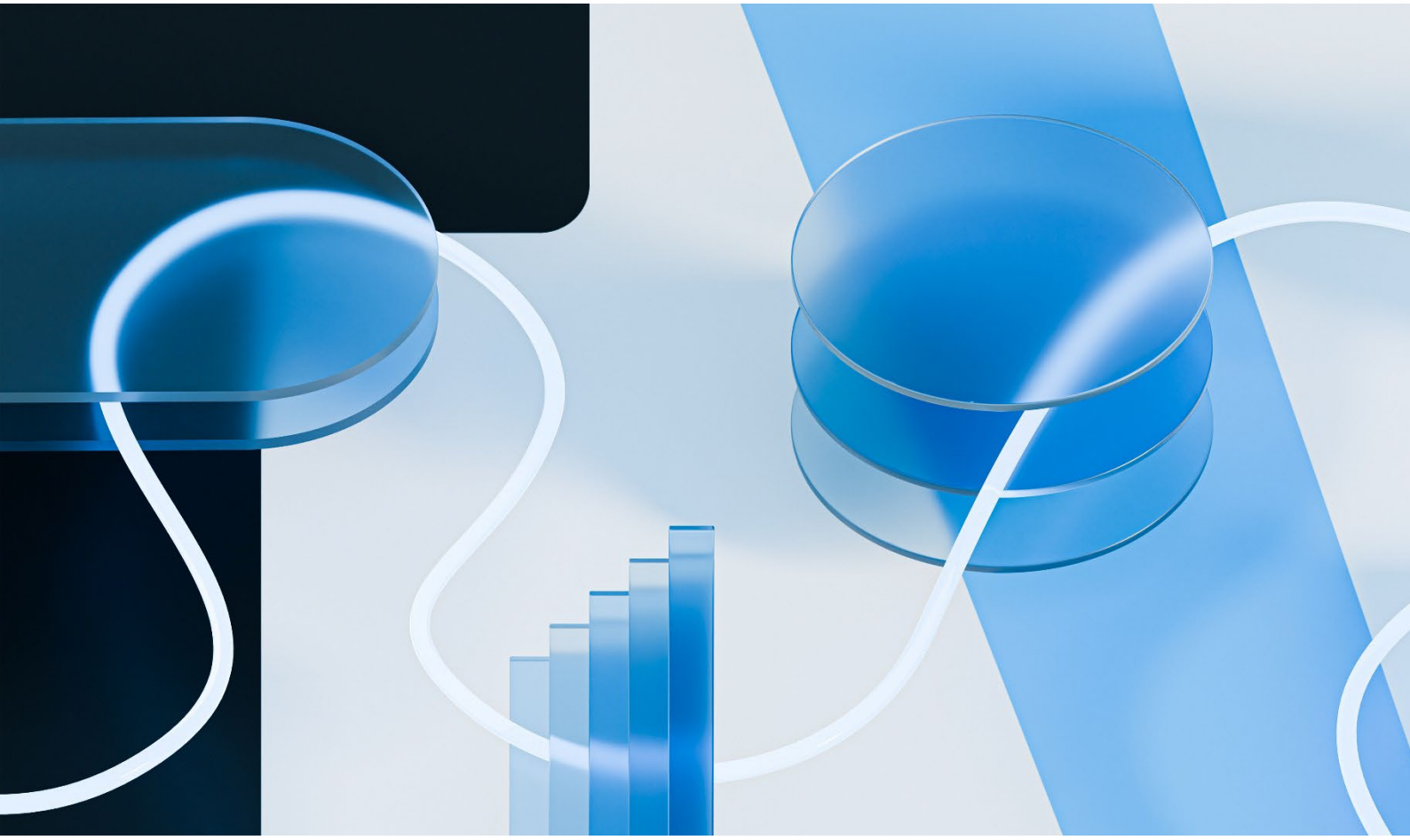
twon
twin of online social networks



TWin of Online Social Networks

Deliverable 6.3

Report About Strategies to Improve Platform Mechanics

Main Authors: COSIMA PFANNSCHMIDT

Funded by
the European Union

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia), Slovenska Tiskovna Agencija (Slovenia), Dialogue Perspectives e.V (Germany).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

Project name	TWin of Online Social Networks
Project acronym	TWON
Project number	101095095
Deliverable number	D6.3
Deliverable name	Report About Strategies to Improve Platform Mechanics
Due date	30 September 2025
Submission date	30 September 2025
Type	R – Document, report
Dissemination level	PU – Public
Work package	WP6
Lead beneficiary	FZI Forschungszentrum Informatik (Germany)
Contributing beneficiaries and associated partners	FZI Forschungszentrum Informatik (Germany), Universiteit van Amsterdam (the Netherlands), Universität Trier (Germany), Institut Jozef Stefan (Slovenia), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia), Slovenska Tiskovna Agencija (Slovenia) and Dialogue Perspectives e.V (Germany)

Contents

1. Introduction

The TWin of Online Social Networks (TWON) project investigates how the architecture and design of online social networks (OSNs) shape democratic debates in Europe and beyond. By building a *digital twin* of an OSN and populating it with generative agents, TWON enables an independent and systematic exploration of platform mechanics – without relying on the goodwill or data access policies of commercial providers such as Meta, X, or TikTok.

The project addresses pressing societal questions: How do algorithmic design choices foster or hinder democratic deliberation? To what extent do phenomena such as filter bubbles, echo chambers, polarization, and disinformation emerge from platform mechanics rather than user preferences alone? And how can evidence-based interventions mitigate harmful effects while strengthening digital citizenship?

This deliverable (D6.3) is part of Work Package 6, which focuses on translating TWON's research into actionable strategies and policy recommendations. Specifically, the report identifies and evaluates measures to improve platform mechanics in ways that reduce polarization, limit addictive design, and empower users to act as informed citizens. By linking state-of-the-art behavioral science with TWON's unique digital twin methodology, this document provides guidance for regulators, civil society, and platform providers on how to align OSNs more closely with democratic and societal goals.

Online social networks (OSNs) such as TikTok, Instagram, X, and WhatsApp are used daily by millions of people for communication, entertainment, and information. Early expectations were high: these platforms were seen as tools for broadening citizen participation and amplifying marginalized voices in public discourse. Today, however, concerns about their societal impact dominate the discussion. The rapid spread of disinformation, the prevalence of online hate speech, and addictive usage patterns have raised urgent questions about their role in democratic societies.

These challenges stem not only from user behavior but also from platform design. The architecture of OSNs determines what content is visible, which interactions are encouraged, and how users' attention is captured. Because these platforms are run by private companies, design decisions are typically optimized for engagement time and monetization through targeted advertising (Bak-Coleman et al., 2021; Narayanan, 2023, Metzler & Garcia, 2024). This commercial logic tends to favor emotionalized and extreme content, filter bubbles, and features that incentivize prolonged and frequent use.

Such dynamics can have detrimental consequences at both individual and societal levels. They contribute to addictive behaviors, deepen polarization, and erode trust in democratic institutions. In light of these risks, providers such as Meta and TikTok must recognize their societal responsibility. Guided by scientific evidence, they should redesign platform mechanisms to foster well-being, support mental health, strengthen social cohesion, and enable nuanced political discussion (Metzler & Garcia, 2024). This deliverable maps the current academic discussion on possible avenues to improve platform mechanics. It offers solutions from systemic changes to behavioral nudges around fostering constructive political debates, limiting addictiveness as well as enabling transparency and user customization.

2. Suggested Measures to Improve Platform Mechanics

2.1. Fostering Constructive Political Debates

Online social networks play a crucial role in shaping public discourse, but their current design often exacerbates polarization, amplifies extreme voices, and facilitates the rapid spread of misinformation. To support democratic deliberation, platforms must adopt mechanisms that reduce division, encourage exposure to diverse perspectives, and promote civil dialogue. Within the TWON project, we have discussed this in-depth in Deliverable D5.2 “Debate Metrics” (Stolwijk, 2025).

2.1.1. Promoting Civility and Reflection

Optimizing algorithms for qualities such as civility has been proposed as a way to strengthen democratic discourse (Lewandowsky & Kozyreva, 2022; Oremus et al., 2021). However, since civility is context-dependent - angry responses can sometimes be legitimate when addressing injustice - platforms might allow users themselves to define the values guiding algorithmic prioritization (Lewandowsky & Kozyreva, 2022; Lorenz-Spreen et al., 2020).

As part of the TWON project, Stolwijk (2025) has defined metrics for the deliberative quality of online debates. These include: exposure to political content, engagement with political content, contributing political content, diversity of exposure and quality of exposure. Additional indicators are (in-)civility, interactivity and novelty. For the design of deliberative online platforms, the authors Stolwijk (2025) suggest to determine the weight of metrics on an individual basis, rather than equally for everybody and to optimize the metrics for the long-run outcome, rather than for simultaneously good scores on every metric. The focus is on fostering diverse exposure to political content on online social networks, seeing platforms as one aspect within the larger media system of deliberative democracy.

Interventions that add friction to online interactions can also reduce impulsive and toxic behavior. Increasing the effort required to share content - such as adding an extra confirmation click, introducing short delays, or prompting reflection before posting - has been shown to decrease the spread of sensationalist and misleading information (Brady et al., 2020; Lorenz-Spreen et al., 2020; Menczer, 2021). Empathic and humanizing prompts have further been demonstrated to reduce affective polarization and harassment, creating conditions for more respectful exchanges (Saveski et al., 2022; Hangartner et al., 2021; Munger, 2017).

2.1.2. Reducing Polarization and Amplifying Moderate Voices

Research suggests that reducing the visibility of strongly partisan or triggering content can help alleviate perceived polarization (Rose-Stockwell, 2018). Algorithmic adjustments that prioritize content with cross-partisan appeal may highlight moderate voices and reveal areas of opinion overlap, encouraging recognition of common ground (Bail, 2021). In parallel, platforms could promote more trustworthy news sources, improving the quality of political information available to users (Bhadani et al., 2022).

Algorithms can also be leveraged to foster gradual intergroup understanding. By presenting users with content that is not too distant from their existing views, algorithms can encourage small steps toward appreciating alternative perspectives without overwhelming them (Levendusky, 2018). Such design choices can support constructive dialogue rather than reinforcing ideological silos.

2.1.3. Harnessing Social Norms and Informational Cues

Another promising direction involves the strategic use of view metrics. Platforms could display metrics about both active and passive user behaviors, such as the number of people who scrolled over a post without reacting (Lorenz-Spreen et al., 2020). These cues may counteract biased perceptions and false-consensus effects, whereby users overestimate the degree of agreement with posts. Similarly, contextualizing engagement by e.g. stating the number of likes a post has in relation to the number of readers, provides a more accurate picture of public opinion than raw like counts (Lorenz-Spreen et al., 2020).

Informational nudges can also raise epistemic quality. Highlighting when content originates from anonymous or state-controlled sources, linking to vetted references, or displaying the cascade of how posts spread encourages critical evaluation (Lorenz-Spreen et al., 2020). Transparent sorting algorithms and clear labeling of content types further improve understanding of why certain posts appear in feeds.

Nudges that emphasize social norms can also reduce the spread of misinformation. Communicating that most users do not share or endorse certain misleading claims has been shown to reduce others' willingness to do so, harnessing the influence of descriptive norms (Kozyreva et al., 2024).

2.1.4. Combating Misinformation Through Inoculation and Refutation

Beyond reactive fact-checking, proactive strategies are crucial. Inoculation interventions - such as brief videos, interactive games, or explanatory texts - can prepare users to recognize misleading tactics before encountering misinformation. By highlighting common strategies used to deceive, inoculation fosters resilience without limiting free expression (Kozyreva et al., 2024). Complementary approaches include informational labels that indicate source credibility or warn users about contested claims. These refutation strategies reliably reduce misconceptions (Kozyreva et al., 2024). Additionally, algorithms can limit the spread of false content directly, by downranking it (Bak-Coleman et al., 2022).

2.2. Limiting Addictiveness

Addictive design in social media platforms has emerged as a central concern for policymakers and researchers, as these mechanisms exploit psychological vulnerabilities to maximize engagement. The European Commission has recognized the urgency of this issue, embedding restrictions on addictive digital design within the forthcoming Digital Fairness Act (EU Commission, 2025).

Social media platforms deliberately implement features that capture and prolong user attention. Reward systems - such as "likes," achievements, progress metrics, and point-based incentives - cultivate a sense of accomplishment and encourage repeated use (Granda et al., 2025). Time-related features, such as tracking and optimizing user log-ins, further enhance immediacy and compulsive interaction. Infinite scrolling and automated content loading exacerbate this tendency by creating an "endless" stream of content, which fosters habitual use without natural stopping cues (Granda et al., 2025; Montag et al., 2019). Personalization algorithms, which recommend content based on behavioral data, intensify user engagement by continuously providing material tailored to individual preferences. While these mechanisms increase satisfaction, they also fuel dependency, encourage social comparison, and deepen emotional reliance on platforms for validation and self-esteem (Granda et al., 2025). This dynamic is reinforced by real-time communication features and usability designs - such as seamless synchronization, responsive interfaces, and aesthetic cues - that facilitate frictionless, prolonged use. A range of interventions has been proposed to counteract these addictive design practices.

2.2.1. Adjusting Content Delivery and Personalization

One effective strategy to counteract compulsive engagement with digital platforms is to reconsider how frequently and how precisely content is delivered to users. Reducing the rate at which new material appears in news feeds – or refreshing feeds only at predetermined intervals (for example, once per hour instead of continuously) – can significantly mitigate the tendency toward endless scrolling (Montag et al., 2019). Such measures disrupt the constant reinforcement cycle that encourages users to check back compulsively in search of novelty. In addition, platforms could experiment with deliberately reducing the precision of personalization algorithms. Instead of offering content exclusively tailored to a user's prior behavior and preferences, they might occasionally introduce posts, articles, or recommendations that fall outside the individual's usual areas of interest. This practice not only helps to interrupt addictive consumption patterns but also plays a critical role in addressing filter bubbles and echo chambers, phenomena that reinforce pre-existing opinions and limit exposure to diverse perspectives (Pariser, 2011; Jamieson & Cappella, 2008; Sunstein, 2007). By integrating greater randomness and variety into content delivery, platforms could encourage more balanced information diets and foster more reflective engagement.

2.2.2. Neutralizing Design Cues and Introducing Interruptions

Another important avenue for promoting healthier use behaviors lies in the visual and interaction design of platforms. At present, many social media and messaging services employ highly stimulating design elements, such as colorful highlights, badges, or dynamic icons, that draw the user's attention to new content and reinforce habitual checking. By shifting toward more neutral, less stimulating designs, platforms could reduce the salience of these cues and help break cycles of compulsive interaction. Moreover, integrating deliberate interruptions into the user experience can serve as an effective safeguard against unreflective engagement. Examples of such flow-interrupting mechanisms include screen dimming after prolonged usage, pop-up prompts encouraging users to take a break, or gentle reminders of elapsed time. These design interventions can act as "speed bumps" that slow down automatic behavior and provide opportunities for users to reconsider whether they wish to continue. Evidence suggests that such mechanisms can discourage continuous engagement and help individuals regain a sense of control over their platform use (Granda et al., 2025; Sindermann et al., 2022).

2.2.3. Rethinking Notifications and Feedback Signals

Notifications and feedback signals are among the most powerful tools platforms employ to capture user attention. Constant push notifications create a persistent sense of urgency, prompting users to re-engage even when the content is not of immediate relevance. Limiting or disabling such notifications can therefore be a highly effective measure to reduce unnecessary engagement. Beyond this, the removal of socially pressuring elements – such as read receipts in messaging apps (e.g., WhatsApp's "blue ticks") – may further alleviate the compulsion to respond instantly. By minimizing these subtle but powerful forms of social pressure, platforms can create a communication environment that prioritizes user autonomy over responsiveness. In addition, restructuring the way platforms provide feedback – for instance, replacing instant acknowledgments with less intrusive alternatives – could further weaken the cycle of constant checking. Together, these measures may reduce the psychological burden associated with digital communication and promote healthier, less stressful interaction patterns (Sindermann et al., 2022).

2.2.4. Redesigning Reward Systems

Finally, addressing the underlying reward mechanisms of digital platforms represents a crucial step toward reducing addictive use. Many current systems are built on variable reinforcement schedules, in which

rewards – such as likes, notifications, or other forms of feedback – are delivered unpredictably. This type of reward structure is known to trigger powerful dopamine-driven cycles of anticipation and response, keeping users hooked in the hope of receiving the next gratifying signal. By moving away from such variable reinforcement schedules and adopting more predictable or transparent feedback mechanisms, platforms could weaken these cycles and reduce compulsive engagement. For example, notifications could be bundled and delivered at regular intervals rather than in an unpredictable stream, or the visibility of engagement metrics such as likes and shares could be reduced or removed. Redesigning reward systems in this way not only disrupts patterns of addictive use but also encourages users to focus on the intrinsic value of content and interactions rather than the extrinsic rewards tied to platform mechanics. Such changes could support a healthier and more intentional relationship between users and digital platforms (Granda et al., 2025).

2.3. Transparency and Enabling User Customization

A promising approach to improving online social networks for the good of society is to empower users with greater control over how content is curated and displayed. Currently, ranking algorithms are largely opaque and optimized for engagement, often at the expense of user autonomy. Allowing users to customize their feeds can strengthen informational autonomy and reduce the risks associated with one-size-fits-all algorithmic design. A prerequisite for user autonomy and customization, however, is transparency about usage patterns and underlying algorithms.

2.3.1. Transparency

Empowering users with information about platform mechanisms can strengthen autonomy, reduce addictive use, and improve the quality of online discourse. Transparency regarding algorithms, time spent online, and the risks of addictive design helps individuals make more deliberate choices about their engagement (Granda et al., 2025). Furthermore, reminder bots and usage statistics can discourage excessive scrolling (Zhang et al., 2022; Metzler & Garcia, 2024), while features such as break prompts or gradual screen dimming can encourage healthier habits (Granda et al., 2025), as discussed in Section 2.2.2.

2.3.2. Customization of Feeds and Algorithms

Users could be given the option to decide which ranking algorithm structures their feed. For example, individuals might prioritize posts from close friends, news content, popular posts, or simply the most recent updates. Such options would not only enhance transparency regarding the mechanisms underlying content delivery but also provide users with greater agency over their online experiences. Research shows that separate topic-focused feeds, rather than a single main feed, may be particularly effective, especially if enhanced with algorithmic suggestions that help users explore relevant areas of interest (Zhang et al., 2022; Metzler & Garcia, 2024).

At the same time, the boundaries of customization need careful consideration. While user or community-level choices in algorithmic ranking can be beneficial, allowing individuals to restrict themselves exclusively to partisan content risks reinforcing polarization and undermining exposure to moderate or opposing perspectives (Lewandowsky & Kozyreva, 2022; Lorenz-Spreen et al., 2020; Metzler & Garcia, 2024). Designing customizable systems that maintain opportunities for intergroup contact is therefore essential.

2.3.3. Boosting Informational Autonomy through Self-Nudging

Customization can also serve as a form of “self-nudging,” where users act as their own choice architects. By enabling individuals to design and sort their own news feeds, platforms can encourage the creation of

information environments tailored toward higher epistemic quality (Lorenz-Spreen et al., 2020). For instance, customization tools could help users downrank low-quality sources, highlight epistemically reliable outlets, and structure feeds in ways that promote intentional rather than reactive engagement. This requires transparent sorting algorithms, clear layouts, and the provision of epistemic cues that help users make informed choices about their digital environments.

Algorithms could also be adapted to prioritize news users explicitly state they want to see, shifting away from engagement-maximization toward preference fulfillment (Rathje et al., 2022; Metzler & Garcia, 2024). This would enable platforms to better reflect user intent rather than exploit behavioral tendencies.

2.3.4. Customization of Notifications and Attention Management

Beyond feed design, customization options can help mitigate compulsive use and attention capture. For instance, allowing users to tailor notification settings enables them to avoid constant interruptions and instead engage with content in a more deliberate manner. Rather than issuing a stream of alerts, platforms could offer digestible summaries at chosen intervals, fostering intentional engagement over reactive browsing (Granda et al., 2025).

3. Conclusion

Improving the societal impact of online social networks requires a careful balance between enhancing user autonomy and ensuring platform accountability. Limiting addictive mechanisms – such as infinite scrolling, highly personalized feeds, and push notifications – can reduce compulsive use and psychological dependency without eliminating engagement. Interventions that promote exposure to diverse perspectives further mitigate both addictive behaviors and ideological isolation, supporting more nuanced public discourse (Rathje et al., 2021).

Customization, transparency, and informational nudges empower users to reclaim control over their online experiences, fostering autonomy, epistemic quality, and healthier engagement patterns. Tools such as feed design options, notification management, and literacy-enhancing cues help users regulate their time, critically assess content, and participate constructively in digital spaces (Granda et al., 2025; Lorenz-Spreen et al., 2020).

These strategies underscore that the dysfunctions of social media are not solely the result of algorithmic curation but also emerge from the fundamental architecture of platforms, which incentivizes emotionally reactive and polarizing interactions (Larooij & Törnberg, 2025). Addressing these structural dynamics, even in the face of commercial incentives that favor engagement over societal good, is essential for aligning platform design with democratic, informational, and relational purposes.

The measures suggested in this deliverable have been discussed in the consortium. The TWON Large Scale Simulations (LSS) can be used to test the effects of some of the suggested measures, including reducing the visibility of false or strongly partisan/triggering content, optimizing algorithms for civility, promoting trustworthy news sources and showing interventions around posts that include mis/disinformation. Within the scope of the TWON project we aim to test for the effect on debate quality and civility of the “truth sandwich approach” to counter misinformation (König, 2023), algorithmic boosting or downweighing of suspected disinformation and of success-driven algorithmic ranking versus chronological or personalized ranking.

To fully harness the potential of online social networks as spaces for democratic and civil debate, platforms should be designed to foster inclusivity, transparency, and meaningful user control. Providers must embrace their societal responsibility by curbing hate speech and disinformation, encouraging pro-social behavior, and reducing the addictive features of their systems. At the same time, lawmakers must regulate these platforms, which now shape much of the public sphere, to ensure accountability and responsible governance.
Important progress has already been made at the EU level through the Digital Services Act (DSA), the Digital Markets Act (DMA), and the AI Act, which aim to restore public oversight of digital infrastructure. However, the real impact of these frameworks depends on their implementation, which must be closely monitored and evaluated to inform future interventions. Crucially, effective access to platform data for researchers, as foreseen in Article 40 of the DSA, remains essential to this process.

4. References

Bail, C. (2022). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press. https://doi.org/10.1515/9780691246499

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Van Bavel, J. J., & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), e2025764118. https://doi.org/10.1073/pnas.2025764118

Bhadani, S., Yamaya, S., Flammini, A., Menczer, F., Ciampaglia, G. L., & Nyhan, B. (2022). Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*, *6*(4), 495–505. https://doi.org/10.1038/s41562-021-01276-5

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., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

EU Commission. (2025). *Digital Fairness Act: Call for Evidence*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14622-Digital-Fairness-Act_en

Granda, M. F., Sarmiento, M.-B., Nuñez, A.-G., Maldonado, R., & Parra, O. (2025). Developing a Design Features Taxonomy of Human-Computer Interaction in Social Media that Affect User Engagement and Addictive Behaviors. In J. Grabis, T. E. J. Vos, M. J. Escalona, & O. Pastor (Eds.), *Research Challenges in Information Science* (pp. 313–330). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92474-3_19

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

König, L. M. (2023). Debunking nutrition myths: An experimental test of the 'truth sandwich' text format. *British Journal of Health Psychology*, 28(4), 1000–1010. https://doi.org/10.1111/bjhp.12665

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Larooij, M., & Törnberg, P. (2025). *Can We Fix Social Media? Testing Prosocial Interventions using Generative Social Simulation* (No. arXiv:2508.03385). arXiv. https://doi.org/10.48550/arXiv.2508.03385

Levendusky, M. S. (2018). When Efforts to Depolarize the Electorate Fail. *Public Opinion Quarterly*, *82*(3), 583–592. https://doi.org/10.1093/poq/nfy036

Lewandowsky, S., & Kozyreva, A. (2022, April 7). *Algorithms, lies, and social media*. Nieman Lab. https://www.niemanlab.org/2022/04/algorithms-lies-and-social-media/

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, *4*(11), 1102–1109. https://doi.org/10.1038/s41562-020-0889-7

Menczer, F. (2021, September 13). *How "engagement" makes you vulnerable to manipulation and misinformation on social media*. Nieman Lab. https://www.niemanlab.org/2021/09/how-engagement-makes-you-vulnerable-to-manipulation-and-misinformation-on-social-media/

Metzler, H., & Garcia, D. (2024). Social Drivers and Algorithmic Mechanisms on Digital Media. *Perspectives on Psychological Science*, *19*(5), 735–748. https://doi.org/10.1177/17456916231185057

Montag, C., Lachmann, B., Herrlich, M., & Zweig, K. (2019). Addictive Features of Social Media/Messenger Platforms and Freemium Games against the Background of Psychological and Economic Theories. *International Journal of Environmental Research and Public Health*, *16*(14), 2612. https://doi.org/10.3390/ijerph16142612

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Narayanan, A. (2023, March 9). *Understanding social media recommendation algorithms*. Knight First Amendment Institute. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Oremus, W., Alcantara, C., Merrill, B., & Galocha, A. (n.d.). How Facebook shapes your feed. *Washington Post*. https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2022). *People think that social media platforms do (but should not) amplify divisive content*. PsyArXiv. https://doi.org/10.31234/osf.io/gmun4

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Rose-Stockwell, T. (2023, September 22). How to Design Better Social Media. *Medium*. https://tobiasrose.medium.com/how-to-fix-what-social-media-has-broken-cb0b2737128

Saveski, M., Gillani, N., Yuan, A., Vijayaraghavan, P., & Roy, D. (2022). Perspective-Taking to Reduce Affective Polarization on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 885–895. https://doi.org/10.1609/icwsm.v16i1.19343

Sindermann, C., Montag, C., & Elhai, J. D. (2022). The design of social media platforms – Initial evidence on relations between personality, fear of missing out, design element-driven increased social media use, and problematic social media use. *Technology, Mind, and Behavior*, *3*(4). https://doi.org/10.1037/tmb0000096

Stolwijk, S. (2025). *TWin of Online Social Networks Deliverable D5.2 Definition of Metrics*. OSF. https://doi.org/10.31235/osf.io/6qv5y_v1

Sunstein, C. R. (2007). Republic.com 2.0. Princeton University Press.

Zhang, M. R., Lukoff, K., Rao, R., Baughan, A., & Hiniker, A. (2022). Monitoring Screen Time or Redesigning It?: Two Approaches to Supporting Intentional Social Media Use. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517722

twon

twin of online social networks

Contact us

Damian Trilling
Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam

Funded by
the European Union