

Societal impact of systemic risks. Democratic processes in the context of disinformation

Ann-Kathrin Watolla, Patrick Zerrer, Jan Rau, Lisa Merten, Matthias C. Kettemann, Cornelius Puschmann

04 July 2025

**Read the study
(in German only):**



About the research study

- Commissioned by the Bundesnetzagentur (German DSC) in 2024
- **Interdisciplinary study** drawing from political science, sociology, communication studies, psychology, computer science, information science, and human computer interaction research
- **Scoping Review** as a „‘mapping’ [to summarise] a range of evidence in order to convey the breadth and depth of a field” (Levac et al., 2010, p. 1)
- Drawing from recent publications, mainly focusing on the German or European context
 - Peer-reviewed research literature
 - Grey literature (pre-prints, reports, etc.)
 - Reports from VLOPs and VLOSEs

Addressing common assumptions about disinformation

Guiding question:

In what ways does disinformation manifest itself on social platforms, and what impact does it have on society and democratic processes?

The study

- provides a **broad overview** of **what we know** and **what we don't know** about the societal impact of systemic risks, especially the role of disinformation for democratic processes.
- refutes many **common assumptions**.
- highlights required **areas of further research**.

A white 3D hashtag symbol (#) is the central focus, standing on a light-colored desk. In the background, there are several papers, some with blue and white patterns, and a red apple is partially visible on the left side. The background is softly blurred, creating a shallow depth of field.

Debunking prevalent assumptions



Debunking assumptions:

#1 – “The term ‘disinformation’ is clearly defined.”

Debunking assumption 1: “The term ‘disinformation’ is clearly defined.”

- The term disinformation has increasingly become **part of public discourse** in recent years
- Commonly used by researchers, politicians, media producers and distributors – often **without a shared interpretation** of the term (Bleyer-Simon & Reviglio, 2024; Dreyer et al., 2021)
- In public dialogue, often referring to individual posts that contain **false or misleading information** (Kessler, 2023)
- **Broad variety of disinformation types** aggravate a shared understanding of the term

Debunking assumption 1: “The term ‘disinformation’ is clearly defined.”

What we know:

- There is no standardised definition of the term “disinformation”
- Disinformation refers to demonstrably false or misleading information
- Creating and spreading disinformation has the potential to cause social harm

What we don't know:

- Clearly distinguish disinformation from fake news, misinformation, conspiracy theory, etc.
- Where does disinformation begin and where does it end



What's next?

More foundational research on **disinformation as comprehensive narratives** needed



Debunking assumptions:

#2 – “Disinformation is the biggest challenge of digital media.”

Debunking assumption 2: “Disinformation is the biggest challenge of digital media.”

“What is new today is the epidemic spread of disinformation on the Internet, the enormous power of digital media, but also the variety of attacks on the use of public reason.”

– Federal President Dr Frank-Walter Steinmeier, Berlin, 21 March 2018

- Most polarising factor in the digital space: the **constant clash between antagonistic political actors**, generated by the internet's networked structure (Bail et al., 2018; Bruns, 2019; Rau & Stier, 2019)
- The digital attention economy, including platform design and platform algorithms, are artificially fueling these conflicts (Arora et al., 2022)

Debunking assumption 2: “Disinformation is the biggest challenge of digital media.”

What we know:

- Digital platforms are amplifying conflict by design
- This constant digital overrepresentation and reinforcement of conflict is undermining social cohesion and democracy

What we don't know:

- The size of the impact of platform algorithms vs. organic factors
- How to enable alternative information distribution logics beyond the attention economy



What's next?

Exploring interventions to enable different attention distribution logics on digital platforms



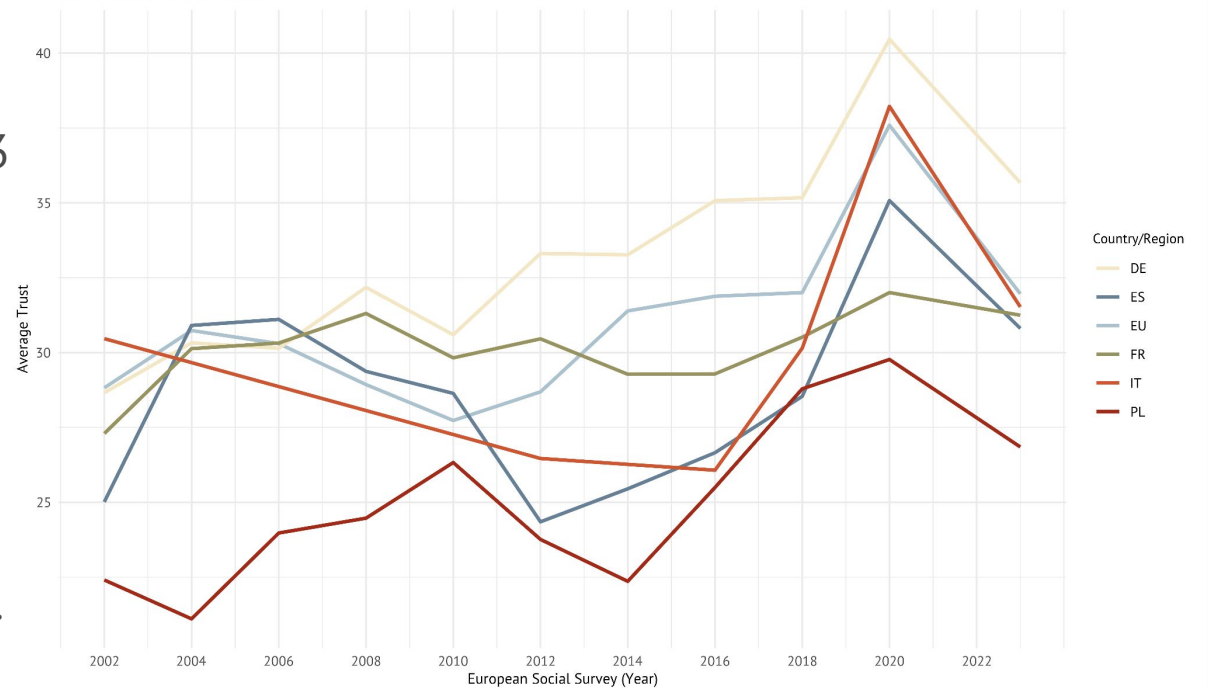
Debunking assumptions

#3 – “Misinformation erodes trust.”

Debunking assumption 3: “Desinformation erodes trust.”

- **Core Role of Trust:** Trust is *essential* for democratic function – connecting citizens to institutions, enabling decision-making, reducing political costs (Levi & Stoker, 2000; Zmerli, 2023)
- **Declining Trust = Instability:** Loss of trust weakens institutions, increases citizen skepticism, hinders compromise & political stability (Citrin & Stoker, 2018)
- **Trust & Participation:** Decreased trust often leads to *less* participation in formal politics (e.g. voting), but can *increase* non-institutionalised political action (e.g. protest) (Hooghe & Marien, 2013)

Development of Trust in Europe:



Debunking assumption 3: “Desinformation erodes trust.”

What we know:

- Digital platforms are central to political communication & opinion formation
- Exposure to disinformation use correlates with declining trust in societal institutions
- Platforms enable rapid spread of disinformation, eroding trust
- Debunking often has limited reach & can backfire

What we don't know:

- Causal relationship between social media use and trust in traditional media
- Causal relationship between social media use and trust in societal institutions



What's next?

Better **access to data** from digital platforms, studies with an **longitudinal & experimental research design**



Debunking assumptions

#4 – “Echo chambers and filter bubbles amplify misinformation.”

Debunking assumption 4: “Echo chambers and filter bubbles amplify desinformation.”

- Users mainly consume **information that aligns** with their political preferences (Aruguete et al., 2023)
- Studies show that, in practice, users access a **variety of information sources**
- User preferences are **reinforced by recommendation algorithms** that prioritise engagement in the context of the attention economy (Figà Talamanca & Arfini, 2022)
- Findings suggest that constant encounters with different political actors in the same digital space **increase social polarisation**, rather than fragmentation (Bail et al., 2018; Bruns, 2019; Rau & Stier, 2019)

Debunking assumption 4: “Echo chambers and filter bubbles amplify desinformation.”

What we know:

- Politically radical users are often more digitally engaged and tend to share content that aligns with their own ideas
- Digital media design and algorithms often expose users to content that reinforces their political views

What we don't know:

- Actual composition of users' information sources
- Strength and effect of recommendation algorithms



What's next?

Data access needed to better understand the amplification effects of algorithms, while taking into account the diversity of platforms and algorithms and the high rate of change over time.



Debunking assumptions

#5 – “AI-generated misinformation has a significant impact on public opinion formation.”

Debunking assumption 5: “AI-generated disinformation has a significant impact on public opinion formation.”

- Increasing use of **AI-generated content is a common tactic** for spreading disinformation (Guess et al., 2023)
- Potential **impact of generative AI** on disinformation can be divided into four categories (Xu et al., 2024):
 - increased quantity
 - improved quality
 - greater personalisation, and
 - 'hallucination', whereby plausible yet false information is generated
- The European Union's AI Act (2024) provides clear requirements and obligations for developers and operators of AI systems. This results in a **labelling obligation** for automated systems.

Debunking assumption 5: “AI-generated disinformation has a significant impact on public opinion formation.”

What we know:

- Volume of AI-generated disinformation is increasing
- Quality of deepfakes is increasing, thus making them appear more credible
- Use of AI tools for political information is still comparatively low, especially among the general population

What we don't know:

- Currently no conclusive empirical proof of mass manipulation/significant influence of public opinion through generative AI
- Impact of increasingly realistic deepfakes on their persuasiveness



What's next?

Research analysing the growing concern about the impact of disinformation on public opinion formation



Debunking assumptions

#6 – “Younger people are more competent in dealing with AI-generated disinformation than older generations.”

Debunking assumption 6: “Younger people are more competent in dealing with AI-generated disinformation than older generations.”

Multiple studies indicate **generational differences** in the use of and familiarity with AI-generated content (Bitton et al., 2024; Hashmi et al., 2024; Sippy et al., 2024):

- Younger users (under 40) are significantly more likely to use chat-based AI tools
- They also demonstrate greater knowledge of deepfake technologies, due to more frequent exposure and stronger digital skills

→ Overall, digital literacy tends to be higher in younger age groups BUT **higher exposure and confidence do not guarantee better detection skills** – many users overestimate their ability to identify deepfakes

Debunking assumption 6: “Younger people are more competent in dealing with AI-generated disinformation than older generations.”

What we know:

- Older people, women and those with at least a bachelor's degree tend to be more concerned about misinformation
- Studies show that many people are unable to identify deepfakes as manipulated material

What we don't know:

- Causality between in the ability to recognise deepfakes and contact with deepfakes as well as an overall more experienced use of digital media
- Extent to which factors such as age, level of education and political orientation can influence how disinformation is assessed



What next?

Empirical research **isolating individual demographic** factors



Debunking assumptions

#7 – “We can train people to detect AI-generated disinformation.”

Debunking assumption 7: “We can train people to detect AI-generated disinformation.”

- Studies show that around 40% of participants are unable to identify deepfakes as manipulated material, but many overestimated their ability to identify them correctly (Birrer & Just, 2024)
- Overall low recognition rate of deepfakes can be attributed to people's fundamental trust in audiovisual content (Hameleers & Marquart, 2023)
- Progressive improvement in the quality of deepfakes poses a significant problem for the human detection of deepfakes (Patel et al., 2023)

Using AI to **automatically recognise disinformation** and **create transparency**:

- Natural language processing: AI models can analyse texts for linguistic patterns and contradictions, statements, and missing references (Tajrian et al., 2023)
- Image and video analysis: AI-supported tools analyse fine details, such as eye movements and lip synchronisation, to detect manipulated content (Ghai et al., 2024; Patel et al., 2023)

Debunking assumption 7: “We can train people to detect AI-generated disinformation.”

What we know:

- People overestimate their ability to identify AI-generated disinformation like deepfakes
- Technical progress eliminates inauthentic facial expressions, unnatural speech rhythms and low image quality, which previously made deepfakes more easily recognisable

What we don't know:

- Future progress of technology and its impact on AI-generated content
- Future ability of AI's capabilities to detect deepfakes



What next?

Strengthening **platforms' responsibility** to identify AI-generated content and labelling it accordingly



Debunking assumptions

#8 – “Disinformation impacts the behaviour of societal and political functionaries.”

Debunking assumption 8: “Disinformation impacts the behaviour of societal and political functionalities.”

- Social and political functionalities form the **backbone of a functional democratic society** through their voluntary work (Cheruiyot, 2024)
- They are increasingly becoming the **target of politically motivated hostility** in both the digital and analogue spaces (Seeger et al., 2024)
- Digital platforms provide a space for extremist and violent actors to spread their ideology, mobilise and share practical information about carrying out violent acts (HateAid et al. 2025)

Debunking assumption 8: “Disinformation impacts the behaviour of societal and political functionaries.”

What we know:

- Increase in political violence and violent extremism, particularly right-wing extremism and conspiracy ideology movements
- contains demonstrably false or misleading information
- creating and spreading disinformation has the potential to cause social harm

What we don't know:

- Extent of ‘chilling effects’
- Impact of other social factors than digital media contributing to extremist violence



What next?

Countermeasures towards chilling effects and support for societal and political functionaries



Conclusion and outlook

Summary of the current state of knowledge on disinformation

1. Exposure is a black box

We lack valid, large-scale data on:

- **Who is actually exposed** to disinformation
- **How often**, through which channels, and in what contexts
 - Without this, impact assessments remain speculative without proper data access

2. Causal effects remain unclear

- No conclusive evidence linking disinformation to **voting behaviour** or **policy attitudes**
 - Methodological challenges in isolating causal pathways (vs. correlation) without proper data access

3. Algorithmic amplification is poorly understood, but what we know is concerning

- How recommendation systems contribute to disinformation spread is **not transparent**
 - Insufficient data access for independent audit of **AI-driven content curation**

Pathways to move forward

Enforcement of the DSA for sound analyses of the impact, effect, and role of disinformation for public discourse:

- Empirical evidence must guide regulation (e.g. DSA Art. 34–35 risk assessments)
- Interventions (DSA Art. 35) must aim at changing the structural conditions of attention distribution on digital platforms (see overrepresentation of conflict)
- Existing transparency reports from VLOPs/VLOSEs are insufficient
- Data access (Art. 40 DSA):
 - Provision of **meaningful** researcher access not just platform-curated data sets
 - Establishment of a standardised, independent and secure research data access framework

→ To build effective regulation and resilience, we need **empirically grounded answers** – not assumptions.

Contact the research team



Dr. Ann-Kathrin Watolla
Alexander von Humboldt
Institute for Internet and Society

ann-kathrin.watolla@hiig.de



Dr. Patrick Zerrer
University of Bremen

pzerrer@uni-bremen.de



Jan Rau
Research Institute Social Cohesion

j.rau@leibniz-hbi.de



Dr. Lisa Merten
Leibniz Institute for Media
Research | Hans Bredow Institute

l.merten@leibniz-hbi.de



Prof. Dr. Matthias C.
Kettemann, LL.M. (Harvard)
Alexander von Humboldt Institute
for Internet and Society
matthias.kettemann@hiig.de



Prof. Dr. Cornelius
Puschmann
University of Bremen

puschmann@uni-bremen.de

References I

- Arora, S. D., Singh, G. P., Chakraborty, A., & Maity, M. (2022). Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183, 121942. <https://doi.org/10.1016/j.techfore.2022.121942>
- Aruguete, N., Calvo, E., & Ventura, T. (2023). News by Popular Demand: Ideological Congruence, Issue Salience, and Media Reputation in News Sharing. *The International Journal of Press/Politics*, 28(3), 558-579. <https://doi.org/10.1177/19401612211057068>
- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216–9221. <https://doi.org/10.1073/pnas.1804840115>
- Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*, 14614448241253138. <https://doi.org/10.1177/14614448241253138>
- Bitton, D. B., Hoffmann, C. P., & Godulla, A. (2024). Deepfakes in the context of AI inequalities: Analysing disparities in knowledge and attitudes. *Information, Communication & Society*, 1–21. <https://doi.org/10.1080/1369118X.2024.2420037>
- Bleyer-Simon, K., & Reviglio, U. (2024). Defining Disinformation across EU and VLOP Policies. European Digital Media Observatory. <https://edmo.eu/wp-content/uploads/2024/10/EDMO-Report-%E2%80%93-Defining-Disinformation-across-EU-and-VLOP-Policies.pdf>
- Bruns, A. (2019). Are filter bubbles real? Polity Press.
- Cheruiyot, D. (2024). Comparing Risks to Journalism: Media Criticism in the Digital Hate. *Digital Journalism*, 12(3), 294–313. <https://doi.org/10.1080/21670811.2022.2030243>
- Citrin, J., & Stoker, L. (2018). Political Trust in a Cynical Age. *Annual Review of Political Science*, 21(Volume 21, 2018), 49–70. <https://doi.org/10.1146/annurev-polisci-050316-092550>
- Dreyer, S., Stanciu, E., Potthast, K. C., & Schulz, W. (2021). Desinformation. Risiken, Regulierungslücken und adäquate Gegenmaßnahmen. Landesanstalt für Medien NRW.
- Figà Talamanca, G., & Arfini, S. (2022). Through the Newsfeed Glass: Rethinking Filter Bubbles and Echo Chambers. *Philosophy & Technology*, 35(1), 20. <https://doi.org/10.1007/s13347-021-00494-z>
- Ghai, A., Kumar, P., & Gupta, S. (2024). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, 37(2), 966–997. <https://doi.org/10.1108/ITP-10-2020-0699>
- Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., ... Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656), 398–404. <https://doi.org/10.1126/science.abp9364>
- Hameleers, M., & Marquart, F. (2023). It's Nothing but a Deepfake! The Effects of Misinformation and Deepfake Labels Delegitimizing an Authentic Political Speech. *International Journal of Communication*, 17, 6291–6311.

References II

- Hashmi, A., Shahzad, S. A., Lin, C.-W., Tsao, Y., & Wang, H.-M. (2024). Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2405.04097>
- HateAid, Koch, L., Voggenreiter, A. & Steinert, J. I. (2025). Angegriffen & alleingelassen. Wie sich digitale Gewalt auf politisches Engagement auswirkt. Ein Lagebild. <https://hateaid.org/neue-studie-politisch-engagierte-und-digitale-gewalt/>
- Hooghe, M., & Marien, S. (2013). A Comparative Analysis of the Relation Between Political Trust and Forms of Political Participation in Europe. *European Societies*, 15(1), 131–152. <https://doi.org/10.1080/14616696.2012.692807>
- Kessler, S. H. (2023). Vorsicht #Desinformation: Die Wirkung von desinformierenden Social Media-Posts auf die Meinungsbildung und Interventionen.
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1), 69. <https://doi.org/10.1186/1748-5908-5-69>
- Levi, M., & Stoker, L. (2000). Political Trust and Trustworthiness. *Annual Review of Political Science*, 3(Volume 3, 2000), 475–507. <https://doi.org/10.1146/annurev.polisci.3.1.475>
- Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, 11, 143296–143323. <https://doi.org/10.1109/ACCESS.2023.3342107>
- Rau, J. P., & Stier, S. (2019). Die Echokammer-Hypothese: Fragmentierung der Öffentlichkeit und politische Polarisierung durch digitale Medien? *Zeitschrift für Vergleichende Politikwissenschaft*, 13(3), 399–417. <https://doi.org/10.1007/s12286-019-00429-1>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024.
- Seeger, C., Frischlich, L., Obermaier, M., Schmid, U., Riesmeyer, C. & Menke, M. (2024). Die dunkle Seite der Wissenschaftskommunikation – Erfahrungen von Kommunikationswissenschaftler:innen mit inzivilen Angriffen. CHARMS Report #1, online verfügbar unter: <https://osf.io/xcrkp/>
- Sippy, T., Enock, F., Bright, J., & Margetts, H. Z. (2024). Behind the Deepfake: 8% Create; 90% Concerned. Surveying public exposure to and perceptions of deepfakes in the UK (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2407.05529>
- Tajrian, M., Rahman, A., Kabir, M. A., & Islam, Md. R. (2023). A Review of Methodologies for Fake News Analysis. *IEEE Access*, 11, 73879–73893. <https://doi.org/10.1109/ACCESS.2023.3294989>
- Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2401.11817>
- Zmerli, S. (2023). Political Trust. In F. Maggino (Hrsg.), *Encyclopedia of Quality of Life and Well-Being Research* (S. 5278–5281). Springer International Publishing. https://doi.org/10.1007/978-3-031-17299-1_2202