Individual Benefit – Collective Harm. The Logic of the Prisoner's Dilemma in Digital Discourse

DOI : 10.46522/S.2025.02.3

József Zoltán MÁLIK PhD
Institute of Social Sciences and International Studies, Budapest Metropolitan University
jmalik@metropolitan.hu

Abstract: *The rapid expansion of social media platforms has profoundly transformed the structure of the public sphere and the dynamics of social discourse. The digital space initially carried the promise of democratizing the expression of opinion; however, the algorithmic logic of platforms—operating according to the rules of the attention economy—has a distorting effect on public debate. Emotionally charged, oversimplified, and polarizing content tends to gain greater visibility, thereby encouraging the instrumental application of individual communication strategies. This mechanism can be described by the logic of the game-theoretic Prisoner's Dilemma: decisions that are rational at the individual level lead to negative consequences at the social level. As a result, the quality of discourse deteriorates, the boundaries of a commonly accepted reality become blurred, and the basic conditions for deliberative democracy are undermined. This study aims to interpret this paradoxical situation in the digital public sphere and to explore what types of intervention might help reverse it. Through a theoretical-analytical approach grounded in interdisciplinary scholarship, it identifies four main areas where constructive change may occur: (1) transparent and proportionate institutional regulation, (2) algorithmic mechanisms that foster content diversity and deliberation, (3) the promotion of critical media literacy and digital ethics in education, and (4) the cultivation of community (social) norms that support respectful dialogue. Rather than offering empirical predictions, the study proposes a conceptual model that reflects and anticipates evolving efforts to re-establish digital environments as spaces of shared understanding and democratic engagement.*

Key words *prisoner's dilemma; algorithms; attention economy; public sphere; digital discourse; collective harm.*

1. Introduction

The emergence of social media and online platforms has fundamentally transformed the structure of the public sphere. The mass communication model dominant in the twentieth century was characterized by editorial offices in the role of gatekeepers, professional content production, controlled distribution, and linear, one-way communication. In contrast, the digital media space of the twenty-first century has foregrounded decentralization and interactivity. The democratization of opinion expression and content creation has enabled broad segments of society to directly shape public discourse. However, this has come at a price. The decentralization of the digital public sphere has gone hand in hand with the fragmentation of common spaces for debate. Media that previously enabled relatively unified social dialogue have been replaced by platforms where users communicate within algorithmically regulated bubbles. These "personalized" media spaces often reinforce existing prejudices and prevent encounters with opposing views (Pariser 2011). The resulting echo chamber effect not only narrows individual worldviews but also undermines social cohesion by eliminating the possibility of dialogue about a shared reality (Sunstein 2001).

The discourse on social media is characterized by its highly visual, fast-paced, and often emotional nature. According to the logic of the attention economy, content is effective if it elicits an immediate reaction; thus, simplified, symbolic messages come to the fore. For example, TikTok offers a format that conveys even political messages in a highly fragmented, visual, and often ironic manner. Instead of unfolding argumentative debates, users respond to each other's opinions in short, snappy videos, often with humor, sarcasm, or strong visual symbols. A similar trend can be observed on Twitter, where the character limit not only provides a formal framework but also pushes discourse toward short, often moralizing, rhetorically exaggerated statements. These media spaces have also transformed the way information is consumed. Netflix's algorithms, for example, offer personalized recommendations, causing users to become acquainted with the same topic—such as climate change or migration—from entirely different perspectives. The platform not only entertains but, by thematizing and contextualizing social issues, actively shapes the viewer's worldview. Meanwhile, reception often functions as an enclave, resulting in parallel realities instead of debates (Benkler, Robert and Hal 2018).

Thus, the democratic potential has not disappeared but faces new challenges. There is a paradox here: the freedom of access and expression does not, in itself, guarantee quality discourse. The structural transformation of the public sphere (fragmentation, algorithmic filtering, visual dominance, and the logic of entertainment) constitutes a fundamentally different arena of debate than the ideal assumed by classic deliberative democracy. In this sense, the digital space not only offers new channels for public communication but also introduces new logics that separate individual and collective interests. Algorithmic logic not only mediates but also structures: it determines the mode of information flow, priorities, and the rhythm of discourse. The filtering mecha-

nisms of platforms—such as Facebook's news feed ranking, YouTube's recommendation system, or TikTok's "For You Page"—are not neutral channels but systems of rules that implicitly assign value to content. As a result, it is not only apparent who says what, but also what we do not hear, because the algorithm does not allow it to reach us (Tufekci 2015; Gillespie 2018).

This study adopts a conceptual distinction between the "public sphere" in its broad, classical sense and the "digital public sphere" as a specific configuration of technologically mediated communication spaces. While the former includes institutionalized, regulated media such as television, radio, and the press, the latter is shaped primarily by algorithmically driven, commercially operated platforms that exert new forms of control over visibility, attention, and interaction. The digital public sphere is therefore not merely an extension of traditional democratic spaces—it is a qualitatively different environment that affects, and in many cases transforms, the broader ecology of public discourse. Importantly, its influence is not confined to online debates: the communicative patterns fostered by social media—emotional intensification, binary framing, and symbolic performance—often spill over into other domains, including broadcast media and face-to-face conversations. For this reason, understanding how the digital public sphere operates is not only essential for analyzing online communication but for grasping broader transformations in contemporary public culture.

Methodologically, this study follows a theoretical-analytical approach, interpreting behavioral and communicative patterns within the digital public sphere through an interpretive framework informed by academic discourse across multiple fields. The theoretical foundation of the investigation is the classic Prisoner's Dilemma model from Game Theory, which is conceptually adapted to describe the incentive structures that characterize distortions in the logic of social media discourse. This analytical thought experiment serves as a communicative framework to understand how individual content strategies are rationally rewarded in the short term, yet undermine the very conditions that make deliberative communication possible at the social level. Platform users typically do not perceive the structural consequences of their actions, as the temporal acceleration of the digital environment obscures long-term effects and feedback.

In addition to this systemic interpretation, the study applies a deductive-analytical lens to examine institutional, algorithmic, and pedagogical interventions that may facilitate the emergence of more constructive forms of discourse. The argument operates at the intersection of communication theory, political science, and digital social research, drawing on international scholarly literature and contemporary digital platform practices, which are treated not as empirical data but as conceptual and normative references.

2. The Prisoner's Dilemma and the Tragedy of the Commons: Strategic Behavior and Collective Loss in the Attention Economy

The Prisoner's Dilemma is one of the most emblematic models in game theory. It illustrates a situation in which two actors, acting in their own rational self-interest, ultimately produce an outcome that is worse for both than if they had cooperated. In the context of social media, the model helps to conceptualize how users—striving for visibility, engagement, and recognition—adopt content strategies that may seem effective individually but lead to the degradation of discourse on a collective level. Instead of promoting deliberation or nuanced discussion, platforms reward content that triggers strong emotional responses, polarization, or simplistic moral positioning. Users on social media typically do not act with the intention of undermining discourse. However, within the logic of algorithmic filtering, those who optimize their content for visibility by sharing provocative, emotionally charged, or sensational posts are disproportionately rewarded. The platform amplifies such content because it drives engagement—clicks, likes, shares, and comments—which, in turn, fuels advertising revenue. This creates a competitive environment in which even those who value thoughtful discussion feel compelled to adopt similar strategies or risk becoming invisible. The result is a race to the bottom in content quality and a gradual erosion of the conditions necessary for constructive public discourse.

Yet while the Prisoner's Dilemma effectively captures the dilemma of strategic individual behavior versus collective harm, it is equally useful—if not more so—to approach this phenomenon through the lens of the Tragedy of the Commons. Garrett Hardin's classic model (1968) describes a scenario in which multiple individuals, acting independently and rationally according to their own self-interest, deplete a shared but limited resource, thereby causing harm to the entire group. While each individual's action may be justified in isolation, the cumulative effect is disastrous. In the digital public sphere, the "commons" is not a pasture or natural resource, but collective attention**,** emotional energy, and the possibility of meaningful discourse**.** These are finite and fragile assets. The competition for attention, intensified by algorithmic design, creates incentives for users to maximize personal exposure at the expense of discursive quality. As Shoshana Zuboff (2019) argues, platforms are not neutral intermediaries—they are commercial systems engineered to extract behavioral data and manipulate user behavior in the service of profit. This transformation turns users not only into content producers but also into *predictable behavioral patterns* subjected to monetization.

Building on Jon Elster's theory of social mechanisms (2007), this paper argues that the dynamics of the attention economy are not just isolated behavioral choices—they constitute a self-reinforcing social mechanism. That is, the structure of platform incentives reproduces and intensifies itself over time. As Evgeny Morozov (2013) notes, these are not accidental by-products of digital technology but consequences of a deeply embedded commercial logic. The more emotional, polarizing, or extreme the content, the more likely it is to be algorithmically amplified—thus reshaping the norms of

acceptable discourse. This trend is visible in how users often feel pressure to share personal experiences, express outrage, or dramatize opinions, not because they value these tactics, but because *not* doing so leads to invisibility. This creates a *feedback loop*: users respond to platform incentives by adapting their communication, which in turn reinforces the platform's logic. Even those critical of this system are not immune. As Benkler, Faris, and Roberts (2018) show, the architecture of digital media systematically favors divisive content, often marginalizing nuanced, fact-based information. The long-term consequence is not only a decline in content quality but also *a corrosion of social trust*, as users become desensitized, polarized, or disillusioned.

This erosion affects not only what is said but how and why it is said. When attention is the primary currency of public communication, the dynamics of discourse shift from collaborative truth-seeking to competitive visibility-seeking. In this environment, even well-intentioned actors may find themselves contributing to the degradation of public reason simply by participating according to the dominant logic of the system. Thus, the tragedy is not merely one of individual failure but of a structurally induced collective vulnerability.

In sum, the dynamics of the digital public sphere cannot be understood solely through the lens of strategic choice (as in the Prisoner's Dilemma) but require the broader sociological insight of the Tragedy of the Commons. Here, the problem is not simply one of coordination failure, but of systemic overuse and depletion of collective cognitive and emotional resources. Addressing this dilemma requires interventions that go beyond individual behavior change, involving structural reform of platform incentives and redefinition of what counts as valuable participation in the digital public sphere.

3. Platform Logics and the Business Model of Engagement

The structural distortions of the digital public sphere are not the result of mere technical flaws, but symptoms of a deeply rooted business logic. Platform design is not value-neutral; it is oriented toward maximizing user engagement, which functions as a proxy for monetizable attention. The longer users stay on the platform, the more advertisements they see, and the more data the platform collects—this is the foundational equation of surveillance capitalism (Zuboff 2019). As a result, platform architecture is systematically optimized not for the quality of discourse, but for its capacity to capture and hold attention.

From this perspective, the algorithm is not simply a content sorter—it is an economic instrument. It is designed to promote content that triggers strong emotional responses, fuels interaction, and keeps users scrolling. Cathy O'Neil (2016) emphasizes that such algorithmic systems often reproduce and amplify existing biases, creating self-reinforcing patterns of distortion. The consequences are both epistemic and social: content is selected not for accuracy or social value, but for virality.

Crucially, this logic is not imposed despite user behavior but co-evolves with it. Platforms adapt to users' preferences and behaviors, while users adapt their communication strategies to platform incentives. This mutual adaptation creates what might be called a second-order Prisoner's Dilemma, in which not only users but platforms themselves are caught in rational strategies that produce collective harm. From the platform's perspective, maximizing engagement is a rational business decision. From the user's perspective, aligning with emotionally intense, polarizing content is often the most effective visibility strategy. The damage—to public discourse, trust, and democratic dialogue—is a shared but unowned cost.

Basic Mechanisms of Algorithmic Operation

Algorithms do not simply "select" content, but fundamentally shape the structure of digital discourse. The main types include:

- **Ranking algorithms:** These determine the order in which posts, videos, and news appear to users. Facebook and Instagram, for example, use weighted variables: a post from a friend that has received many reactions will appear higher than less interactive content from an unknown source (Bakshy, Solomon and Lada 2015).
- **Recommendation systems:** YouTube, Netflix, and TikTok recommendations are based on machine learning models that "predict," based on a user's prior interactions, what will maintain their interest for the longest time. This predictive logic is prone to gradual radicalization (Ribeiro et al. 2020).
- **Content curation and trending mechanisms:** Twitter's "Trending Topics" algorithm not only shows the most frequently mentioned topics but also weights them based on the user's location, interests, and behavior. This can contribute to the development of so-called "rapid response panic-button discourses" (Matias 2019).

These algorithms are not transparent to users. Their operation is not public, and in many cases, even the platforms themselves do not fully understand all the effects of their AI-based systems. This technological unaccountability is one of the main obstacles to the democratization of digital discourse (Gillespie 2018).

Distortive Effects of Algorithms

Algorithmic operation is not inherently "malicious," but its structural effects deeply influence how we think and debate collectively. The main distortive effects include:

- **Confirmation bias:** Recommendation systems tend to favor content that reinforces users' existing views, as these are more likely to elicit positive interactions. This strengthens the formation of opinion bubbles and reduces the chance of encountering differing perspectives (Pariser 2011).
- **Polarization and radicalization:** Algorithms on platforms such as YouTube or TikTok tend to promote more extreme content that elicits stronger reactions. This

creates structural incentives for sensationalist, simplistic, "clickbait" narratives (Tufekci, 2015; Ribeiro et al., 2020).

- **Emotional inflation:** Since algorithms primarily reward interactions, content that triggers strong emotional responses (anger, fear, outrage) moves to the forefront. This leads, in the long term, to an exhausting, emotionally charged discourse devoid of rational debate (Guess et al. 2019).
- **Content distortion and informational asymmetry:** Because platforms do not disclose how their algorithms work, users cannot see the criteria by which they receive certain information. This can result in informational inequality and indirect manipulation (Napoli 2019).

The transparency and accountability of algorithms are fundamental democratic issues. In the current situation, most users do not know the criteria by which content is ranked, why they see what they see, or why certain posts disappear from their feed. The European Union's Digital Services Act (DSA), effective from 2024, has introduced legal instruments to make algorithmic decision-making more transparent (Klein 2023; Klein 2024). Large platforms must provide researchers with access to their operations and will also have audit obligations (European Commission 2022). These initiatives are first steps toward algorithmic public policy, but further norms and institutions are necessary for real societal control to develop. The concept of algorithmic accountability is not only a technical matter but also an ethical and political one. The quality of the public sphere is closely linked to how visibility is generated on platforms. The questions of "what is seen" and "what is not seen" are crucial for shaping the social agenda; therefore, these cannot remain solely at the level of business decisions.

Majority Illusion and Perceived Consensus

The perception of what constitutes a "popular" or "dominant" opinion is also distorted by the architecture of social media networks. The phenomenon known as the majority illusion (Lerman et al. 2016) describes how users may perceive a certain view as widely accepted simply because highly connected individuals promote it. Due to the unequal distribution of visibility and influence, fringe views may appear mainstream, while moderate positions remain invisible. This illusion is not just epistemic: it affects behavior, incentivizing users to align with seemingly dominant narratives even if they are not widely held. This compounds the cycle of amplification and reinforces polarizing or extreme discourses.

4. Regulation, Responsibility, and New Gatekeepers

The rise of social media platforms has fundamentally transformed the role of gatekeeping in the public sphere. Traditional gatekeepers—editors, journalists, and institutional media actors—once held the authority to filter, curate, and contextualize information before it reached the public. The digital public sphere, by contrast, is often framed as post-gatekeeping: a decentralized space where users can bypass institutional control and speak directly to audiences. This development was initially

seen as a democratizing force, allowing for more diverse voices and perspectives[1]. However, the absence of traditional gatekeepers has not eliminated gatekeeping itself; it has merely transformed it.

First of all, traditional gatekeepers and their associates (i.e., journalists), in their role as content editors—due in no small part to their institutionally prominent function within democratic public discourse—benefit from a distinct level of fundamental rights protection. This protection is accompanied by heightened legal responsibility under press law. In contrast, new gatekeepers enjoy only the blessings of the right to freedom of expression. However, lacking actual content production and editorial activity, they do not share in the editorial liability imposed on media content providers (cf. Klein 2016; Klein 2020). In the digital environment, platforms themselves have become the new gatekeepers, not by controlling the content that can be produced, but by determining what becomes visible, amplified, or marginalized. This algorithmic gatekeeping is quantitative rather than qualitative: it operates through ranking, recommendation, and amplification metrics, not editorial judgment. As a result, content is filtered not based on its accuracy or relevance, but on its capacity to generate engagement. The question is no longer only *who speaks*, but *who is heard*, and under what conditions.

This shift raises fundamental questions about responsibility and accountability. Unlike traditional media institutions, which are subject to professional norms and public scrutiny, platforms often claim neutrality. Yet their design choices, ranking algorithms, moderation policies, and interface structures directly shape public discourse. They determine which voices are amplified, which topics trend, and which perspectives remain invisible. In this sense, platforms are infrastructural actors in the public sphere, wielding power not through content creation but through content circulation. The central paradox of the digital public sphere is that more speech does not necessarily produce better discourse. In classical liberal theory, freedom of expression is a foundational value, essential to individual autonomy and democratic debate. Yet in the context of algorithmic amplification, unrestricted speech can produce systemic distortions: noise overwhelms signal, emotional extremity drowns out nuance, and polarizing narratives dominate over deliberative engagement.

This leads to what some scholars describe as the *"paradox of tolerance"*: a situation in which the unregulated proliferation of speech undermines the very conditions necessary for democratic dialogue. In extreme cases, disinformation, hate speech, and coordinated manipulation campaigns exploit platform logics to gain visibility, disrupt public trust, and silence vulnerable voices. The normative dilemma is clear: how can societies protect the freedom to speak without enabling the degradation of public reason? Furthermore, the asymmetrical visibility created by algorithmic systems challenges the assumption that all voices have equal access to public attention. While anyone *can* speak, not all speech is heard. The structural features of platforms privilege emotionally provocative, high-

[1] The legal assessment of the new gatekeepers has been discussed by Klein (2020).

velocity content over slow, reasoned argument. This asymmetry creates a hierarchy of attention in which commercial logic trumps democratic deliberation.

Regulatory Responses and Options

In response to these challenges, regulatory initiatives have begun to emerge. The European Union's Digital Services Act (DSA), effective from 2024, marks a significant step toward institutional oversight of platform operations. The DSA imposes transparency obligations on large platforms, including requirements to explain how algorithms work, share data with independent researchers, and submit to external audits (European Commission 2022). These are essential foundations for algorithmic accountability—the principle that the mechanisms shaping public discourse should be subject to democratic oversight. However, legal regulation alone is insufficient. Platform architectures are highly dynamic, often opaque, and shaped by commercial incentives that evolve faster than legislation. This has led some scholars to advocate for "regulation by design"—technological frameworks that embed normative goals into platform functionality. Examples include algorithmic diversity mechanisms, friction-increasing interface features (e.g., prompts before sharing), or content labeling systems that encourage deliberation over reaction.

The European Union's regulation clearly responds to the recognition that traditional legal solutions can only be partially effective, and therefore, it applies a new regulatory concept to platforms. While it does rely on platforms' self-regulation, it intends to define the framework of such regulation itself. The DSA, which can be understood as the EU's general platform regulation, follows the *precautionary principle* and applies *ex ante rules* instead of traditional *ex post legal norms*. Rather than focusing on restoring or compensating for damage after it occurs, the regulation concentrates on preventing individual harm or minimizing risk. Klein refers to this regulatory concept as *"vaccine law"* (see Klein 2024a; Klein 2024b; Klein 2024c), contrasting it with the traditional legal model's *ex post facto* structure. Vaccine law, as Klein argues in several of his studies, does not seek to restore individual harm or even collective damage (to the legal order), but rather employs specific tools aimed at ensuring such harm does not occur in the first place. In terms of its effect mechanism, regulation thus works similarly to vaccination: its goal is to establish immunity and prevent the occurrence of harm (or disease).

A key feature of this form of regulation is the regulation of technology and regulation by technology. In the latter case, legal norms are embedded into algorithmic code, and it is this code—functioning like a vaccine—that assumes responsibility for minimizing the risk of harm. Another noteworthy element of the regulation is the *fundamental rights mechanism* introduced for cases of individual rights violations, which—according to Klein—may mark the beginning of a new era in fundamental rights protection (Klein 2023; Klein 2024c). The DSA contains provisions, both in its general clause on fundamental rights and in the risk mitigation measures required of very large platforms, that can be interpreted as fundamental rights tests applicable to legal relationships between private parties. In determining contractual terms, the intermediary service provider—as part of its duty of care—may only impose objective and proportionate

restrictions on freedom of expression, which corresponds to a fundamental rights test applying the principles of necessity and proportionality (Klein 2024a).

Despite these efforts, a deeper tension remains: platforms are not public utilities but private corporations, whose primary obligation is to shareholders, not to democratic society. As long as engagement remains the key performance metric, distortive content will retain structural advantages. Addressing this requires not only institutional and technical reform, but also a rethinking of the economic model underpinning digital communication. Just as environmental regulation must shift incentives to discourage pollution, digital regulation must make *discourse quality* a factor that platforms cannot afford to ignore. The future of the digital public sphere depends not only on how societies defend freedom of speech, but also on how they structure the architecture of listening[2]. Visibility, amplification, and interaction are not neutral processes — they are designed, optimized, and monetized. Ensuring that these processes align with democratic values is one of the central governance challenges of the twenty-first century.

5. Interventions and Possibilities: Reclaiming the Digital Public Sphere

If the dynamics of the digital public sphere are structurally distorted by platform logics, commercial incentives, and algorithmic design, the solution cannot rely solely on individual goodwill. Nevertheless, agency remains possible—both at the level of individuals and institutions. Rebuilding a space for meaningful, pluralistic public discourse requires a multi-layered approach that addresses the pedagogical, institutional, and structural dimensions of digital communication.

Pedagogical Interventions: Awareness, Literacy, and Resistance

A crucial step toward improving digital discourse is fostering what may be called critical awareness. Before critical thinking can be exercised, individuals must recognize that the information they consume is filtered**,** personalized, and economically structured. This meta-level understanding—awareness of being in a system—precedes any specific evaluative skills. It is not enough to spot misinformation or identify fallacies; users must understand the *logic* by which content reaches them in the first place.

Building on this foundation, media analysis skills and critical thinking can be developed to interrogate content, context, and intent. These include the ability to deconstruct visual rhetoric, identify manipulative framing, recognize confirmation bias, and trace sources of information. Educational initiatives in schools, universities, and civic spaces must therefore shift from content-focused to structure-aware approaches, training users not only to decode messages but to decode the systems that circulate them.

Institutional Roles: Platforms, Educators, and Public Actors

Institutions—both public and private—have a vital role in modulating the conditions of digital discourse. Educational institutions can integrate platform studies, digital

[2] See also Klein (2025) in this issue of *Symbolon*.

ethics, and civic digital literacy into their curricula to enhance student learning. Public broadcasters and independent media can model deliberative formats and slow content, offering alternatives to platform-driven narratives.

Importantly, platforms themselves must be addressed not only as causes of the problem but also as potential arenas for intervention. Features that slow down the pace of interaction (e.g., read-before-share prompts, time-delay posting, friction elements), mechanisms that flag or contextualize content, and recommendation systems that diversify sources can all contribute to a healthier discourse environment.

However, the adoption of such measures often runs counter to the platforms' commercial interests. Here lies a fundamental tension: slow content is valuable for democratic discourse, but it is less profitable in an engagement-driven model. This disjunction means that voluntary reform is unlikely to succeed unless external pressures—regulatory, reputational, or economic—shift the incentive structure.

Structural Challenges: Reimagining the Platform Incentive Model

At the heart of the problem lies a structural dilemma: platforms optimize for engagement, not deliberation. The current algorithmic logic is not a technological inevitability, but the product of a business model that treats attention as a resource to be mined. Consequently, any serious reform must address this model directly. One possible analogy is with environmental regulation. Companies will not reduce emissions voluntarily if polluting remains more profitable. Similarly, platforms will not deprioritize outrage-driven virality if it remains the most lucrative path. Regulation must therefore create conditions in which it becomes more beneficial for platforms to promote pluralism, accuracy, and discourse quality than to exploit division and emotional manipulation.

This raises a fundamental question: who is the prisoner in the digital dilemma? Is it the user, caught in a system they cannot escape? Or the platform, bound to a logic that maximizes profit while degrading public goods? In many ways, both are trapped. The solution must therefore be systemic: a reconfiguration of incentives, norms, and architectures that rebalances the relationship between freedom, responsibility, and profit.

6. Conclusion: Toward a Sustainable Digital Public Sphere

The transformation of the public sphere in the digital age is neither linear nor neutral. The same platforms that promised democratization have also enabled fragmentation, polarization, and distortion. This study has argued that these developments are not merely the result of poor individual choices or toxic subcultures but emerge from the structural logic of the digital public sphere itself—a logic governed by algorithmic design, economic incentives, and attention-maximizing architectures.

By drawing on models such as the Prisoner's Dilemma and the Tragedy of the Commons, we have shown that the current digital environment incentivizes behavior that is rational at the individual level but collectively harmful. In this dynamic, engagement becomes the currency, and emotional intensity becomes the strategy**.** The result is

a public discourse that is not only less deliberative but also less democratic, as visibility is determined not by merit, truth, or civic value, but by algorithmic relevance.

Furthermore, the logic of surveillance capitalism entrenches this dilemma: the more emotional, controversial, and addictive the content, the more valuable it is in economic terms. This commercial model is not just a background condition—it is the infrastructure of the public sphere in the twenty-first century.

Yet this condition is not immutable. Interventions are possible, and indeed necessary. At the pedagogical level, building critical awareness and media literacy can empower users to understand and resist manipulative dynamics. At the institutional level, educators, journalists, public institutions, and even platform designers can create counter-infrastructures that support slow, reflective, and pluralistic discourse. And at the structural level, regulatory reforms and economic re-alignments must challenge the assumption that virality is value.

Ultimately, the challenge is not simply to fix the digital public sphere but to reimagine it. What would it mean to design platforms that prioritize understanding over engagement, curiosity over certainty, pluralism over polarization? What would it take to make *slow content* not just possible, but desirable? These are not technical questions alone—they are political, ethical, and cultural challenges that cut to the heart of democratic life in the digital era.

The future of public discourse depends on our ability to resist not only misinformation and manipulation but also the very systems that make them profitable. If we wish to preserve the promise of the public sphere in a digital age, we must commit to building infrastructures—technological, institutional, and intellectual—that make deliberation, not distortion, the default logic of communication.

Acknowledgments: I would like to thank the following colleagues for their constructive critical comments during the preparation of the final version of the text: Viktor Friedmann, Tamás Klein, Erzsébet Németh, and Eszter Vizler.

REFERENCES

BAKSHY, Eytan, SOLOMON Messing, and LADA A. Adamic, 2015, "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239): 1130–32. https://doi.org/10.1126/science.aaa1160.

BENKLER, Yochal, ROBERT, Faris, and HAL, Roberts, 2018, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford: Oxford University Press.

ELSTER, Jon, 2007, *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

EUROPEAN COMMISSION, 2022, *Digital Services Act (DSA): Regulation (EU) 2022/2065*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

GILLESPIE, Tarleton, 2018, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

GUESS, Andrew M., JONATHAN Nagler, and JOSHUA Tucker, 2019, "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5(1): eaau4586. https://doi.org/10.1126/sciadv.aau4586.

HARDIN, Garrett, 1968, "The Tragedy of the Commons." *Science* 162(3859): 1243–48. https://doi.org/10.1126/science.162.3859.1243.

KLEIN, Tamás, 2016, "A tárhelyszolgáltató 'omnipotens' felelőssége, mint alkotmányjogi problematika [The 'omnipotent' liability of hosting providers as a constitutional issue]." In KOLTAY, András and TÖRÖK, Bernát ed., *Sajtószabadság és médiajog a 21. század elején. 3. kötet* [*Freedom of the Press and Media Law at the Beginning of the 21st Century, Vol. 3*]. Budapest: Wolters Kluwer, 349–374.

KLEIN, Tamás, 2020, *Sajtószabadság és demokrácia* [*Freedom of the Press and Democracy*]. Budapest: Gondolat – Dialog Campus.

KLEIN, Tamás, 2023, "Az újmédia szabályozásának új iránya? – A Digital Services Act alapjogvédelmi mechanizmusa [A new direction in the regulation of new media? – The fundamental rights mechanism of the Digital Services Act]." In TÖRÖK, Bernát and ZŐDI, Zsolt, eds., *Digitalizálódó társadalom* [*Digitalising Society*]. Budapest: Ludovika, 117–140.

KLEIN, Tamás, 2024a, "A DSA alapjogvédelmi mechanizmusa, mint alkotmányjogi nóvum – A magánjogviszonyokban érvényesülő horizontális alapjogvédelem legutóbbi példája [The fundamental rights mechanism of the DSA as a constitutional novelty – The latest example of horizontal fundamental rights protection in private legal relations]." In KOLTAY, András, SZIKORA, Tamás and LAPSÁNSZKY, András eds., *A vadnyugat vége? Tanulmányok az Európai Unió platformszabályozásáról* [*The End of the Wild West? Studies on the European Union's Platform Regulation*]. Budapest: Orac, 260–294.

KLEIN, Tamás, 2024b, "Új jogterület a jogtudomány horizontján? [A new 'area of law' on the jurisprudential horizon?]." *Jogtudományi Közlöny* 2024(3): 134–146.

KLEIN, Tamás, 2024c, "Platform Regulation and the Protection of Fundamental Rights." *Symbolon* 2024(2): 7–24.

KLEIN, Tamás, 2025, "Constitutional Assessment of Untrue Statements of Fact. Disinformation Campaigns, Controlled Public Discourse, and Filter Bubbles as Threats to the Democratic Public Sphere – Constitutional Theoretical Approaches from Milton's Pursuit of Truth to the Restriction of Fake News" *Symbolon*, no. 49: 7-77, https://doi.org/10.46522/S.2025.02.1

LERMAN, Kristina, XIAORAN, Yan, and XIN-ZENG, WU, 2016, "The 'Majority Illusion' in Social Networks." *PLOS ONE* 11(2): e0147617. https://doi.org/10.1371/journal.pone.0147617.

MATIAS, J. Nathan, 2019, "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions." *Proceedings of the National Academy of Sciences*, 116(20): 9785–9789. https://doi.org/10.1073/pnas.1813486116 .

MOROZOV, Evgeny, 2013, *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.

NAPOLI, Philip M., 2019, *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. New York: Columbia University Press.

O'NEIL, Cathy, 2016, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

PARISER, Eli, 2011, *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

MANOEL HORTA, Ribeiro, OTTONI, Rapahel, WEST, Robert, ALMEIDA, Virgílio A. F., and MEIRA JR., Wagner, 2020, "Auditing Radicalization Pathways on YouTube." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT'20)**, 131–41. https://doi.org/10.1145/3351095.3372879.

SUNSTEIN, Cass R., 2001, *Republic.com*. Princeton: Princeton University Press.

TUFEKCI, Zeynep, 2015, "Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency." *Colorado Technology Law Journal* 13(203): 203–18.

ZUBOFF, Shoshana, 2019, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

.