Check for updates

ESSAY

Conspiracy beliefs and social media: Addressing a systemic risk

[version 1; peer review: 3 approved with reservations]

Maxime Lebrun[1], Hanne Dumur-Laanila[2], Gwenda Nielen[3], Pablo Hernández Escayola[4]

[1]The Centre of Excellence for Countering Hybrid Threats, Helsinki, 00230, Finland
[2]The Centre of Excellence for Countering Hybrid Threats, Helsinki, 00230, Finland
[3]Tilt, Utrecht, The Netherlands
[4]Maldita.es, Madrid, 28026, Spain

v1 **First published:** 08 Mar 2024, **4**:53
https://doi.org/10.12688/openreseurope.17043.1
Latest published: 08 Mar 2024, **4**:53
https://doi.org/10.12688/openreseurope.17043.1

Abstract

Conspiracy beliefs are an important lever for hybrid threat activity. It presents a risk for liberal democratic governance at multiple levels. The widespread circulation of conspiracy beliefs on social media contributes to discrediting sources of authority, science, and expertise. This article addresses the circulation of conspiracy beliefs on social media as a systemic risk. Because social media companies operate within a legal loophole of facilitating illegal and harmful content, they have become the primary medium for the spreading of conspiracy beliefs. This article suggests closing the legal loophole since it provides a space for potential hybrid threat activity. Social media companies should be made responsible for content published through their services in a similar fashion to that of regular news media. It explores the extent to which regulation of social media could be a systemic response to the challenge of conspiratorialism as a space for hybrid threat activity directed at our democratic societies.

Keywords

conspiracy beliefs, social media regulation, social networking, online politics, AI models, systemic risk

H2020 This article is included in the Horizon 2020 gateway.

Open Peer Review

Approval Status ? ? ?

	1	2	3
version 1 08 Mar 2024	? view	? view	? view

1. **Valentina Pitardi**, University of Surrey, Guildford, UK
2. **Niko Pyrhönen**, University of Helsinki, Helsinki, Finland
3. **Thomas Wold**, North University Stjordal Campus, Stjordal, Norway

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Sociology gateway.

Corresponding authors: Maxime Lebrun (maxime.lebrun@hybridcoe.fi), Hanne Dumur-Laanila (hanne.dumur-laanila@hybridcoe.fi)

Author roles: Lebrun M: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Dumur-Laanila H**: Investigation, Project Administration, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Nielen G**: Writing – Original Draft Preparation, Writing – Review & Editing; **Hernández Escayola P**: Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883054 (Empowering a Pan-European Network to Counter Hybrid Threats [EU-HYBNET]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Lebrun M *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lebrun M, Dumur-Laanila H, Nielen G and Hernández Escayola P. **Conspiracy beliefs and social media: Addressing a systemic risk [version 1; peer review: 3 approved with reservations]** Open Research Europe 2024, **4**:53 https://doi.org/10.12688/openreseurope.17043.1

First published: 08 Mar 2024, **4**:53 https://doi.org/10.12688/openreseurope.17043.1

Introduction: the legal loophole of social media, a space for hybrid threat activity

Hybrid threat activity consists in combining a series of harmful events, risks, and threats towards undermining the stability and strength of a liberal democracy. Countering such activity implies reducing the margin of manoeuvre it can exploit. It is about decreasing the potential leverage upon democratic societies. Conspiracy beliefs and conspiratorial mindsets constitute important levers in creating systemic risks to civic discourse. Conspiratorialism and the virality models of social media work hand in hand. Social media are the key spaces within which conspiracy communities grow and where their members radicalize. The 2016 attack on Comet Pizza in Washington DC; the attack on the US capitol in January 2021; or refusal to implement COVID19 protective measures hampered societal security on a series of levels.

Belief in conspiracy is what gives concrete effect to conspiracies. Belief in conspiracy has historically been one of the key drivers of social and political events. Groups and individuals think that their world is ruled by a clique of plotters engaged in a concerted, secret, and fraudulent plan (Reichstadt, 2019). Conspiracies do exist and have existed throughout history but have seldom managed to reach their goal. Machiavelli warned that conspiracies have an extremely thin chance of success. Max Weber's sociology of unintended consequences demonstrates that political action generally fails to reach its initial purpose.

The policy field is ripe with calls to create a range of measures for straightening the facts as a way to correct misinformation and counter disinformation, which often rely on conspiratorial mindsets. Conspiratorialism, the belief in conspiracies, is a strong identification factor for individuals. Despite that the purpose of fact checking is to prevent conspiracies to spread, social scientist Christine Abdalla Mikhaeil argues that the usual measures to combat misinformation such as fact checking and debunking false claims not only fail to reach their purpose but can actually contribute to this identification process.

Conspiratorial radicalization through identification with like-minded individuals is an inherent consequence of the business model of social media, from the presentation of content reinforcing and matching existing beliefs to the ability to share, post quickly and effortlessly. The problem with these measures is twofold. They risk reinforcing conspiracy beliefs; and secondly, the responses or reactions are *post ante* while the circulation of conspiracy beliefs on social media platforms is an acute problem that causes security, safety and social cohesion problems which take very concrete and dangerous forms and manifestations. For these reasons, the problem must be faced *ex ante*. Self or shared regulation on social media has shown promising results but corporate discretion and a self-congratulating narrative for the platforms has hampered its credibility. Social media companies argue that they are not responsible for contents posted online and are reluctant to provide transparency over their technical regulation of contents, i.e. algorithms. The difficulty with online conspiracy dynamics is that if they fall short of hate speech and incitement to violence, those contents must be considered harmful but not illegal.

The widespread circulation of conspiracies and reinforced belief in them is a symptom of the legal loophole social media companies benefit from. They were able to develop their business models and modes of operation free from liability over the effects of their services. This article explores the policy problem of conspiracy belief circulation on social media by questioning the role of social media companies as information publishers. It explores the extent to which regulation of social media could be a systemic response to the challenge of conspiratorialism as a space for hybrid threat activity directed at our democratic societies.

The article first delves into the notion of conspiracy beliefs as a tool for contesting established authorities and expertise. It then examines a few conspiratorial belief circulation manifestations on Telegram, X, and instant messaging apps. The article goes on to examine the question of social media platform responsibilities and regulations as publishers, exploring how a shift in regulation may be able to address the systemic risks to civic discourse that conspiratorialism circulation presents to democratic societies.

Belief in conspiracies: a matter of rationality or power?

The formation of conspiracy beliefs

A conspiracy belief holds that a situation is the result of a hidden and deceptive plot of a group of people. Conspiracy theorists assert about the idea that explanations of certain types of events and situations are either false or deliberately misleading (Uscinski *et al.*, 2016). This is causing *conspiratorial* thinking. Conspiracy beliefs are a central component of our culture and politics (Enders & Smallpage, 2019) varying from mild suspicions to extremely strong beliefs. Those beliefs are emotionally charged, and evoke strong emotions and feelings, which accounts for the appeal for these beliefs (Wojczewski, 2022). A conspiracy belief differs from scientific beliefs as it essentially makes an unfalsifiable claim (Sunstein & Vermeule, 2009). A conspiracy belief is a baseless accusation which does not follow the essential standards of the administration of the proof or scientific reasoning. Conspiratorialism is an abusive—not proven—attribution of the causes of major events to a permanent, ongoing, and never-ending conspiracy. Conspiracies have been part of human sensemaking throughout history (van Prooijen & Douglas, 2017). They fulfil part of the fundamental need for social belonging and a sense of belonging to a social group. This need is rooted in human social history. Conspiracy beliefs may provide psychological comfort and sense-making.

Researchers Abdalla Mikhaeil and Baskerville have identified four steps of how conspiracy beliefs are forming on social media: confirmation, affirmation, protection, and enactment. The first one, *identity confirmation* is a phase when individuals are actively seeking to verify one's self-views. In this phase,

content in social media and mainstream media is actively browsed on several platforms to locate content, elaborate beliefs and to connect the dots that would allow to sustain relevant interactions between individuals and to create mutually verifying context. The second phase, *affirmation*, is when individuals select or disassociate the information from their original sources. In the *protecting* phase, individuals are safeguarding and further protecting their information environment, or the integrity of the groups they feel they belong to. Safeguarding one's own environment also implies discrediting other individuals' views by presenting counterevidence. *Enactment* is the final phase, where individuals seek broader social approval for their views. In this phase, recruitment and call for violent actions can take place, leveraging the community userbase. The role of social media—implied by their logic of audience aggregation—is to breed the idea of a shared identity by acting as an *echo chamber* to build and reinforce conspiracy beliefs, that in the extreme case, can escalate to violence through radicalization. (Abdalla Mikhaeil & Baskerville, 2023).

Conspiracy beliefs as discursive power contestation

Distrust in authorities, science, and expertise must be paralleled with the massive spread of conspiracy beliefs on social media. Trust is eroded by scandals, corruption and misinformation that create distrust and lead to seek information from elsewhere. Crises, conflicts, and rapid changes in everyday life provoke uncertainty and fear that can lead to belief in conspiracies. Situations of large societal crisis are amplifiers of spreading conspiracy beliefs. This is the 'function' of conspiracy beliefs in reducing uncertainty. Many of the conspiracies denounced by "conspiracy theorists" cast doubt on the intentions, abilities, or choices of public and private authorities.

Conspiracy beliefs are strongly associated to political stances, to the consumption of political information (Lee *et al.*, 2022) and presidential campaigns (Bond & Neville-Shepard, 2023). Research results outline that exposure to conspiracy beliefs reduce the intention to vote, participation into politics, trust in government services and institutions (Butler *et al.*, 1995; Einstein & Glick, 2015; Enders & Smallpage, 2019; Jolley & Douglas, 2014). Political ideology determines political perceptions and attitudes and serves as an information filter that is closely connected to the habit of accepting information that confirms own ideas. For politicians, embracing conspiracies can become one of the strategies to increase their personal support and undermine their political opposition, paving their way for power. The use of conspiratorial language by politicians may be viewed as an attempt to reform the existing political system (Lee *et al.*, 2022; Pasek *et al.*, 2015).

Conspiracy beliefs can be considered as tools for discursive power contestation. The rhetoric of the "red pill" is revealing of such mindset. "Redpilling" is a recurring discursive tool in conspiracy belief-based communication and exchanges online. It means that for one to understand how the world is ruled, it is necessary to take a "red pill" to reveal the world how it truly is beyond the artificial power structures that constrain individual freedom. Aim is to offer shared set of beliefs that contradict facts proposed and accepted by society at large. Hence, the "red pill" stands for an anti-democratic master plot narrative in which each microlevel narrative become part of the greater macro-level narrative giving the "red pill" a visible concrete form, rendering external conviction inefficient (Aikin, 2019).

Conspiracy beliefs amplify political and social polarization by reinforcing pre-existing biases and creating "us *versus* them" mentalities. This became evident during Donald Trump's 2020 presidential campaign. Widespread anxiety and distrust that undermines confidence in political figures and the whole system, lead to doubts about the very foundations of the social contract. Conspiracy beliefs thrive in politics because of the instability and uncertainty that democratic politics inherently creates. Trump's invocation of threats to American democracy from internal and external sources and his online communication have normalized forms of hate-speech (Valcore *et al.*, 2023). His conspiratorial discourse, denigrating and deprecating language bred hate speech on social media. The strategic use of media systems allowed him to constitute an effective and winning political organization (Eddington, 2018). These groups became an *echo chamber* of these beliefs, mobilising his supporters to the use of violence. The role and attention granted to QAnon during his campaign, an oftentimes messianic tone and religious saviour narrative show that Trump attempted to connect right-wing evangelical politics with anti-government conspiracy beliefs (Ibid.). This is example of weaponized conspiracy beliefs against liberal democratic politics.

The red pill rhetoric, mixing facts and fiction, greater acceptance of political violence, and the spread of these beliefs' bursts, create particularly dangerous atmosphere in society putting the democratic foundations under direct attack. Violent thoughts, accompanied with memes and jokes blurs the line between posturing and supporting violence, normalizing ideologies and acts that have been usually considered beyond the bounds of politics. The consequences of political violence stemmed from these beliefs may lead to democratic decline and open a profitable space for hybrid threat activity.

Conspiratorial strategies: Telegram, instant messaging, and X

The strategy followed by conspiracists on social media is adapted to the type of platform they use. Two factors influence what type of platform is used: the ability to reach a very broad audience and the platform's level of content moderation. Conspiracists feel important to organize themselves in other types of digital spaces as well with fewer restrictions for their messages. This strategy allows to maintain cohesion in their communities, transmit extreme messages that can be problematic on other networks and use those communities as a refuge to maintain their activity if accounts on the largest social networks are closed. For example, when broadcasting audio-visual content on YouTube, it is common to have parallel accounts on other platforms such as Odysee or Rumble, with a lower level of content moderation. If YouTube

closes a channel or bans one of the videos, content is then generated on other alternative platforms, that allows to be in contact with the hard core of their community.

The role of Telegram

Within this strategy, the role of Telegram in spreading conspiracy beliefs is very relevant. Telegram has a private and a public part. In the public part, the two main collective messaging formats are Groups and Channels. These are distinguished by who can send messages. In Groups anyone can post messages while in Channels only administrators have the possibility to do it. Groups can have up to 200,000 members, unlike other platforms like Signal or Whatsapp that limit to 1000 and 256 respectively. Channels have no limitation on users. The entire private part of the platform works with encrypted messages that are outside any moderation policy of content. On its FAQ page the company states that in private groups and chats they do not process requests for illegal content. The content of the channels and groups can be publicly reported, although Telegram defends a policy of non-interference in the content of the application and the limit their interventions to a minimum, arguing that they would not block those "peacefully express alternative opinions." Anyone can report content in serious cases such as the appearance of messages related to terrorism, violation of copyright or pornography. Individual users, other organizations or even governments can request the removal of content. For example, the frequently asked questions section on its website explains the process to send requests to the authorities of the European Union (EU) for the removal of terrorist content under Regulation (EU) 2021/784.

There are numerous investigations that warn of how the public spaces within Telegram have become an instrument that is used to spread disinformation and extremist speeches. Several studies (Rogers, 2020) confirm that after the deletion of Donald Trump's accounts on Facebook, Instagram, YouTube, Reddit, and X (formally known as Twitter), there was a mass migration of users to Telegram. This growth was focused on far-right accounts that either created new channels or grew exponentially (Bryanov *et al.*, 2022) Even Pavel Durov, the company's founder, celebrated the arrival of 25 million new users in 72 hours.

Furthermore, these channels allow them to achieve a high impact among the community of loyal followers. According to a study by the University of Lvov (Plakhta, 2020), posts on Telegram channels are considerably more effective at reaching followers than Facebook and Instagram community pages. This predilection of extremists for Telegram has been analysed in studies (Walther & McCoy, 2021) that indicate that it is a growing environment for a variety of hate ideologies, also for all types of conspiracy beliefs or for far-right movements.

The implementation of these discourses has also investigated extremes in Telegram channels in different geographical areas. It has been demonstrated how networks are formed that share disinformation Netherlands, how these channels are used to attack Ukrainian refugees in Poland or how radical content is disseminated in Eastern European countries.

In Spain, Maldita recently investigated the contents and operation of a group of Telegram channels specialized in spreading disinformation and conspiracy beliefs. The main groups that primarily spread disinformation about COVID-19 during the pandemic were selected. The researchers were struck by the fact that when the Russian invasion of Ukraine began, these channels changed and started to fill with pro-Russian content. The investigation found that, although most of its content is related to COVID-19, the themes could be easily adapted depending on the public debates that occurred at any given time but always maintaining a conspiratorial approach in which a powerful elite is moving the threads in secret to harm ordinary citizens.

These Telegram channels function as a "disinformation supermarket." They collect a large amount of content that reinforces their conspiracy beliefs from a wide variety of sources and make them available to their audience, the consumers of those beliefs. The sources used by channel administrators could be disinformation websites, blogs, social network content, media articles or anything published in Spain or abroad that fits with their conspiracy beliefs. These conspiracy contents are spread in a new space, Telegram. Some of the traffic is also directed to the original sources because the link is usually shared along with the headline or a screenshot of the contents. The users of these channels—that work like echo chambers—receive a constant flow of content that reinforces their conspiracy ideas. It becomes very easy to share it in private conversations. There is a part of this disinformation dissemination circuit, the private part, that cannot be monitored. However, it must be considered to calculate the magnitude of the problem posed by the way these disinformative Telegram channels act.

In the fight against disinformation, Telegram has hardly done anything. Its few actions were the result of pressure from various authorities. For example, in early 2022, the Federal Supreme Court in Brazil decided to suspend the operation of the application in the country after repeated rejection of Telegram to freeze accounts that spread disinformation as it has been requested judicially. The order was reversed after the adoption of measures that included monitoring the 100 most popular channels in the country, and the removal of classified content shared by the account of President Jair Bolsonaro including the restriction of several profiles that had spread disinformation. Telegram also had to act in Germany when the Interior Minister stated that if the platform did not comply with the rules regarding hate speech, it would be banned and would have to pay fines of up to EUR 55 million. As a response, the company finally decided to block 60 channels that share disinformation and conspiracy beliefs.

Disinformation in instant messaging applications

Other messaging apps have taken some measures against disinformation. For example, WhatsApp has limited the mass

dissemination of messages and only allows forwarding to a maximum of five chats at a time to try to make it harder for hoaxes to go viral on the platform. In addition, on its website, the application recommends not forwarding messages when in doubt. It has included specific advice on how to prevent the spread of disinformation, including consulting multiple sources of information or fact-checking organizations. The instant messages applications are designed to hold conversations private but when these platforms are used massively, the structure of chats and direct contacts between individuals ends forming a network in which it is possible to reach large audiences. In this way, applications such as WhatsApp or parts of Telegram, with millions of users around the world, have become an instrument for disinformation.

One of the key characteristics of messaging applications is privacy. Unlike in public platforms like Twitter and Facebook, where public content is being spread, in messaging applications, messages are encrypted. This makes the content moderation very difficult. This private environment also influences the type of conversations held in these applications. While on public social networks, comments are largely directed to strangers, in messaging applications people contact others in their environment, such as family or friends. That conditions the behaviour of users. In these spaces they usually speak more freely because they are in touch with their community, with people they trust (Valeriani & Vaccari, 2018). Studies have analysed how people who have extreme ideological positions take advantage of these characteristics. Like-minded people and groups can express themselves more comfortably than in open social networks.

These groups form alternative spheres for political debate. They know that their views can cause rejection if they are expressed in open networks, that is why they prefer intimate and closed environments where they can control who receives their messages. That feeling of closeness with their public together with the difficulties of moderating content make messaging applications a common resource for extremists and proponents of conspiracy beliefs (Theocharis *et al.*, 2021) In addition to the activity within messaging applications, it should not be forgotten what some studies call the "effect cascade" (Gursky *et al.*, 2022) which consists of disinformation jumping between various platforms and social networks. This effect works in two ways. On the one hand, upward contamination occurs when the disinformation content, that has begun to spread in private spaces, jump to public social networks. On the other hand, there are contents that want to be removed from public scrutiny and they end up taking refuge in applications where they can be shared in an intimate environment.

X

Content moderation policies of large platforms have been effective to some degree in stemming the flow of conspiracy beliefs online and reducing the extremist rhetoric. After Elon Musk took control of Twitter (now X) in October 2022, studies have shown immediate effects in the diffusion of problematic content in the social network: an immediate spike in tweets using hate terms (Benton *et al.*, 2022) or a dramatic rise in hate speech upon Musk purchasing Twitter (Hickey *et al.*, 2023). The most radical or controversial accounts significantly increased their engagement, such as accounts active in far-right networks, or with the most popular untrustworthy accounts.

Elon Musk's content moderation strategy goes almost exclusively through Community Notes[1], which could be problematic. While carefully devised groups of users might come to similar conclusions as professional fact-checkers more quickly and at scale, their insights on a particular content do not necessarily rely on solid evidence, nor is that evidence always available for other users to make their own judgements. Also, particularly on highly salient political issues, studies have shown that coordinated action might distort the results (Allen *et al.*, 2022). Shifting the task of fact-checking to regular users, acting without any methodology, professional standards, or safeguards for independence, can be counterproductive and does not align with the definition of fact-checking established in the Code of Practice on Disinformation and committed to by Twitter, particularly as the requirements the company set to join the program do not prioritize authoritative sources or experts in the field.

Social media and the risk to civic discourse

The web 2.0, which is characterized by the ease with which individuals produce and share content, has made the internet social and seemingly participative. Social media epitomizes this dynamic. Social media platforms typically have very little public oversight and democratic governance although the market of attention can be deeply disruptive. This part addresses the systemic risk to civic discourse that social media poses in brutalising and radicalizing parts of the communication space. In order to address this trend, the EU's Digital Services Act (DSA) establishes that very large online platforms, including thus the major social media companies, need to conduct an assessment of the systemic risks that their operations create. One of those systemic risks is that to civic discourse. Civic discourse makes democracy possible, by solving issues through dialogue and rejecting violence. Social media companies host and publish content that their users produce, exchange, diffuse. That content tends to be of a more violent, emotional, and raw nature than offline communication or regular news media content. The key contention of this article is to suggest social media must undergo more thorough regulation since it relies on a vastly unregulated market of human attention and that the working logics of social media and especially their business model pose an inherent systemic risk to civic discourse, and therefore to the integrity and functioning of liberal democratic governance.

[1] X users can leave notes on any posts that can be potentially misleading. Notes become public only when there are enough contributors who rate the note helpful. Community notes are not edited nor modified by X teams. A post with a Community Note is not labelled, removed, or addressed unless it violates the X rules, terms of service of their privacy policy.

Social media in a legal loophole of content responsibility: not just publishers

The development of social media companies happened under a legal regime that pre-dates the emergence of machine-learning techniques and the use of suggestion algorithms on websites. The Communications Decency Act Section (CDA) 230 C(1) (Chander, 2022) is the governing benchmark under which social media companies have been able to develop their operations globally. Section 230 has become the "international law" of the various social media platforms. It protects and grants them immunity regarding decisions of content upkeep and removal on their systems. It protects platforms from being liable to the speech on their sites and regarding their moderation of that speech. Determining the type of liability for social media companies regarding the information communicated through and by them is a fundamental question for taming the circulation of conspiracy beliefs and the real-world safety and security consequences they may imply. It is important to consider the regulatory landscape of social media as the key enabler of systemic risks and especially towards civic discourse. Social media companies have regularly interpreted that section 230 shields them from liability over content produced and shared by its users on their systems. Indeed section 230 C(1) provides that "*no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider"* (Ibid.). An expansive interpretation of section 230 has been upheld as recently as in two US Supreme Court Cases this year. Majority opinions in a series of similar cases held that section 230 effectively shields social media companies from liability in cases where plaintiffs had argued that the content recommendation algorithms and suggestions effectively aided terrorist speech and radicalisation of terrorist offenders, included during the Paris November 2015 terrorist attacks. The Courts considered that algorithmic recommendation between users and targeting of third-party content did not make the platforms publishers of content, therefore immune from liability.

Legal and scholarly controversy over the extent of immunity is worth noting. First in terms of context. The original spirit of the section 230 was to correct the legal system following the *Stratton-Oakmont 1995* decision that held as a publisher a website that had decided to remove indecent material from its website, and thus become liable for defamation regarding other objectionable content. This decision was seen by Congress showing "*a legal system that provides massive disincentive for the people who might best help us control the internet to do so*". Section 230 of the CDA was therefore introduced as a way to secure a due diligence space for interactive service computer providers to remove indecent content – in that case to reduce child exposure to adult content. Second, late Judge Robert Katzmann's partly dissenting opinion in the Dyroff v. The Ultimate Software Group, No. 18-15175 (9th Cir. 2019) case offers an insightful contest to the broad immunity interpretation of Section 230. He argues that Section 230 does not protect Facebook's content suggestion and friend suggestion algorithms. Such algorithms actually go far beyond user's consumption of content as it makes the users to become part of unique communities based on what Facebook thinks the user might like and pay attention to. The creation and sustainment of a real-world network of affinities even surpasses the publisher or editorial function, pointing to a legal loophole. The third argument in the controversy over social media immunity goes that if they cannot be held liable for information provided by a third party, they could be held liable for the production of their own content. The result of suggestion algorithms can be construed as a message / content provided to the users of Facebook: the message is that they might like or connect with such message and such person. Suggesting, ranking content in a presentation that might increase engagement of its user could render a social media platform a liable publisher since the hierarchisation of content is very similar to an editorial process. Late Judge Katzmann argued that when the content becomes in such way independent of third-party information provision, then DCA immunity should not apply, and platforms should be held responsible for that content. This is a nuanced dent into the blanket immunity from liability that platforms argued they enjoy based on the fact that they don't provide filtering or editorializing of the content produced by their users and shared by them and that their general conditions provide that users remain owners of the content they choose to share on the platform. Put simply, "Facebook may be immune under the CDA from plaintiffs' challenge to its allowance of Hamas accounts, since Facebook acts solely as the publisher of the Hamas users' content. That does not mean, though, that it is also immune when it conducts statistical an"

The incentives not to regulate their contents or avoid editorial responsibility are strong for the social media platforms. Conspiratorial misinformation is a key engagement driver for the platforms, but it produces polarization, decreases the readiness of groups and individuals to solve their political differences through debate at the benefit of violence. This social media virality business model poses an obvious systemic risk to civic discourse. The circulation of the narrative of a stolen election in the US in 2020, massively abounded by group memberships on Facebook and user suggestions and the feedback loop of engagement led to the readiness of a crowd of a majority of regular citizens to ransack the US capitol. Contents predicated on indignation, hate, or fear leads to strong engagement on social media platforms is at the core of the attention soliciting business model of social media platforms. Regulation—meaning the reduction of the circulation of content or their banning altogether—runs structurally counter the virality business model upon which the platforms are predicated.

The notion of coordinated inauthentic behaviour

Effective regulation of social media platforms should make it difficult for conspiratorial, misinformation content to circulate and reach large audiences. The notion of "coordinated inauthentic behaviour" (CIB), a term invented by Facebook in 2018, is a standard concept for social media platforms in tackling disinformation and misinformation on their systems. The problem with CIB is that it lacks clarity, and that the

notions of coordination and inauthenticity of behaviour are not neatly defined thresholds. Most importantly, CIB is partly inadequate in grasping the changing modes of political mobilization on the internet. A series of coordinated and inauthentic behaviours were differently appraised by social media companies underlining that the standards for qualifying CIB are more than blurry.[2] CIB occurrences are often endemic to the political systems / democratic societies they originate from. CIB originating clearly from foreign governments is a somewhat clearer standard as there is an element of hostile foreign interference. But the same clarity of judgement cannot be afforded for domestic political questions and issues which can more often than not give rise to conspiratorial beliefs circulation. The distinction between a political actor legitimately using the affordances of digital services—coordinated and inauthentic—and the notion of illegitimate CIB is extremely sensible to make.

Many platforms are using bot detection software to determine if there are bots used to spread content. These bot detection models assess the behaviour of an account to identify if the account is a bot. But this centric way of detecting bots has reached its limits with the possibilities of bot software enabling them to behave as an account that is represented by a human. With a network of authentic looking bots, possibly being facilitated by botnet SaaS, it is easy to inauthentically spread any narrative over social media without being detected by a bot software. A new methodology to find these prohibited manipulation efforts is by using similarity models. With this method, a proportional database of content related to a conspiracy belief is run through a similarity model. The model is language agnostic and detects identical messages across the different languages in the dataset. The model also identifies similar messages with a similarity score. Using this method, botnets are identified more effectively, and evidence of CIB is easier gathered and presented in a clear overview. This provides additional possibilities for reporting and moderation, making it more efficient and effective to limit these botnet activities.

Normative Europe: what to regulate?

Social media regulations in Europe and North America have different undertones. While freedom of speech affords wide protection to content published on social media with virtually no control in terms of decency, correctness and undesirable outcomes in the US, regulation in the European Union focuses on user privacy and have more stringent rules on what counts as acceptable lawful content. Hate speech and incitement to violence are for instance forbidden in the EU. European regulations consider, that freedom of speech should not be the condition *sine qua non* but that preserving plurality is an equally important aspect of social media. The European area in general regards freedom of speech as a relative right, which shall not veer for instance into hate speech or incitement to violence. The EU's DSA mandates very large online platforms (VLOPs)—including therefore social media companies like Facebook, X and Google's YouTube—to remove illegal content on their platform and to engage in a systemic risk analysis posed by their operations. Spearheading social media regulation around the world, the DSA's provisions face objective difficulties, starting from the varying definition of illegal content and the wide private discretion that the VLOPs will keep enjoying.

Conspiracy beliefs can be likened to "lawful but awful" content on social media. Spreading conspiracy beliefs on social media does not necessarily fall under the category of illegal content. The baseline reports of VLOPs on their fulfilment of the Code of Practice on Disinformation show some positive aspects and different levels of commitment, but they generally come short of fulfilling what they committed to, with some of them showing no progress at all since becoming signatories. Telegram, one of the most critical actors in the dissemination of disinformation and conspiracy beliefs in the EU, is not yet part of this mechanism.

The circulation of conspiracy beliefs on social media falls under the risk categories of Article 34 of the DSA. In particular, conspiracy beliefs endanger civic discourse, electoral processes and public security as specified in article 34. The difficulty of countering conspiracy beliefs circulation is that this type of content is harmful without necessarily being illegal, which reduces what regulators can do. Conspiracy circulation must be correctly defined as a research object. The words, word combinations, specific hashtags, excluded terms or symbols need to be understood to build any kind of automated detection system such as an AI-based model. This is the case with hate and call-to-violent-action kinds of content. The issue of science denial is a complex one. Social media platforms have their own regulations on the use of the online environment. To prevent actors with harmful intentions exploiting the platforms to mislead and manipulate users, there are varying degrees over platforms, limitations to what users (accounts) are allowed to share or how to share content. It is crucial, that the different national authorities responsible for the implementation of the DSA are working closely with platforms and organizations reporting about harmful conspiracy beliefs. It is important that workflows and agreements on reporting, responsibilities and accountability are made specifically in the topic of harmful conspiracies beliefs spreading online.

The discussion in the section "Social media in a legal loophole of content responsibility: not just publishers" could develop in the US, as immunity over contents published on their platform must be completed. It is possible to build a case for social media regulation by considering the VLOPs as information gatekeepers. Social media has emerged as a key public space which would need to be based on the same regulation as information and media gatekeepers (Napoli, 2019). Being excluded from them or not using them comes

[2] "*Platforms' policies around "coordinated" and/or "inauthentic" behaviour are, like most of their policies, vaguely written, flexibly interpreted, and inconsistently applied. In the light of real fears about coordinated foreign efforts to interfere with our elections, the lack of clarity on these policies leaves platforms open to charges of bias in moderating speech, further undermining faith in the democratic process".*

indeed with a significant cost to market entry on a growing number of market segments. It comes also with a cost in terms of access to information. The question at stake here is that social media platforms exercise *de facto* content decisions and should thus be considered to fall under publisher's liability. Whether content decisions are made by humans, algorithms or artificial intelligence, those decisions determine access to information, determine the order of information presented and in fact restrict or expand the range of information that any individual can access. For this reason, as (Fukuyama & Grotto, 2020) have argued, the role of social media companies is very similar to that of large broadcasting and press corporations in the 1950s to 1970s when laws were imposed on them based on publisher liability. Social media companies sort, arrange and present information to fulfil their virality model and produce maximum engagement of users in order to sell advertisement space. This arrangement amounts to content decision and therefore editorial role for the VLOPs as to the information presented to users. Social science research has shown that engagement is likelier to be effective on contents that generate or appeal to hate, indignation, fear. The spread of conspiracy beliefs narratives is usually responding to those categories of emotions.

Conclusions

The way social media companies work is relevant to tame the potential for hybrid threat activity. Disinformation, misinformation, conspiracy belief circulation, brutalisation, and polarization of societies are phenomena deeply detrimental to civic discourse and more broadly to the soundness of liberal democratic governance. The business model of social media—capitalizing on human attention *i.e.*, user engagement—encourages those detrimental phenomena structurally. Conspiracy belief is more a matter of power and contestation than a matter of rationality. In short, it is very difficult to convince conspiracy believers away from their conspiratorial mindset. The circulation of conspiracy beliefs runs wide on various social media platforms, with the notable examples of Telegram and X under Elon Musk. The article advances that social media exploit a legal loophole that makes them non responsible publishers since they are historically and durably shielded by a blanket immunity provided by the US legislation. This legal loophole must be fixed considering that the platforms may actually produce their own content via the message their suggestion algorithms deliver to their users. The legal loophole is also wide in terms of responsibility since social media platforms also clearly work beyond publication of content: they also sustain real-world networks of affinities. For this series of reasons as outlined in the article, countering hybrid threats implies taming the circulation of conspiracy beliefs on social media by getting the platforms to become responsible of the contents that go through their services. This would place an appropriate business incentive for regulation and would likely preserve the profitability of social media companies as businesses. Most critically, it would undermine a lot of the conspiratorial material online and structurally reduce polarization and violence in political debates in our democracies.

Ethics and consent

Ethical approval and consent were not required.

Data availability

The data for this article consists of bibliographic references, which are included in the References section.

References

Abdalla Mikhaeil C, Baskerville RL: **Explaining online conspiracy theory radicalization: A second-order affordance for identity-driven escalation.** *Inform Syst J.* 2023; 1–25.
Publisher Full Text

Aikin SF: **Deep Disagreement, the Dark Enlightenment, and the Rhetoric of the Red Pill.** *J Appl Philos.* 2019; **36**(3): 420–435.
Publisher Full Text

Allen J, Martel C, Rand DG: **Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program.** In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* 2022; Retrieved October 10., 2023.
Publisher Full Text

Benton B, Choi JA, Luo Y, *et al.*: **Hate Speech Spikes on Twitter After Elon Musk Acquires the Platform.** *School of Communication and Media Scholarship and Creative Works.* 2022; 33.
Reference Source

Bond EB, Neville-Shepard R: **The Rise of Presidential Eschatology: Conspiracy Theories, Religion, and the January 6th Insurrection.** *Am Behav Sci.* 2023; **67**(5): 681–696.
Publisher Full Text

Bryanov K, Vasina D, Pankova Y, *et al.*: **The Other Side of Deplatforming: Right-Wing Telegram in the Wake of Trump's Twitter Ouster.** In: D. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova, I. Musabirov, & S. Pashakhin (Ed.). *Digital Transformation and Global Society. DTGS 2021. Communications in Computer and Information Science.* Springer, 2022; 417–428.
Publisher Full Text

Butler LD, Koopman C, Zimbardo PG: **The Psychological Impact of Viewing the Film *JFK:* Emotions, Beliefs and Political Behavioral Intentions.** *Polit Psychol.* 1995; **16**(2): 237–57.
Publisher Full Text

Chander A: **Section 230 and the International Law of Facebook.** *Yale Journal of Law & Technology.* 2022; 24.
Reference Source

Eddington SM: **The Communicative Constitution of Hate Organizations Online: A Semantic Network Analysis of "Make America Great Again".** *Soc Media Soc.* 2018; **4**(3).
Publisher Full Text

Einstein KL, Glick DM: **Do I Think BLS Data Are BS? The Consequences of Conspiracy Theories.** *POLIT BEHAV.* 2015; **37**(3): 679–701.
Publisher Full Text

Enders AM, Smallpage SM: **Who Are Conspiracy Theorists? A Comprehensive Approach to Explaining Conspiracy Beliefs.** *SOC SCI QUART.* 2019; **100**(6): 2017–2032.
Publisher Full Text

Fukuyama F, Grotto AJ: **Comparative Media Regulation in the United States.** *Social Media and Democracy.* 2020; 199–219.
Publisher Full Text

Gursky J, Riedl MJ, Joseff K, *et al.*: **Chat Apps and Cascade Logic: A Multi-Platform Perspective on India, Mexico, and the United States.** *Soc Media Soc.* 2022; **8**(2).
Publisher Full Text

Hickey D, Schmitz M, Fessler D, *et al.*: **Auditing Elon Musk's Impact on Hate Speech and Bots.** 2023; **17**: 1133–1137.
Reference Source

Jolley D, Douglas KM: **The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions.** *PLoS One.* 2014; **9**(2): e89177.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lee H, Andreu Perez L, Kim JN: **A duality of belief in conspiracy theories amplification: how active communicative actions work differently by trust in the Trump and Biden administrations.** *Online Media and Global Communication.* 2022; **1**(3): 524–550.
Publisher Full Text

Napoli PM: **Social Media and the Public Interest: Media Regulation in the Disinformation Age.** Columbia University Press, 2019.
Reference Source

Pasek J, Stark TH, Krosnick JA, *et al.*: **What motivates a conspiracy theory? Birther beliefs, partisanship, liberal-conservative ideology, and anti-Black attitudes.** *Elect Stud.* 2015; **40**: 482–489.
Publisher Full Text

Plakhta D: **Telegram as a tool for political manipulation.** *TV and Radio Journalism.* 2020; (19): 88–94.

Reichstadt R: **L'opium des imbéciles. Essai sur la question complotiste.** Éditions Grasset, 2019.
Reference Source

Rogers R: **Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media.** *Eur J Commun.* 2020; **35**(3): 213–229.
Publisher Full Text

Sunstein CR, Vermeule A: **Conspiracy theories: causes and cures.** *J Political Philos.* 2009; **17**(2): 202–227.
Publisher Full Text

Theocharis Y, Cardenal A, Jin S, *et al.*: **Does the platform matter? Social media and COVID-19 conspiracy theory beliefs in 17 countries.** *New Media Soc.* 2021; **25**(12): 3412–3437.
Publisher Full Text

Uscinski JE, Klofstad C, Atkinson MD: **What Drives Conspiratorial Beliefs? The Role of Informational Cues and Predispositions.** *Polit Res Q.* 2016; **69**(1): 57–71.
Publisher Full Text

Valcore J, Asquith NL, Rodgers J: **"We're led by stupid people": Exploring Trump's use of denigrating and deprecating speech to promote hatred and violence.** *Crime Law Soc Change.* 2023; **80**: 237–256.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Valeriani A, Vaccari C: **Political talk on mobile instant messaging services: a comparative analysis of Germany, Italy, and the UK.** *Inf Commun Soc.* 2018; **21**(11): 1715–1731.
Publisher Full Text

van Prooijen JW, Douglas KM: **Conspiracy theories as part of history: The role of societal crisis situations.** *Mem Stud.* 2017; **10**(3): 323–333.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Walther S, McCoy A: **US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism.** *Perspect Terror.* 2021; **15**(2): 100–124.
Reference Source

Wojczewski T: **Conspiracy theories, right-wing populism and foreign policy: the case of the Alternative for Germany.** *J Int Relat Dev.* 2022; **25**: 130–158.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 13 September 2024

https://doi.org/10.21956/openreseurope.18417.r39799

© **2024 Wold T.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Thomas Wold
Business School, North University Stjordal Campus, Stjordal, Norway

Thank you for this article on a very interesting topic. Conspiracy beliefs and misinformation on social media is indeed and important topic.
I find the text difficult to follow. It jumps from topic to topic. From legal loopholes, to conspiracies, and various social media platforms.
I think a more thorough definition of what conspiracy theories are would make it easier for the reader to follow the logic of the text. Also, a clearer discussion on who beliefs in conspiracy theories, and how these mechanisms work psychologically. Then it would be easier to give a description on how conspiracy theories thrive on social media, and how the different platforms functions.
There is a lot of references to research here, but it would be better to also include some words on the actual findings and conclusions from the various research. Now there are some claims, like the claim that belief in conspiracy has been one of the key drivers to social and political events, that seems unfounded.

Is the topic of the essay discussed accurately in the context of the current literature?
Partly

Is the work clearly and cogently presented?
No

Is the argument persuasive and supported by appropriate evidence?
Partly

Does the essay contribute to the cultural, historical, social understanding of the field?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Social media.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 20 June 2024

https://doi.org/10.21956/openreseurope.18417.r41082

© 2024 Pyrhönen N. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Niko Pyrhönen
University of Helsinki, Helsinki, Finland

This article makes a convincing point concerning the necessity to further develop measures through which social media companies would bear more responsibility over publishing or channeling harmful content to its users. The pertinent social problems are accurately and lucidly described, contextualized and based on state-of-the-art research.

The main shortcoming of the article is that it does not discuss in any detail the measures with which social media companies could be forced or encouraged to undertake any substantial action against the spreading of harmful content, not how these measures could be implemented. Typically, at least in an academic text, this would not be a major shortcoming as it often makes sense to focus on specific research problems and leave peripheral problems outside the scope of inquiry. In this case, however, this delimitation carries more severe consequences for the main argument and the supporting analytical narrative. Firstly, as the problems the article (otherwise laudably) focuses on are comparatively well known by researchers and practitioners, at least on the descriptive level, this delimitation challenges the originality of the contribution. Secondly, this delimitation allows only very abstract and superficial addressing of how any effective control measures for these companies can be implemented and enforced, particularly in a manner that does not undermine some fundamental tenets on freedom of expression or allow flawed democracies or authoritarian regimes to benefit from these measures. Thirdly, and linked to the second point, it would be very important to discuss, at least briefly, the potential limitations and unintended consequences the suggested control measures may carry, and how the identification potential problems might help in pre-empting them.

Finally, and not related to the aforementioned considerations, there are some issues with the language, many pre-positions are incorrect and some arguments are cut mid-sentence, which makes some passages relatively difficult to grasp. To summarize my statement on the otherwise valuable and topical contribution, there are avenues, including even some low-hanging, for further developing the main argument in a manner that makes it is easier to follow, exemplifies the suggested control measure in a little more detail, and illustrates an awareness of some of the major challenges implementing and enforcing these measures may carry.

Is the topic of the essay discussed accurately in the context of the current literature?
Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by appropriate evidence?
Partly

Does the essay contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: social media, conspiracy theories, alternative and stigmatized knowledge, political mobilization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 09 May 2024

https://doi.org/10.21956/openreseurope.18417.r39807

© 2024 Pitardi V. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Valentina Pitardi
Department of Marketing, University of Surrey, Guildford, England, UK

This research explores the policy problem of conspiracy belief circulation on social media. I believe this is an interesting piece of research and the topic is timely and relevant. There are few areas of improvements however that I detail below.
First, while there is relevant literature mentioned in the paper, important studies from psychology and consumer behaviour that can help clarify how and why conspiracy beliefs are formed are currently missing. For instance, important factors driving circulation and formation of conspiracies are degree of knowledge and expertise as well as levels of trust in the system. Without appropriate discussions of these two factors, I believe the research provides only one single view of how conspiracy beliefs form and develop. At the end of this report I provide details about useful resources that can help further develop this section.
Second, the three case studies presented are particularly interesting but I find them quite descriptive. For instance, is not clear what are the differences between these three platforms and what are the factors that enable (or not) the spread of misinformation more on some and less in other. A more systematic approach in the presentation of the case studies would improve the

robustness of the evidence provided.
Moving to the social media discussion, I find sometimes that the tone used in the research leans toward a description of SM merely as instruments of conspiracies creation. While I agree that they play a major role in the spread of misinformation and I fully support the need for a regulation, I do not think that this is the only side of the story. The paper should describe social media and their functions more objectively and explain how certain functionalities enable and facilitate conspiracies beliefs and circulation. This would provide a fairer and more comprehensive account of the complexities around the topic.

As for the missing references, I believe the papers published in this recent special issue on Current Psychology can help authors to further develop their research.

https://www.sciencedirect.com/journal/current-opinion-in-psychology/special-issue/10XFDXCG968

Wish you good luck in your endeavour

References
1. Gordon Pennycook • Lisa K Fazio: The Psychology of Misinformation •. *sciencedirect. CTHS*. 2024.

Is the topic of the essay discussed accurately in the context of the current literature?
Partly

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by appropriate evidence?
Partly

Does the essay contribute to the cultural, historical, social understanding of the field?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Consumer psychology and behavior

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.