State Responsibility for Non-State Actors: Past, Present and Prospects for the Future* (Bloomsbury, 2022, reissued in paperback 2024).

Executive Summary

Human interaction mediated by social media introduced a different component to the supply and consumption of information: the tailoring of content to individuals via algorithmic curation. This personalization made possible by data profiling shapes the distribution and effects of false content. A related risk concerns the potential impact of misbelief on users, specifically with respect to their freedom of thought. The purpose of this paper is to explain this risk, and to outline discussion points on how it may be mitigated by technical interventions designed with reference to the human right to freedom of thought, while highlighting the implementation challenges and limitations of each intervention.

To help frame and update the current discourse about why false content is problematic, the concept and harm of "regressive thought" is introduced, which involves a misbelief restricting the thinking of a person. In explaining the risk of misbelief leading to regressive thought in social media users, it becomes clearer how such conversion can prevent or limit free thinking. Being more narrow-minded due to an attachment to a misbelief harms the inner self of a person. It closes the mind off to existing and potential avenues of thought, including those that are capable of considering alternative ideas, narratives and perspectives, especially thoughts that do not align with false content supporting a particular misbelief. Such thinking may also result in external harm, should it be acted upon, both to the misbeliever and to others.

In forming part of efforts to mitigate this risk and its connected effects, leveraging the relationship between online expression and thought holds promise. The remainder of the paper explores the potential of doing so by deploying technical measures that moderate the consumption of false content (in contrast to those focused on its supply). The measures considered here are aimed at protecting against misbelief leading to regressive thought while respecting user agency and choice. They consist of labelling false content, watermarking content generated by computer applications, presenting alternative informational sources via interstitial pop-ups linked to false content and collectivizing user reporting. Each one is intended to encourage the real-time exercise of the human right to freedom of thought when individuals encounter and engage with false content — aiming to disrupt quick, reactive thinking and prompt slower, deliberative thinking.

The paper recommends these measures receive adequate public scrutiny, particularly because it is currently unsettled whether and when each should be deployed, as well as in what states. Additional unanswered questions include how these measures should be implemented, whether usage should vary across different platforms and who should make the decisions connected to these matters. The significance of deliberating such constraints concerns four key issues:

- → The effectiveness of each measure in enabling freedom of thought while moderating false content;
- → The risk of users becoming dependent on platform support when navigating information online;
- → Securing accountability for the decisions taken regarding the design, adoption, oversight or non-implementation of each measure; and
- → The inability of these measures to treat the root causes of the supply, consumption and impacts of false social media content, meaning they are — at best — a complement to regulatory approaches capable of engaging with the systemic factors underpinning this problem.

Introduction

The uptake of social media ushered in a new component of human interaction. A distinctive feature of this digital technology is how it mediates communication and information across the globe, namely, by personalizing access and contribution to the online information environment via data-driven algorithmic curation. Content presented to users varies based on their data profile. Undertaken to maximize quantitative user engagement for the purpose of extracting, processing and exploiting personal data, the bespoke tailoring of content to specific individuals results in consumption and sharing experiences differing across users. Algorithms rank and supply content based on this data, with the effect that what one user accesses

on a particular platform, another user may never encounter there. The reach afforded to what users contribute may also be dissimilar, with content from one account attaining outsized reach beyond its immediate network, while content from another permeates only a few user feeds. The management of these supply-consumption exchange cycles consists of decision making about what content receives whose attention and when. Various factors affect this process, including the regularity of user interaction on a platform; the levels of engagement content receives (for example, in the form of reactions, comments, shares or saves); its recency; and its relevance to the inferred interests of users.

Social media feeds can be understood as customized murals of information, with each constantly changing according to the user perceiving it. As a result of this practice, forms of online expression are unevenly distributed and weighted across networks within the confines of a particular platform. Regardless of the actors that initiate or cultivate the related conduct (whether a company, government, content farm or bot), when false information is introduced and spreads across a platform, related impacts are disparate. A reason for this heterogeneity is that the supply of false content does not necessarily mean it will be consumed. Even if it is consumed, that does not necessarily mean that a particular user will believe it. And even if they do, that does not necessarily mean their behaviour will correspond to that misbelief.

However, although misbehaviour does not always transpire as a result of misbelief, that does not necessarily mean there has been no harm. A consistent component connected to the supply and consumption of false content concerns the human right to freedom of thought. Content personalization on social media takes advantage of human cognition, meaning user interactions with false content risk generating, developing and consolidating misbelief. By consuming false information, there exists a risk that individuals will believe it. Such misbelief in and of itself shapes something deeply connected to human agency: the state of the inner self. It is the establishment of falsehood within the mind of a user that can impair their freedom of thought, should "regressive thought" develop. This new concept and harm is introduced to help frame and inform discourse about precisely *why* false content is problematic. It concerns a misbelief restricting the thinking of a person, limiting or preventing free thinking. The harm itself exists *independently* of whether and what manifestations arise as a result of the related thinking. Regressive thought is thus a separate consideration from those regarding the linkage between stimuli and thought, on the one hand, and decision making and conduct, on the other.

This paper addresses two matters. First, it offers insight into the relationship between online expression and thought, specifically by providing an understanding of how misbelief can impair freedom of thought when regressive thought occurs and why user exposure to false social media content increases this risk. Second, it examines how to enable freedom of thought when social media users navigate false content. The approach considered here consists of supporting users by deploying technical measures on platforms that engage with issues connected to the consumption of false content. Each measure is considered primarily regarding its potential to enable freedom of thought, while highlighting the implementation challenges and limitations they present. Their overall aim is to reduce the risk of user misbelief transforming into regressive thought, while respecting user decision making and their right to access and assess information, but without reaching a point where responsibilization occurs, meaning users bear most of the burden for navigating false content. This balance between maintaining user choice and providing user support is a difficult one to strike when managing the problem of false content.

The measures considered here engage with content across audio, image, text and video formats. They consist of labelling false content, watermarking content generated by computer applications, presenting alternative sources of information via interstitial pop-ups linked to false content and collectivizing user reporting. Each one involves human-computer interaction grounded in a human rights-based approach to moderating false content. All of them should receive adequate public scrutiny, particularly because they raise various questions that remain unsettled. These include what measures should or should not be deployed, in what states and when — whether constantly or during particular periods that involve issues such as heightened affective polarization, or spikes in concentrations of false content (such as those surrounding conflicts, disasters and elections). Further matters concern how these measures should be deployed,

whether their implementation should change depending on the platform and who should make the decisions relating to all these matters.

The significance of deliberating usage constraints also concerns four key issues:

- → - clarifying the efficacy of each measure in its ability to enable freedom of thought while moderating false content;
- → - providing opportunities to mitigate the risk of users becoming dependent on platform support when navigating information online;
- → - securing accountability for decisions taken regarding the design, adoption, oversight or non-implementation of each measure, which is necessary to delivery answerability for any adverse consequences connected to their use; and
- → - recognizing that these measures cannot treat the root causes of the supply and consumption of false social media content.

It is also important to acknowledge that the insights considered here attempt to work with the current business models underpinning social media platforms that aim to increase revenue by maximizing quantitative user engagement. Their utility is thus limited because of this approach. The supply and consumption of false information on social media is a socio-psychological problem, exacerbated by technical details that arise from the setting and the execution of corporate priorities. The measures considered here cannot treat the root causes of these problems; they can only manage its symptoms — and how effectively remains debatable. It therefore requires re-emphasizing that these measures should receive adequate public scrutiny, even though, or perhaps especially because, those working for social media platforms can design and deploy them directly without the involvement of state organs and the public.

In order to effectively protect freedom of thought in the context of addressing the supply, consumption and impacts of false social media content, state-led regulations shepherded by democratic processes are necessary. Corporate preferences under capitalist systems remain more or less constant, focused on driving profit through wealth extraction. While such a priority need not necessarily be at odds with public interests, state organs have a responsibility to ensure conflicts and tensions between the two are resolved in favour of the public. Aligning governance with public preferences on the matter of social media regulation has the potential to support a sociocultural pivot toward user empowerment — one grounded in operationalizing human rights, promoting civic virtue and securing the common good. This long-term policy sits in stark contrast to the current situation of user exploitation, where human rights are routinely undermined and social discord and democratic decay continue.

The Risk of Regressive Thought Arising from Misbelief

The relationship between freedom of expression and freedom of thought forms a crucial part of grasping why the supply and consumption of false social media content poses risks. Providing further understanding of it can inform content moderation practices that aim to protect these human rights, but the links between the two require clarification (Ligthart et al. 2022). The connection of concern here is that between content supplied on social media, its consumption by users and how this impacts freedom of thought. Considered from this perspective, thought is partly a social process (Reisman 2023). Elements of it include exposure to information, interactions with interlocutors and reflecting on exchanges (Mackenzie-Gray Scott 2023b), which can improve thought. Social media users can potentially benefit from consuming content precisely because it can expand their thinking. Sharing information and engaging with the expressions of others stimulates mental activity and can help improve awareness, knowledge and understanding (Van Gulick 2012).

However, exposure to the expressions of others also brings the risk of undermining freedom of thought. Expressive conduct is capable of generating changes in thought that can harm individuals. Examples include subliminal messaging, psychologically disturbing content and providing information importunately (Swaine 2022). A further proposed instance is when false information consumed by an individual converts into

misbelief in a way that constricts their thinking. The significance of such misbelief should not be understated. Even if "currently available evidence does not support a clear link between beliefs generated or reinforced through misinformation and aberrant behavior" (Adams et al. 2023), harm is not limited to that which occurs from user behaviour. The onset of misbelief prompted by consuming false social media content may initiate a process toward harming the inner self of a person. There are different stages within this process, combining both mental and social activity, which encompass the onset, development and consolidation of misbelief (Ariely 2023). The thought that forms as a consequence risks being impacted in terms of regression. When a person becomes more narrow-minded due to an attachment to a misbelief, regressive thought ensues and freedom of thought is impaired. This outcome is principally due to the fact that this human right entails the freedom to form, hold, and change or develop thoughts, and to keep thoughts private (O'Callaghan and Shiner 2025). Regressive thought inhibits this freedom, particularly by discouraging the formation, development or change of thoughts that are *incompatible* with the related misbelief.

Should such thinking be acted upon, external harm may arise as well, both to the person undertaking the related conduct and to others. In addition to the harm to the inner self of a person, there still lingers the potential for harm connected to the occurrence of misbehaviour because of the regressive thought stemming from a misbelief grounded in false information. This risk exists, even if the likelihood of its prevalence varies across individuals. When considering the relationship between thought and behaviour, the severity of consuming false content is arguably heightened, because in addition to the harm to the inner self (should the potential harm occur as well), then both internal and external harms occur. The false content has not only led to regressive thought but to misbehaviour as well. Yet regardless of external impact, the shackling of thought to a position derived from misbelief diminishes freedom of thought, speaking to the difference between thinking and free thinking (Munro 2024).

Related impacts on the mind can be better understood when distinguishing between positive and negative misbelief. Positive misbelief is that which does not harm the inner self, or create harms externally, and may even generate benefits for the misbeliever and others (for example, misbelieving that a merry old man pops into every home one night per year to deliver a present to each child not only brings excitement and joy to many children, but also allows many adults to have fun with storytelling and share in the wonder with those children). The thinking attached to positive misbelief can also instill a sense of hope, encourage people to imagine things differently, be more open-minded or consider alternative perspectives. Behaviour corresponding to a misbelief also ties back into the nature of false information. Not all falsehoods, even if acted upon, will result in harm, whether to the misbeliever or to anyone else (for example, if someone will not drink tea while playing chess because they believe doing so accelerates the aging of their children).

By contrast, negative misbelief harms at least the inner self of a person and has the potential to externally harm the misbeliever, along with others. The related harm of regressive thought arises when thinking is bound to a misbelief, where a person becomes mentally tethered to positions associated with it. Part of their mental activity gets warped. Negative misbelief transforming into regressive thought closes the mind off to existing and potential avenues of thought, including those that are capable of considering alternative ideas, narratives and perspectives, particularly thoughts that do not align with the substance contained in the false information supporting a particular misbelief. This internal harm is intensified, should regressive thought disrupt freedom of the will, because a person risks becoming "a helpless bystander" to the information influencing them (Frankfurt 1971).

Related impacts on the inner self are potentially multiple. Deep-seated misbelief may result in a person losing their sense of self, including when this sense is inflated to a point where they slip into denial about what is objectively true (McKay and Dennett 2009). Alternative views based on truth may become less plausible, even if they had been before the shift toward misbelief. The ability to reflect on the possibility that a particular stance may be mistaken erodes. This reduction in considering information that conflicts or is in tension with an understanding settled upon in the mind may, in turn, foster confirmation bias and motivated reasoning. Users can end up interpreting content — including accurate information — in a way that confirms their

misbelief, seeking out more material that aligns with it and resisting contrary information (Wolff and Taddicken 2024). Perhaps the final stage of regressive thought is indoctrination, whereby multiple sources containing false information reinforce the misbelief at issue, including to a point where it is so embedded in the mind that trying to address it becomes counterproductive (Government Communication Service Behavioural Science Team 2022). The mental filter used to assess information may become corrupted as a result of one, some or all of these impacts on mental activity. Users undergoing such regressive change within the mind may therefore suffer a diminishing capacity to recognize and retain true information, and extract and discard false information. The connected abilities to make concessions and self-correct may also corrode as a result. Reality can get conflated with truth. In the extreme, such change may invert truth and falsity — what is objectively true becomes subjectively false or vice versa.

Given these risks of regressive thought arising from misbelief, it is concerning how frequently social media users encounter false content, especially because freedom of thought includes being free to forget information (O'Callaghan 2023). Especially when combined with platform design choices that create usage addiction, algorithmic curation can prevent a respite from exposure to false content. Not being provided space and time to forget falsehoods could lead to the establishment of misbelief. Content being curated according to the inferred preferences and interests of users means that information presented to them on platforms may align within the confines of a particular partisan framing, even if users receive information from many different sources. This data-driven circumscription shaping the ranking and supply of content can create challenges for users to forget false content that is tailored according to their inferred preferences and interests. The continued consumption of false content along a single partisan line may increase exposure to sources that, in turn, help fortify misbelief, while potentially also reinforcing a specific partisan perspective.

Intensifying these challenges is affective polarization, which is driven by partisan sorting on platforms that appears to foster intra-group agreement and inter-group conflict (Törnberg 2022). Made possible by data profiling, customized content curation can issue constant reminders about a particular overarching stance on a subject, which a user may not be able to avoid when connected to a particular platform (Bisbee et al. 2022). Continued intrusion of this sort from content being provided importunately may hinder users in terms of their ability to decide what information to focus on, remember, dismiss or challenge.

With increases in exposure to falsehoods and decreases in self-directed attention, users face more instances capable of prompting and solidifying misbelief. While exposure to false content may not necessarily result in the formation of misbelief in every instance — particularly because users have the ability to actively evaluate and interpret information (as opposed to being passive recipients of it) — increased exposure to false claims appears to generally increase the inclination to believe them (Li and Yang 2024). Users whose social media experiences are saturated with false content are therefore likely placed at greater risk of regressive thought than users exposed to less of such content. In what is known as the "illusory truth effect," the perceived truthfulness of information can increase when it is encountered more frequently (Hassan and Barber 2021), indicating that the "more we are exposed to the same false claims, the more our brains begin to mislead us into believing them" (Farahany 2023).

When reflecting on the relationship between consuming false content, thought and the related impacts both internal and external to the self, user behaviour corresponding to misbelief rather than truth can certainly be problematic. But so potentially is misbelief when taken alone. Regressive thought harms the inner self of a person with adverse impacts on freedom of thought. Due to the cyclic process of information shaping thought, thought shaping behaviour and behaviour shaping expression, the ability to filter thoughts is key. When connected to social media platforms, where attempts to capture attention are constant, it is important that users can notice, extract and retain what is beneficial to them, and ignore, discard or actively oppose what is harmful. Supporting users with this task of information filtration is thus key to enabling their freedom of thought when navigating false content, while guarding against the risk of misbelief leading to regressive thought.

Technical Measures Aiming to Protect Against Regressive Thought

Public demand for effective measures to address false social media content appears to be increasing, providing further impetus for better policy on the matter, especially when considering that the majority of social media users across states may want more content moderation, even if that means further modulating freedom of expression (Theocharis et al. 2025). While the technical measures considered below are aimed at helping meet this demand, particularly in order to enable freedom of thought when users navigate false content, three caveats must be mentioned.

First, this approach works with the current business models underpinning the largest social media platforms. It is thus of limited utility with respect to operationalizing the human right to freedom of thought. The measures considered below should not be read as solutions to the problems posed to freedom of thought by false social media content. The supply and consumption of false information on these platforms, any corresponding negative misbelief in users and harmful behaviour based on it — even if exacerbated by technical details present on platforms — is a socio-psychological problem. It demands adopting approaches beyond those of a technical nature. As a matter of long-term policy, robust legislative and socio-cultural changes are needed to adequately address the root causes of this problem. A key dilemma here is the lobbying prowess of the largest multinational technology conglomerates skewing policy and regulation in their favour to serve corporate interests rather than promote innovation, serve the public and protect human rights (Khanal, Zhang and Taeihagh 2025).

Second, although integrating human rights considerations into the digital design mechanisms of content moderation has the potential to reduce negative human rights impacts (Penagos 2024), safeguards should be erected to avoid legitimizing an "algorithmic leviathan," where platform algorithms disproportionately manage human rights (Gomez et al. 2024). State organs have a leading role to play here. A proposed element of the human right to freedom of thought is "states fostering an enabling environment for freedom of thought," which is separate from other elements currently part of this right, namely, those concerning what constitutes an unlawful interference with it, whether that is being forced to reveal thoughts, being punished for thoughts or thoughts being impermissibly altered.[1] Although those working for social media platforms can design and deploy the below measures directly, states providing a legislative mandate to do so would not only demonstrate that they are taking their human rights obligations seriously, but would also offer the opportunity for the conceptions and digital design underpinning these measures to receive adequate public scrutiny. The oversight of digital technologies used to engage with matters of public interest is essential — an undertaking that should involve democratically accountable institutions.

Third, the measures considered below are not exhaustive and focus on issues connected to the consumption of false content. Further research has the opportunity to clarify how other measures, including those that focus more on the supply of false content, can or cannot secure particular human rights while moderating false content (Avram et al. 2020). Those not considered here include algorithmic downgrading, content removal, deplatforming and shadow banning. While these measures alter the supply of false content, they do not maintain user agency and choice with respect to having the opportunity to access and consume information. Nor do they necessarily address the consumption of false content. Users can switch platforms to access false content on alternative platforms, meaning removing or withholding content on one platform could result in an increase in its consumption beyond that platform (Rogers 2020). The switching cost of interested users is also reduced, as lack of access to the desired content makes changing platforms an easier decision (Mackenzie-Gray Scott 2023c). The Streisand effect should also be kept in mind, where attempts to suppress information can lead to it receiving more attention and credence than it might have otherwise attracted (The Royal Society 2022). As to user agency and choice, respecting both demands users being permitted to navigate false content with support, not being removed from this process — the significance of which lies in appealing to the human "capacity for conscious deliberation and

1 *Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed*, UNGAOR, 76th Sess, UN Doc A/76/380 (2021).

choice" (Susser, Roessler and Nissenbaum 2019). Users should be involved in deciding what content they encounter and engage with on social media, even though platforms should lessen individual responsibility for navigating false content by implementing adequate content moderation measures. A balance must be struck in content moderation practice that respects individual agency and choice without responsibilizing users.

The below measures concern the relationship between online expression and thought and are aimed at reducing the risk of negative misbelief, or at least countering it adequately enough to avert regressive thought. They engage with content that is not only text-based, on which there appears to be excessive emphasis, despite it being crucial to address the different forms in which false content appears on social media (Murphy et al. 2023). Each measure also taps into how humans perceive and process online content, particularly so as to aid information filtration, and to prompt in users "a slower and more deliberative mode of thinking" (Celadin, Panizza and Capraro 2024).

Labelling False Content

Functional Description

Including a visible label with content containing false information to notify users that what they are engaging with does not depict a factual occurrence.

Significance for Freedom of Thought

- Assists users in assessing the veracity of content, which is particularly significant given the fast-paced nature of informational supply and consumption on social media, especially when combined with users not having the time to investigate every piece of content they encounter (Swire and Ecker 2018).
- Demarcation helps determine credibility, providing context to users before they choose to engage further. This pre-emption can discourage users from engaging with sources of information that contain false content.
- Transparency regarding whether content has captured a factual occurrence provides users clarity before deciding if they want to lend the related information credence.
- Confronts false content capable of leading to regressive thought at the point of exposure, instead of attempting to address this risk later, by which time negative misbelief may have had the chance to more fully develop.
- Guards against users being manipulated into misbelief from them genuinely not knowing that a piece of content contains false information.

Implementation Challenges and Limitations

- The technical capabilities to accurately recognize false content are still developing (Al Ajmi et al. 2024), a challenge compounded by the diversity of such content, which in turn requires adaptable detection systems (Rustam et al. 2024), the language limitations in data sets, and the discrepancy between day-to-day social media usage and controlled environments for research purposes, which impact findings regarding efficacy (Müller et al. 2024).
- Software designed for the purpose of detecting false content in audio files in particular not only needs to cover different languages but also changes in the accent, rhythm and tone of human speakers versus those created or manipulated by a computer (Ballesteros et al. 2021). This challenge speaks to the significance of localizing content moderation, especially considering differences across cultures regarding matters such as parody or satire (Hatano 2023).
- With respect to detecting false images, detection software needs to be capable of distinguishing between images that are real, even when both images share similar — if not the same — properties, such as semantic content, aspect ratio, resolution and file format (Rajan et al. 2024).
- Software for false content detection also raises matters concerning explainability, so that users can understand how they work. This difficulty, combined with current technical capacities, raises the question of whether false audio detection in particular can generalize, both across platforms and across states, even if

image- and text-based detection systems can do so with high accuracy rates (Uppada, Patel and Sivaselvan 2022).

- Even if technical capabilities develop with high and trusted accuracy rates for detecting false content, labelling should not be expected to prevent the risk of misbelief translating into regressive thought without the implementation of other measures, including those that are not technical in nature (Vinhas and Bastos 2022).

- Labelling as a form of content moderation involves making value judgements about content that may generate distrust in some users (Stewart 2021). Implementing platforms may also be perceived as being arbiters of truth instead of powerholders that are responsible for guarding against the harms capable of arising from user misbelief in false content.

- Usage of this measure may create an "implied truth effect" for some users, where labelling content as false contributes to the perceived accuracy of content that is not labelled as false, but which may still be false or misleading (Pennycook, Bear et al. 2020). Therein lies a risk that labels may not catch some false content, resulting in users misbelieving it. Similarly, the label may be incorrect, which could lead to users misbelieving that accurate content is false, or to users distrusting the process because of such mistakes. In addition, there is a risk of labels applying to so much content that their efficacy as a distinguishing tool is reduced below adequacy.

- Credibility of a source informing a label is crucial, as sources perceived by users as more credible have been shown to be more effective at countering misbelief (Liu et al. 2023). Labels are thus perhaps best designed with the inclusion of a note briefly and clearly stating the source that made the determination, with supplementary information made available to users regarding why the decision was made to attach a label to the content at issue. Providing users with explanations and insight into better understanding the process and its ultimate determination may help build trust.

- Labelling that consists of fact-checking undertaken by a platform or third parties without accompanying context may be considered less trustworthy than measures that supplement fact-check labels with annotations (whether created by users, the platform, third parties or a mix of all three), which provide an explanation about why the content at issue has been labelled (Drolsbach, Solovev and Pröllochs 2024).

- Despite fact-check labels appearing to be effective at reducing misbelief across states (van Erkel et al. 2024; Koch, Frischlich and Lermer 2023; Porter and Wood 2021; Walter et al. 2020), even for users who are most distrusting of this process (Martel and Rand 2024), it can sometimes backfire depending on the user. Users who are highly partisan on a specific issue may reject the label and actively seek out content that supports their misbelief (Weeks et al. 2017). Users undertaking their own evaluations of false content by relying on online search results may also increase its perceived veracity, as individuals can end up consuming corroborating information from other sources containing falsehoods (Aslett et al. 2024).

- The design, wording and placement of labels impact efficacy (Martel and Rand 2023; Moravec, Collis and Wolczynski 2023). Ambiguous messages should be avoided so as to avert user misperceptions that the label itself is false information (Park et al. 2021). Labels should be integrated on platforms in ways that users easily notice them. Numerous questions of digital design are raised here, which need to factor in differences across users, such as in hearing and sight.

- Labelling can reduce user engagement with false content (Gruzd, Mai and Soares 2024), including lowering the likelihood that it is shared (Nekmat 2020), highlighting a financial incentive for owning companies not to deploy labels on their platforms, if they depended on maximizing quantitative user engagement to increase firm revenue.

- It is currently unclear whether this measure should extend to the use of audio and image filters, which are widely used to make humans look and sound different. This problem of scope also concerns audio and images that do depict a factual occurrence but have been edited by users (for example, memes and podcast excerpts).

Watermarking Content Generated by Computer Applications

Functional Description

→ Embedding visible information into outputs generated by computer applications that consist of image and/or text, so as to inform users that the content does not originate from a human.

Significance for Freedom of Thought

→ Individuals rely on others to provide "cues on the basis of which to filter" information (Levy 2023). Such evaluations provide indications toward credibility, allowing individuals to distinguish between different claims to knowledge, which help people decide what information and sources of it to trust or not. One such cue that arguably needs to be provided to social media users is the source origin of content with respect to it being a computer or a human.

→ It is important that users understand the provenance of content they encounter on social media. Content production is no longer limited to humans. Some computer applications mimic the ways in which humans produce and disseminate information. When a user is exposed to content on a platform, knowing that it was created by a human, generated by a computer or a mix of the two can help individuals decide whether and why they want to engage with it further and whether they should believe it.

→ It is a matter of transparency that users are informed in these choices about who and what they spend their time interacting with while connected to social media platforms.

→ By including watermarks that are visible to users on content generated by computer applications, users are immediately informed that they are not encountering forms of human expression but outputs generated by a computer.

→ While requiring further discussion and action, it is a separate matter whether content generated by computers should be prohibited on social media platforms intended for human communication, because, for instance, computers are not capable of human expression (only informational reproduction devoid of thought and feeling [Noë 2024]) and may therefore be considered to have no place on platforms ostensibly designed for humans to connect and communicate with each other.

Implementation Challenges and Limitations

→ Should the complexity of computer applications that can generate social media content continue to increase, accurately detecting such outputs may become harder over time.

→ The question arises as to whether legislation should be passed requiring every company releasing software capable of generating social media content to simultaneously release software capable of detecting it, which could aid watermarking practice (Fraser, Dawkins and Kiritchenko 2024).

→ The process of distinguishing between content generated by computer applications, content modified or co-produced with a human using it and content created by a human (particularly considering users share content that they have been involved in producing using digital tools, including to critique the associated processes and outputs) remains to be determined. Deciding how to clearly differentiate between these three types of content so as to accurately inform users about whether they are engaging with content from a computer, a human or a mix of both, requires further debate.

→ It remains unclear whether computers, humans or a mix of both are better at accurately detecting whether content has been generated by a computer or created by a human. Questions arise here regarding operational capacity, resource distribution and scalability, including with respect to whether users themselves should be involved in detecting computer-generated content for platforms to watermark (Cui et al. 2023).

→ The design, wording and placement of a watermark impact perceptibility, presenting a trade-off between ensuring the watermark is easily visible and that such placement does

not negatively interfere with the consumption experiences of users. Striking a balance involves ensuring users recognize watermarks without them being discouraged from consuming watermarked content.

Presenting Alternative Sources of Information to Accompany False Content

Functional Description

- Inserting a digital nudge to accompany false content that triggers an in-platform interstitial pop-up providing alternative sources of accurate information that are trusted by the specific user, and without the pop-up commenting on the truth content of either the original false content or the alternative content.

Significance for Freedom of Thought

- Affective polarization, outrage and exposure to extreme content are intertwined issues tied to the supply and consumption of false content on social media, which the dynamics and design on some platforms encourage and facilitate. Users may be "much more likely to see content from like-minded sources than they are to see content from cross-cutting sources" (Nyhan et al. 2023), with some platforms being ideologically segregated (González-Bailón et al. 2023). At the same time, users are exposed to diverse content on platforms (Yang et al. 2020). This feature can be leveraged to enable freedom of thought when users interact with false content by providing them with additional sources of information (Reuben et al. 2024).
- The "marketplace of ideas" is a misnomer in the context of the social media information environment (Bambauer 2006). True information can only compete with false information in the mind of a user if both are known to that person and then considered alongside prior knowledge and mis/beliefs (Mackenzie-Gray Scott 2023a).
- Algorithmic curation and the cognitive attributes of users combine in ways whereby people are not aware of different knowledge claims and informational sources, and if made aware, process and respond to them differently (Ecker et al. 2022). Accurate information, no matter the amount, is thus not capable of reducing concentrations of false content within user feeds, unless it permeates them — a process described as "informational osmosis" (Mackenzie-Gray Scott 2023a).
- A delivery method capable of acting as a conduit for this transfer involves utilizing interstitial pop-ups that are triggered if a user engages with false content. Also, the pop-up containing alternative content with true information could be designed so that it is presented to users within the platform, either before or after they engage with the original content containing false information and especially if they proceed to share it further.
- This measure presents trusted alternative sources of information on the same topic as the content containing falsehoods. Levels of trust toward informational sources can be inferred from the data profile of each user and utilized to inform alternative source selection. Considering the significance of establishing user trust in true information and encouraging doubt about false information (Organisation for Economic Co-operation and Development 2023), tailoring the presentation of alternative content to user trust levels is significant.
- The delivery method of an in-platform pop-up also adds friction to the consumption and supply of false content, but without prompting the user to leave the platform, and does so without having to label content regarding its relationship to truth according to a fact-checker. Introducing tailored alternative informational sources without commenting on their truth content provides inlets to suggest new information to users, while delivering nuance on divisive issues. Such sources offer opportunities for users to consider and distribute information as part of a shared reality, thus providing overlap between their individual realities.
- This measure may aid the practice of users finding common ground across different positions, which are embedded in trust levels toward diverse informational sources containing

truth instead of falsity. Increasing this overlap in viewpoints is particularly significant in light of the dangers of consensus reality, where a group of people all accept and hold the same perceptions about a particular matter.

- One-sided content that ignores a plurality of views and opinions is countered. Addressing the harm of regressive thought entails encouraging a willingness to consider information beyond that supporting a particular standpoint. This readiness to reflect on something from a different perspective is not only possible, including because "humans have always been able to change their beliefs" (Harari 2024), but it may well also form part of the essence of freedom of thought.

- Measures targeting user trust in reliable sources demonstrate greater effects in combatting misbelief than measures aimed at reducing acceptance of false information (Acerbi, Altay and Mercier 2022), underscoring the importance of exposing users to sources that they trust and that contain accurate information.

- As this particular digital nudge aims to prompt the type of thinking that is more conscious, analytical and guided by reasoning, it avoids the problems and risks linked to behavioural interventions that only engage thinking that is automatic, unconscious and guided by heuristics — principally that of bypassing the human capacity to reason, which undermines rational decision making and perhaps crosses the line between permissible influence and manipulation.

- Engaging conscious mental activity in decisions about what content a person consumes on social media is crucial to respecting the interplay between freedom of thought, human agency and choice (Mackenzie-Gray Scott 2023a). Awareness of the thought process involved in selecting content to engage with leaves open the possibility for users to reject the alternative sources presented in pop-ups.

Implementation Challenges and Limitations

- Despite the promise this digital nudge offers to reduce both the consumption and the spread of false content (Thornhill et al. 2019; Pennycook, McPhetres et al. 2020; Pennycook and Rand 2022; Salomon-Ballada et al. 2023), users may not trust it. There remain unsettled controversies linked to nudging regarding human agency, as well as doubts surrounding the promise of nudges being capable of effectively steering people toward predetermined outcomes while simultaneously preserving their capacity for choice. It may be that nudges can do one or the other, but not both. Nudges also appear to prioritize decision-making outcomes over decision-making processes, assuming a "better ability in relation to complex and deeply individual choices" (Schramme 2024), while failing to account for outcome preferences varying across individuals.

- Encouraging doubt and critical thought in users can also contribute to them slipping into cynicism and distrust, including about information that is grounded in truth. A careful balance needs to be struck in content moderation practice that empowers users to engage with social media content discerningly without convincing them to distrust all information they encounter online (Bateman and Jackson 2024).

- The interstitial pop-up including the alternative sources needs to be designed so that users can easily return to the original false content and can consume and share the alternative sources without being redirected to third-party web pages beyond the platform. This feature of the measure could be designed in many different ways, but it needs to incorporate the consideration that keeping users connected to the implementing platform is key. Prompting users to consider engaging with external content is in tension with the preferences of owning companies seeking to maximize user engagement on their platforms.

- The content of the data sets used to inform the algorithm presenting alternative sources is based on inferences being drawn about what informational sources each user trusts. These inferences may be inaccurate, while potentially attracting accusations of bias regarding the sources that are considered reputable enough for inclusion in the database.

- Users are already exposed to large volumes of content when connected to social media platforms, raising the concern that presenting yet more content via interstitial pop-ups could result in informational overload. Related impacts on mental activity may mean the

information contained in alternative sources remains ignored or underappreciated, even if it is trusted. Limiting the number of alternative sources within the pop-up is crucial to not overwhelming users with options. Encouraging consumption of alternative sources depends on that information being manageable to navigate and the content containing it being straightforward to engage with so as to catch and hold user attention (for example, interactive swipe-through slides containing salient messaging).

Collectivizing User Reporting

Functional Description

- Allowing users to report and vote on those reports for the purpose of annotating content that may contain false information, providing a collectively agreed-upon viewpoint that adds perspective to the potentially false content.

Significance for Freedom of Thought

- This approach asks users who contribute to collectively settle on what information should accompany potentially false content. Providing users with access to potentially false content — and then allowing those who wish to participate to arrive at a collectively agreed-upon summary that speaks to the truth content of the potentially false content — provides an opportunity for users to reason in light of the information before them.
- Fostering conscious deliberation and choice when users engage with content forms part of respecting their freedom of thought. This measure asks that users engage with each other to reach consensus about what information should be shared to accompany potentially false content, providing an opportunity for users to interact with a common purpose, with people holding diverse views, and to reflect on the communication and information part of this process.
- Mental activity can be enhanced through expression, in part, because attempting to articulate thoughts can help clarify them. By commenting on the truth content of information, while being aware that the supporting reasoning impacts the extent to which a comment is considered helpful by others, users are required to consider how best to express their perspective. Such a process can be facilitated by digital tools that make it more structured (Sleigh et al. 2024), indicating its potential to slow thinking down and make it less reactive, resulting in a more thoughtful, iterative and/or self-correcting expression. The related interaction with other users pursuing the same undertaking also challenges their mental ability to adapt to new information.
- This process encourages users to think about other perspectives and take them into consideration when forming their mis/beliefs. As one study notes, participants were of the view that a "'collective thought process was better than everyone's individual opinions'. They emphasized that hearing everyone's opinions would enable them to understand other perspectives, which would either help them recognize misunderstandings, confront their own biases, or, if they're right, find more support for their worldview" (Agarwal, Shahid and Vashistha 2024).
- Truth content of information also changes over time. Information considered to be true during a particular period can be proven to be false at a later time, or vice versa. This changeability speaks to the significance of providing space and time for alternative views to be considered in the present. Free thinking can be fostered or diminished depending on the boundaries of this space and time expanding or receding. In the context of consuming and supplying content on social media, this process involves breaking through, and being provided breaks from, the informational silos created by algorithmic categorization (Frost 2023).

Implementation Challenges and Limitations

- A key difference between this measure and those above is that it relies on the voluntary work of users. While a legitimate way to involve users in moderating false content, it should be adopted as a complement — and not as a replacement —

to foundational content moderation measures. This is particularly important because crowdsourcing-based measures of this sort that users opt into can be framed as "empowering" users. Yet the extent to which such measures do so is questionable in light of responsibilization, which in this context involves shifting responsibility from states and platforms (which are the actors with the most power to effectively address false content) to social media users (who are the actors with considerably less power to address this problem).

- Instead of investing more in moderating false content, platforms rely on users contributing to this effort without having to finance the time users spend creating and agreeing upon the best annotation to accompany potentially false content.
- This shift away from contracting or employing workers to carry out necessary content moderation provides an avenue for platforms that implement this measure to shroud responsibility, should it prove counterproductive, ineffective or harmful. Platforms can disavow mistakes arising through the use of this approach, and perhaps even blame users or bad actors that manipulate it, meaning ownership of wrongdoing becomes disjointed. Blame or praise can be disproportionally allocated to users or platforms.
- As this measure can be framed as one that "empowers" users, the optics involved risks platforms adopting the measure to make it look like they have adequate content moderation in place without actually having it, and thus undermining freedom of expression and freedom of thought while giving the appearance that these human rights are being respected. This can be achieved by involving users in content moderation but without implementing the foundational measures necessary for adequately addressing false content.
- Marketing this measure can take priority over ensuring it forms part of adequate content moderation practice. Some platforms may use this measure "as a smokescreen for neglect of their responsibilities" to the information environment (Stafford 2025). That said, there is the potential for future development of this measure, especially because research can support the improvement of such community-based models to effectively moderate false content while protecting expression and thought (Seering 2020).
- It should not be viewed as a corrective measure to users consuming false content. Instead, this measure ideally functions as a harmonizing tool that supports users agreeing upon something that raises doubts about potentially false content, highlighting the depolarizing potential of this approach.
- Deployment guides users to arrive at common ground across individual perspectives that may vary considerably, depending on the issues at hand. Yet such "consensus can be reached on anything, even if it is nowhere close to being true" (Mackenzie-Gray Scott 2023a). Throughout history, people have collectively arrived at acceptable stances on subjects that were part of a shared reality that was detached from the truth (O'Connor and Weatherall 2019).
- Users can coordinate in a deliberate attempt to manipulate the information shared in annotations, so as to supply further false content along with the potentially false content. This type of conduct could also occur "for the purpose — or at least with the effect — of undermining the credibility of accurate sources" (Mackenzie-Gray Scott 2023a).
- Users may show partisan bias in how they evaluate user reports. Even if different perspectives are provided to create the content of an annotation, if they do not fall within a particular partisan stance on the given subject under scrutiny, they may be considered unhelpful by users, even if their truth content is robust and reasoned with rigour. This highlights the importance of carefully considering the ramifications of partisanship when implementing crowdsourcing-based user reporting (Allen, Martel and Rand 2022).
- Annotations may not appear alongside enough false content, and those that do appear may be too slow in doing so. This raises questions regarding how effective this measure is at reducing the likelihood of regressive thought developing in users exposed to false content without being made aware of that exposure, or not in sufficient time (Chuai et al. 2024).

Conclusion

The precision and frequency with which false content is presented to social media users creates challenges for protecting their freedom of thought. Although misbehaviour may not occur as a result, harms connected to misbelief are not limited to those involving thoughts being manifested in user conduct. Should regressive thought occur as a result of misbelief stimulated by false content, freedom of thought is impaired, meaning this human right also risks being undermined when individuals interact with false content mediated by social media. Yet the design features of these platforms can be altered to confront this problem, so that users are supported in a way that enables freedom of thought when navigating false content. Social network dynamics can be harnessed to mitigate the market-driven incentives influencing the owning companies of social media platforms, with the potential to reduce the net negative impacts of false content through the very goal of user engagement that some companies seek to maximize. Platform design can be optimized to decrease the consumption of false content without removing users from this process. By leveraging the relationship between online expression and thought, human rights-based technical measures can be deployed on platforms to protect against the consumption of false content that risks leading to regressive thought. This approach not only aims to protect the human right to freedom of thought, but also reconciles that protection with the collective interests concerning the supply, consumption and impacts of false content.

Even so, the measures considered here should receive adequate public scrutiny, particularly by democratically accountable institutions that can inspect and debate their validity, utility and value, including in comparison to other approaches and measures. Technical responses form only part of tackling the existence and impacts of false social media content. As a matter of long-term policy, individual users should be empowered without being responsibilized. Public institutions across states — providing opportunities for the public to provide their input — should discuss and enact regulatory approaches capable of addressing the root causes of the supply, consumption and impacts of false social media content, while ensuring the human right to freedom of thought remains central in such efforts, especially because it is essential to human flourishing. Securing this human right should be on the agenda of every policy maker wanting to effectively address false social media content. Deference to the corporate power underlying social media platforms has gone far enough. Choices regarding content moderation cannot occur at the whims of company executives. These decisions need to be taken by state organs, particularly so as to provide for democratic accountability, public deliberation and securing trust in policy. States have a responsibility to adequately regulate social media platforms, not place corporate priorities above the public interest by leaving these platforms to govern as they please. Corporate self-governance is not the answer to the problem of false social media content. Instead, its shortcomings provide further confirmation for why states need to introduce, debate and enact quality regulations that protect human rights and promote innovation in the global effort of preventing and reducing the harms connected to false social media content.

Works Cited

Acerbi, Alberto, Sacha Altay and Hugo Mercier. 2022. "Research note: Fighting misinformation or fighting for information?" *Harvard Kennedy School Misinformation Review* 3: 1–15. https://misinforeview.hks.harvard.edu/article/research-note-fighting-misinformation-or-fighting-for-information/.

Adams, Zoë, Magda Osman, Christos Bechlivanidis and Björn Meder. 2023. "(Why) Is Misinformation a Problem?" *Perspectives on Psychological Science* 18 (6): 1436–63. https://doi.org/10.1177/17456916221141344.

Agarwal, Dhruv, Farhana Shahid and Aditya Vashistha. 2024. "Conversational Agents to Facilitate Deliberation on Harmful Content in WhatsApp Groups." Preprint, arXiv. https://doi.org/10.48550/arXiv.2405.20254.

Al Ajmi, Sahar Abdullah, Khizar Hayat, Alaa Mohammed Al Obaidi, Naresh Kumar, Munaf Salim Najim AL-Din and Baptiste Magnier. 2024. "Faked speech detection with zero prior knowledge." *Discover Applied Sciences* 6: 288–302. https://doi.org/10.1007/s42452-024-05893-3.

Allen, Jennifer, Cameron Martel and David G. Rand. 2022. "Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 245: 1–19. https://doi.org/10.1145/3491102.3502040.

Ariely, Dan. 2023. *Misbelief: What Makes Rational People Believe Irrational Things*. New York, NY: Harper Collins.

Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler and Joshua A. Tucker. 2024. "Online searches to evaluate misinformation can increase its perceived veracity." *Nature* 625: 548–56. https://doi.org/10.1038/s41586-023-06883-y.

Avram, Mihai, Nicholas Micallef, Sameer Patil and Filippo Menczer. 2020. "Exposure to social engagement metrics increases vulnerability to misinformation." *Harvard Kennedy School Misinformation Review* 1 (5): 1–11. https://doi.org/10.37016/mr-2020-033.

Ballesteros, Dora M., Yohanna Rodriguez-Ortega, Diego Renza and Gonzalo Arce. 2021. "Deep4SNet: deep learning for fake speech classification." *Expert Systems with Applications* 184: 115465. https://doi.org/10.1016/j.eswa.2021.115465.

Bambauer, Derek E. 2006. "Shopping Badly: Cognitive Biases, Communications, and the Fallacy of the Marketplace of Ideas." *University of Colorado Law Review* 77: 649–709. https://lawreview.colorado.edu/wp-content/uploads/2025/08/Bambauer.pdf.

Bateman, Jon and Dean Jackson. 2024. *Countering Disinformation Effectively: An Evidence-Based Policy Guide*. January 31. Washington, DC: Carnegie Endowment for International Peace. January 31. https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en.

Bisbee, James, Megan A. Brown, Angela Lai, Richard Bonneau, Joshua A. Tucker and Jonathan Nagler. 2022. "Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden." *Journal of Online Trust & Safety* 1 (3): 1–65. https://doi.org/10.54501/jots.v1i3.60.

Celadin, Tatiana, Folco Panizza and Valerio Capraro. 2024. "Promoting civil discourse on social media using nudges: A tournament of seven interventions." *Proceedings of the National Academy of Sciences of the United States of America Nexus* 3 (10). https://doi.org/10.1093/pnasnexus/pgae380.

Chuai, Yuwei, Haoye Tian, Nicolas Pröllochs and Gabriele Lenzini. 2024. "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?" *Proceedings of the ACM on Human-Computer Interaction* 8 (CSCW2): 1–52. https://doi.org/10.1145/3686967.

Cui, Wanyun, Linqiu Zhang, Qianle Wang and Shuyang Cai. 2023. "Who Said That? Benchmarking Social Media AI Detection." Preprint, arXiv. https://arxiv.org/abs/2310.08240.

Drolsbach, Chiara Patricia, Kirill Solovev and Nicolas Pröllochs. 2024. "Community notes increase trust in fact-checking on social media." *Proceedings of the National Academy of Sciences of the United States of America Nexus* 3 (7). https://doi.org/10.1093/pnasnexus/pgae217.

Ecker, Ullrich, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou et al. 2022. "The psychological drivers of misinformation belief and its resistance to correction." *Nature Reviews Psychology* 1: 13–29. https://doi.org/10.1038/s44159-021-00006-y.

Farahany, Nita A. 2023. *The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology*. New York, NY: St. Martin's Press.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20. https://doi.org/10.2307/2024717.

Fraser, Kathleen C., Hillary Dawkins and Svetlana Kiritchenko. 2024. "Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods." Preprint, arXiv. https://arxiv.org/abs/2406.15583.

Frost, Neli. 2023. "The global 'political voice deficit matrix.'" *International Journal of Constitutional Law* 21 (4): 1041–68. https://doi.org/10.1093/icon/moad084.

Gomez, Juan Felipe, Caio V. Machado, Lucas Monteiro Paes and Flavio P. Calmon. 2024. "Algorithmic Arbitrariness in Content Moderation." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2234–53. https://doi.org/10.1145/3630106.3659036.

González-Bailón, Sandra, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio et al. 2023. "Asymmetric ideological segregation in exposure to political news on Facebook." *Science* 381 (6656): 392–98. https://doi.org/10.1126/science.ade7138.

Government Communication Service B ehavioural Science Team. 2022. *The Wall of Beliefs: A toolkit for understanding false beliefs and developing effective counter-disinformation strategies*. September. www.communications.gov.uk/wp-content/uploads/2022/09/Wall_of_Beliefs_-publication.pdf.

Gruzd, Anatoliy, Philip Mai and Felipe B. Soares. 2024. "To Share or Not to Share: Randomized Controlled Study of Misinformation Warning Labels on Social Media." In *Disinformation in Open Online Media,* MISDOOM 2024, Lecture Notes in Computer Science, vol. 15175, edited by Mike Preuss, Agata Leszkiewicz, Jean-Christopher Boucher, Ofer Fridman and Lucas Stampe, 46–69. Cham, Switzerland: Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-71210-4_4.

Harari, Yuval Noah. 2024. *Nexus: A Brief History of Information Networks from the Stone Age to AI.* London, UK: Fern Press.

Hassan, Aumyo and Sarah J. Barber. 2021. "The effects of repetition frequency on the illusory truth effect." *Cognitive Research: Principles and Implications* 6: 38–49. https://doi.org/10.1186/s41235-021-00301-5.

Hatano, Ayako. 2023. "Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation." *Australian Year Book of International Law Online* 41: 127–56. https://doi.org/10.1163/26660229-04101017.

Khanal, Shaleen, Hongzhou Zhang and Araz Taeihagh. 2025. "Why and how is the power of Big Tech increasing in the policy process? The case of generative AI." *Policy and Society* 44 (1): 52–69. https://doi.org/10.1093/polsoc/puae012.

Koch, Timo K., Lena Frischlich and Eva Lermer. 2023. "Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media." *Journal of Applied Social Psychology* 53 (6): 495–507. https://doi.org/10.1111/jasp.12959.

Levy, Neil. 2023. "Fake News: Rebuilding the Epistemic Landscape." In *The Oxford Handbook of Digital Ethics*, edited by Carissa Véliz, 103–20. Oxford, UK: Oxford University Press.

Li, Jinhui and Xiaodong Yang. 2024. "Does exposure necessarily lead to misbelief? A meta-analysis of susceptibility to health misinformation." *Public Understanding of Science*. https://doi.org/10.1177/09636625241266150.

Mackenzie-Gray Scott, Richard. 2023a. "Managing Misinformation on Social Media: Targeted Newsfeed Interventions and Freedom of Thought." *Northwestern Journal of Human Rights* 21 (2): 109–84. https://scholarlycommons.law.northwestern.edu/njihr/vol21/iss2/1.

———. 2023b. "The Algorithmic Management of Misinformation That Protects Liberty." Tech Policy Press, August 23. www.techpolicy.press/the-algorithmic-management-of-misinformation-that-protects-liberty/.

———. 2023c. "YouTube Updates its Policy on Election Misinformation." *Verfassungsblog*, June 8. https://verfassungsblog.de/youtube-updates-its-policy-on-election-misinformation/.

Ligthart, Sjors, Christoph Bublitz, Thomas Douglas, Lisa Forsberg and Gerben Meynen. 2022. "Rethinking the Right to Freedom of Thought: A Multidisciplinary Analysis." *Human Rights Law Review* 22 (4): 1–14. https://doi.org/10.1093/hrlr/ngac028.

Liu, Xingyu, Li Qi, Laurent Wang and Miriam J. Metzger. 2023. "Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness." *Communication Research* 52 (6): 719–46. https://doi.org/10.1177/00936502231206419.

Martel, Cameron and David G. Rand. 2023. "Misinformation warning labels are widely effective: A review of warning effects and their moderating features." *Current Opinion in Psychology* 54: 101710. https://doi.org/10.1016/j.copsyc.2023.101710.

———. 2024. "Fact-checker warning labels are effective even for those who distrust fact-checkers." *Nature Human Behaviour* 8: 1957–67. https://doi.org/10.1038/s41562-024-01973-x.

McKay, Ryan T. and Daniel C. Dennett. 2009. "The evolution of misbelief." *Behavioral and Brain Sciences* 32 (6): 493–510. https://doi.org/10.1017/S0140525X09990975.

Moravec, Patricia L., Avinash Collis and Nicholas Wolczynski. 2023. "Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook." *Information Systems Research* 35 (3): 1435–47. https://doi.org/10.1287/isre.2022.0305.

Müller, Nicolas M., Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga et al. 2024. "MLAAD: The Multi-Language Audio Anti-Spoofing Dataset." Preprint, arXiv. https://arxiv.org/abs/2401.09512.

Munro, Daniel. 2024. "Restoring Freedom in Freedom of Thought." Opinion, Centre for International Governance Innovation. www.cigionline.org/articles/restoring-freedom-in-freedom-of-thought/.

Murphy, Gillian, Constance de Saint Laurent, Megan Reynolds, Omar Aftab, Karen Hegarty, Yuning Sun and Ciara M. Greene. 2023. "What do we study when we study misinformation? A scoping review of experimental research (2016–2022)." *Harvard Kennedy School Misinformation Review* 4 (6). https://doi.org/10.37016/mr-2020-130.

Nekmat, Elmie. 2020. "Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media." *Social Media + Society* 6 (1). https://doi.org/10.1177/2056305119897322.

Noë, Alva. 2024. "Rage against the machine." *Aeon*, October 25. https://aeon.co/essays/can-computers-think-no-they-cant-actually-do-anything.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott et al. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature* 620: 137–44. https://doi.org/10.1038/s41586-023-06297-w.

O'Callaghan, Patrick. 2023. "Enlightened remembering and the paradox of forgetting: From Dante to data privacy." *Law and Humanities* 17 (2): 210–27. https://doi.org/10.1080/17521483.2023.2223806.

O'Callaghan, Patrick and Bethany Shiner, eds. 2025. *The Cambridge Handbook of the Right to Freedom of Thought*. Cambridge, UK: Cambridge University Press.

O'Connor, Cailin and James Owen Weatherall. 2019. *The Misinformation Age: How False Beliefs Spread.* New Haven, CT: Yale University Press.

Organisation for Economic Co-operation and Development. 2023. *OECD Skills Outlook 2023: Skills for a Resilient Green and Digital Transition.* Paris, France: OECD Publishing. https://doi.org/10.1787/27452f29-en.

Park, Sungkyu, Jaimie Yejean Park, Jeong-han Kang and Meeyoung Cha. 2021. "The presence of unexpected biases in online fact-checking." *Harvard Kennedy School Misinformation Review* 2 (1): 1–11. https://doi.org/10.37016/mr-2020-53.

Penagos, Emmanuel Vargas. 2024. "ChatGPT, can you solve the content moderation dilemma?" *International Journal of Law and Information Technology* 32: 1–27. https://doi.org/10.1093/ijlit/eaae028.

Pennycook, Gordon, Adam Bear, Evan T. Collins and David G. Rand. 2020. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings." *Management Science* 66 (11): 4944–57. https://doi.org/10.1287/mnsc.2019.3478.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu and David G. Rand. 2020. "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science* 31 (7): 770–80. 1–11 https://doi.org/10.1177/0956797620939054.

Pennycook, Gordon and David G. Rand. 2022. "Nudging Social Media toward Accuracy." *The ANNALS of the American Academy of Political and Social Science* 700 (1): 152–64. https://doi.org/10.1177/00027162221092342.

Porter, Ethan and Thomas J. Wood. 2021. "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom." *Proceedings of the National Academy of Sciences of the United States of America* 118 (37): e2104235118. https://doi.org/10.1073/pnas.2104235118.

Rajan, Anirudh Sundara, Utkarsh Ojha, Jedidiah Schloesser and Yong Jae Lee. 2024. "On the Effectiveness of Dataset Alignment for Fake Image Detection." Preprint, arXiv. https://arxiv.org/html/2410.11835v1.

Reisman, Richard. 2023. "From Freedom of Speech and Reach to Freedom of Expression and Impression." Tech Policy Press, February 14. www.techpolicy.press/from-freedom-of-speech-and-reach-to-freedom-of-expression-and-impression/.

Reuben, Maor, Lisa Friedland, Rami Puzis and Nir Grinberg. 2024. "Leveraging Exposure Networks for Detecting Fake News Sources." *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5635–46. https://doi.org/10.1145/3637528.3671539.

Rogers, Richard. 2020. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." *European Journal of Communication* 35 (3): 213–29. https://doi.org/10.1177/0267323120922066.

Rustam, Furqan, Wajdi Aljedaani, Anca Delia Jurcut, Sultan Alfarhood, Mejdl Safran and Imran Ashraf. 2024. "Fake news detection using enhanced features through text to image transformation with customized models." *Discover Computing* 27 (54). 1–11 https://doi.org/10.1007/s10791-024-09490-1.

Salomon-Ballada, Sebastian, Paloma Bellatin, Lila Tublin, Mónica Wills Silva and Federica Demergasso. 2023. "Behavioral science can help prevent the spread of fake news." *Behavioural Insights Team* (blog), March 17. www.bi.team/blogs/behavioral-science-can-help-prevent-the-spread-of-fake-news/.

Schramme, Thomas. 2024. "Why Health-enhancing Nudges Fail." *Health Care Analysis* 32 (1): 33–46. https://doi.org/10.1007/s10728-023-00459-7.

Seering, Joseph. 2020. "Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 1–28. https://doi.org/10.1145/3415178.

Sleigh, Joanna, Shannon Hubbs, Alessandro Blasimme and Effy Vayena. 2024. "Can digital tools foster ethical deliberation?" *Humanities and Social Sciences Communications* 11, 117. https://doi.org/10.1057/s41599-024-02629-x.

Stafford, Tom. 2025. "Do Community Notes work?" *Impact of Social Science* (blog), January 14. https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/14/do-community-notes-work/.

Stewart, Elizabeth. 2021. "Detecting Fake News: Two Problems for Content Moderation." *Philosophy & Technology* 34: 923–40. https://doi.org/10.1007/s13347-021-00442-x.

Susser, Daniel, Beate Roessler and Helen Nissenbaum. 2019. "Online Manipulation: Hidden Influences in a Digital World." *Georgetown Law Technology Review* 4 (1): 1–45. https://philarchive.org/archive/SUSOMHv1.

Swaine, Lucas. 2022. "Does Hate Speech Violate Freedom of Thought?" *Virginia Journal of Social Policy & the Law* 29: 1–30. https://vasocialpolicy.org/wp-content/uploads/2023/03/29.1-3.pdf.

Swire, Briony and Ullrich Ecker. 2018. "Misinformation and Its Correction: Cognitive Mechanisms and Recommendations for Mass Communication." In *Misinformation and Mass Audiences*, edited by Brian G. Southwell, Emily A. Thorson and Laura Sheble, 195–212. Austin, TX: University of Texas Press.

The Royal Society. 2022. *The online information environment: Understanding how the internet shapes people's engagement with scientific information*. January. London, UK: The Royal Society. https://royalsociety.org/news-resources/projects/online-information-environment/.

Theocharis, Yannis, Spyros Kosmidis, Jan Zilinsky, Friederike Quint and Franziska Pradel. 2025. *Content Warning: Public Attitudes on Content Moderation and Freedom of Expression.* Content Moderation Lab, Technical University of Munich Think Tank. https://doi.org/10.17605/OSF.IO/F56BH.

Thornhill, Calum, Quentin Meeus, Jeroen Peperkamp and Bettina Berendt. 2019. "A Digital Nudge to Counter Confirmation Bias." *Frontiers in Big Data* 2: 1–9. https://doi.org/10.3389/fdata.2019.00011.

Törnberg, Petter. 2022. "How digital media drive affective polarization through partisan sorting." *Proceedings of the National Academy of Sciences of the United States of America* 119 (42): e2207159119. https://doi.org/10.1073/pnas.2207159119.

Uppada, Santosh Kumar, Parth Patel and B. Sivaselvan. 2022. "An image and text-based multimodal model for detecting fake news in OSN's." *Journal of Intelligent Information Systems* 61: 367–93. https://doi.org/10.1007/s10844-022-00764-y.

van Erkel, Patrick F. A., Peter van Aelst, Claes H. de Vreese, David N. Hopmann, Jörg Matthes, James Stanyer and Nicoleta Corbu. 2024. "When are Fact-Checks Effective? An Experimental Study on the Inclusion of the Misinformation Source and the Source of Fact-Checks in 16 European Countries." *Mass Communication and Society* 27 (5): 851–76. https://doi.org/10.1080/15205436.2024.2321542.

Van Gulick, Robert. 2012. "Consciousness and Cognition." In *The Oxford Handbook of Philosophy of Cognitive Science*, edited by Eric Margolis, Richard Samuels and Stephen P. Stich, 19–40. Oxford, UK: Oxford University Press.

Vinhas, Otávio and Marco Bastos. 2022. "Fact-Checking Misinformation: Eight Notes on Consensus Reality." *Journalism Studies* 23 (4): 448–68. https://doi.org/10.1080/1461670X.2022.2031259.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37: 350–75. https://doi.org/10.1080/10584609.2019.1668894.

Weeks, Brian E., Daniel S. Lane, Dam Hee Kim, Slgi S. Lee and Nojin Kwak. 2017. "Incidental Exposure, Selective Exposure, and Political Information Sharing: Integrating Online Exposure Patterns and Expression on Social Media." *Journal of Computer-Mediated Communication* 22 (1): 363–79. https://doi.org/10.1111/jcc4.12199.

Wolff, Laura and Monika Taddicken. 2024. "Disinforming the unbiased: How online users experience and cope with dissonance after climate change disinformation exposure." *New Media & Society* 26 (5): 2699–720. https://doi.org/10.1177/14614448221090194.

Yang, Tian, Sílvia Majó-Vázquez, Rasmus K. Nielsen and Sandra González-Bailón. 2020. "Exposure to news grows less fragmented with an increase in mobile access." *Proceedings of the National Academy of Sciences of the United States of America* 117 (46): 28678–83. https://doi.org/10.1073/pnas.2006089117.

CIGI

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org