Rau, Jan; Harfst, Jan-Ole; Mast, Tobias

Article

Platform badges for civic communication: An interdisciplinary discussion of a risk mitigation measure pursuant to Art. 35 DSA

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Rau, Jan; Harfst, Jan-Ole; Mast, Tobias (2025) : Platform badges for civic communication: An interdisciplinary discussion of a risk mitigation measure pursuant to Art. 35 DSA, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 4, pp. 1-32, https://doi.org/10.14763/2025.4.2054

This Version is available at:
https://hdl.handle.net/10419/336198

INTERNET POLICY REVIEW

Volume 14 **Issue** 4

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Platform badges for civic communication: An interdisciplinary discussion of a risk mitigation measure pursuant to Art. 35 DSA

Jan Rau *Research Institute Social Cohesion (RISC) | Sub-Institute Hamburg*
Jan-Ole Harfst *Leibniz Institute for Media Research | Hans-Bredow-Institut*
Tobias Mast *Leibniz Institute for Media Research | Hans-Bredow-Institut*

DOI: https://doi.org/10.14763/2025.4.2054

Published: 8 December 2025
Received: 24 February 2025 **Accepted:** 20 June 2025

Funding: This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under the grant number 01UG2450IY ("RISC – Hamburg"). The responsibility for the content of this publication lies with the authors. This work was partially funded by the Mercator Foundation ("DSA Research Network"). The responsibility for the content of this publication lies with the authors.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Rau, J., Harfst, J.-O., & Mast, T. (2025). Platform badges for civic communication: An interdisciplinary discussion of a risk mitigation measure pursuant to Art. 35 DSA. *Internet Policy Review*, *14*(4). https://doi.org/10.14763/2025.4.2054

Keywords: Platform governance, Fundamental rights, Hate speech, Risk mitigation measure, Digital Services Act

Abstract: To promote civility in public discourse on online platforms while respecting user freedoms, this paper explores a voluntary user badge system. By asking users to commit to civic norms of discourse, this system rewards their commitment with increased visibility, offering a potentially effective but underutilised approach to fostering constructive engagement. We examine different design options for these badges from an interdisciplinary perspective, assessing their potential benefits and risks in terms of discourse quality and users' fundamental rights, particularly within the context of the attention economy and the challenges posed by disinformation as well as conflict-driven and emotionally negatively charged content. The badge system aims to introduce a new logic of attention distribution to digital platforms, strengthening the structural conditions for civic communication. Through this analysis, we aim to contribute to platform governance from the perspectives of both political communication studies and legal studies.

Introduction

Initially celebrated as a catalyst for democracy, public discourse on the internet and digital media has, over the last decade, taken a more dystopian turn. Digital media is now frequently regarded as both a venue for and contributor to disinformation, extremism, and societal polarisation, particularly in the context of major political shifts in Western societies (Henrich et al., 2010; Lorenz-Spreen et al., 2023). While scholars caution against attributing these developments solely to digital media, noting that underlying social and political conflicts play a major role, there is broad support that digital media amplifies and accelerates them (Lorenz-Spreen et al., 2023; Schroeder, 2018).

In response to these developments, questions of platform governance have gained prominence. The term refers to the regulation of digital platforms by states, platforms, users, and advertisers, and to the interaction between these entities (Gorwa, 2019). Platforms govern user behaviour not only through rules such as terms of service (ToS) and their enforcement (Mast, 2023; Mast & Ollig, 2023; Wielsch, 2019), but also through regulation-by-design, encompassing architecture, features, and affordances that shape user behaviour and communication (Kim et al., 2022; Yeung, 2017).

In addition, platforms themselves are increasingly subject to governance. Building on national initiatives, the European Union continues to expand the platform economy and its 'Digital Single Market' through a growing set of legal frameworks. At the centre of this regulatory landscape is the Digital Services Act (DSA), which complements and intersects with private regulatory measures, forming a complex, 'hybrid' network of governance (Schulz & Ollig, 2023). The DSA aims to create a safe, predictable online environment that protects fundamental rights, applying the most stringent requirements to *very large online platforms* (VLOPs) and *very large online search engines* (VLOSEs) serving over 45 million EU users.

Under the DSA's risk-based regulatory approach (Efroni, 2021; Roth-Isigkeit, 2024; Husovec, 2024, pp. 270-284, pp. 316-346), providers of VLOPs must independently manage systemic risks through a process of assessment and mitigation (Art. 34, 35 DSA). Art. 34 DSA mandates that VLOP providers "diligently identify, analyse, and assess *any* systemic risks in the Union" originating from their services on at least an annual basis. Subsequently, Art. 35 DSA requires VLOP providers to implement "reasonable, proportionate and effective mitigation measures" tailored to these identified risks. In doing so, providers must take "particular consideration" of the potential effects on fundamental rights. At the time of writing, the Commission has

designated 23 services as VLOPs, of which the following are mainly used for general communication between users: Facebook, Instagram, Snapchat, TikTok, X, LinkedIn, Pinterest, and YouTube.[1]

This paper proposes and assesses a user badge for civic communication as a regulation-by-design governance mechanism and risk mitigation measure in line with Art. 35 DSA. The badge, offered as an opt-in feature by the platform, is linked to user behaviour requirements and provides greater visibility to compliant users. Users acquiring the badge commit to norms of civic communication, such as refraining from spreading dis- and misinformation (Zeng & Brennen, 2023) and engaging in deliberative exchange (Esau et al., 2020). In return, their content receives enhanced visibility within the platform's design, aiming to counter the distortive dynamics of the current attention economy and to instead introduce a new logic of attention distribution to digital platforms. By improving the structural conditions for civic communication, the badge creates opportunities for norm-abiding communicators to gain attention, incentivises constructive engagement, and reduces the visibility of problematic communicators.

Bridging perspectives from political communication science and legal studies, we first introduce the badge and its functions as a governance mechanism (Section I) and then explain how it can address systemic risks in the context of Art. 34, 35 DSA (Section II). Next, we provide a more detailed evaluation of the badge's effects and potential challenges from a political communication science perspective (Section III) and a legal perspective (Section IV), before concluding with a brief discussion (Conclusion).

Section I: User badges for civic communication as a risk mitigation measure for digital platforms

This section introduces the key components of the badge mechanism. While platforms will ultimately adapt implementation to their specific challenges, several essential design choices deserve discussion.

Badges as a platform governance tool

Badges that categorise specific user groups and visibly distinguish them from others are a common mechanism used on digital platforms. They are often part of a tiered governance approach, where different users are treated differently and are

1. Constantly updated overview at https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

held to varying standards (Caplan & Gillespie, 2020).

Many of the subsample of the VLOPs mainly used for general communication between users already implement some form of badge system, most prominently verification badges that indicate a user's identity has been confirmed by the platform or a verification partner. These appear in various forms across Facebook, Instagram, LinkedIn, YouTube, X, Pinterest, and TikTok (Meta, 2025a; Meta, 2025b; LinkedIn Corp., 2025; Google, 2025; X Corp., 2025; TikTok, 2025), with some platforms allowing users to purchase verification, others offering application-based approval, and some requiring users to meet specific criteria, such as a subscriber threshold.

Beyond verification, several platforms already offer badges that link user status to behaviour or performance. Snapchat's Snap Star programme, for instance, recognises creators with substantial reach and adherence to standards of originality, safety, and content appropriateness (Snap Inc., 2025). X Premium, the successor to Twitter's former blue check system, is a paid service that confers additional functionalities and boosts user ranking in some algorithmic systems. Together, these examples illustrate the two central elements of our proposal: linking eligibility to specific user behaviour and rewarding it with enhanced visibility. Crucially, unlike trusted flaggers under Art. 22 DSA, our badge would not grant users content moderation powers but solely signal their communicative status.

Visibility as a platform governance tool

The visibility component of the proposed badge follows a regulation-by-design logic, recognising that platform architecture and recommender systems strongly shape platform affordances and user behaviour (Yeung, 2017; Kim et al., 2022; Grüning et al., 2024). The discoverability of user content is largely determined by ranking algorithms, trending lists, and recommendation features, which function as gatekeepers to public attention. Research and whistleblower disclosures have repeatedly shown that these systems often privilege emotionally charged, sensational, or divisive content, thereby amplifying harmful communication such as extremism and disinformation (Hao, 2021; Horwitz, 2021; Lewis, 2018; Ribeiro et al., 2019; Whittaker et al., 2021). Platforms have responded to these issues, partly by downranking content they classify as borderline (Gillespie, 2022) and incorporating visibility governance as a dimension of content moderation (Goldman, 2021).

However, visibility governance is not only about reducing harmful content; there have also been efforts to increase the visibility of content deemed desirable. For

example, Facebook adjusted its algorithms and platform design to prioritise trusted news by allowing users to rank sources (Frenkel & Maheshwari, 2018; Meta, 2018). Later, it emphasised original reporting by rewarding transparent authorship (Meta, 2020). While these initiatives were subsequently scaled back—such as reducing visibility for political content (Meta, 2021) and phasing out dedicated news tabs (Meta, 2023)—they demonstrate how platform design can be leveraged to enhance the visibility of desirable content and users.

The proposed badge follows a similar logic by making commitment to civic norms a ranking signal. Users who meet the badge criteria would receive preferential placement across algorithmically curated spaces such as content feeds, search results, and community recommendations. Importantly, the badge should not override all other ranking signals but rather complement them, creating a meaningful yet proportionate visibility boost that rewards norm-compliant communication without distorting the overall information ecosystem.

Duties of care: Adhering to civic communication norms

Defining civic communication norms is inherently challenging, as interpretations of functional discourse in a democracy vary widely. Habermas's deliberative model for example prioritises consensus through rational-critical debate (Habermas, 2021; Wessler, 2018), whereas Mouffe and Laclau (Mouffe, 2000; Mouffe & Laclau, 1985) argue for an agonistic model that preserves conflict as a constitutive part of politics. Suppressing conflict, they warn, risks obscuring structural inequalities and power imbalances. The DSA itself references "civic discourse" in Article 34 and recital 82, but does not provide a definition, leaving room for interpretation.

For the purpose of this paper, we propose two variants of norms of civic communication that are based on an extended systemic risk analysis in the logic of Art. 34 DSA (which will be elaborated in detail in the next section) and which could be integrated in the duties of care of the badge.

Tackling dis- and misinformation

Although disinformation is already a key focus of governance and moderation efforts (Saurwein & Spencer-Smith, 2020), the badge could serve as an additional layer, imposing duties of care on users to avoid spreading dis- and misinformation. Such duties align with the principle of balancing conflicting legal interests. Under Article 10 of the European Convention on Human Rights (ECHR), while untrue statements of fact are protected by fundamental rights, they can be more readily restricted in favour of competing legal interests than expressions of opinion (EC-

tHR, 2009, para 29; ECtHR, 2021, para 51). The European Court of Justice (ECJ), which has not yet applied similarly differentiated case law to the EU fundamental right of freedom of opinion and expression, typically orients itself on such ECtHR case law (Art. 52(3) Charter of Fundamental Rights (CFR)). Comparable obligations exist in §§ 6, 19 German Interstate Media Treaty (Medienstaatsvertrag) and Section 2 of the German Press Codex, a tool of voluntary self-regulation for German media outlets (Deutscher Presserat, 2017), which impose due diligence obligations on media actors and journalists. While defining and assessing truthfulness remains challenging, platforms have already developed content moderation mechanisms to address this issue (Stewart, 2021).

Definition: While definitions of dis- and misinformation as well as relevant subcategories and/or related phenomena vary widely (Zeng & Brennen, 2023), the European Code of Conduct on Disinformation (European Commission, 2025) understands dis- and misinformation as follows:

- disinformation is false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm;
- misinformation is false or misleading content shared without harmful intent, though the effects can still be harmful, e.g., when people share false information with friends and family in good faith.

Implementation: Badge users would commit to (and be held responsible for) not engaging in any intentional dissemination of disinformation as well as to avoid disseminating misinformation whenever possible. To ensure the latter, users could commit to a particular due diligence process towards their communication. An example of such a diligence process could be the already mentioned German Press Codex. Section 2 of the German Press Codex requires media outlets (among other things) to check information intended for publication for accuracy and to reproduce it truthfully, whereby its meaning must not be distorted or falsified. Symbolic photos must be clearly marked or identifiable as such, and when publishing survey results, relevant information such as the number of respondents, the time of the survey, or the representativeness must be highlighted (Deutscher Presserat, 2017). While these due diligence requirements should be adjusted to the specifics of social media communication, they can serve as a general role model for a similar process in the context of the badge.

Enhancing deliberation

To address the digital amplification of conflict-driven and negative-emotionally

charged content, which can contribute to societal hyperpolarisation and authoritarian takeover (see section II for more details), a self-commitment to a deliberative communication style could be an effective countermeasure. Deliberation, often originating in Habermas's ideas about deliberative democracy (Habermas, 2021; Wessler, 2018), emphasises the exchange of reasons, critical reflection, and mutual respect among participants.

Definition: For the purpose of this badge, a classic understanding of deliberation, including dimensions such as rationality, reciprocity, respect, and constructiveness (Esau et al., 2020), could be a meaningful way to achieve the set goals of reducing the amplification of conflict-driven and negative-emotional content. Rationality refers to the degree to which participants stay focused on the topic and support their claims with reasons, reflecting a structured and thoughtful engagement. Reciprocity captures the extent to which participants engage with one another's arguments, indicating dialogical responsiveness and a willingness to consider alternative views. Respect is seen in the maintenance of civil discourse, where participants avoid hostility and treat others' contributions with dignity. Constructiveness highlights the orientation toward problem-solving and compromise, showing that deliberation can aim not just at expression but at generating workable solutions (Esau et al., 2020).

It should, however, be underlined that in the rich and diverse body of literature on deliberation, there are many additional and potentially differing perspectives on how to understand and operationalise it. Next to the proposed classic understanding, more inclusive variants, for example, also value emotion, storytelling, and everyday political talk (Esau et al., 2020; Friess & Eilders, 2015).

Implementation: To support the usage of deliberation (especially given the complexity of the concept), users should not only be educated about the concept (see next subsection, Badge acquisition), but in addition could be supported by AI chatbot interventions. New studies show the potential of AI in educating and supporting users in adjusting their language to meet certain communicative norms (without undermining or altering the users substantive argument) (Argyle et al., 2023; Gelovani et al., 2025).

A mechanism to be potentially utilised by many norms

While we in this text focus on dis-/misinformation avoidance and deliberation as examples, the proposed mechanism (user commitment to duties of care, rewarded by platform algorithmic amplification) is relatively norm-agnostic and could accommodate alternative or additional standards depending on platform context and

systemic risk profiles. Concepts such as incivility (Kim et al., 2022), democratic listening (Dobson, 2014), bounded informational diversification (Brady et al., 2023), or PRIME (Brady et al., 2023) may offer further guidance on what constitutes beneficial or harmful discourse. The badge should thus be understood as a flexible governance instrument, adaptable to evolving risks and scholarly insights.

Badge acquisition

We suggest a layered approach for badge acquisition: users may commit to one or several proposed norms (e.g., avoiding the dissemination of dis- or misinformation), but not to the other. Especially deliberation represents a specific community style that may not suit every user's content, allowing commitment to only one norm provides flexibility and still enables partial benefit from the badge. The amplification of user communication can then vary according to the number of norms they commit to.

Users should be informed about the norms they commit to, for instance through short texts or videos explaining how to avoid misinformation or practice deliberation. Clear examples and a summary of potential consequences for violations should be included.

Eligibility criteria could be implemented to prevent misuse by spam accounts, for example by requiring accounts to have been active for a specified duration (e.g., a week or month) or verified as belonging to a real person. Another option would be a platform-administered vetting process that could also consider users' past communication and activities when deciding on badge allocation. While potentially more effective against abuse, such vetting is more resource-intensive and raises fundamental rights concerns, which will be outlined below.

Compliance, badge loss, and recovery

Compliance: To assess compliance with the duties of care, several approaches (depending on the concrete duty) are possible connecting already existing as well as additional content moderation layers:

Many platforms already use a mix of measures to detect disinformation, including AI-driven content moderation, automated flagging, and human review teams who respond to reported content and patterns. These review teams often rely on user-based notice-and-takedown procedures or trusted flagger models. Similarly, the compliance of badge users could be assessed automatically through moderation algorithms and complemented by human review based on algorithmic or user

flags.

For users committing to deliberative principles, algorithmic assessments could evaluate the presence or absence of these norms in their communication. These concepts have already been operationalised for quantitative research (Esau et al., 2020; Behrendt et al., 2024; Esau et al., 2023; Steenbergen et al., 2003), providing a basis for platform assessments. Challenges remain, however, as many operationalisations have been developed for English and other Western languages, complicating a potential global application (if wanted). These frameworks are also vulnerable to *biases* and other limitations that may affect fairness or accuracy (Kathirgamalingam et al., 2024; Kim et al., 2022; Kim, 2023; see Section III).

Building on existing and/or additional but mostly automated content moderation efforts, the implementation of a compliance assessment for the badge could be introduced with relatively low additional costs for the platforms.

Compliance should be measured over a user's overall communication within a defined period (for example, a month) or across a set number of posts, rather than per individual item. Occasional mistakes – such as accidental misinformation – should not result in immediate badge loss (see next subsection on warnings). Likewise, occasional departures from deliberative norms, such as incivility, may not be inherently problematic unless they dominate a user's style. Focusing on overall communication also mitigates measurement errors on individual posts, since algorithmic misclassifications are likely to balance out over a larger sample (Kathirgamalingam et al., 2024; Kim et al., 2022; Kim, 2023). Overall, compliance assessments should allow for some divergence but penalise systematic violations.

Loss and recovery: To maintain the badge's integrity, a system should address users who fail to comply with its norms. Existing enforcement models from major VLOPs provide inspiration: most use a tiered "strike" system that escalates penalties with repeated violations, while severe breaches may trigger immediate sanctions. Variants of such systems are employed by YouTube (Google, 2025), TikTok (TikTok, 2025a), Facebook and Instagram (Meta, 2025), Snapchat (Snap Inc., 2025), and Pinterest (Pinterest, 2024, 2025). The strength of this approach lies in matching the severity of the penalty to the gravity of the violation. Research shows that not only severe penalties such as permanent deplatforming can reduce ToS-breaking behaviour (Chandrasekharan et al., 2017), but also lighter sanctions like warnings (Yildirim et al., 2021). For badges, the strongest sanction would involve losing the badge and its benefits. Recovery options should also be considered: while recidivism risk exists, a temporary suspension combined with a requirement to meet ac-

quisition criteria again could suffice to promote future compliance. Evidence from online discussion platforms indicates that users who experience a temporary suspension are not more likely to face another suspension than those who have never been suspended (Tangwaragorn et al., 2025, p. 9).

Finally, some of the DSA's requirements relating to reporting systems and moderation decisions would also apply to a badge system. Particularly relevant, the DSA provides for an internal complaint-handling system, which would cover the badge system presented here: Art. 20 letter a) DSA obliges providers of online platforms to provide access to such a system that enables recipients of the service to lodge complaints, electronically and free of charge, amongst others against decisions regarding the restriction of visibility of information. The concept of visibility restriction poses greater difficulties of interpretation than the simple deletion of information or blocking of accounts, because it is hard to determine a general "status quo" of content visibility among the algorithmically individualised feeds of different recipients (Fertmann, 2025, para. 30-33). In the present scenario, however, Art. 20 letter a) DSA is applicable because the withdrawal of the badge is a comparatively safe sub-case of "delisting." The badge system and the extension of the complaint-handling system would have to be pointed out in advance by the platform in its ToS in accordance with Art. 14(1) sent. 2 DSA.

Section II: Subsumption under Art. 34, 35 DSA – How can the badge address systemic risks?

This section examines how the proposed badge integrates into Art. 34 and 35 DSA. The systemic risk framework has been described as "extremely broad, vague and abstract" (Griffin, 2023, p. 77) and "by far the most open-ended DSA provision" (Husovec, 2024, p. 316), with few explicit examples (Janal, 2021, p. 266). Articles 34 and 35 function as umbrella norms covering technical design, terms of service, and content moderation. We first outline how digital media has distorted public discourse, then show how these developments constitute systemic risks under Art. 34, and finally discuss how the badge could serve as a mitigation measure under Art. 35.

The digital distortion of the public arena

Digital media has reshaped the public sphere, introducing new actors and shifting discursive power (Klinger & Svensson, 2015; Schroeder, 2018). Radical and extreme actors, in particular, excel at exploiting these new communication opportunities (Jungherr et al., 2019; Miller-Idriss, 2020; Schroeder, 2018). In addition, this

shift has intensified the attention economy, where communicators and platforms compete for limited human attention amidst a flood of information (Franck, 1998; Webster, 2014). As the competitiveness of the attention economy has substantially increased, both, communicators and platforms, utilise digital tracking systems which provide precise content optimisation for communicators to align with user and algorithm preferences and maximise engagement (Jungherr et al., 2020).

The amplification of dis- and misinformation: As the former gatekeeping role of traditional media weakens, communicators face fewer constraints when spreading dis- and misinformation. Although prevalence and impact remain debated (Adams et al., 2023; Ecker et al., 2024; Simon et al., 2023), studies indicate such content often gains disproportionate visibility on digital platforms (Luo et al., 2022; Vosoughi et al., 2018).

The amplification of conflict-oriented and negative-emotionally charged content: Studies in social psychology show that people are drawn to content favouring their ingroups and are especially responsive to negative, emotionally charged moralised information (Brady et al., 2023). While these biases originally evolved to foster cooperation and problem-solving, they may be ill-suited for the hyper-competitive digital attention economy as outlined above.

Research shows that content emphasising group identity, emotional conflict and antagonism—"us vs. them" rather than solution-oriented debate—achieves high visibility online (Brady et al., 2019, 2021; Kaiser & Rauchfleisch, 2020). Contrary to the echo-chamber hypothesis, polarisation is frequently driven by continuous, unmoderated confrontation across political groups, enabled by interconnected digital spaces and the prominence of conflict-oriented communication (Bail, 2021; Törnberg, 2022). Several studies also show that emotion and outrage amplify dis- and misinformation (Bago et al., 2020; Bakir & McStay, 2018; McLoughlin et al., 2024; Vosoughi et al., 2018).

Because conflict-laden, negative-emotion communication is an effective communication strategy and can create favourable climates for certain political agendas (Mau et al., 2024), many actors deliberately intensify this hostile and antagonistic discourse. For authoritarian actors in particular, such styles are central to communicative success (Freistein et al., 2022; Heiss & Matthes, 2020; Miller-Idriss, 2020). The role of digital platforms in this context is especially problematic, as they, driven by the logic of the digital attention economy, often seem to reinforce this engagement-generating discursive style through their recommendation algorithms (see, for example, Brady et al., 2020, 2023; Hao, 2021; Horwitz, 2021; Lewis, 2018;

Ribeiro et al., 2019; Whittaker et al., 2021).

Art. 34 DSA: What systemic risks can be addressed?

Building on the above, both the amplification of dis- and misinformation and of conflict-driven, negative-emotional content pose systemic risks on many VLOPs primarily used for interpersonal communication. Particularly relevant is the systemic risk under Art. 34(1)(c) DSA, which covers "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security." Other systemic risks can also be addressed, including protection of fundamental rights under letter b) (especially human dignity and privacy), prevention of illegal content under letter a) (notably criminal hate speech), and protection of minors under letter d), as civil discourse is consumed by and influences them.

Dis- and misinformation have been central to platform governance debates over the past decade (Jungherr & Schroeder, 2021; Saurwein & Spencer-Smith, 2020). Their spread raises concerns that democracies may lose the ability to maintain shared understandings of truth and falsehood – an essential precondition for accountability, elections, and policy agreements (Herzog, 2024). Combating disinformation requires a variety of complementary measures (Hägle et al., 2024), but the DSA addresses its negative effects on democratic processes through Art. 34(1)(c), classifying them as systemic risks to civic discourse, electoral integrity, and public security. It explicitly targets disinformation coordinated through campaigns (Beyerbach, 2024), a focus already reflected in formal proceedings against Facebook and Instagram (European Commission, 2024b).

Compared to disinformation, the amplification of conflict-driven, emotionally charged content has received less attention in platform governance debates. Yet growing evidence links its prevalence to what Levitsky and Ziblatt term hyperpolarisation – progressive erosion of democratic norms and mutual toleration (Brady et al., 2023; Levitsky & Ziblatt, 2018; Lorenz-Spreen et al., 2023). Hyperpolarisation fosters political gridlock, weakens responses to societal challenges, creates openings for authoritarian encroachment, and may escalate into political violence (Levitsky & Ziblatt, 2018; Kleinfeld & Sedaca, 2024; Oittinen & Molokach, 2023). Given the digital success of authoritarian actors outlined earlier and the negative effects elaborated within this paragraph, amplification of such content constitutes a systemic risk to civic discourse, electoral processes, and public security.

In addition to systemic risks highlighted so far, dis- and misinformation and the amplification of conflict-driven and negative-emotionally charged content can also

contribute to "actual or foreseeable negative effects for the exercise of fundamental rights" (Art. 34(1) letter b)). The problematic effects of online platforms described above are echoed in legal scholarship: Given that VLOPs play an essential role in the public sphere (Augsberg & Petras, 2022, p. 102), excessive hate speech and agitation—often intensified by poor communication standards and prevalent in the context of dis- and misinformation and antagonistic discourse—can threaten users' freedom of expression and information (Art. 11 CFR). Such a hostile environment might lead users to self-censor, disengage from online discourse, or even delete their accounts, ultimately resulting in a loss of valuable contributions to public debate—a phenomenon known as the *silencing effect* (Gelber & McNamara, 2016; Lüdemann, 2019, p. 282 et seq.; Markard & Bredler, 2021, p. 865; Völzmann, 2021, p. 621).

Art. 35 DSA: How can the badge address these systemic risks?

Badges as a risk mitigation measure pursuant to Art. 35 DSA

Art. 35 DSA lists several examples of suitable mitigation measures, and the badge fits well within these requirements. Displaying badges on user profiles and providing an application process constitute "adapting the design, features or functioning" of VLOP services under Art. 35(1)(a). The rules of conduct accepted during application amount to an adaptation of platforms' "terms and conditions and their enforcement" (Art. 35(1)(b)). Finally, granting badge users increased visibility requires "adapting algorithmic systems, including recommender systems," allowing platforms to prioritise badge-compliant content (Art. 35(1)(d)).

The badge system could be offered globally as a voluntary measure. However, since the EU's regulatory competence is limited to the EU's single market, including systemic risks as defined in Articles 34 and 35 of the DSA, the badge system can only function as a risk mitigation measure within the EU under the legal act. To avoid fuelling the narrative of internationally contentious sovereignty conflicts and spillover effects, platforms should either explicitly clarify that they are adopting an approach that extends beyond the EU on a purely voluntary basis or make the badge system available exclusively to users from the EU. As platforms generally offer different technical versions of their services for different regulatory areas and regularly roll out or test new features in limited contexts, a territorially differentiated approach should not present any technological challenges. It seems unproblematic in this context that badge-bearing people from the EU may then communicate with those outside the EU.

Expected effects of the badge as an intervention mechanism

Art. 35 DSA requires mitigation measures to be tailored, reasonable, proportionate, and effective, with particular regard to fundamental rights. While the badge's effectiveness and proportionality depend on concrete design—especially specific, enforceable rules of conduct—there are strong indications that it could address the outlined systemic risks.

As outlined above, we propose two duties of care connected to the badge mechanism: first, avoiding the dissemination of dis- and misinformation, and second, engaging in deliberative communication. By introducing a new logic of attention distribution to digital platforms and enhancing the structural conditions for civic communication, the goal of the badge is to:

1. Create an opportunity space for communicators committed to civic communication norms to gain visibility.
2. Incentivise all communicators to engage within this space according to these norms.
3. Undermine distortive communication strategies by problematic communicators.
4. As a result of 1) - 3), increase the overall visibility of civic communication.

The proposed new attention distribution logic aims to reshape and enhance the structural conditions for civic communication by reducing the communicative advantage currently enjoyed by disinformation and conflict-driven and negative-emotionally charged content. Communicators who adhere to civic norms would no longer face systemic disadvantages, making civic engagement a competitive and viable strategy. This adjustment also creates new incentives for those indifferent to civic communication to align with these norms, as doing so could enhance their visibility and success. Problematic actors who rely on disinformation and emotionally charged, divisive tactics might be disincentivised, as their strategies would no longer provide the same rewards. If they persist, their communication would likely be less visible and effective compared to before. On a broader platform level, the volume and visibility of dis- and misinformation and conflict-oriented content could decrease. This might mitigate the negative effects of hostile discourse climates and could enhance the protection of fundamental rights. Additionally, users engaging with content would benefit from clearer signals, such as indicators showing that a communicator has committed to civic communication norms. These markers would provide users with insights into the probable reliability or potential bias of the content they consume, fostering a more informed and balanced digital environment.

As a general limiting factor, it should be emphasised once again that issues such as hyperpolarisation and authoritarian encroachment are highly complex phenomena. As indicated in the introduction, while digital media discourse contributes to these challenges in a meaningful way, it is far from being their only source (Jungherr et al., 2020; Schroeder, 2018). Any risk mitigation measure targeting digital platforms can therefore only address the aspects of these issues that are directly linked to platform dynamics. However, as the analysis above has shown, these aspects are substantial and relevant.

Overall, this initial assessment suggests the badge is a suitable measure for mitigating several systemic risks. The next sections examine effects, benefits, challenges, and limitations in more depth—Section III from a political communication perspective and Section IV from a legal perspective.

Section III: Extended discursive evaluation of the badge

While the badge seems to be a suitable measure for addressing systemic risks as outlined in Art. 34, 35 DSA, its potential side effects and limitations should be carefully examined and, where possible, mitigated before implementation.

Importance of political conflict

A key concern when implementing measures to reduce antagonistic discourse is recognising the essential role of political conflict. As Section I highlighted, conflict and disagreement are not inherently problematic in a pluralistic democracy. Agonist scholars such as Mouffe and Laclau regard in- and out-group conflict as fundamental to politics (Mouffe, 2000; Mouffe & Laclau, 1985). Efforts to suppress conflict – such as enforcing deliberative settings that prioritise consensus – often fail to resolve underlying tensions and risk reinforcing existing domination and injustice. Allowing political contestation instead creates space to challenge these injustices (Mouffe, 2000; Scudder & White, 2023).

Kreiss and McGregor (2024) similarly stress that moral claims, collective identities, and emotions are vital for marginalised groups seeking unity and visibility. Strong moral positions and emotional expression help form communities and advocate for their struggles. Negative emotions toward political opponents can be justified. From this perspective, polarisation is not necessarily harmful but an inevitable result of deep societal conflict and legitimate grievances. Digital media has expanded the opportunity for excluded voices to express these grievances, but attempts

to curb polarisation without addressing its roots risk merely suppressing the issues rather than resolving them.

It is crucial to acknowledge these concerns before implementing a measure like the proposed badge without careful consideration. At the same time, digital media has already created space for political newcomers, raising the visibility of oppositional and marginalised voices. Although the badge may slightly slow this process by reducing conflict and polarisation, it is unlikely to change its overall trajectory. Even Habermas' critics concede that excessively antagonistic discourse can harm democracy (Mouffe, 2000), fostering hyperpolarisation and openings for authoritarian actors (Levitsky & Ziblatt, 2018). The badge is intended to mitigate these risks accompanying the digital transformation of the public sphere—not to suppress or reverse its plurality. Rather, it seeks to give democracies an opportunity to develop mechanisms to meaningfully integrate this expanding diversity of perspectives and enable a conversation and competition between them without overheating in the process.

Definitional power, its potential abuse, and biases

Like established content-moderation practices, implementing the badge requires defining and enforcing what counts as dis- or misinformation and incivility. This necessarily concentrates power that can be abused and is prone to subjectivity and bias. Authoritarian regimes illustrate the danger: accusations of spreading falsehoods are frequently used to silence opposition (Funk et al., 2024; Truong et al., 2018). Without robust checks and safeguards, the badge could unintentionally enable comparable abuses.

Beyond deliberate manipulation, what counts as truth or falsehood—especially in political contexts—is often situational and shaped by intersecting factors such as gender, class, and race (Herzog, 2024). These dynamics influence what is accepted as truth, reproducing existing marginalisation. The risk is heightened by moderation practices that prioritise cost-efficiency, which may reinforce biases. Perceptions of civility are similarly context-dependent and contested, complicating the creation of universally accepted standards (Kathirgamalingam et al., 2024; Kim et al., 2022; Kim, 2023).

Appropriate safeguards are essential to prevent abuse and limit bias. The badge should be embedded in existing moderation frameworks, particularly the DSA-mandated appeal mechanisms (see Section I), to reduce arbitrary decisions on badge acquisition, loss, and recovery. Decentralising decision-making – for exam-

ple through social media or community councils (Kettemann & Schulz, 2023) – can broaden representation, improve diversity, and better protect marginalised users.

Abuse by problematic actors

Some problematic communicators may still commit to the badge's duties of care, often for strategic reasons. Authoritarian and extremist actors already moderate their content on stricter platforms to gain visibility and funnel audiences to less regulated spaces (Fielitz & Schwarz, 2023); similar tactics could emerge with the badge. Many harmful ideas are difficult to capture through binary truth tests or deliberative criteria, as they are often expressed in evasive, sarcastic, or nuanced ways that evade detection, especially by automated systems (Baider, 2023). Consequently, such actors might still qualify for the badge and use the resulting visibility to spread harmful ideologies.

However, the impact of such exploitation may be less severe than it appears. Authoritarian actors depend heavily on conflict-oriented, emotionally charged communication to gain traction online. To qualify for the badge, they would need to adopt more civil styles, losing a core element of their strategy. Even if they obtained the badge, this shift would constrain their ability to distort discourse and likely reduce their visibility. As an additional safeguard, enhanced vetting procedures for badge acquisition – as suggested in Section I – could further limit abuse.

Section IV: Extended legal evaluation of the badge

This section assesses the proposed civic communication badge measure from a legal perspective. It begins by analysing how the badge measure aligns with the overarching principles of the DSA. The second part examines the implications for fundamental rights, highlighting the challenges that must be addressed when designing and implementing the badge system.

Compatibility with overarching DSA principles

As shown above, the badge system proposed here can be effectively subsumed under the management of systemic risks but is also in line with the overarching logics of the DSA. At first glance, one might have doubts about this, as the DSA aims to create harmonised requirements, i.e., a level playing field, as stated in Art. 1(2) DSA. Many provisions in the DSA are designed to empower platform users and safeguard them from arbitrary decisions made by platform operators. Art. 14(4) DSA, for instance, functions as "a sort of general clause against unfair treatment by

providers" (Husovec, 2024, p. 254) and therefore mandates that operators act objectively when moderating content, thereby prohibiting arbitrariness. Additionally, according to Art. 21(3) letter f) DSA, out-of-court dispute settlements must meet the requirement of fairness. Furthermore, under Art. 25(1) DSA, online interfaces must also not be designed in such a way that they manipulate users or otherwise prevent them from making free decisions.

However, a badge system that treats groups of users on a platform differently does not inherently contradict this concept. While it is true the relative reduction in visibility suffered by non-privileged users is disadvantageous for them, this does not necessarily imply unacceptable discrimination. On the contrary, by recognising the general terms and conditions set by platforms as the foundation for content moderation, the DSA in principle respects the platforms' self-determined communication guidelines (Husovec, 2024, p. 327). Many provisions, including those on the management of systemic risks, consider the ToS to be a formative governance instrument.[2]However, it can only function as such if it is variable.

This impression is reinforced by the fact that the DSA itself grants special rights to certain user groups on platforms and privileges them if they meet certain requirements. In doing so, it differentiates between the notions of need for protection and the worthiness of protection. For example, reports by trusted flaggers (Art. 22(1), Art. 16 DSA) and reports from organisations commissioned by users (Art. 86(2), Art. 20 DSA) are given preferential treatment in content moderation. Certain reports from media services are also given preferential treatment under an amendment to the DSA by Art. 18 of the European Media Freedom Act (EMFA) to strengthen their position on online platforms. Furthermore, minors are granted special protections under Art. 28 DSA, establishing a regulatory regime that goes beyond the standard protections for other user groups. In this context, a differentiated badge system aligns with the principles outlined in the DSA, as it would elevate civil communication that is deemed especially worthy of protection while downgrading communication that may require additional safeguards.

Fundamental rights assessment

The design of the badges must be shaped by the fundamental rights enshrined in the CFR in two critical ways. First, these rights serve as the foundation for addressing one of the systemic risks the badges aim to mitigate: the potential of "actual or foreseeable negative effects for the exercise of fundamental rights" as outlined in

2. See Art. 14, Art. 20(1), Art. 21, Art. 27, Art. 34(2)(b), Art. 35(1) let. b DSA.

Art. 34(1) subparagraph 2 sentence 2 letter b DSA. Second, under Art. 35(1) DSA, as a risk mitigation measure, the badges themselves must be designed with consideration of their potential impact on fundamental rights. VLOP providers are therefore required to take precautions to avoid unintended systemic risks, such as the suppression of freedom of expression (recital 86 DSA).

To achieve the optimal effect on fundamental rights, the badges should be crafted in a way that promotes a secure, democratic online environment while minimising any self-restrictive consequences they might cause. In short, it's about distinguishing between opinions and factual claims, considering the degree of users' autonomy, and ensuring the procedure and decisions are transparent and safeguarded.

VLOPs are not states or EU actors

In literature on platform governance and business and human rights (BHR), the issue is often raised that, while the services are highly significant in terms of fundamental rights and have great decision-making power over how freedom is used, they are not formally bound by the fundamental rights catalogues of state constitutions or international law (York & Zuckmann, 2019; Callamard, 2019). The same was in principle true for EU law, as Article 51(1) CFR stipulates that the fundamental rights outlined in the Charter are binding only on EU institutions and member states when implementing European Union law. Private companies, such as platform operators, are not included in this provision. As interpreted by the ECJ, the conventional cases of horizontal effect under Union law do not extend to the content moderation activities of private entities (Mast & Ollig, 2023, p. 464 et seq.). In spite of the moral dimension of fundamental rights, platform operators such as Meta were legally free to decide whether to adhere to human rights such as the United Nations' General Principles on Business and Human Rights (UNGPs) when drafting the charter of their Oversight Board. However, this was without the formal enforcement power of a human rights organisation and was therefore deficient (Douek, 2021; Tiedeke & Fertmann, 2024).

The European legal acts of platform law now remedy this situation. Both the Terrorist Content Online (TCO) Regulation and the DSA refer to fundamental rights and require platform operators to respect those when moderating content (Art. 5(1) TCO Regulation, Art. 14(4) DSA). However, it is crucial to note that platform operators do not wield sovereign power. Although the ECJ has not yet provided a precise interpretation of Art. 14(4) DSA, it appears that privately run platforms—capable of asserting their own fundamental rights as the freedom to conduct a business (Art. 16 CFR)[3]—are subject to fewer constraints on fundamental

rights in content moderation practices compared to state or EU actors. This supports the feasibility of risk mitigation measures like the proposed badges, which could help manage 'awful but lawful' content by establishing moderation standards that align with fundamental rights while still allowing platforms the discretion to define their own community guidelines.

As shown above, the DSA maintains this sensitivity to fundamental rights concerning the CFR when assessing and mitigating systemic risks in Art. 34(1) letter b) and Art. 35(1). Online platforms provide nearly unrestricted communication on virtually any topic and in any style, meaning that regulating discourse could implicate a broad range of fundamental rights (Mast, 2024 p. 613). As such, instead of focusing on specific hypothetical scenarios, it is smarter to establish general considerations related to fundamental rights that should guide the implementation of a badge system. The following section outlines key concepts and dimensions that warrant particular attention in this context.

The idea of autonomy

Fundamental rights, as freedoms, inherently protect individual autonomy, including the right to abstain from or consent to an infringement of these rights. According to fundamental rights theory, no such right is deemed impaired if an individual knowingly and willingly consents to or waives its protection (Ehlers & Germelmann, 2023, § 2.2 para. 122; Jarass, 2021, para. 18). This is consistent with long-established legal principles, such as the Latin maxim *volenti non fit iniuria* ("to a willing person, no harm is done"). The CFR relies on this tradition by taking free consent explicitly into account in the right to integrity of a person (Art. 3(2) letter a) and the protection of personal data (Art. 8(2)).

In addition, the ECJ's interpretation of fundamental rights is influenced by the ECHR and the ECtHR's case law (cf. Art. 52(3) CFR), while the latter has acknowledged that individuals can waive their right to privacy (ECtHR, 2010, para. 71).

This overarching idea of consent applies to the badge system if platform users can freely choose whether to communicate with or without adhering to stricter language requirements tied to the badge. In this framework, users should be able to make an informed, rational decision about whether to adopt the badge's norms. However, the legitimacy of autonomy as a factor decreases if the consequences of

3. Whether the provision and design of a communication platform constitutes an expression of freedom of expression or media (Art. 11 CFR) itself is yet to be clarified at a European level. The German Federal Court of Justice (BGH) ruled in this sense in its Facebook judgement of 29 July 2021 – III ZR 179/20, §§ 69, 70.

this decision become overwhelming or if users experience significant pressure—either from the platform or societal forces—to choose a particular option. Therefore, the badge system must clearly outline the pros and cons of opting in or out, ensuring that users' decisions are not unduly coerced. In this sense, the autonomy of users is further increased by the information strategies and the layered acquisition approach, both outlined above.

Maintaining neutrality of opinion

Similar to Article 10 ECHR, Article 11 CFR protects all opinions and statements of fact, regardless of their content, quality, or stylistic form (ECJ, 2001, para. 39; Jarass, 2021, Art. 11 para. 11). A key component of freedom of expression is the neutrality of opinion by state and EU institutions, as it supports undistorted public discourse and individuals' free character development. While platform operators are likely not bound by the same strict neutrality as state actors, certain implementations of the badge system could challenge this principle if not carefully designed. In this respect, the legal assessment must distinguish between the dissemination of disinformation and misinformation and the obligation to engage in deliberative communication behaviour.

The ECJ has yet to significantly clarify its interpretation of freedom of expression, often relying on the case law of the ECtHR for guidance. As described above, the ECtHR distinguishes between value judgements and factual assertions in its assessments of fundamental rights. When factual statements are deemed false, the conflicting legal interests, such as the personal rights of those affected, typically take precedence over freedom of expression. In other words, European fundamental rights do not afford a freedom of facts equivalent to the freedom of opinion. Further differentiation can be made that intentional disinformation carries less weight than unconscious misinformation in terms of fundamental rights. Conceptually, a badge system should take this into account.

It is at least as difficult to evaluate deliberative communication obligations. As a general guideline, badges pose significant challenges to freedom of expression when they are linked to specific opinions. While aggressive or derogatory styles can be addressed quite clearly through the badges, especially when they aim to safeguard the personal rights of third parties and uphold other legal interests, the freedom of expression under Art. 11 CFR protects in principle the linguistic style and (non-)deliberative manner of participation in the discourse. There tends to be friction here, which a badge design that respects autonomy should compensate for.

Equality of communication opportunities

The concepts of democracy and anti-discrimination are intimately tied to the notion that citizens who wish to participate in public discourse should have comparable opportunities to express themselves and engage in debates. In German legal thought, this principle is referred to as "communicative equality of opportunity" (Hoffmann-Riem, 1990; Schulz, 1998); however, it has not yet been fully explored by the ECtHR or the ECJ. Within the platform economy and broader commercial contexts, this principle is further enriched by the idea that large companies offering goods and services to a diverse user base should treat all users equally, regardless of their identity—an idea that also characterises the law of ToS.

A badge designed to enhance the reach or visibility of individual users can significantly improve their acceptance and effectiveness in public discourse. On the face of it, this creates unequal treatment between users who have a badge and those who do not. However, it should be noted that the badge system provides equal opportunities in that it is generally open to all users. It aligns with the overall logic of online platforms, where users who demonstrate high engagement and spend considerable time on the service are often recognised as power users, accumulating followers and gaining increased visibility through various means. As long as all users have the option to activate the badge, they can choose to enhance their platform privileges similarly, maintaining a level of agency in the process. Furthermore, the option for users to complain about platform decisions, as well as the possibility of recovering lost badges, has a positive effect on equal opportunities. Ultimately, the badge system is less problematic when it targets the tone or style of statements instead of their substantive content. While it is admittedly challenging to draw a sharp line between style and content – given that style frequently communicates or amplifies content, and content can be expressed through stylistic choices – it remains legally viable in most cases to classify statements as primarily concerning one of both.

Proportionality considerations with regard to individual design options

The assessment of a badge system's alignment with fundamental rights hinges on its specific design and functionality. Nonetheless, several key issues emerge: When balancing legal interests, it is crucial to recognise that badges can protect both individual legal interests—such as personal rights, honour, human dignity, religion, physical health, and the exercise of freedom without deterrent effect—and the public discourse essential for democracy. Moreover, the badge system represents a comparatively 'soft' intervention, as the platform's functionality is partially determined by user choice. This approach helps avoid more intrusive 'hard' measures

that could more directly infringe on fundamental rights.

The more accessible the choice between the badge system and the 'normal' communication mode is for users, along with the speed at which they can regain privileges after breaching badge rules, the more favourable the system is for autonomy. However, it is essential to recognise that a badge system designed to maximise autonomy might be less effective in protecting the quality of discourse and the legal interests of third parties compared to a more restrictive approach. In this regard, it appears that there is a qualitative correlation between rigidity and effectiveness, while a system that preserves autonomy could potentially engage a broader user base quantitatively.

In summary, platform operators must navigate several fundamental rights challenges when implementing a badge system, though these challenges are not insurmountable. The inherent tension between autonomy and effectiveness, as well as the balance between individual freedom and the protection of minorities, is a common theme in discussions of fundamental rights and is also evident in the context of this badge system.

Conclusion

In platform governance, the unique dynamics of digital environments and the complex interplay of normative and technological factors require innovative approaches to influencing both platform and user behaviour for the public good (Gorwa, 2019). Voluntary user badges for civic communication could serve as such an innovative as well as effective approach, as we have outlined.

The proposed measure has the potential to address the structural biases of digital platforms and create an opportunity space for civic discourse and incentivise user participation. This approach could enhance the visibility of deliberative communication while, depending on the badge's implementation, potentially reducing the prominence of disinformation and systematically conflict-oriented and emotionally charged discourse. Challenges such as the potential negation of conflict, abuse, and biases in the context of the badge implementation and abuse by problematic actors can be limited.

Legally, the measure aligns well with the concept of risk mitigation outlined in the DSA, as the anticipated discursive improvements address many of the systemic risks specified in Art. 34 DSA. This is achieved through a blend of established content moderation practices, which can be mapped to the measures listed in Art. 35

DSA. In addition, the measure can be designed in a way that respects fundamental rights while maintaining effectiveness. Although there are fundamental rights challenges to consider—such as ensuring neutrality of opinion—these challenges are manageable and not insurmountable.

In addition to the discursive improvements and legal alignment with DSA risk mitigation measures and fundamental rights, the proposed badges for civic communication could also benefit VLOPs. It would be too short-sighted to interpret them as a simple prevention of the natural but potentially harmful business model of online platforms. The badges offer comparatively easy-to-implement and relatively cost-effective options for systemic risk mitigation, as they build on existing content moderation practices widely utilised on most VLOPs. No completely new control system would have to be created, but the regular ToS-based system would have to be differentiated. Its implementation should not be compared with the status prior to the DSA but rather with the measures that would have to be taken in its absence to counter systemic risks in a manner that satisfies Art. 34 and 35 of the DSA. Once implemented, the badges are expected to foster a more positive communication climate, potentially leading to reduced content moderation costs and increased attractiveness in the advertising market.

It is important to acknowledge that the badge's impact remains speculative until it is implemented and systematically assessed. Its effectiveness will likely depend on specific design choices, such as the required norms, the extent of the visibility boost, the platform type, its user community, and the political and media environment in each country. Future research should explore these variables, either through empirical evaluations of actual implementations in digital platform contexts or through simulation and agent-based modelling approaches.

Overall, we believe that the proposed badge could serve as a starting point to disrupt the current distortive attention economy dynamics on digital platforms. By introducing a new logic for attention distribution, it holds the potential to significantly address systemic risks associated with these platforms.

ACKNOWLEDGEMENTS

We sincerely thank Ahrabhi Kathirgamalingam, Thivitha Himmen, Vincent Hofmann, Mike Karst, Cornelius Puschmann, Jan-Hinrik Schmidt, and Patrick Zerrer for their valuable feedback and insightful discussions throughout the development of this work. Their guidance significantly strengthened our research and

manuscript. We also extend our appreciation to the additional colleagues and friends whose input and support contributed to this project.

References

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) is misinformation a problem? *Perspectives on Psychological Science*, *18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, *120*(41), e2311627120. https://doi.org/10.1073/pnas.2311627120

Augsberg, I., & Petras, M. (2022). Deplatforming als Grundrechtsproblem—Die Sperrung durch soziale Netzwerke. *Juristische Ausbildung*, *62*(2), 97–108.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bail, C. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press. https://doi.org/10.1515/9780691216508

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Behrendt, M., Wagner, S. S., Ziegele, M., Wilms, L., Stoll, A., Heinbach, D., & Harmeling, S. (2024). *AQUA - Combining experts' and non-experts' views to assess deliberation quality in online discussions using LLMs* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2404.02761

Beyerbach, H. (2024). Art. 34 DSA. In R. Müller-Terpitz & M. R. Köhler (Eds), *Digital Services Act: Gesetz über digitale Dienste: Kommentar*. C.H. Beck.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947–960. https://doi.org/10.1016/j.tics.2023.06.008

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*(10), 1802–1813. https://doi.org/10.1037/xge0000532

Callamard, A. (2019). The human rights obligations of non-state actors. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 191–226). MIT Press.

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 2056305120936636. https://doi.org/10.1177/2056305120936636

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society*, *20*(11), 4255–4274. https://doi.org/10.1177/1461444818769689

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–22. https://doi.org/10.1145/3134666

Dahlberg, L. (2001). The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society*, *4*(4), 615–633. https://doi.org/10.1080/13691180110097030

Dobson, A. (2014). *Listening for democracy: Recognition, representation, reconciliation* (First). Oxford University Press.

Douek, E. (2021). The limits of international law in content moderation. *UC Irvine Journal of International, Transnational, and Comparative Law*, *6*(37), 37–65.

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., Van Der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2025). Why misinformation must not be ignored. *American Psychologist*, *80*(6), 867–878. https://doi.org/10.1037/amp0001448

Efroni, Z. (2021). The Digital Services Act: Risk-based regulation of online platforms. *Internet Policy Review*. https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606

Ehlers, D., & Germelmann, C. F. (Eds). (2023). *Europäische Grundrechte und Grundfreiheiten* (5th edn). De Gruyter.

Esau, K., Fleuß, D., & Nienhaus, S. (2021). Different arenas, different deliberative quality? Using a systemic framework to evaluate online deliberation on immigration policy in Germany. *Policy & Internet*, *13*(1), 86–112. https://doi.org/10.1002/poi3.232

Esau, K., Wilms, L., Baleis, J., & Keller, B. (2023). For deliberation sake, show some constructive emotion! How different types of emotions affect the deliberative quality of interactive user comments. *Javnost - The Public*, *30*(4), 472–495. https://doi.org/10.1080/13183222.2023.2171217

European Commission. (2024). *Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act*. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2373

European Commission. (2025). *The code of conduct on disinformation*. https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

European Court of Human Rights. (2008). *Decision no. 36109/03, 10.2.2008—Leroy/France*.

European Court of Human Rights. (2009). *Decision no. 17089/03, 23 June 2009—Sorguç v. Turkey*.

European Court of Human Rights. (2010). *Decision no. 1620/03, 9.23.2010—Schlüth* (Issue ion no.

1620/03, p. 9 23 2010-).

European Court of Human Rights. (2021). *Decision no. 36537/15 and 36539/15, 9 March 2021—Benitez Moriana et al. V. Spain*.

European Court of Justice. (2001). *Judgment C-274/99, 3.6.2001—Connolly*.

Fertmann, M. (2025). Art. 20 DSA. In T. Mast, M. C. Kettemann, S. Dreyer, & W. Schulz (Eds), *Digital Services Act / Digital Markets Act (DSA / DMA* (p. 1602). C.H. Beck.

Fielitz, M., & Schwarz, K. (2023). *IDZ Jena: Hate not Found. Das Deplatforming der extremen Rechten*. https://www.idz-jena.de/forschung/hate-not-found-das-deplatforming-der-extremen-rechten

Franck, G. (1998). *Ökonomie der Aufmerksamkeit: Ein Entwurf*. Hanser.

Freistein, Katja, Gadinger, Frank, & Unrau, Christine. (2022). It just feels right. Visuality and emotion norms in right-wing populist storytelling. *International Political Sociology*, *16*(4), olac017. https://doi.org/10.1093/ips/olac017

Frenkel, S., & Maheshwari, S. (2018). *Facebook to let users rank credibility of news*. The New York Times. https://www.nytimes.com/2018/01/19/technology/facebook-news-feed.html

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339. https://doi.org/10.1002/poi3.95

Funk, A., Vesteinsson, K., Baker, G., Brody, J., Grothe, C., Agarwal, A., Barak, M., Loldj, M., Masinsin, M., & Sutterlin, E. (2024). *Freedom on the Net (No. 2024; Freedom on the Net)*. Freedom House. https://freedomhouse.org/sites/default/files/2024-10/FREEDOM-ON-THE-NET-2024-DIGITAL-BOOKLET.pdf

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341. https://doi.org/10.1080/13504630.2015.1128810

Gelovani, S., Müller, P., Meyer, H., Stoll, A., Pröschel, L., Ziegele, M., & Wesseler, H. (2025). Enhancing democratic listening in online discussions: The role of reaction buttons. *Discussion Moderation, and Large Language Models. 75th Annual Conference of the International Communication Association (ICA*.

German Federal States. (2024). *Interstate Media Treaty (Medienstaatsvertrag)*. https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/Interstate_Media_Treaty_en.pdf

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, *28*(1). https://doi.org/10.2139/ssrn.3810580

Google. (2025a). *Community Guidelines strike basics on YouTube*. Google Support. https://support.google.com/youtube/answer/2802032

Google. (2025b). *Verification badges on channels—YouTube Help*. YouTube Help. https://support.google.com/youtube/answer/3046484

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Griffin, R. (2023). The law and political economy of online visibility: Market justice in the Digital Services Act. *Technology and Regulation*, *2023*, 69–79. https://doi.org/10.71265/qqcy5853

Grüning, D. J., Kamin, J., Panizza, F., Katsaros, M., & Lorenz-Spreen, P. (2024). A framework for promoting online prosocial behavior via digital interventions. *Communications Psychology*, *2*(1), 6. https://doi.org/10.1038/s44271-023-00052-7

Hägle, O., Escher, S., Heil, R., & Jahnel, J. (2025). Structuring different manifestations of misinformation for better policy development using a decision tree-based approach. *Policy & Internet*, *17*(2), e420. https://doi.org/10.1002/poi3.420

Haman, M., & Školník, M. (2025). The unverified era: Politicians' Twitter verification post-Musk acquisition. *Journal of Information Technology & Politics*, *22*(2), 167–171. https://doi.org/10.1080/19331681.2023.2293868

Hao, K. (2021). *The Facebook whistleblower says its algorithms are dangerous. Here's why*. MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Heiss, R., & Matthes, J. (2020). Stuck in a nativist spiral: Content, selection, and effects of right-wing populists' communication on Facebook. *Political Communication*, *37*(3), 303–328. https://doi.org/10.1080/10584609.2019.1661890

Herzog, L. (2023). *Citizen knowledge: Markets, experts, and the infrastructure of democracy* (1st edn). Oxford University PressNew York. https://doi.org/10.1093/oso/9780197681718.001.0001

Hoewe, J., Brownell, K. C., & Wiemer, E. C. (2021). The role and impact of Fox News. *The Forum*, *18*(3), 367–388. https://doi.org/10.1515/for-2020-2014

Hoffmann-Riem, W. (1990). Kommunikationsfreiheit und Chancengleichheit. In J. Schwartländer, E. H. Riedel, & W. Brugger (Eds), *Neue Medien und Meinungsfreiheit im nationalen und internationalen Kontext: Interdisziplinäre Kolloquien, Tübingen und Marburg 1985 bis 1988* (pp. 27–58). N.P. Engel.

Horwitz, J. (2021). *The Facebook files*. Wall Street Journal. https://www.wsj.com/articles/the-facebook-files-11631713039

Husovec, M. (2024). *Principles of the Digital Services Act* (1st edn). Oxford University Press. https://doi.org/10.1093/law-ocl/9780192882455.001.0001

Jarass, H. D. (2021). *Charta der Grundrechte der Europäischen Union: Unter Einbeziehung der sonstigen Grundrechtsregelungen des Primärrechts und der EMRK*. C.H. Beck.

Jungherr, A., Rivero, G., & Gayo-Avello, D. (2020). *Retooling politics: How digital media are shaping democracy* (1st edn). Cambridge University Press. https://doi.org/10.1017/9781108297820

Jungherr, A., & Schroeder, R. (2021). Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media + Society*, *7*(1), 2056305121988928. https://doi.org/10.1177/2056305121988928

Jungherr, A., Schroeder, R., & Stier, S. (2019). Digital media and the surge of political outsiders: Explaining the success of political challengers in the United States, Germany, and China. *Social Media + Society*, *5*(3), 2056305119875439. https://doi.org/10.1177/2056305119875439

Kaiser, J., & Rauchfleisch, A. (2020). Birds of a feather get recommended together: Algorithmic homophily in YouTube's channel recommendations in the United States and Germany. *Social Media + Society*, *6*(4), 2056305120969914. https://doi.org/10.1177/2056305120969914

Kalathil, S. (2020). The evolution of authoritarian digital influence: Grappling with the new normal. *PRISM*, *9*(1), 33–50.

Kathirgamalingam, A., Lind, F., Bernhard, J., & Boomgaarden, H. G. (2024). *Agree to disagree? Human and LLM coder bias for constructs of marginalization*. SocArXiv. https://doi.org/10.31235/osf.io/agpyr

Kettemann, M. C., & Schulz, W. (2023). *Platform://Democracy: Perspectives on platform power, public values and the potential of social media councils*. https://doi.org/10.21241/SSOAR.86524

Kim, J., McDonald, C., Meosky, P., Katsaros, M., & Tyler, T. (2022). Promoting online civility through platform architecture. *Journal of Online Trust and Safety*, *1*(4). https://doi.org/10.54501/jots.v1i4.54

Kim, J. Y. (2023). *Machines do not decide hate speech: Machine learning, power, and the intersectional approach* (C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe, Eds). Freie Universität Berlin. https://doi.org/10.48541/DCR.V12.21

Kleinfeld, R., & Sedaca, N. B. (2024). How to prevent political violence. *Journal of Democracy*.

Klinger, U., & Svensson, J. (2015). The emergence of network media logic in political communication: A theoretical approach. *New Media & Society*, *17*(8), 1241–1257. https://doi.org/10.1177/1461444814522952

Kreiss, D., & McGregor, S. C. (2024). A review and provocation: On polarization and platforms. *New Media & Society*, *26*(1), 556–579. https://doi.org/10.1177/14614448231161880

Levitsky, S., & Ziblatt, D. (2018). *How democracies die* (First). Crown.

Lewis, B. (2018). *Alternative Influence*. Data & Society. https://datasociety.net/library/alternative-influence/

LinkedIn Corp. (2025). *Verifications on your LinkedIn profile*. LinkedIn Help.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. https://doi.org/10.1038/s41562-022-01460-1

Lüdemann, J. (2019). Grundrechtliche Vorgaben für die Löschung von Beiträgen in sozialen Netzwerken—Private Ordnung digitaler Kommunikation unter dem Grundgesetz. *MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung*, *22*(5), 279–284.

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, *49*(2), 171–195. https://doi.org/10.1177/0093650220921321

Markard, N., & Bredler, E. M. (2021). Grundrechtsdogmatik der Beleidigungsdelikte im digitalen Raum. *JuristenZeitung*, *76*(18), 864–872. https://doi.org/10.1628/jz-2021-0274

Mast, T. (2023). AGB-Recht als Regulierungsrecht. *JuristenZeitung*, *78*(7), 287–296. https://doi.org/10.1628/jz-2023-0096

Mast, T., & Ollig, C. (2023). The lazy legislature: Incorporating and horizontalising the charter of fundamental rights through secondary union law. *European Constitutional Law Review*, *19*(3), 462–486. https://doi.org/10.1017/S1574019623000238

Mau, S., Westheuser, L., & Lux, T. (2024). *Triggerpunkte: Konsens und Konflikt in der Gegenwartsgesellschaft* (7. Auflage). Suhrkamp.

McKelvey, F., & Hunt, R. (2019). Discoverability: Toward a definition of content discovery through platforms. *Social Media + Society*, *5*(1), 2056305118819188. https://doi.org/10.1177/2056305118819188

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991–996. https://doi.org/10.1126/science.adl2829

Meta. (2018). *Helping ensure news on Facebook is from trusted sources*. About Facebook. https://about.fb.com/news/2018/01/trusted-sources/

Meta. (2020). *Prioritizing original news reporting on Facebook*. About Facebook. https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/

Meta. (2021). *Reducing political content in news feed*. About Facebook. https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/

Meta. (2023). *An update on Facebook news in Europe*. About Facebook. https://about.fb.com/news/2023/09/an-update-on-facebook-news-in-europe/

Meta. (2025a). *Request a verified badge on Facebook*. Facebook Help Center. https://www.facebook.com/help/1288173394636262/

Meta. (2025b). *Transparency Center: Counting strikes*. Meta Transparency. https://transparency.meta.com/enforcement/taking-action/counting-strikes/

Meta. (2025c). *Verified badges on Instagram*. Instagram Help Center. https://help.instagram.com/733907830039577

Miller-Idriss, C. (2020). *Hate in the homeland: The new global far right*. Princeton University Press.

Möller, J., Hameleers, M., & Ferreau, F. (2020). *Typen von Desinformation und Misinformation. Die medienanstalten*. ALM GbR.

Mouffe, C. (2000). *The democratic paradox*. Verso.

Mouffe, C., & Laclau, E. (1985). *Hegemony and socialist strategy: Towards a radical democratic politics*. Verso.

Oittinen, E., & Molokach, B. (2023). *"Vicious, hateful, and divisive" partisans: Understanding and countering antidemocratic political polarization*. MediaWell. https://mediawell.ssrc.org/?post_type=ramp_review&p=78174

Park, J., & Singh, V. K. (2022). How background images impact online incivility. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–23. https://doi.org/10.1145/3555545

Pinterest. (2024). *Digital Services Act risk assessment and mitigation report 2024*. Pinterest Help. https://help.pinterest.com/sites/pinhelp/files/dsa/2024-Pinterest-DSA-RIsk-Assessment-and-Mitigation-Report.pdf

Pinterest. (2025). *Enforcement*. Pinterest Policy. https://policy.pinterest.com/en/enforcement

Presserat (German Press Council). (2017). *German Press Code*. https://www.presserat.de/files/presserat/dokumente/download/Press%20Code.pdf

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). *Auditing radicalization pathways on YouTube* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1908.08313

Saurwein, F., & Spencer-Smith, C. (2020). Combating disinformation on social media: Multilevel governance and distributed accountability in Europe. *Digital Journalism*, *8*(6), 820–841. https://doi.org/10.1080/21670811.2020.1765401

Schroeder, R. (2018). *Social theory after the Internet: Media, technology, and globalization*. UCL Press. https://doi.org/10.2307/j.ctt20krxdr

Schulz, W. (1998). *Gewährleistung kommunikativer Chancengleichheit als Freiheitsverwirklichung* (1.). Nomos-Verl.-Ges.

Schulz, W., & Ollig, C. (2023). Hybrid speech governance—New approaches to govern social media platforms under the European Digital Services Act? *Journal of Intellectual Property, Information Technology, and Electronic Commerce Law*, *14*(4). https://www.jipitec.eu/jipitec/article/view/22/26

Scudder, M. F. (2022). Measuring democratic listening: A listening quality index. *Political Research Quarterly*, *75*(1), 175–187. https://doi.org/10.1177/1065912921989449

Scudder, M. F., & White, S. K. (2023). *The two faces of democracy: Decentering agonism and deliberation*. Oxford University Press.

Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-127

Snap Inc. (2025). *Snapchat Moderation, enforcement, and appeals | Community Guidelines Explainer*. Snap Values. https://values.snap.com/privacy/transparency/community-guidelines/moderation?lang=en-US

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, *1*(1), 21–48.

Stewart, E. (2021). Detecting fake news: Two problems for content moderation. *Philosophy & Technology*, *34*(4), 923–940. https://doi.org/10.1007/s13347-021-00442-x

Tangwaragorn, P., Khern-am-nuai, W., & Kar, W. (2025). The implications of account suspensions on online discussion platforms. *Decision Support Systems*, *189*, 114389. https://doi.org/10.1016/j.dss.2024.114389

Tiedeke, A. S., & Fertmann, M. (2023). A love triangle? Mapping interactions between international human rights institutions, Meta and its oversight board. *European Journal of International Law*, *34*(4), 907–938. https://doi.org/10.1093/ejil/chad062

TikTok. (2025a). *Content violations and bans*. Tiktok Support. https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans

TikTok. (2025b). *Verified accounts on TikTok*. Tiktok Support. https://support.tiktok.com/en/using-tiktok/growing-your-audience/how-to-tell-if-an-account-is-verified-on-tiktok

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

Truong, M., White, J., & Funk, A. (2018). *The rise of digital authoritarianism* (No. No. 2018; Freedom on the Net). Freedom House. https://freedomhouse.org/sites/default/files/2020-02/10192018_FOTN_2018_Final_Booklet.pdf

Twitch Interactive. (n.d.). *About account enforcements and chat bans*. Twitch Help. https://help.twitch.tv/s/article/about-account-suspensions-dmca-suspensions-and-chat-bans

Völzmann, B. (2021). Freiheit und Grenzen digitaler Kommunikation—Digitale Gewalt als Herausforderung der bisherigen Meinungsfreiheitsdogmatik. *MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung*, *24*(8), 619–624.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Webster, J. G. (2014). *The marketplace of attention: How audiences take shape in a digital age*. MIT Press.

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1565

Wielsch, D. (2019). Private law regulation of digital intermediaries. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3369592

X Corp. (2025). *About profile labels and checkmarks on X*. X Help. https://help.x.com/en/rules-and-policies/profile-labels

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

Yildirim, M. M., Nagler, J., Bonneau, R., & Tucker, J. A. (2023). Short of suspension: How suspension warnings can reduce hate speech on Twitter. *Perspectives on Politics*, *21*(2), 651–663. https://doi.org/10.1017/S1537592721002589

York, J. C., & Zuckmann, E. (2019). Moderating the public sphere. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 137–162). MIT Press.

Zeng, J., & Brennen, S. B. (2023). Misinformation. *Internet Policy Review*, *12*(4). https://doi.org/10.14763/2023.4.1725

Published by

Alexander von Humboldt Institute for Internet and Society

CAIS Research for the Digital Age

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies