

TWin of Online Social Networks

Deliverable D6.5

Policy Handouts

Main Authors: COSIMA PFANNSCHMIDT, DR. JONAS FEGERT



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), DialoguePerspectives 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



**DIALOGUE
PERSPECTIVES
E.V.**

Project name	TWIn of Online Social Networks
Project acronym	TWON
Project number	101095095
Deliverable number	D6.5
Deliverable name	POLICY HANDOUTS
Due date	31 JANUARY 2025
Submission date	23 JANUARY 2025
Type	DEC – Websites, patent fillings, videos, etc.
Dissemination level	PU – Public
Work package	WP6
Lead beneficiary	4. FZI Forschungszentrum Informatik (Germany)
Contributing beneficiaries and associated partners	DialoguePerspectives e.V. (DIA), Universiteit van Amsterdam (UvA), Universität Trier (UT), Institut Jozef Stefan (JSI), Karlsruher Institut für Technologie (KIT), Robert Koch Institut (RKI), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (UoB), Slovenska Tiskovna Agencija (STA)

Inhaltsverzeichnis

- Tables 1**
- Figures 1**
- Abbreviations 1**
- 1. Policy Recommendations 2**
 - 1.1. The Process of Developing the First Policy Brief 2**
 - 1.2. The First Policy Brief 4**
- 2. The Ethics Policy Brief 8**
 - 2.1. TWON Policy Brief: On the Ethics of Using Twins of Online Social Networks (TWONs) 8**
 - 2.2. An Interactive Ethics Demonstrator11**
- 3. TWON Feedback on EU Legislation12**
 - 3.1. Opinion on the Delegated Act on Art. 40 DSA by the Research Projects DigitS EU and TWON and the Digital Law Institute Trier12**
 - 3.2. Multi-Stakeholder Consultation for Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act18**
- 4. Stakeholder Meetings24**

- Attachment A: Preliminary Policy Brief – Output of the first Dialogue Perspectives Citizen Lab on 16-19th September 2024 in Karlsruhe, Germany25**

Tables

Table 1: Benefits and risks of possible TWON governance modes. (Page 9)

Figures

Figure 1: Sketch of the interactive ethics demonstrator (work in progress). (Page 11)

Abbreviations

AI – Artificial Intelligence

CERN – European Organization for Nuclear Research

DSA – Digital Services Act

ELW – Dialogue Perspectives European Leadership Workshop

OSN – Online Social Network

TWON – Twin of Online Social Networks

1. Policy Recommendations

1.1. The Process of Developing the First Policy Brief

TWONs offer crucial possibilities to measure and simulate the effects of online social networks. However, even the basic communication of the concept of a digital twin of an online social network, or, in short, a TWON, presents challenges, let alone translating interdisciplinary findings into actionable insights for decision-makers in politics and industry. To make an impact with our research, we want to face this challenge and develop policy recommendations based on scientific results. Additionally, we want to foster digital citizenship and the public debate on the role Online Social Networks should play in our society – and we want to take up citizen perspectives in the process of developing policy recommendations. This is why Citizen Labs play an important role in our project. In discussions between scientists from our project and citizens, we developed a draft of the first policy brief, which we then reviewed with our experts from the consortium in workshop in Dubrovnik, as well as in a later feedback process.

At the first Citizen Lab in Karlsruhe in September 2024, participants and the public engaged in workshops, lectures, a BarCamp and a World Café session on the influence of online social networks on democracy. With input from experts, they explored topics such as the dynamics of social media, the EU's Digital Services Act and media literacy. Participants worked together to produce a policy brief on necessary regulatory action, combining their learning with practical insights. The process was guided by experts from the TWON consortium, including Prof. Dr. Damian Trilling (workshop on “The Limits of Research on Social Media Dynamics”), Prof. Dr Achim Rettinger (workshop on “The Role of Social Media and AI in the Confluence of Real-Life Crisis and Digital Democracy - A Technical Perspective”) and Dr. Eugen Pissarskoi (workshop on “Can Two Wrongs make a Right? – An Ethical Reflection on the Idea of Creating Twins of Online Social Networks?”). The workshops also benefited from the insights of external collaborators from “ISD Institute for Strategic Dialogue” such as Melanie Döring, Project Coordinator "Digital Policy Lab", and Marisa Wengeler, Senior Educator Business Council for Democracy, who held workshops on “The Possibilities and Limitations of the EU’s Digital Services Act” as well as “Pre- & De-bunking in the Online Realm”, ensuring a well-rounded approach to the development of recommendations.

Following the Citizen Lab, the recommendations were then discussed extensively within the TWON consortium during a workshop at the consortium meeting in Dubrovnik in October 2024. By doing so, we ensured that the recommendations were comprehensive and practical and covered ideas from our diverse research fields. The workshop was supported by Judith Peterka, Member of the TWON Advisory Board and Advisor at the German Federal Chancellery for AI.

The combination of academic perspectives with the insights of DialoguePerspectives participants and the public enriched the policy recommendation process. While scientific research provides empirical and theoretical foundations, DialoguePerspectives' participatory approach ensures that the recommendations are comprehensible, linked to the ongoing public debate and reflect the lived experiences of the diverse European communities. This interplay enhances the relevance and applicability of our policy advice, which we will communicate to policymakers and industry leaders in the process of our project.

The final version of the resulting policy brief can be found in the following section 1.2. The original version resulting from the first Citizen Lab can be found in Attachment A.

The Citizen Labs are conducted by TWON consortium member “DialoguePerspectives. Discussing Religions and Worldviews e.V.” within the framework of the DialoguePerspectives program. DialoguePerspectives trains young European leaders in the sciences, culture, politics and business to become experts in a new, society-oriented interreligious-worldview dialogue. DialoguePerspectives contributes significantly to understanding and cooperation in Europe, strengthens and defends European civil society and strives to shape a pluralistic, democratic and solidary Europe. The program brings together participants from diverse communities and backgrounds, encompassing individuals with 19 different religions and beliefs across 25 European countries. Through their unique perspectives and expertise, they contribute to fostering understanding, cooperation, and a pluralistic, democratic, and cohesive Europe.

Within its commitment to TWON, DialoguePerspectives integrates its established expertise in fostering pluralistic dialogue and combating societal polarization with a focused emphasis on digital democracy, hate speech, and disinformation. The program has a proven track record in formulating actionable calls to action and comprehensive policy briefs, as demonstrated through its European Leadership Workshops on topics such as “Plurality & Anti-Discrimination in the Workplace” and events like “Entering the Engine Room: Policy Briefs as a Means of Forging a Pluralistic Europe.” These initiatives include developing policy recommendations aimed at advancing a cohesive and pluralistic Europe, methods and skills it employs within TWON. In this context, DialoguePerspectives has prioritized educating participants on the dynamics of online platforms, the role of AI, and strategies like pre-bunking and de-bunking to combat disinformation — an essential step toward strengthening and promoting a pluralistic and democratic European society. These efforts underscore the program's expertise and capacity to develop actionable policy recommendations for TWON as a part to contribute to shaping a cohesive Europe. A podcast episode on democracy in the digital age with Dr. Jonas Fegert (FZI Forschungszentrum Informatik) was recorded on the role of platforms in European democracies.

1.2. The First Policy Brief

Executive Summary

In the digital age of individualization and digitalization, communities have grown further apart and digital communication has partly replaced connections within one's physical neighborhood. The promise of wider networks facilitated by social media is often at odds with the aim of generating in-depth, meaningful deliberation where people sincerely listen and learn from each other. Democracies face significant challenges, such as the spread of disinformation, hate speech, and polarization on digital platforms. As in the case of the 2024 presidential election in Romania, foreign actors such as the Russian government, are furthermore suspected of influencing elections by supporting a candidate's campaign on TikTok. It has been warned that these issues erode public trust, undermine the inclusiveness of democratic decision making and the learning potential of our societies. Despite these threats, digital platforms also present opportunities for increased democratic participation. This policy brief outlines strategies to build a resilient digital democracy that can mitigate the risks posed by digital platforms driven exclusively by economic motives, while enhancing opportunities for engagement. Key areas include the regulation of platforms, establishment of public/independent participatory platforms, and improvement of media literacy.

Underregulated Digital Platforms are Endangering Democracy

The Arab Spring demonstrated the democratic power of online social networks. However, changing digital media environments have created dependence on a few dominant platforms (Luca and Bazerman 2020) and their engagement-driven business models (Srnicsek 2016). For example, Twitter, now X, was once known for content moderation but has seen a rise in disinformation and hate speech, such as antisemitism, following its takeover by Elon Musk (Miller et al. 2023). Despite boycotts by advertisers over the placement of ads near harmful content, X responded with a lawsuit rather than with content moderation.

While the empirical evidence is mixed, it has been warned that social media news feeds, driven by recommendation systems, contribute to polarization by reinforcing ideological bubbles and amplifying negative emotions, which lead to affective polarization. While platforms are aware of these effects, they resist changes to their algorithms, as this might endanger their business model (Ludwig et al. 2023; Bojic, 2024). The dependency on platforms, the spread of hate speech, and algorithmic bias underscore the need for political action.

The European Commission (2018) recognized the urgency of combating disinformation, especially as AI tools further amplify the problem. How can we address disinformation and polarization driven by platform providers?

Policy Options:

1. **Self-Regulation by Platforms:** Self-regulation has had limited success, and challenges remain in enforcing content moderation and addressing privacy concerns related to bot and user verification.
2. **Regulation of Platform Algorithms:** Platforms must make their algorithms transparent and subject to external audits to ensure accountability and prevent harmful feedback loops. To address the challenges of content diversity and balance in online algorithmic recommendations, algorithms should be intentionally designed to deliver a curated mix of content.
3. **Public/Independent Participatory Platforms:** Decentralized platforms like Mastodon could provide more inclusive discourse, relying on pro-democratic and transparent algorithms. Publicly funded non-state entities, such as the Wikimedia Foundation, could manage these platforms to avoid state misuse.
4. **Improving Media and Data Literacy:** With AI-generated disinformation rising, targeted campaigns to improve media literacy are essential. Explainable AI tools can help users assess content validity in real time.
5. **Support for Independent Media:** EU-level funding mechanisms could support independent journalism, which is key to countering misinformation and fostering public discourse.

Policy Recommendations

1. Fund Research on OSN

To understand the complex mechanisms and effects of OSN, researcher access to platform data (such as under the DSA) is necessary, but also sufficient funds to conduct the research. We need research quantifying undesired effects like opinion polarization, affective polarization, falsehood dissemination and the impact of foreign powers. Additionally, we need to investigate how design decisions of OSN can lead to undesired effects on citizens and societies.

To enable this research, the researcher access to platform data under the DSA needs to be fully implemented and simplified, as stated more in detail in section 3.1..

2. Fund the Development of Public Platforms

Allocate funding to support the development of independent digital platforms that are aligned with EU standards and promote inclusive participation.

Alternatively, create mandatory public safe spaces within existing platforms, where only authenticated and high-quality regional content is displayed. These spaces could be regulated and maintained by public, non-profit actors, while at the same time using the existing infrastructure, where users profit from network effects.

Moreover, e-participation tools should be used to enable meaningful participation of citizens in local debates and actual decision-making at different levels of the state.

3. Enforce Interoperability Between Platforms

Similar to industry standards, force big platforms to ensure interoperability. This way, people are free to engage with their platform of choice, not forced towards the one with the biggest user base. This is an important base for building successful alternatives to existing platforms.

4. Increase Transparency and Accountability

Platforms must be mandated to publish detailed reports on their content moderation practices and provide external access to their algorithms to ensure they are not promoting disinformation and emotionalized content. Art. 40 DSA is a good basis, but researcher access needs to be simplified in practice (see section 3.1.).

5. AI-Driven Content Verification

Invest in the development of technologies such as explainable AI to classify and debunk disinformation in real time, ensuring that users are informed about the credibility of the content they engage with.

6. Promote Media Literacy

Launch media literacy campaigns targeting various age groups, with a particular focus on empowering individuals to recognize and counteract disinformation and manipulation by AI tools.

7. Support Independent Journalism

Create a European fund to support independent media outlets that adhere to high-quality standards. This is important to foster fact-based information even on OSN and to alleviate the economic pressure on media outlets caused by the shift of advertisement budgets away from press and towards OSN. This fund should be managed by an independent body to ensure transparency and accountability.

8. Promote Content Diversity through Algorithm Design

Algorithms should be designed to deliver a curated mix of content that balances emotional tones, to avoid negative bias, and introduces users to various topic areas and political viewpoints, fostering a more inclusive and creative digital environment. To enhance personalization while preventing echo chambers, users should be empowered with customizable settings that allow them to adjust their content diversity preferences. This promotes personalized digital autonomy while ensuring a baseline exposure to differing perspectives. Tech companies, as key players in this model, should adhere to agreed standards in algorithm design and transparently demonstrate their contributions to collective decision-making processes (Bojic, 2024).

Sources

Bojic, L. (2024). AI alignment: Assessing the global impact of recommender systems. *Futures*, 160, 103383. <https://doi.org/10.1016/j.futures.2024.103383>

European Commission (2018). Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Action Plan against Disinformation.

House of Participation (2023). A Taxonomy for Involvement Projects. <https://hop.fzi.de/taxonomy>.

Luca, M., & Bazerman, M. H. (2020). Want to Make Better Decisions? Start Experimenting. *MIT Sloan Management Review*, 61(4), 67-73.

Ludwig, K., Grote, A., Iana, A., Alam, M., Paulheim, H., Sack, H., Weinhardt, C. & Müller, P. (2023). Divided by the algorithm? The (limited) effects of content- and sentiment-based news recommendation on affective, ideological, and perceived polarization. *Social Science Computer Review*, 41(6), 2188-2210.

Miller, C., Weir, D., Ring, S., Marsh, O., Inskip, C., & Chavana, N. P. (2023). Antisemitism on twitter before and after Elon Musk's acquisition. Institute for Strategic Dialogue.

Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.

Srnicek, N. (2017). Platform capitalism. John Wiley & Sons.

Taylor, C., Nanz, P., & Taylor, M. B. (2020). *Reconstructing Democracy: How citizens are building from the ground up*. Harvard University Press. TWONs

2. The Ethics Policy Brief

The TWON which is being developed in our project is a potentially highly effective and influential tool. Therefore, it is crucial to make an ethics assessment on the development and further usage of the tool. This has been done in “Deliverable D7.2 – Impact Assessment and Ethical Guidance Handbook” (Main Authors: Dr. Eugen Pissarskoi and Prof. Dr. Michael Mäs, KIT).

However, we came to the conclusion, that the key findings of the report need to be summarized in a well-digestible form for policymakers. Therefore, we created an Ethics Policy Brief, which can be found in the section below.

2.1. TWON Policy Brief: On the Ethics of Using Twins of Online Social Networks (TWONs)

Many warn that online social networks such as Facebook, X (Twitter) or Telegram are contributing to worrying social dynamics such as the polarization of opinion, the spread of fake news, conspiracy theories, discrimination and large-scale collective outrage. However, demonstrating that online social networks have contributed to the emergence of the undesired outcomes has proven elusive. Scientific reviews of research on the impact of filter bubbles have indeed yielded inconclusive findings, with arguments and evidence supporting both sides of the debate. Tech companies, therefore, find it easy to sidestep all allegations.

There is a way to **overcome this responsibility ping-pong** by creating an analogous technology: Digital TWins of ONline social networks, **TWONs**. These highly advanced and realistic computer models mimic an original online social network as closely as possible. This makes it possible to quantify the extent to which an online social network, as well as specific algorithms, yield undesirable outcomes. Furthermore, they offer a means to optimize the design of online social networks with respect to social, ethical, and epistemic objectives. Accordingly, TWONs might be a **powerful tool for regulating online social networks**.

On the other hand, taming one technology by creating another can give rise to a number of risks of its own. With regard to TWONs, **societal risks** are immediately conceivable: By leveraging vast datasets about users and by intricately representing user behavior, TWONs have the potential to be used in ways that are detrimental to the interests of individuals and societies alike. This is why we must carefully examine ethical implications of different modes of regulating TWONs, so that decision makers get the tools at hand to make a well-founded decision. There needs to be a public discussion on the usage and regulation of TWONs.

Outcomes in a Nutshell

Our ethical analysis of TWONs, based on the currently available outlook on its benefits and risks, has demonstrated that:

1. There is a plenitude of options for regulating the technology of TWONs, ranging from
 - a. unlimited public access: analogously to available free web-search engines or publicly accessible LLM-chatbots
 - b. to a strictly controlled usage: analogously to the governance of sensible technologies (e.g., CERN, applications in nuclear physics) or of sensible data panels (e.g. SOEP).

2. The risks and benefits of the TWON hinge upon the manner and extent to which access to this technology is regulated.
3. Each mode of governance brings with it distinct societal benefits and risks.

Governance Models

Unlimited Public Access		Approved Researchers Access Only	
Possible Benefits	Risks	Possible Benefits	Risks
- High economic gains	- Intensified undermining of individual autonomy from freely available tools for manipulation, mis- and disinformation	- Control of online social networks	- Turns out to be ineffective for control of online social networks
- Public control of online social networks: prevention or disclosure of manipulation, mis- and disinformation	- Collapse of institutions necessary for democratic governance	- Knowledge gains from better measurement of social realities (though smaller than in free public access)	- Restriction of access to TWONs turns out to be infeasible (the access to a certain group turns out not to be restrictable): hidden undermining of individual autonomy
- Unrestricted knowledge gains from better measurement of social realities	- Reinforcement of existing inequalities in financial and political power		

Table 1: Benefits and risks of possible TWON governance modes.

Given what is now known about the possible consequences of the use of TWON, none of the governance modes discussed in this report – from unrestricted access to access only with the authorization of an authority – can be rejected on sound grounds.

a) We cannot currently rule out the scenario that a free public access to TWONs is the sole means by which deployment of manipulative, mis- or disinformative algorithms on online social networks can be revealed and thereby publicly controlled. If this proves to be the case, this would provide a weighty reason for a free public access to TWONs.

b) Based on current knowledge, it is also possible that restricting access to TWONs to approved researchers is sufficient to control the algorithms of online social networks. This, in turn, would be a weighty reason to restrict TWON's availability to a group of researchers.

Since it is unclear which of these scenarios is more realistic, it is impossible to reject one of the governance modes.

Empirical and Ethical Uncertainties

To enable an informed decision on the use of TWONs, future research and deliberation are needed to resolve uncertainties in the evaluation of different governance modes.

Empirical Uncertainties:

a) Feasibility of regulating access to TWONs: It is currently uncertain if it will be practically possible to restrict access to TWONs. This depends on the complexity of underlying technology (if, after the blueprint for the underlying models has been developed, anybody with sufficient financial and/or computational resources can set up TWONs, it is unlikely that access restrictions will become enforceable) and on the amount of personal data from an online social network needed for a reliable simulation.

b) Availability of alternative means for safeguarding democratic values and enforcing legislation on online social networks: the recently adopted legislation in the EU (namely, the Digital Services Act (DSA) and the Artificial Intelligence Act (AIA)) is designed to regulate the activities of large online platforms. Nevertheless, it is currently unclear to what extent these acts can be enforced. It may be the case that an instrument such as a TWON is required for the two acts to become legally effective.

Ethical Uncertainties:

a) Quantification of the extent to which norms, values, and rights worthy of protection (such as democratic self-determination, individual autonomy, the right to informational self-determination) are jeopardized by online social networks: the decision regarding the governance of TWONs is contingent upon the extent to which these values are threatened by the prevailing online social networks and the diverse governance modes of TWONs. At the moment, no comparisons of the threats are available.

b) How should the differences in the potential benefits and risks of different modes of governance of TWONs be weighed? Unrestricted public access promises the highest economic benefits from TWONs, but this mode of governance is also associated with the highest risks. The stricter the regulation of TWONs is, the lower are as the expected economic benefits as the risks. Assessments of societal risks and benefits are often highly controversial within a society.

Key Takeaways

1. If TWONs turn out to be a technology that requires strict regulation, the research and development process must also be subject to regulatory oversight.
2. At the moment, it is unclear how much a TWON's ability to inform the regulation of online social networks hinges on detailed personal data about individual users. By means of modeling of fictional reality, however, the reliance on personal data can be rigorously quantified. We recommend conducting such an analysis.
3. Additionally, the report lays out the methodology used for the ethical analysis. This methodology – reconstruction and analysis of arguments – allows developers as well as interested stake-holders to reflect on ethical controversies in TWON's research and societal governance.

This policy brief is published on the TWON website under <https://www.twon-project.eu/on-the-ethics-of-using-twons-twon-policy-brief-1/>.

2.2. An Interactive Ethics Demonstrator

Ethical reasoning can be complex. In order to facilitate reflections on the ethical implications of a TWON and possible regulation of a such, we are working on a demonstrator for our website (twon-project.eu) with visualizations of the reasonings. Here, citizens and policymakers can test, which regulation would be suitable under which assumptions. A sketch can be found below.

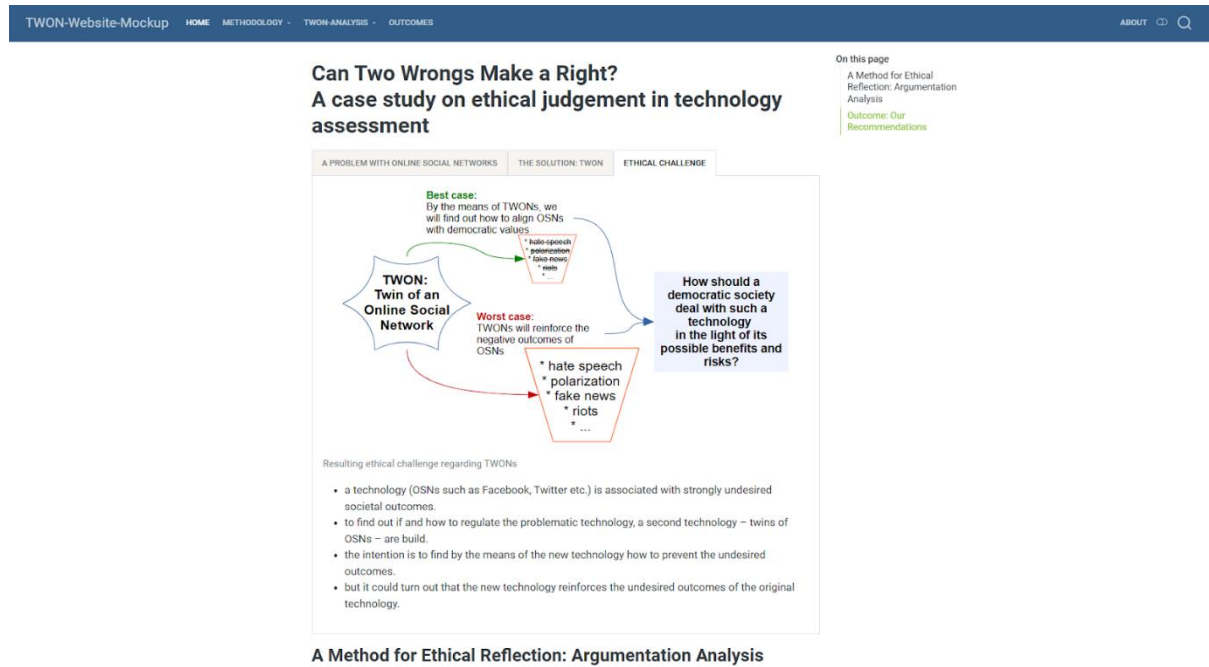


Figure 1: Sketch of the interactive ethics demonstrator (work in progress).

3. TWON Feedback on EU Legislation

Besides formulating policy briefings for policy makers, our consortium also participated in two consultations by the EU Commission: one on the Delegated Regulation on data access provided for in the Digital Services Act (Art. 40 DSA) and one on the Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act. Both submissions took place in December 2024.

3.1. Opinion on the Delegated Act on Art. 40 DSA by the Research Projects DigitS EU and TWON and the Digital Law Institute Trier

[EU Public Consultation on the Delegated Regulation on data access provided for in the Digital Services Act](#)

The opinion has been prepared by the two interdisciplinary research projects DigitS EU and TWON and the Digital Law Institute (IRDT) of Trier University (Germany).

'Digital Sovereignty of Europe' (DigitS EU) is an interdisciplinary research project of Trier University, funded by the research initiative Rheinland-Pfalz 2024-2028. It aims to accompany the implementation of the EU's new digital legal order. The interdisciplinary research network draws on a wide range of expertise from the fields of law, political science and media science as well as computational linguistics, business administration and sinology.

The project 'Twin of Online Social Networks (TWON)' is a Horizon Europe project that develops methods to study the impact of platform mechanics on the quality of public debate. We are developing a so-called digital twin to create a simulation of online social networks and to study such effects. However, to calibrate such a digital twin and to verify conclusions, access to platform data is of paramount importance.

Access to data from VLOPs and VLOSEs is essential for research. It is also essential for providing guidance on how to deal with systemic risks. DigitS EU, TWON, and the Digital Law Institute Trier expressly welcome the access to data granted by Art. 40 DSA and the adoption of a Delegated Act. We appreciate the opportunity to comment on the draft Delegated Act.

In order for Art. 40(4) et seqq. DSA to be effective, it is necessary that the barriers to data access are not artificially raised and that the specific needs of the scientific community are taken into account.

We strongly recommend providing

- simplified procedures to access data for peer review and follow-up research (see 1)
- more specific details to prevent misunderstandings in the application process (see 2-8).

1. Peer Review and Repeatability

Sciences requires the possibility to review and repeat the original research that was made based on access to VLOP/VLOSE data. Thus, it should be possible for other researchers to access the relevant data, for example for the purpose of a peer review process or to review research results in general. A simplified procedure should be open to reviewers in a peer review process and researchers who want to repeat the vetted research project. [1]

The scientific community needs to be able to peer review and replicate the original research that was conducted based on access to VLOP/VLOSE data. It should be possible for other researchers to access the relevant data, for example for the purpose of a peer review process or to review research results in general. A simplified procedure should be available to reviewers in a peer review process and to researchers who wish to repeat the research project that has been reviewed.

Scientific publications go through a peer review process to ensure that scientific standards have been met and that the results are reproducible. In order to check this reproducibility, reviewers in particular need access to the data on which the scientific work is based. Scientific experiments must also be repeatable by other researchers and allow results to be verified, modified or falsified. A large amount of studies show the importance of access to raw data, as researchers can independently come to very different conclusions. Therefore, other researchers need access to the original data to verify the results.

This is particularly important because the data base of VLOPs and VLOSEs changes extremely quickly. Research results are not comprehensible if the subsequent data access concerns a new data base.

Neither Art. 40 DSA nor the Delegated Act provide a specific framework for peer review or the use of follow-up data. We suggest that a new article be created for this purpose.

Such an article could be based on the following principles: Research projects that have already been approved should be published in the DSA data access portal. Peer reviewers and other research groups should be given the opportunity to link themselves to a particular research project, i.e. they can apply to carry out the same or similar research. The DSC should not be required to re-examine the research project as such and should only check that the investigators or new follow-on investigators meet the other requirements. Persons who can demonstrate that they are acting within the framework of a peer review process should be regularly approved, unless there are indications of a risk of abuse. It is also important to ensure that the names of reviewers are not visible in the DSA data access portal, to allow for blind review.

In the Q&A of 19 November, it was reported that if all safeguards are met, access to the data may be possible at a later date. This should be clarified in the Delegated Act.

We therefore propose to insert the following article:

Article 9a

- (1) Reviewers in a peer review process shall be granted access to the data of a vetted research project upon request if they can demonstrate that this is necessary for a peer review procedure regarding the research project of the vetted researcher.
- (2) The Digital Service Coordinator of establishment shall verify that the request includes the following elements:
 - a. the identity and contact details of the reviewer;
 - b. the proposed safeguards to mitigate possible risks in terms of confidentiality, data security and personal data protection corresponding to the data requested, including as regards the modalities of access to and processing of the data;
 - c. an estimation of the required duration of the access.
- (3) The request can be rejected if there are indications of a risk of abuse.

2. Renewed Access for Vetted Researchers

Vetted researchers also need renewed access to the data, even if the research project has been provisionally closed. It should be possible to reopen access to the data.

It may also be necessary for authorized researchers, who have previously accessed data for their original research, to access the data again at a later date, for example if changes are proposed in a peer review process. Even when research projects are (provisionally) closed it may also be necessary to re-access the data, for example if research results are challenged. Therefore, it must be possible to re-access the data again at a future date.

We therefore propose to insert the following article:

Article 9b

A vetted researcher shall be granted renewed access to the data of an already approved research project if this is necessary in the context of the research. The vetted researcher shall submit a request to the Digital Service Provider of establishment.

3. Enabling Pre-Inquiries on Preconditions to Research

Since researchers do not know in advance what kind of data the platforms actually possess researchers should have the opportunity to ask the VLOP/VLOSE for specific information.

Systemic risks and their potential causes cannot always be specified with sufficient clarity in an application under Art. 40 DSA, even if they seem very likely and plausible. For example, it may be reasonable to assume that foreign powers are conducting disinformation campaigns to undermine and manipulate democratic opinion forming in the EU, even if the methods remain unknown (e.g. dissemination of illegal or harmful content, boosting of popularity through inauthentic behavior). Very often, it will be difficult to assess whether a systemic risk exists and what its causes and consequences are.

In order to formulate well-founded inquiries to assess particular systemic risks, researchers require some a priori knowledge of the problem they want to examine and the relevant data sets. Identifying which data is required, however, may require insight into broader dynamics and trends derived from social scientific theories which require data.

It can be problematic for researchers to make a sufficiently specific request to the DSC in order to gain access to this type of data. This is all the more relevant because researchers do not know which parameters the platforms change and which groups are affected. It should therefore be possible to link the research project to questions that the platforms would have to answer.

To take an example from foreign interference in elections: Malevolent actors may try to prevent specific groups of citizens from voting. Such a suspicion is in and of itself not sufficient to formulate an Art. 40 DSA request, because both the messaging and target population remain unclear. Targeted messages may, for example, aim to suppress voter groups through threats, deception, persuasive messages, polarization or other, even currently unknown strategies. They could also attempt to render official communication inefficient through denial-of-service attacks or large volumes of irrelevant content (“noise”, as employed in the PRC). Target populations could consist of ordinary voters, opinion leaders, election officials etc. If Art. 40

DSA is to be effective in enabling research on systematic risks, it needs to facilitate exploratory research aimed at understanding the defining features that are necessary for sufficiently clear requests (including, but not limited to message content, involved user groups, and interaction patterns).

4. Verification of the Data

For reliable research results, it is essential that the data is correct. A control mechanism is therefore needed to ensure that the data has not been manipulated.

Neither the DSA nor the Delegated Act explicitly entail a control mechanism which makes sure that the data provided is correct, i.e. complete, accurate and without errors (e.g. illegible data). Mistakes when providing data have already occurred in the past.[2] As a minimum requirement, VLOPs/VLOSEs should be obliged to provide corrected datasets to all researchers involved if it is verified that the relevant data was wrong. Researchers should be allowed to check the data provided themselves, for example via scraping, and this right should be stipulated in an explicit provision.

5. Systemic Risks

The reference to systemic risks in Art. 34 DSA should not be understood too narrowly. Only a broad understanding can enable basic research capable of determining whether a VLOP/VLOSE concept leads to a systemic risk in the long term.

Research only falls within the scope of Art. 40 DSA if it refers to systemic risks as understood in Art. 34(1), Art. 35 DSA. However, throughout the process of identifying systemic risks, it may occur that data needs to be accessed to enable the verbalization of a particular systemic risk in the first place. Hate and incivility, especially online, are complex and multi-layered. As a result, it is difficult to find new variations of systemic risks if they have not been previously recognized. This would require that more general enquiries could be made.[3] It should be clarified in the form of a recital that basic research on systemic risks is also covered by the provision, as well as research based on initial suspicion that a systemic risk may arise.

Proposal for a new recital

(XX) It is not always possible to predict in advance which algorithms, interfaces or behaviors could lead to a systematic risk. However, corresponding basic research is particularly important for understanding systematic risks. As a consequence, research projects that aim to investigate systematic risks that were not previously recognized should also be authorized.

6. Research Groups

Research groups should be provided with data access for every research member.

It is not clear from the Delegated Act whether only one access to the data must be provided for a research group or whether there is simultaneous access for all researchers. In the Q&A on 19 November, it was

reported that all members of the research group would be granted access. This should be clarified in the form of a recital.

Proposal for a new recital

(xx) Research is mainly carried out in groups. To enable effective research, not only the principal researcher, but all members of the research group should have their own access to the data.

7. Scope of the Obligation to Justify According to Art. 8

It should be clarified in a recital that no excessive requirements are placed on the individual proofs within the framework of Art. 8.

Researchers must meet certain requirements when applying for data access. It has already been criticized in the context of Art. 40 DSA that there are no clear specifications as to how extensive the individual proofs must be.[4] Neither does Art. 8 explain how extensive or detailed the justifications or evidence needs to be. A recital should provide examples of what evidence may be sufficient and emphasize that the requirements are not set too high. Excessively high requirements would only result in unnecessary bureaucracy. A certain basic trust should be placed above all in state or publicly funded research organizations.

Proposal for a New Recital

(XX) The evidence that researchers must provide in accordance with Article 8 should not be unreasonably demanding. The identity of a researcher can usually be proven by a digital copy of an ID or a passport. To prove a formal relationship between the researcher and the research organisation, a simple declaration by the research organisation that the researcher is associated is sufficient.

8. A/B testing

Researchers need access to (results of) A/B tests.

To assess systemic risks on VLOPs/VLOSEs, researchers do not only need observational data but also experimental data. VLOPs/VLOSEs continuously conduct so-called A/B tests, in which a specific feature is changed for one group but not another. In doing so, they can learn how changes in platform mechanics effect outcomes like user engagements. Although A/B tests are often the only way to identify influencing factors that put users at risk, the results of such tests are almost never published. It should be explicitly possible for researchers to request access to lists of the tests that are conducted, the description of these tests, and their results. Preferably, going beyond this, an additional route should be created that allows researchers to initiate a request to have future specific A/B tests conducted, as far as this is ethically and practically feasible.

It should be noted that due to specific affordances and user bases, insights into tests on one VLOP/VLOSEs cannot be generalized for others. It is therefore mandatory to gain access to the specific VLOP/VLOSEs under investigation.

The opinion was prepared on behalf of DigitS EU, TWON and IRDT by:

- Dr. Max Dregelies, DigitS EU, IRDT, University of Trier (Law)
- Pia Diemath, DigitS EU, IRDT, University of Trier (Law)
- Julian Gschneidner, DigitS EU, University of Trier (Media Studies)
- Fabian Hofmanns, DigitS EU, IRDT, University of Trier (Law)
- Prof. Dr. Pascal Jürgens, DigitS EU, University of Trier (Media Studies)
- Prof. Dr. Christian Nuernbergk, DigitS EU, University of Trier (Media Studies)
- Cosima Pfannschmidt, TWON, FZI Karlsruhe (Sociology)
- Prof. Dr. Benjamin Raue, DigitS EU, IRDT, University of Trier (Law)
- Prof. Dr. Achim Rettinger, DigitS EU/TWON, University of Trier (Computational Linguistics)
- Prof. Dr. Damian Trilling, TWON, University of Amsterdam (Communication Sciences)
- Prof. Dr. Antje von Ungern Sternberg, DigitS EU, IRDT, University of Trier (Law)

Sources

[1] Also Klinger, U., & Ohme, J. (2023). What the Scientific Community Needs from Data Access under Art. 40 DSA: 20 Points on Infrastructures, Participation, Transparency, and Funding, p. 6.

[2] For example <https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists>.

[3] Thomas, K., Akhawe, D., et al. (2021). SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. 2021 IEEE Symposium on Security and Privacy (SP), 247–267.

[4] Denga, in Denga/Heinze/Steinrötter, EU Platform Regulation, § 6 mn. 107.

3.2. Multi-Stakeholder Consultation for Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act

Submitted by: 11 December 2024

Questionnaire

Section 1. Questions in Relation to the Definition of an AI System

The definition of an AI system is key to understanding the scope of application of the AI Act. It is a first step in the assessment whether an AI system falls into the scope of the AI Act.

The definition of an 'AI system' as provided in Article 3(1) AI Act is aligned with the OECD definition: 'AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'

Recital 12 provides further clarifications on the definition of an AI system.

The following seven elements can be extracted from the definition:

- 1) 'a machine-based system'
- 2) 'designed to operate with varying levels of autonomy'
- 3) 'may exhibit adaptiveness after deployment';
- 4) 'for explicit or implicit objectives';
- 5) 'infers, from the input it receives, how to generate outputs'
- 6) 'predictions, content, recommendations, or decisions'
- 7) 'can influence physical or virtual environments'

Question 1: Elements of the definition of an AI system

The definition of the AI system in Article 3(1) AI Act can be understood to include the above mentioned main elements. The key purpose of the definition of an AI system is to provide characteristics that distinguish AI systems from 'simpler traditional software systems or programming approaches'. A key distinguishing characteristic of an AI system is its capability to infer, from the input it receives how to generate outputs. This capability of inference, covers both the process of obtaining output in the post-deployment phase of an AI system as well as the capability of an AI system to derive models or algorithms or both from inputs or data at the pre-deployment phase. Other characteristics of an AI system definition such as the system's level of autonomy, type of objectives, and degree of adaptiveness, help to define main elements of the AI system as well as to provide clarity on the nature of the AI system but are not decisive for distinguishing between AI systems and other type of software systems. In particular, AI systems that are built on one of the AI techniques but remain static after deployment triggered questions related to the scope of the AI Act, understanding of the concept of

inference and the interplay between the different characteristics of the AI system definition. The guidelines are expected to provide explanation on the main elements of the AI system definition.

Question: Explain why one or more of these elements require further clarification and what part of this element needs further practical guidance for application in real world applications?

Instead of a technical definition, we argue in favor of a risk-based definition of AI. Relevant criteria are, whether the system is fully controllable, fully explainable, has non-complex behavior and whether the behavior of the system can be traced back to human decisions that can be made accountable.

Additionally, it needs to be considered whether the function in which the tool is used, is risky, as every system can potentially be used for a risky purpose. Also potential unintended consequences must be part of the risk assessment, as AI systems are capable of large-scale personalization and influence in human-like quality.

The explicit exclusion of ‘simpler traditional software systems or programming approaches’ is important, in order to not overregulate tools such as statistical models.

Question 2: Simple software systems out of scope of the definition of an AI system

The AI Act does not apply to all software systems but only to systems defined as 'AI systems' in accordance with Article 3(1) AI Act. According to recital 12, the notion of AI system should be distinguished from ‘simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations’. In particular the use of statistical methods, such as logistic regression, triggered questions related to the conditions under which certain software systems should be considered out of the scope of AI system definition. The Commission guidelines are expected to provide methodology for distinguishing AI systems from simpler traditional software systems or programming approaches and thus would help define systems that are outside the scope of the AI Act.

Question: Please provide examples of software systems or programming approaches that does not fall under the scope of the AI system definition in Article 3(1) AI Act and explain why, in your opinion, the examples are not covered by one or more of the seven main elements of the definition of an AI system in Article 3(1) AI Act.

If the behavior of the system is easily understood by a programmer and/or end-user, especially if it behaves in a deterministic way, the system should not fall within the scope of AI systems in this act. This is in fact nicely illustrated by the enumeration in the text: “In particular the use of statistical methods, such as logistic regression, triggered questions related to the conditions under which certain software systems should be considered out of the scope of AI system definition.” Some examples could be popularity based ranking on social media or in news platforms, the use of collaborative filtering, or similar standard recommendation approaches.

We also believe that systems that are easily recognizable as such should be treated differently from subtle or hidden systems. For instance, if AI is used to distort content while the user does not expect it, this should be different from a system where it is completely obvious that AI is used, and where the user is aware of this.

Section 2. Questions in Relation to the Prohibitions (Article 5 AI Act)

Article 5 AI Act prohibits the placing on the EU market, putting into service, or the use of certain AI systems that can be misused and provide novel and powerful tools for manipulative, exploitative, social control and/or surveillance practices.

The Commission guidelines are expected to include an introductory section explaining the general interplay of the prohibitions with other Union legal acts, the high-risk category and general-purpose AI systems as well as relevant specifications of some horizontal concepts such as provider and deployer of AI systems, 'placement on the market', 'putting into service' and 'use' and relevant exceptions and exclusions from the scope of the AI Act (e.g. research, testing and development; military, defense and national security, personal non-professional activity).

Pursuant to Article 5(1) AI Act, the following practices are prohibited in relation to AI systems:

Article 5(1)(a) – Harmful subliminal, manipulative and deceptive techniques -> relevant for TWON

Article 5(1)(b) – Harmful exploitation of vulnerabilities

Article 5(1)(c) – Unacceptable social scoring

Article 5(1)(d) – Individual crime risk assessment and prediction (with some exceptions)

Article 5(1)(e) – Untargeted scraping of internet or CCTV material to develop or expand facial recognition databases

Article 5(1)(f) – Emotion recognition in the areas of workplace and education (with some exceptions)

Article 5(1)(g) – Biometric categorisation to infer certain sensitive categories (with some exceptions)

Article 5(1)(h) – Real-time remote biometric identification (RBI) in publicly accessible spaces for law enforcement purposes (with some exceptions)

This section includes questions on each of the aforementioned prohibitions separately and one final question pertaining to all prohibitions alike and the interplay with other acts of Union law.

A. Questions in relation to harmful subliminal, manipulative or deceptive practices

The prohibition under Article 5(1)(a) AI Act targets AI systems that deploy subliminal techniques, purposefully manipulative or deceptive techniques that materially influence behaviour of people or aim to do so in significantly harmful ways. The underlying rationale of this prohibition is to protect individual autonomy and well-being from manipulative, deceptive and exploitative AI practices that can subvert and impair individuals' autonomy, decision-making, and free choice.

Proposed structure of the guidelines

It is proposed that the Commission guidelines would cover the following aspects regarding Article 5(1)(a) AI Act:

- *Rationale and objectives of the prohibition*
- *Main elements of the prohibition*
 - *AI systems deploying subliminal, purposefully manipulative and deceptive techniques*
 - *with the objective or the effect of materially distorting behaviour*
 - *in a manner (reasonably likely to) cause significant harm*
- *AI systems out of scope of the prohibition*
- *Interplay with other Union law (e.g. data protection, consumer protection, digital services regulation, criminal law)*

Main elements of the prohibition

Several cumulative elements must be in place at the same time for the prohibition in Article 5(1)(a) AI Act to apply:

1) The activity must constitute 'placing on the market' (Article 3(9) AI Act), 'putting into service' (Article 3(11) AI Act), or 'use' of an AI system (Article 3(1) AI Act). The prohibition applies to both providers and deployers of AI systems, each within their own responsibilities.

2) The AI system must 'deploy subliminal techniques beyond a person's consciousness (e.g. deploying imperceptible images or audio sounds), purposefully manipulative (e.g. exploiting cognitive biases, emotional or other manipulative techniques) or deceptive techniques' (e.g. presenting false and misleading information to deceive individuals and influence their decisions in a manner that undermines their free choices). These techniques are alternative, but they can also apply in combination.

3) The techniques deployed by the AI system should have the objective or the effect of materially distorting the behaviour of a person or a group of persons. The distortion must appreciably impair their ability to make an informed decision, resulting in a decision that the person or the group of persons would not have otherwise made. This requires a substantial impact whereby the technique deployed by the AI system does not merely influence a person's (or group of persons) decision, but should be capable of effectively undermining their individual autonomy and ability to make an informed and independent free choice. This suggests that 'material distortion' involves a degree of coercion, manipulation or deception that goes beyond lawful persuasion that falls outside the ban.

4) The distorted behaviour must cause or be reasonably likely to cause significant harm to that person, another person, or a group of persons. In this context, important concepts that will be examined in the guidelines are the types of harms covered, the threshold of significance of the harm and its reasonable likelihood from the perspective of the provider and/or the deployer. 'Significant harms' implies sufficiently important adverse impacts on physical, psychological health or financial interests of persons and groups of persons that can be compound with broader group and societal harms. The determination of 'significant harm' is fact and context specific, necessitating careful consideration of each case's individual circumstances.

For the prohibition to apply, all elements must be in place and there must be a causal link between the techniques deployed, the material distortion of the behaviour of the person and the significant harm that has resulted or is reasonably likely to result from that behaviour.

Question 3: Taking into account the provisions of the AI Act, what elements of the prohibition of harmful manipulation and deception do you think require further clarification in the Commission guidelines?

Please select all relevant options from the list

- placement on the market, putting into service or use of an AI system
- deploying subliminal, purposefully manipulative or deceptive techniques
- with the objective or the effect of materially distorting behaviour of a person or groups of persons
- in a manner that causes or is reasonably likely to cause significant harm
- none of the above

Question: Please explain why the elements selected above require further clarification and what needs to be further clarified in the Commission guidelines?

We recommend clarifying, what falls under “manipulation”, as one of the problems that AI systems in the context of media have, is that they can distort perceptions of public opinion: If a recommender system or algorithmic filtering system consistently overemphasizes a specific viewpoint, a conspiracy theory, a contested (scientific) fact, and hide another, then there is a substantial risk that users get a distorted perception of reality. They can be made to believe that an (extreme) minority viewpoint is widely shared. This can have strong influences on real-world behaviour, such as voting. Does this fall under “manipulation”?

We worry that “purposefully” does not sufficiently cover unintended, yet severe side effects. One example of such being the case of a teenager, who committed suicide after forming a strong bond to a chatbot in Character.ai.

We suggest to re-consider the specific emphasis on “subliminal” techniques. The concept of so-called subliminal messages is highly contested, the empirical evidence is weak and conflicting. The two other points, “deploying subliminal, purposefully manipulative” and/or “deceptive” techniques sufficiently cover the risks. Specifically highlighting “subliminal” only leads to less clarity, given that it the concept itself is contested.

Question 5: Do you have or know concrete examples of AI systems where you need further clarification regarding certain elements of this prohibition to determine whether the AI system is in the scope of the prohibition or not?

Yes

No

Question: Please specify the concrete AI system, how it is used in practice as well as the specific elements you would need further clarification in this regard

In 2019, it emerged that the Netherlands Tax Administration used a biased algorithm detecting childcare benefits fraud, wrongly labeling people as fraudsters. The system targeted those with "a non-Western appearance", low incomes, or dual nationality, penalizing families over a mere suspicion. Thousands of victims faced poverty, suicides occurred, and children were taken into foster care.

Amazon used AI to review job applicants' resumes since 2014. In 2015, they realized its new system was rating candidates for software developer jobs and other technical posts in a gender-biased way. Amazon's AI models were trained on applications submitted to the company over a 10-year period, where most of the applications came from men. The algorithm concluded that male applicants were preferred and penalized resumes that indicated that the applicant was female.

Facial recognition technology helps police identify individuals by comparing images to a database and suggesting potential suspects. However, studies reveal biases in these systems: Algorithms perform poorly at identifying people besides white men due to biased training data and magnified police prejudices, leading to false arrests. In 2020, Robert Williams, a black man, was wrongfully detained by Detroit police, based solely on AI, sparking scrutiny.

In all 3 cases, the AI tools are not meeting all 4 criteria, as they are not purposefully manipulative or use subliminal techniques. Nevertheless, they are very harmful.

4. Stakeholder Meetings

In order to identify relevant stakeholders in the field of Online Social Network regulation, FZI has mapped stakeholders from government (national and supranational level), NGOs, tech-industry, scientific community and media. The consortium feedbacked and supplemented this stakeholder overview during a workshop session at the Dubrovnik consortium meeting in October 2024. The key stakeholders identified were at the EU Commission (DG Connect and the AI Office), Members of the European Parliament in the IMCO and LIBE committees, as well as NGOs in the field of digital policy and German representatives of government and parliament.

Meetings have taken place with Sergey Lagodinsky (MEP, Greens/EFA), Matthias Spielkamp and Oliver Marsch (AlgorithmWatch), Svea Windwehr (Electronic Frontier Foundation), Josef Holnburger (CeMAS), Josephine Ballon (HateAid), Petra Olschowski (Minister of Science, Research and Arts of the State of Baden-Württemberg) as well as with representatives of the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth. The Federal President of Germany's new stance on the digital public sphere was developed with the support of the TWON consortium and informed by the presented policy recommendations.

Additionally, our consortium member Prof. Dr. Achim Rettinger is part of the Forum on Information and Democracy's "[Working Group on Artificial Intelligence and its Implications for the Information and Communication Space](#)".

Prof. Achim Rettinger, Prof. Christof Weinhardt and Dr. Jonas Fegert have jointly written an opinion-editorial on AI regulation in the German Newsletter Der Tagesspiegel ("[Macht und Machenschaften der KI: „Technologien dürfen gesellschaftliche Konflikte nicht verschärfen“](#)").

Consortium member Dr. Ljubisa Bojic is member of an independent [Foresight Expert Panel](#) established by the United Nations Environment Programme (UNEP) in cooperation with the International Science Council, to support identify and evaluate global emerging issues and signals of change, one of them being technology related mental health crisis. He has written for [Human Futures magazine](#), discussing topics such as AI Observatory, complex testing in virtual reality (CERN for AI), and regulated diversity in recommender systems as necessary steps towards safe, robust, and accountable AI. Finally, Dr. Bojic is member of a working group on AI Act formed by the Serbian Ministry for Science, Technological Development and Innovation.

Attachment A: Preliminary Policy Brief – Output of the first Dialogue Perspectives Citizen Lab on 16-19th September 2024 in Karlsruhe, Germany

Preliminary Policy Brief: Strategies for a Resilient Digital Democracy – From Disinformation to Engagement

Executive Summary

In the digital age, democracies face significant challenges from disinformation, hate speech, and polarization on digital platforms. These issues erode public trust and exacerbate societal divisions. Despite these threats, digital platforms also present opportunities for increased democratic participation. This policy brief outlines strategies to build a resilient digital democracy that can mitigate the risks posed by unregulated digital platforms while enhancing opportunities for engagement. Key areas include the regulation of platforms, establishment of public/independent participatory platforms, and improvement of media literacy.

Unregulated digital platforms are endangering democracy

The Arab Spring demonstrated the deliberative power of online social networks. However, changing digital media environments have created dependence on a few dominant platforms (Luca and Bazerman 2020) and their data-driven business models (Srnicek 2016). For example, Twitter, now X, was once known for content moderation but has seen a rise in fake news and hate speech, such as antisemitism, following its takeover by Elon Musk (Miller et al. 2023). Despite boycotts by advertisers over the placement of ads near harmful content, X responded with a lawsuit rather than regulatory measures.

Social media news feeds, driven by recommendation systems, contribute to polarization by reinforcing ideological bubbles and amplifying negative emotions, which lead to affective polarization. While platforms are aware of these effects, they resist changes to algorithms that benefit their business (Ludwig et al. 2023). The resulting dependence on platforms, the spread of hate speech, and algorithmic bias underscore the need for political action.

The European Commission (2018) recognized the urgency of combating disinformation, especially as AI tools further amplify the problem. How can we address disinformation and polarization driven by platform providers?

Policy Options:

1. **Self-Regulation by Platforms:** Self-regulation has had limited success, and challenges remain in enforcing content moderation and addressing privacy concerns related to bot and user verification.
2. **Public/Independent Participatory Platforms:** Decentralized platforms like Mastodon could provide more inclusive discourse, free from algorithmic bias. Publicly funded non-state entities, such as the Wikimedia Foundation, could manage these platforms to avoid state misuse.
3. **Improving Media and Data Literacy:** With AI-generated disinformation rising, targeted campaigns to improve media literacy are essential. Explainable AI tools can help users assess content validity in real time.

4. **Support for Independent Media:** EU-level funding mechanisms should support independent journalism, which is key to countering misinformation and fostering public discourse.
5. **Regulation of Platform Algorithms:** Platforms must make their algorithms transparent and subject to external audits to ensure accountability and prevent harmful feedback loops.

Policy Recommendations

1. **Fund the Development of Public Platforms**
Allocate funding to support the development of independent, regional-level digital platforms that are aligned with EU standards and promote inclusive participation.
2. **Increase Transparency and Accountability**
Platforms must be mandated to publish detailed reports on their content moderation practices and provide external access to their algorithms to ensure they are not promoting disinformation.
3. **Promote Media Literacy**
Launch media literacy campaigns targeting various age groups, with a particular focus on empowering individuals to recognize and counteract disinformation and manipulation by AI tools.
4. **Support Independent Journalism**
Create a European fund to support independent media outlets that adhere to high-quality standards. This fund should be managed by an independent body to ensure transparency and accountability.
5. **AI-Driven Content Verification**
Invest in the development of technologies such as explainable AI to classify and debunk disinformation in real time, ensuring that users are informed about the credibility of the content they engage with.

Conclusion

A resilient digital democracy requires a multi-faceted approach that balances regulation, public platform creation, and enhanced media literacy. By fostering transparency in digital spaces and supporting independent journalism, democratic societies can mitigate the risks posed by disinformation while empowering citizens to engage meaningfully in digital discourse.

Author: Lotta Badenheuer, Participants from DialoguePerspectives ELW 2024

Sources

European Commission (2018). Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Action Plan against Disinformation. Join(2018) 36 Final.

House of Participation (2023). A Taxonomy for Involvement Projects. <https://hop.fzi.de/taxonomy>.

Luca, M., & Bazerman, M. H. (2020). Want to Make Better Decisions? Start Experimenting. *MIT Sloan Management Review*, 61(4), 67-73.

Ludwig, K., Grote, A., Iana, A., Alam, M., Paulheim, H., Sack, H., Weinhardt, C. & Müller, P. (2023). Divided by the algorithm? The (limited) effects of content-and sentiment-based news recommendation on affective, ideological, and perceived polarization. *Social Science Computer Review*, 41(6), 2188-2210.

Miller, C., Weir, D., Ring, S., Marsh, O., Inskip, C., & Chavana, N. P. (2023). Antisemitism on twitter before and after Elon Musk's acquisition. Institute for Strategic Dialogue.

Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.

Srnicek, N. (2017). *Platform capitalism*. John Wiley & Sons.


Taylor, C., Nanz, P., & Taylor, M. B. (2020). *Reconstructing Democracy: How citizens are building from the ground up*. Harvard University Press.





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