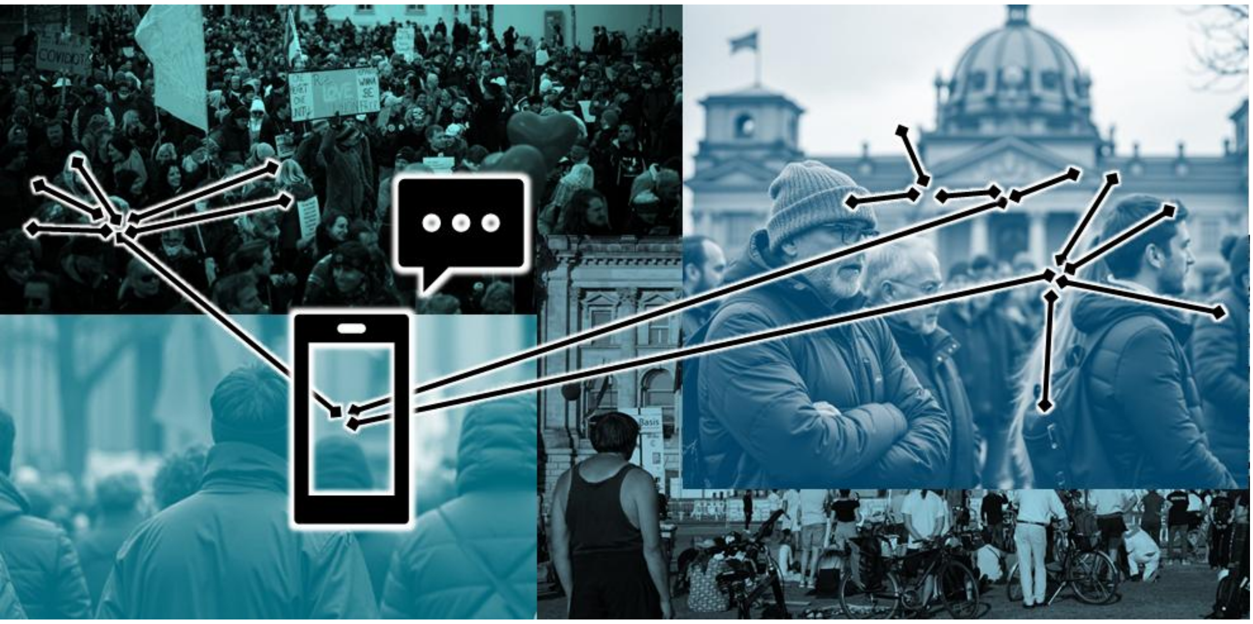
DYNAMO

Policy Paper

Disinformation in Messenger Services

Current Challenges and Recommendations for Legal and Social Measures



SPONSORED BY THE

Federal Ministry
of Education
and Research

Imprint

Authors:
Tahireh Panahi, Carolin Jansen, Amancay Ancina, Katarina Bader, Jeong-Eun Choi, Gerrit Hornung, Nicole Krämer, Lars Rinsdorf, Karla Schäfer, Inna Vogel, York Yannikos and Martin Steinebach

https://www.dynamo.sit.fraunhofer.de/

Contact/Project Coordinator:
Prof Dr Martin Steinebach

Phone: +49 6151 869-349
E-mail: martin.steinebach@sit.fraunhofer.de

Fraunhofer SIT
National Research Centre for Applied Cybersecurity ATHENE
Rheinstraße 75
64295 Darmstadt
Germany

Publication details:
DOI: 10.17185/duepublico/83259

Publishing institution: University of Duisburg-Essen, University Library,
DuEPublico, Universitätsstraße 9-11, 45141 Essen, Germany
https://duepublico2.uni-due.de

1st edition, March 2025

Summary

The spread of disinformation in digital media poses a growing challenge and a serious threat to social peace and democratic processes. Messenger services such as Telegram and WhatsApp, in particular, have developed into platforms on which false information is disseminated on a massive scale. So far, there has been a lack of effective counter-strategies tailored to the complex dynamics of disinformation dissemination on messenger services. In this policy paper, researchers from the fields of computer science, law, psychology and journalism (funded by the german Federal Ministry of Education and Research project DYNAMO) present joint recommendations for action and highlight the need for further research.

Firstly, the paper describes how messenger services are used to create anti-constitutional and state-sceptical counter-publics. In the subsequent analysis of the current legal framework, it is found that existing legal acts, such as the EU's Digital Services Act (DSA), do not address the specific challenges of disinformation in messenger services adequately. This underscores the necessity for supplementary measures to effectively curb the spread of disinformation in messenger services in line with fundamental rights. The policy paper makes concrete proposals for supplementing the DSA that are tailored to the specific requirements of messenger services.

The policy paper centers on the interdisciplinary analysis and evaluation of an approach to prevent the spread of disinformation in messenger services, so-called prebunking. Compared to conventional fact checks, prebunking starts before the actual spread of disinformation, either by providing specific information on individual disinformation content (narrow-spectrum) or by training general media literacy (broad-spectrum). Taking into account current research findings in psychological and communication science, prebunking is critically assessed in terms of its technological and legal feasibility. Although prebunking can be regarded as a strategy that is easy to implement overall and broad-based approaches represent a technically easy-to-implement option for regulating messenger services while protecting fundamental rights, psychological studies suggest that content-specific approaches could be more effective. Further research on the empirical effectiveness and practical applicability in line with fundamental rights is therefore required.

1 Introduction

In the digital world, an increase in the spread of disinformation has been observed in recent years. Disinformation is defined as false factual claims that are disseminated with the intention of causing harm. It is frequently employed to incite mistrust and animosity towards specific groups, organizations, or states, thereby creating social unrest. Establishing effective counter-strategies has therefore become increasingly urgent. Simply holding public social networks accountable does not suffice. Disinformation actors and their followers have long been gathering in messenger services such as Telegram or WhatsApp. With their public and private (group) chats and channels, messenger services offer an opportunity to form large state-sceptical counter-publics and to disseminate ideologies.

This is where the DYNAMO project comes in. We have investigated how messenger services are used to spread disinformation, how disinformation can be detected there, and which approaches exist to counteract disinformation. We have analysed the problem from different scientific perspectives by incorporating and combining expertise from the fields of computer science, law, psychology and journalism.

Based on our analyses and studies, this policy paper presents findings on the dissemination channels of disinformation, actors' strategies, the role of emotions and possible countermeasures. Building on these findings, we show the existing challenges in terms of legal regulation and how important yet difficult it is to differentiate between public and private communication in messenger services. In particular, we refer to the Digital Services Act (DSA), an EU regulation designed to curb the risks posed by the spread of disinformation. The DSA faces increasing criticism from the US for restricting free speech. However, it is crucial that European legal experts — who evaluate it based on EU law — are the key voices in shaping its development. We explain why this regulation has so far been insufficiently effective against disinformation within the hybrid communication structures of messenger services. We then formulate proposals for supplementary legal measures and suggest further strategies for combating disinformation. We focus on so-called prebunking.[1] This refers to preventive measures that are taken before the spread of disinformation and either provide specific information on individual types of disinformation or train general media skills. Finally, we assess the suitability of prebunking from our various scientific perspectives and highlight the need for further research.

[1] Roozenbeek, J., & van der Linden, S. (2019): Fake News Game Confers Psychological Resistance Against Online Misinformation. *Palgrave Commun 5*(1), pp. 1-10. Available at https://www.doi.org/10.1057/s41599-019-0279-9 (accessed on 17/07/2024) and Lewandowsky, S., & van der Linden, S. (2021) Countering Misinformation and Fake News Through Inoculation and Prebunking, *European Review of Social Psychology, 32*(2), pp. 348-384. Available at https://www.doi.org/10.1080/10463283.2021.1876983 (accessed on 19/07/2024).

2 Problem Analysis

2.1 Use of Messenger Services as Infrastructure for Communitising State-Sceptical Counter-Publics

Messenger services are an ideal environment for the formation of anti-constitutional and state-sceptical counter-publics[2], for example in terms of spreading disinformation, real-world networking and the radicalisation of groups with an affinity for disinformation.[3] This poses a challenge to social cohesion and representative democracy, particularly due to its compartmentalised nature and frequent lack of counter-speech, as it is not possible to mediate between different interests or reach compromises without common spaces for discourse. On the one hand, state-sceptical actors spreading disinformation use messenger services because some service providers such as Telegram impede law enforcement due to their passivity. On the other hand, messenger services' technical communication features (Affordances) - especially those by Telegram - are ideally suited to building state-sceptical counter-publics.[4] This includes the fluid transition between individual, group and mass communication:

Individual communication (one-to-one) takes place when only two users communicate with each other (private conversation, comparable to communication via SMS). In the Telegram example, group communication can be limited to small groups, such as private family chats, but it can also take place in much larger spaces, in both private and publicly accessible groups of up to 200,000 members (few-to-few, many-to-many). While communication between members is possible and desired in groups, the channel feature represents a further dimension of specific communication: Telegram channels allow channel operators to reach an unlimited number of subscribers (one-to-many) and spread their own narratives without the option of counter-speech. Other messenger services such as WhatsApp and Signal have similar features, although they differ in some respects (e.g. limiting the number of group members).

Curating content on Telegram channels has proven to be a key practice of those who network in counter-publics on Telegram: Channel operators blend their own content with forwarded content and often employ disinformation as a central element of complex alternative constructions of reality. Divergent realities from quality media and "alternative" media are often combined and thus, links are created between different ideologies and milieus.

[2] Anders: Jungherr, A., & Schroeder, R. (2021). Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy. *Social Media + Society, 7*(1), pp. 1-13. Available at https://doi.org/10.1177/2056305121988928 (accessed on 23/07/2024).

[3] Counter-publics are considered here when they act in an anti-constitutional and disinformational manner (see also Schulze, H., Hohner, J., Greipl, S., Girgnhuber, M., Desta, I., & Rieger, D. (2022). Far-Right Conspiracy Groups on Fringe Platforms: A Longitudinal Analysis of Radicalisation Dynamics on Telegram. *Convergence: The International Journal of Research into New Media Technologies, 28*(4), pp. 1103-1126. Available at https://doi.org/10.1177/13548565221104977 (accessed on 25/07/2024).

[4] See also Schulze, H., Greipl, S., Hohner, J., & Rieger, D. (2024). Social Media and Radicalisation: An Affordance Approach for Cross-Platform Comparison. *M&K 72*(2), pp. 187-212. Available at https://www.nomos-elibrary.de/10.5771/1615-634X-2024-2-187/social-media-and-radicalization-an-affordance-approach-for-cross-platform-comparison-jahrgang-72-2024-heft-2?page=1 (accessed on 25/07/2024).

Such actors use both locally networked small groups, i.e. channels that serve the organisation of real-world networking (few-to-few communication) and large groups (many-to-many communication) as well as the channel feature (one-to-many communication). Telegram thus allows actors to form local communities. These (often also local) groups contribute to the construction of (state-sceptical) counter-publics by promoting topics and discussions that are relevant in their thematic and/or geographical context.[5] These state-sceptical counter-publics emerging through the dissemination of disinformation thus lead to an increasing loss of trust in state institutions and an increasing polarisation of society via messenger services, both in the digital space and through the networking that takes place in the real world.

Their broadcasting feature deems channels to have a public impact per se,[6] especially if they attain a high reach. However, this is not essential in order to have a public impact.[7] Focusing on quantitative metrics alone is therefore not sufficient since the formation of counter-publics on messenger services can also depend on other factors arising from the affordances of the platforms: From follow-up communication (reach) within the messenger services as well as from links to alternative platforms and/or beyond the platforms.

To further determine the extent to which communication through channels can unfold public relevance, the motives and practices of group admins and channel operators must be considered. For this, it is useful to focus on the content that is effectively disseminated. In order to spread disinformation on messenger services such as Telegram and to achieve publicity, journalistic production routines are often adopted and used strategically.[8] This includes, in particular, emotionally charged representations and content suitable for triggering community-building effects. Our analyses[9], focused

[5] In this context, so-called "engagement farming" is also relevant - a strategy that aims to artificially increase one's presence on social media through various tactics in order to increase likes, comments, shares and other forms of interaction. This can take the form of users participating in multiple discussions without offering much added value, tagging numerous users to attract attention or using controversial content to provoke reactions.

[6] The Telegram FAQs on channels state the following: "Channels are a way to send public messages to a large audience, as channels can have an unlimited number of members."

[7] Bader, K.; Müller, K., & Rinsdorf, L. (2023): Zwischen Staatsskepsis und Verschwörungsmythen. Eine Figurationsanalyse zur kommunikativen Konstruktion von Gegenöffentlichkeiten auf Telegram [Between State Scepticism and Conspiracy Myths: A Figuration Analysis of the Communicative Construction of Counter-Publics on Telegram]. In: *M&K, 71* (3-4), pp. 248-265. Available at https://doi.org/10.5771/1615-634X-2023-3-4-248%20 and Rinsdorf, L., Bader, K., & Jansen, C. (2024). Telegram als Plattform für staatsskeptische Akteur:innen Telegram as a Platform for State-Sceptical Actors. In: C. Nuernbergk, J. Haßler, J. Schützeneder, & N. Schumacher (Eds.), *Politischer Journalismus. Konstellationen – Muster – Dynamiken [Political Journalism: Constellations - Patterns – Dynamics]* (pp. 97-108). Baden-Baden: Nomos Verlag. Available at https://www.nomos-elibrary.de/10.5771/9783748939702-131/telegram-als-plattform-fuer-staatsskeptische-akteur-innen?page=1 (accessed on 29/07/2024).

[8] See also Eisenegger, M. (2021). Dritter, digitaler Strukturwandel der Öffentlichkeit als Folge der Plattformisierung [Third, Digital Structural Change of the Public Sphere as a Consequence of Platformisation] (p. 31). In M. Eisenegger, M. Prinzing, P. Ettinger, & R. Blum (Eds.), *Digitaler Strukturwandel der Öffentlichkeit: Historische Verortung, Modelle und Konsequenzen [Digital Structural Change in the Public Sphere: Historical Localisation, Models and Consequences]* (1st edition 2021, pp. 17-39). Springer Fachmedien Wiesbaden. Available at https://www.springerprofessional.de/digitaler-strukturwandel-der-oeffentlichkeit/19030718 (accessed on 29/07/2024) and Hepp, A., & Coudry, N (2023). Necessary Entanglements: Reflections on the Role of a "Materialist Phenomenology" in Researching Deep Mediatisation and Datafication. *Sociologica, 17*(1), 137-153. Available at https://doi.org/10.6092/issn.1971-8853/15793 (accessed on 29/07/2024).

[9] Bader, K.; Müller, K., & Rinsdorf, L. (2023): Zwischen Staatsskepsis und Verschwörungsmythen. Eine Figurationsanalyse zur kommunikativen Konstruktion von Gegenöffentlichkeiten auf Telegram [Between State Scepticism and Conspiracy Myths: A Figuration Analysis of the Communicative Construction of Counter-Publics on Telegram]. In: *M&K, 71* (3-4), pp. 248-265. Available at https://doi.org/10.5771/1615-634X-2023-3-4-248%20 (accessed on 29/07/2024) and Rinsdorf, L., Bader, K., & Jansen, C. (2024). Telegram als Plattform für staatsskeptische Akteur:innen [Telegram as a Platform for State-Sceptical

on the Telegram platform, show, along with many other studies[10], that state-sceptical counter-publics do indeed form on Telegram, which can be categorised into different types of messenger communication. These include (1) **communicating worldviews with a monothematic focus on a key issue**, such as Covid-19 or the Ukraine war, (2) **fostering communities through cultivation of conspiracy narrative perspectives** by communicating current topics from an insider's perspective using QAnon,[11] (3) **generating credibility through seriousness** by means of journalistic communication, (4) **political influencing**, in which a central opinion-driven actor portrays their own personality and creates proximity, (5) **generating attention and reach through linking and referencing** by disseminating content and advertising channels and blogs, and (6) **conspiracy narratives for beginners** in channels that develop conspiracy narrative worldviews and carry out persuasion work.

2.2 The Regulatory Framework and its Legal Challenges

Considering the communication science analyses, effective countermeasures appear to be urgently needed to protect democratic discourse and social cohesion. However, legislation of countermeasures against disinformation in messenger services cannot be implemented easily. There are significant challenges in terms of fundamental rights alone. Messenger services' public and private communication features are subject to different fundamental rights requirements.

2.2.1 EU-Primary Law and Fundamental Rights

Freedom of expression can be particularly relevant with regard to the disseminators of content that (allegedly) constitutes disinformation (Art. 10 para. 1 ECHR, Art. 11 para. 1 GRCh). It is controversial if untrue factual claims that are deliberately made or proven to be untrue are generally afforded protection by freedom of expression (e.g. the German Federal Constitutional Court argues against this). However, freedom of expression does apply when false factual claims are blended with subjective value judgements. Any government measure that prohibits or hinders the expression of opinions could potentially constitute an infringement of the freedom of expression, such as legal obligations to delete and block (alleged) disinformation on messenger services. Even state surveillance of communication in messenger services may infringe freedom of

Actors]. In: C. Nuernbergk, J. Haßler, J. Schützeneder, & N. Schumacher (Eds.), *Politischer Journalismus. Konstellationen – Muster – Dynamiken [Political Journalism: Constellations - Patterns – Dynamics]* (pp. 97-108). Baden-Baden: Nomos Verlag. Available at https://www.nomos-elibrary.de/10.5771/9783748939702-131/telegram-als-plattform-fuer-staatsskeptische-akteur-innen?page=1 (accessed on 29/07/2024).

[10] e.g. Holnburger, J. (29/03/2023). *Chronologie einer Radikalisierung. Wie Telegram zur wichtigsten Plattform für Verschwörungsideologien und Rechtsextremismus wurde [Chronology of a radicalisation: How Telegram became the most important platform for conspiracy believers and right-wing extremism]*. Report by the Centre for Monitoring, Analysis and Strategy (CeMAS). Available at https://cemas.io/publikationen/telegram-chronologie-einer-radikalisierung/ (accessed on 28/06/2024).

[11] QAnon is not a fixed organisation, but rather an idea or legend that has formed as a loose movement on the internet and is becoming increasingly visible offline. The so-called Q-texts are often cryptic and difficult to understand, usually consisting of sentence fragments or questions. A central theme is the myth of a dark, secret elite that allegedly controls the USA through the "Deep State". These statements often contain hidden anti-Semitic insinuations. Real events are often interpreted as evidence for these claims. In Germany, the QAnon theories were initially spread primarily by right-wing extremists and supporters of the Reichsbürger movement. However, with the protests against the coronavirus measures, they also found favour in parts of this new movement.

expression, as opinions can no longer be expressed impartially. Additionally, freedom of expression protects against being forced to disseminate someone else's opinion as one's own as can be the case with (automated) labelling of suspected disinformation, for example.

Recipients of (alleged) disinformation can invoke the fundamental right to **freedom of information** (Art. 11 para. 1 GRCh and Art. 10 para. 1 sentence 2 ECHR). This guarantees the right to obtain information from generally accessible sources without hindrance. Public messenger functions can fall within the scope of this fundamental right. Even sources of information (e.g. channels or groups) that are proven to predominantly disseminate disinformation[12] are generally covered by the broad scope of the protection of freedom of information. Any prevention or significant impediment of access to information sources infringes this fundamental right. Protection is also provided against imposed information, which may include the mandatory labelling of disinformation.

The fundamental right to **private communications** also plays a major role for messenger services (Art. 7 GRCh, Art. 8 ECHR). It protects the confidentiality of communications and the context of communication. Any reading, filtering and evaluation of private communications infringes the secrecy of telecommunications. This fundamental right therefore protects private chats, but not public channels and groups.

The fundamental right to **data protection** includes the right to determine the collection and processing of personal data (Art. 8 para. 1 GRCh and Art. 8 para. 1 ECHR). Legal measures against disinformation in messenger services can infringe the fundamental right in various ways, e.g. by storing, forwarding or otherwise processing data or metadata.

At the same time, the fundamental rights of messenger service providers must be respected, particularly the **freedom to conduct a business** (Art. 16 GRCh). This includes the protection of entrepreneurial activity, which encompasses the disposal of technical resources, among other things. Freedom of contract, e.g. the stipulated general terms and conditions of messenger services, is also fundamentally protected.

In addition to fundamental rights issues, the impact on democracy must be considered (Art. 2, 9-11 TEU). Disinformation can impair the free opinion-forming process, which is considered an integral prerequisite for democracy. Conversely, measures against disinformation can also impair the free opinion-forming process and thus democracy. Practical challenges in the application of the law lie in **proving a harmful intent**.

Furthermore, possible negative **reciprocal effects** of regulation must be considered. It can be problematic if users resort to less regulated areas of services or services that are less willing to co-operate in the event of stricter legal regulation.[13] Such an effect was

[12] e.g. Holnburger, J. (29/03/2023). *Chronologie einer Radikalisierung. Wie Telegram zur wichtigsten Plattform für Verschwörungsideologien und Rechtsextremismus wurde [Chronology of a radicalisation: How Telegram became the most important platform for conspiracy believers and right-wing extremism].* Report by the Centre for Monitoring, Analysis and Strategy (CeMAS). Available at https://cemas.io/publikationen/telegram-chronologie-einer-radikalisierung/ (accessed on 28/06/2024).

[13] Panahi, T., & Zurawski, P. (2023), Messenger & Co: Das Unsichtbare regulieren? [Messenger & Co: Regulating the Invisible?]. In Kemmesies, U., Wetzels, P., Austin, B., Büscher, C., Dessecker, A., Hutter, S., & Rieger, D. (Eds.), MOTRA-Monitor 2022 (pp. 410-429). Available at https://doi.org/10.53168/ISBN.978-3-9818469-6-6_2023_MOTRA [accessed on 18/07/2024).

evident, for example, when distributors of illegal content or content that violates terms and conditions retreated from social networks such as Facebook to messenger services such as Telegram.[14]

Today, there are several specific legal measures intended to combat disinformation (see 2.2.2) that must be assessed in the light of EU primary law and fundamental rights. Whether any **encroachments on fundamental rights** that may be caused by these provisions can be **justified** depends on the drafting of the specific provisions and on the appropriate consideration of relevant fundamental rights.

2.2.2 The new EU Legislation

Combating disinformation has become the subject of new legal acts in recent years. One example is **The Strengthened EU Code of Practice against Disinformation 2022**, a co-regulation,[15] which signatories are asked to comply with primarily on a voluntary basis. In addition, the **regulation on the transparency and targeting of political advertising** and the **Artificial Intelligence Act** cover certain aspects of the topic of disinformation. Arguably the EU's strictest action against disinformation consists of **restrictive measures** (sanctions) against several Russian state media, which were temporarily banned from broadcasting and disseminating due to their persistent disinformation activities.

Above all, however, the **Digital Services Act (DSA)** is intended to serve as a central act for platform regulation in the EU and, as such, to counteract disinformation. Under the new US government, the DSA and its enforcement have faced growing criticism for limiting free speech (e.g., Vice President Vance at the 2025 Munich Security Conference). Some major online platform leaders have also voiced concerns. However, European legal experts, who assess it based on EU-primary law and fundamental rights, should be the key voices in shaping its development.

3 Combating Disinformation in the Digital Services Act

The DSA aims to harmonise the legal framework for intermediary services across the EU and to contribute to a safe, predictable and trusted online environment. In the following, we will analyse the applicability of the DSA to messaging services and its effectiveness against disinformation. We will outline the current legal situation as well as the associated issues and propose solutions.

[14] Jünger, J., & Gärtner, C. (2020). *Datenanalyse von rechtsverstoßenden Inhalten in Gruppen und Kanälen von Messengerdiensten am Beispiel Telegram [Data analysis of Illegal Content in Groups and Channels of Messenger Services Using the Example of Telegram]*. Düsseldorf: Medienanstalt NRW, p. 6. Available at https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Zum_Nachlesen/Telegram-Analyse_LFMNRW_Nov20.pdf (accessed on 18/07/2024).

[15] In the case of co-regulation, the legislator sets targets, but their realisation is left to non-state actors. The strengthened EU Code of Practice against Disinformation is based on guidelines from the EU Commission, which have been implemented by commercial enterprises as well as fact-checking organisations. The current Disinformation Code was published on 16 June 2022 and has so far been signed by 44 parties, including Meta and TikTok.

3.1 Applicability to Messenger Services

An important question is whether messenger services fall within the scope of the DSA. The DSA contains various provisions that are aimed at different types of intermediary services (Art. 3 g) DSA).

3.1.1 Online Platforms

Provisions that may be relevant for combating disinformation apply primarily to hosting services that disseminate information to the public (“online platforms”, Art. 3 i) DSA). This requires making information available to a potentially unlimited number of third parties (Art. 3 k) DSA).

The recitals[16] of the DSA expressly clarify that "private messaging services" do not fall within the scope of the definition of online platforms,”as they are used for interpersonal communication between a finite number of persons determined by the sender of the communication" (recital 14). However, this does not mean that all messenger services fall outside the scope of the provisions for online platforms, since the recitals provide for a **function-related categorisation** of online platforms (recital 15). Accordingly, a distinction must be made between the individual services of a provider as to whether they fall under the provisions of the Regulation (e.g. individual chats, groups, channels). This therefore allows a messenger service to be considered an online platform, at least in part. For example, a Telegram channel without access restrictions with 500,000 subscribers could be considered an online platform, whereas a chat between two people or a closed group with 30 members would not.

However, the application of the DSA to very large but closed groups remains problematic, as there is no "potentially unlimited number" of third parties. Yet, it would be inappropriate not to assume a public sphere if, for example, groups with up to 200,000 members exist on Telegram.[17]

3.1.2 Very Large Online Platforms

Very large online platforms (VLOPs) are subject to special obligations, some of which may relate to the handling of disinformation (Art. 34, 35 DSA). A VLOP is an online platform with an average number of average monthly active recipients of the service in the Union is equal to or higher than 45 million, and which has been designated as VLOP by the EU Commission (Art. 33 para. 1 DSA). In the case of messenger services, determining the required number of users is particularly problematic as it is not yet clear whether only users utilising the service for the public dissemination of information should be counted (Art. 3 i DSA). In this case, each user would have to be checked individually to determine whether the service is used publicly or privately. This would

[16] Recitals are not part of the legally binding legislative text but serve as a legal justification and can be used to interpret the law.

[17] Telegram groups allow up to 200,000 members, so-called Giga groups work completely without a limit on the number of members, see Telegram.org (n.d.): *Questions and answers.* Available at https://telegram.org/faq/de#f-was-ist-der-unterschied-zwischen-gruppen-und-kanalen (accessed on 12/03/2024) and https://core.telegram.org/api/channel (accessed on 28/06/2024).

involve collecting masses of personal data, for which there is currently no legal basis in this constellation.[18] Additionally, usage patterns are subject to considerable fluctuations.[19]

3.1.3 Proposal: Adaption of the Criteria for the "Dissemination to the Public"

We propose adapting the criteria for the "dissemination to the public" (Art. 3 k) DSA). Restrictive and unambiguous criteria must be adopted to avoid any infringement of the fundamental right to private communication (see section 2.2.1). One option would be to set a **high numerical threshold** (e.g. 10,000 members). In addition, the **accessibility of joining** a group or channel should be considered. Allowing users to join a group/channel without access restrictions, e.g. for every user via a publicly accessible invitation link, is an indication of public communication.

3.1.4 Proposal: Specification of Calculation Methods

We also propose adapting the calculation methods for very large online platforms e.g. by means of a delegated act of the EU Commission (Art. 33 para. 3 DSA). As it is currently unclear when a messenger service can be classified as a "very large online platform", the methods for calculating the number of users of messenger services must be specified.

One possibility would be for the calculation criteria to exclude users who do not use the public functions from the calculation as far as possible. But in this scenario a large amount of personal data on user behaviour would have to be processed (automatically), which would not only be challenging practically but also in terms of data protection law (see above). Instead, to avoid breaches of data protection law, it should be made clear that the extensive calculation methods set out in recital 77 DSA are not transferable to messenger services. In contrast, regular user surveys based on declarations of consent under data protection law would be more data protection-friendly, but not a reliable instrument if the information obtained is not verified.

Another simpler option would be to clarify that all users who are registered for the service in a specific timeframe must be included as relevant users in the calculation.[20] This approach would lead to clear and consistent results. At first glance, such an approach seems to contradict the wording of the regulation, which only refers to online platforms and thus to public communication (Art. 33 Abs. 1 DSA). However, most hybrid messenger services allow all users to use the public functions, which is why these

[18] The obligation to grant data access to certain institutions and researchers also only applies if it is clear that the online platform is very large (Art. 40 DSA).

[19] Panahi, T., Hornung, G., Schäfer, K., Choi, J.-E., Steinebach, M., & Vogel, I. (2023), Desinformationserkennung anhand von Netzwerkanalysen – ein Instrument zur Durchsetzung der Pflichten des DSA am Beispiel von Telegram [Disinformation Detection Based on Network Analyses - a Tool for Enforcing the Obligations of the DSA Using the Example of Telegram]. In Friedewald, M., Roßnagel, A. Neuburger, R., Bieker, F., & Hornung, G. (Eds.), *Datenfairness in einer globalisierten Welt [Data fairness in a Globalised World]* (pp. 343-370). Baden-Baden. Available at https://www.nomos-elibrary.de/10.5771/9783748938743-343/desinformationserkennung-anhand-von-netzwerkanalysen-ein-instrument-zur-durchsetzung-der-pflichten-des-dsa-am-beispiel-von-telegram?page=1 (accessed on 18/07/2024).

[20] This does not mean that the provisions of Art. 34 et seq. DSA should then also apply to private communication functions. The proposal only refers to the categorisation of a service as a very large online platform.

should be included in the calculation. This is also in line with the legal definition of the user according to Art. 2 lit. b) DSA, which refers to public communication, but not only. Ultimately, this approach would interfere least with the fundamental rights to data protection and private communication, given that such a calculation could be made by means of an anonymized registration count and would therefore generally require no or little personal data to be processed. Additionally, it would probably be the most resource-efficient option from the companies' perspective.

3.2 Effectivity against Disinformation

The DSA is intended – at least according to the recitals – to counter disinformation (recital 9). Despite this objective, the DSA does not contain a definition of disinformation and does not mention the term disinformation in any provision. Nevertheless, there are some provisions that can (indirectly) counter disinformation to a limited extent.

The DSA contains several repressive provisions relating to the **moderation of illegal content** (e.g. reporting and remedial procedures, Art. 16 DSA, suspension obligation, Art. 23 para. 1 DSA). Disinformation is not illegal in itself. However, certain forms of disinformation may be illegal under EU and Member State law. German criminal law, for example, contains criminal offences that may include disinformation and are therefore relevant under the provisions of the DSA (e.g. defamation under Section 187 StGB, incitement to hatred under Section 130 StGB, in each case in the variant of "denial").

However, the obligations to **assess and minimise risks** set out in the DSA are not only directed against illegal content, but also against "systemic risks" (Art. 34, 35 DSA), which includes disinformation (Recital 84). At least once a year, very large online platforms and search engines must identify, analyse and assess all systemic risks arising from the design or operation of their services and related (algorithmic) systems, or the use of their services. To minimise these risks, they must, if necessary, revise their technical features, algorithms and general terms and conditions.

The DSA also contains preventive measures that can have an indirect effect on combating the spread of disinformation, such as the **transparency** of algorithmic recommendation systems and online advertising (Art. 26, 27 DSA). In addition, general terms and conditions must be transparent and contain information on how content is moderated (Art. 14 para. 4 DSA).

One of the biggest weaknesses of the DSA is that it contains numerous **vague formulations**, which lead to legal uncertainty. On the one hand, this can lead to the passivity of service providers, e.g. by not taking sufficient account of the hybrid structure of their services. On the other hand, service providers might feel obliged to collect masses of personal data and monitor private communications, which must also be prevented. This can be seen, for example, in the open-ended wording of Art. 34 and 35 DSA (e.g. "negative effects on civic discourse”).

3.2.1 Proposal: Clarification of Terms and Definition of Disinformation

As the DSA only contains a few provisions that can (indirectly) act against disinformation, and these provisions include numerous vague formulations, the DSA needs to be adapted and further specified in order to increase its effectiveness against disinformation in messenger services.

In view of the vague wording of Art. 34 and 35 DSA, clarification should be provided as to the negative effects on civic discourse. Since the recitals refer to a threat to civic discourse through disinformation, a definition of the term should be added to the DSA.

The EU Commission could further utilise the option provided by law to issue special guidelines including **best practice examples for messenger services,** which can serve as orientation (Art. 35 para. 3 DSA). The guidelines should clarify that messenger service providers must analyse the significance of their different communication features regarding the systemic risks as part of the risk assessment. For example, they should be obliged to analyse how the potential number of members and membership options could affect the use of the service, and which technical features could favour the formation of problematic counter-publics and radicalisation (See section 2.1).

3.2.2 Interim Conclusion

The legal analysis shows that public communication features of messenger services fall within the scope of the DSA (see 3.1.1). However, communication in very large but closed groups is excluded from its scope, although it can contribute to the formation of problematic counter-publics and radicalisation. Further solutions tailored to the hybrid communication functions of messenger services are therefore required. In addition, the DSA is primarily a risk-based regulation that can have reactive and repressive (indirect) effects against disinformation rather than contributing directly to the media literacy of individual users in the long term. We also see potential for reform here.

4 Prebunking as a Preventive Intervention

One promising measure frequently discussed in the fight against disinformation is **prebunking**. Josep Borrell, the former EU High Representative for Foreign Affairs, mentioned prebunking as a possible strategy to combat disinformation in the election year 2024.[21] In this section, we will explore what prebunking entails and assess its potential to mitigate disinformation in messenger services.

Prebunking is a **preventive measure** that aims to protect individuals from the influence of false information. Unlike debunking, which involves correcting false information retrospectively, prebunking seeks to counter false information in advance.

[21] Heise.de (24/01/2024). EU-Außenbeauftragter: Vergiftete Informationen untergraben die Demokratie [EU High Representative for Foreign Affairs: Poisoned information undermines democracy]. Available at: https://www.heise.de/news/EU-Aussenbeauftragter-Vergiftete-Informationen-untergraben-die-Demokratie-9606550.html (accessed on 27/08/2024).

The concept of prebunking is rooted in McGuire's inoculation theory.[22] [23] According to this theory, people can be "immunised" against false information in a similar way to vaccinations against diseases. A prior confrontation with a weakened version of false information with simultaneous argumentative refutation is intended to activate "mental antibodies" without altering underlying beliefs.[24] This means that peripherally existing beliefs are "attacked" by the microdoses of false information and lead to recipients having to actively confront both their own beliefs and the "false microdoses". Prebunking therefore aims to increase an individual’s **resistance** to disinformation campaigns by reinforcing fact-based beliefs.

4.1 Understanding the Prebunking Approach

The term *prebunking* is used in various contexts and can describe different measures. These measures vary in their emphasis and content and can be tailored specifically (narrow-spectrum) or broadly (broad-spectrum).[24]

4.1.1 Narrow-Spectrum Prebunking

Narrow-spectrum prebunking involves presenting a version of the specific false information that is later encountered. This approach aligns closely with the principles of classic inoculation, which consists of two core components:[25] Firstly, recipients are warned of false messages that could challenge their beliefs. They are then presented with a "weakened (micro) dose" of the specific false information. The presentation of the weakened version of the false message is referred to as **refutational preemption** or **prebunking.** This dual approach intends to activate resistance mechanisms such as perceiving a threat and generating counterarguments.

4.1.2 Broad-Spectrum Prebunking

Broad-spectrum prebunking takes a more general approach, in which recipients are informed about common **manipulation techniques and strategies.**[26] One example of a common manipulation strategy is the invocation of alleged researchers, experts or institutions to enhance the credibility of shared false information, although they lack the scientific knowledge or expertise.[27] Another example involves the use of manipulative rhetoric, such as emotional language or scapegoating, where blame is assigned to a

[22] McGuire, W. J. (1961). The Effectiveness of Supportive and Refutational Defenses in Immunising and Restoring Beliefs Against Persuasion. *Sociometry, 24(2),* p. 184. Available at https://doi.org/10.2307/2786067 (accessed on 29/07/2024).

[23] McGuire, W. J. (1964). Inducing Resistance to Persuasion: Some Contemporary Approaches. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, 1*, pp. 191-229. New York, NY: Academic Press. Available at https://doi.org/10.1016/S0065-2601(08)60052-0 (accessed on 27/08/2024).

[24] Lewandowsky, S., & Van Der Linden, S. (2021). Countering Misinformation and Fake News Through inoculation and Prebunking. *European Review of Social Psychology, 32*(2), pp. 348-384. Available at https://doi.org/10.1080/10463283.2021.1876983 (accessed on 29/07/2024).

[25] Compton, J. A., & Pfau, M. (2005). Inoculation Theory of Resistance to Influence at Maturity: Recent Progress in Theory Development and Application and Suggestions for Future Research. *Communication Yearbook, 29*(1), pp. 97-145. Available at https://doi.org/10.1080/23808985.2005.11679045 (accessed on 29/07/2024).

[26] Lewandowsky, S., & Van Der Linden, S. (2021). Countering Misinformation and Fake News through Inoculation and Prebunking. *European Review of Social Psychology, 32*(2), pp. 348-384. Available at https://doi.org/10.1080/10463283.2021.1876983 (accessed on 29/07/2024).

[27] Cook, J. (2020). Deconstructing Climate Science Denial. In D. C. Holmes, & L. M. Richardson (Eds.), *Research Handbook on Communicating Climate Change* (pp. 62-78). Edward Elgar Publishing. Available at https://doi.org/10.4337/9781789900408.00014 (accessed on 29/07/2024).

Prebunking as a Preventive Intervention

group or individual without addressing actual solutions to the problem.[28] Broad-spectrum prebunking, therefore, does not expose recipients to a "micro-dose" of false information but instead focuses on education and the promotion of critical thinking skills. By explaining common strategies before individuals encounter disinformation, this approach aims to make manipulation attempts more recognisable and to strengthen resistance to disinformation campaigns. While this newer prebunking approach does not address specific content, it can still reduce belief in false information.[29] When combined with a forewarning about specific disinformation, broad-spectrum prebunking can also be understood as a form of inoculation.[30]

In the following, we use the term *prebunking* as a general term for narrow-spectrum and broad-spectrum methods.

4.2 Evaluating Prebunking within Disciplines

4.2.1 Psychological Perspective

Previous research shows that classic inoculation methods can enhance cognitive processing (i.e., the process of thinking and understanding) of the presented content.[31] They also indicate that triggering feelings of threat can increase both the motivation to engage with and to resist false information.[32] More recent studies suggest that inoculation stimulates reflection on the topic.[33] Additionally, several field studies have demonstrated the effectiveness of broad-spectrum prebunking messages on platforms like YouTube, such as short videos explaining common manipulation techniques used in disinformation campaigns.[34] This indicates that **broad-spectrum prebunking** in particular can help individuals recognize manipulation techniques and foster a critical attitude towards false information. A 2023 meta-analysis further shows that **narrow-spectrum prebunking** can **reduce the credibility of false information** and have positive effects on the sharing of true information.[35] However, the impact of inoculation on the

[28] Roozenbeek, J., Van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological Inoculation Improves Resilience Against Misinformation on Social Media. Science Advances, 8(34), eabo6254. Available at https://doi.org/10.1126/sciadv.abo6254 (accessed on 29/07/2024).

[29] Cook, J. (2020). Deconstructing Climate Science Denial. In D. C. Holmes, & L. M. Richardson (Eds.), *Research Handbook on Communicating Climate Change* (pp. 62-78). Edward Elgar Publishing. Available at https://doi.org/10.4337/9781789900408.00014 (accessed on 29/07/2024).

[30] Roozenbeek, J., Van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological Inoculation Improves Resilience Against Misinformation on Social Media. Science Advances, 8(34), eabo6254. Available at https://doi.org/10.1126/sciadv.abo6254 (accessed on 29/07/2024).

[31] Pfau, M., Tusing, K. J., Lee, W., Godbold, L. C., Koerner, A. F., Penaloza, L., Hong, Y. H., & Yang, V. S. H. (1997). Nuances in Inoculation: The Role of Inoculation Approach, Ego-Involvement, and Message Processing Disposition in Resistance. *Communication Quarterly, 45*(4), pp. 461-481. Available at https://doi.org/10.1080/01463379709370077 (accessed on 29/07/2024).

[32] Compton, J., & Pfau, M. (2004). Use of Inoculation to Foster Resistance to Credit Card Marketing Targeting College Students. Journal of Applied *Communication Research, 32*(4), pp. 343-364. Available at https://doi.org/10.1080/0090988042000276014 (accessed on 29/07/2024).

[33] Compton, J., Van Der Linden, S., Cook, J., & Basol, M. (2021). Inoculation Theory in the Post-truth Era: Extant Findings and New Frontiers for Contested Science, Misinformation, and Conspiracy Theories. Social and Personality *Psychology Compass, 15*(6), e12602. Available at https://doi.org/10.1111/spc3.12602 (accessed on 29/07/2024).

[34] Roozenbeek, J., Van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological Inoculation Improves Resilience Against Misinformation on Social Media. Science Advances, 8(34), eabo6254. Available at https://doi.org/10.1126/sciadv.abo6254 (accessed on 29/07/2024).

[35] Lu, C., Hu, B., Li, Q., Bi, C,. & Ju, X. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, 25*, e49255. Available at https://www.jmir.org/2023/1/e49255 (accessed on 29/07/2024).

sharing of false information is less clear. The findings suggest that inoculation reduces the spread of health-related disinformation but no other types of false information. Data of one of our current studies within the DYNAMO project shows that prebunking significantly reduces the credibility of disinformation only when it is **highly specific (narrow-spectrum)**. Further research is needed to thoroughly analyze and improve the long-term effectiveness of prebunking measures.

In order to assess prebunking as a measure against disinformation, both its opportunities and risks must be considered. Its greatest advantage is that false information can be refuted before it is processed, potentially preventing the ***Continued Influence Effect***. This effect describes how disinformation, once processed, can be plausibly integrated with an individual's existing knowledge and continues to influence behaviour, thinking and attitude even after correction.[36] A broad prebunking message that provides information about common manipulation techniques could also help to expand **digital media and information literacy**. This could encourage recipients to critically evaluate and analyse new information. However, the effect of prebunking is limited in duration and diminishes over time if not reinforced.[37] Additionally, prebunking messages may be rejected or misinterpreted if they conflict with a person's preexisting beliefs, reflecting the psychological mechanism known as *confirmation bias*.[38] The impact of prebunking on trust in **true information** is also **not yet fully understood.**[39] For instance, there is a risk that trust in information, even if credible and accurate, might be undermined.

Prebunking measures can strengthen resistance to disinformation and reduce belief in false information. However, the influence on sharing false information remains unclear, although mitigating the sharing of false information is critical in the fight against it. Thus, despite a lower belief in false information, individuals may still share it. Moreover, it is uncertain whether prebunking measures affect trust in true news or how personal factors and attitudes shape their effectiveness. From a psychological perspective, prebunking remains a generally suitable method. Nevertheless, further research is needed to better understand the underlying psychological mechanisms of different prebunking measures and to improve their effectiveness in preventing the sharing of false information.

[36] Johnson, H. M., & Seifert, C. M. (1994). Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), p. 1420. Available at https://doi.org/10.1037/0278-7393.20.6.1420 (accessed on 29/07/2024).

[37] Maertens, R., Roozenbeek, J., Basol, M., & Van Der Linden, S. (2021). Long-Term Effectiveness of Inoculation Against Misinformation: Three Longitudinal Experiments. *Journal of Experimental Psychology: Applied, 27*(1), 1-16. Available at https://doi.org/10.1037/xap0000315 (accessed on 29/07/2024).

[38] Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*, pp. 175-220. Available at https://doi.org/10.1037/1089-2680.2.2.175 (accessed on 29/07/2024).

[39] On the one hand, study results suggest that both true and false information may be perceived as less credible (see Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General, 152*(9), pp. 2411-2437. Available at https://doi.org/10.1037/xge0001395 (accessed 29/07/2024), while other studies do not confirm this (see Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, 25*, e49255. Available at https://www.jmir.org/2023/1/e49255 (accessed on 29/07/2024).

Prebunking as a Preventive Intervention

4.2.2 Communication Science Perspective

From a communication science perspective, prebunking measures appear to be useful in combating the spread of disinformation. Despite the initial confirmations mentioned above that prebunking can be helpful,[40] corresponding measures must be planned carefully. Several existing findings can be used **for the design of prebunking measures**. For example, they should be adapted to typical forms of content preparation that we were able to identify in our empirical research (see section 2.1). Broad spectrum measures could, for example, shed light on particularly emotionalised writing styles, on the focus on a key topic frequently used to convey disinformation, the communication of alternative constructions of reality and the provision of isolated communities. However, these findings could also be used to recognise disinformation campaigns, which can then be responded to with a narrow-spectrum measure. It must be taken into account that curation (see section 2.1) is a practice that can indicate the spread of disinformation: Reality is constructed on disinforming Telegram channels not only through posts written specifically for this purpose, but also to a significant extent through selection and commenting.

Interviews conducted in 2024 with a total of nine **experts** working in various specialist disciplines in the field of combating disinformation (media law, political consulting, science, state actors) confirm the positive effect on combating the spread of disinformation. In terms of practicability, prebunking measures are very well suited from the perspective of those experts who apply intervention measures in their work context and bring them to the attention of civil society. At the same time, however, doubts are expressed as to who the target group of prebunking measures should most likely be and how they can be reached. This is linked to the argument that at least broad-spectrum prebunking measures could only be effective in the long term but not in the short term. It must be emphasised that the target group of countermeasures, especially those located in state-sceptical counter-publics, can be difficult to reach, both in terms of content and technology. Finally, there are considerable doubts about the willingness of service providers to cooperate (above all the messenger service Telegram).

4.2.3 Technical Perspective

From a technical perspective, it is important to differentiate between the broad and narrow spectrum when analysing prebunking measures. Implementing the presentation of common manipulation techniques and strategies (broad spectrum) is technically straightforward. The utilisation of **pop-up messages, information feeds, or channels** facilitates the dissemination of predefined information to recipients with minimal technical complexity. Moreover, **playful media skills training** (gamification) can enhance the engagement and retention of the audience.

For the specific "vaccination" in the narrow spectrum area information about the respective disinformation must first be collected and analysed. If the dissemination of

[40] Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological Inoculation Improves Resilience Against Misinformation on Social Media. *Science Advances, 8*(34), eabo6254. Available at https://doi.org/10.1126/sciadv.abo6254 (accessed on 29/07/2024).

such information is to be achieved via messages focusing on the specific channel or group, it is essential to **continuously monitor**[41] the messenger service. The classic inoculation (two-part structure) would therefore be technically difficult to implement. Providing general information about current disinformation campaigns would not require continuous monitoring and therefore be easier.

For example, to determine facts about a specific message in a messenger service, a **database** would have to be **set up** with which the content of this message can be compared. This database would have to be continuously expanded with additional and up-to-date information. Even if the database were to be set up successfully, it would still be very challenging from a technical perspective to systematically extract data from the database and the messenger services. In this context, the question arises as to which messages should be subject to prebunking measures, given the inefficiency of checking all messages for possible false information.

Careful consideration must also be given to the organisation of the prebunking messages. Should they be adapted as closely as possible to the misinformation or should general, topic-related facts be presented? Depending on the design, the difficulty of technical implementation would vary. If, for example, a prebunking message with similar content to the false message is preferred, this similarity must be made measurable and it must be decided how high this similarity should be. A further challenge is the **response time**, which depends on the design of the database created. The response time increases the more tailored the prebunking is to the current disinformation.

4.2.4 Legal Assessment

Prebunking measures **have not yet been enshrined in law**. Although prebunking measures can optionally be used as a possible risk mitigation measure under the DSA (Art. 35 DSA), prebunking is not mentioned in the catalogue of exemplary measures, so it can hardly be understood as an obligation for providers. At the level of voluntary commitments under the **2022 Strengthened Code of Practice on Disinformation**, at least one of its measures can be understood to include prebunking. In line with the broad-spectrum method, the signatory parties commit to building and implementing features or initiatives that empower users to think critically about information they receive and help them to determine whether it is accurate (Commitment No. 25).

However, a **legal obligation** for service providers to implement prebunking measures that goes beyond a mere voluntary commitment appears to be a promising approach. From a legal perspective, the obligation to prebunk is **more favourable to fundamental rights** than many other measures (see section 2.2.1.). As described, the deletion and blocking of content or accounts in particular can severely impair the fundamental rights to freedom of expression and information. But also, the labelling of suspected

[41] Panahi, T., Hornung, G., Schäfer, K., Choi, J. E., Steinebach, M., & Vogel, I. (2023). Desinformationserkennung anhand von Netzwerkanalysen – ein Instrument zur Durchsetzung der Pflichten des DSA am Beispiel von Telegram [Disinformation Detection Based on Network Analyses - a Tool for Enforcing DSA Obligations Using the Example of Telegram]. In Friedewald, M., Roßnagel, A., Neuburger, R., Bieker, F., & Hornung, G. (Eds.), *Daten-Fairness in einer globalisierten Welt [Data Fairness in a Globalised World]* (Vol. 2, pp. 343-370). Nomos-elibrary, 28, pp. 343-370. Available at https://www.nomos-elibrary.de/10.5771/9783748938743-343.pdf (accessed on 29/07/2024).

Prebunking as a Preventive Intervention

disinformation (debunking, flagging) can interfere with freedom of expression (see 2.2.1). Many of the possible measures would require the filtering of private messages, which would impair the fundamental rights to private communication and data protection.

However, it cannot be ruled out that prebunking measures can lead to the impairment of these fundamental rights. In principle, the following applies: a prebunking measure is more compatible with fundamental rights the more opinion-neutral it is, the more self-determined user behaviour remains possible, the less external intervention in the communication process takes place and the less personal data is processed.

Prebunking measures that generally promote **media literacy through the technical design** of the messenger services and that do not require real-time interventions therefore appear most compatible with fundamental rights. Ideally, from a fundamental rights perspective, a prebunking intervention should not start with a specific factual assertion, but rather provide generally accessible, abstract offers to expand media literacy. Instead of filtering or labelling content, e.g. with certain keywords for the purpose of advance warning, prebunking measures should be generally available to users by design (e.g. through highlighted information feeds/channels/stories, gamification). It should also be borne in mind that labelling individual content would conflict with the fundamental right to freedom of expression (see section 2.2.1.). In contrast, the **broad-spectrum** method can guarantee neutrality of opinion and is preferable as a milder means regarding **freedom of expression**. The proposal for a legal formulation could be based on voluntary commitment no. 25 of the 2022 Strengthened Code of Practice on Disinformation, given that a consensus between many companies, fact-checking organisations and the EU Commission has already been established.

Prebunking measures would be a suitable solution for combating disinformation, especially for the private communication features of messenger services (one-to-one, few-to-few), which have as yet been scarcely regulated, given that many of the possible prebunking measures, especially in the broad spectrum, do not interfere with the fundamental right to private communication. Prebunking measures also fit in with the hybrid structure of many services, which combine public and private communication features. An obligation to implement prebunking measures would be a way of **regulating** the **entire messenger service** and thus curbing disinformation in entire user networks.[42]

In addition, service providers should be obliged to **design** the prebunking measures used **in line with fundamental rights.** The monitoring and filtering of private messages should be explicitly excluded.

Service providers should also be obliged to provide **transparent general terms and conditions** that provide information on the concrete prebunking mechanisms used, among other things. Art. 14 DSA is not yet sufficient for this, as it only requires the transparency of restrictions on the service, which does not include prebunking.

[42] If the proposal were to be taken into account in an amendment to the DSA, the new provision would consequently not have to be included in the catalogue of obligations for online platforms, but in the general provisions for hosting services, so that the private communication functions of messenger services would also be covered.

In principle, the **wording of the law** should be as **technology-neutral** as possible in order to promote creativity and thus diverse innovations of service providers in the development of prebunking measures. In other words, no specific prebunking measures should be prescribed by law, meaning that new technologies could fulfil the legal requirements through open formulations. In doing so, the law would interfere less with service poviders' Freedom to conduct a business under Article 16 of the CFR.

Finally, messenger service providers should be obliged to **(externally) evaluate** their advances. Although data from the private communication features of messenger services would probably only be available via voluntary data donations and voluntary experimental settings due to the fundamental rights to data protection and private communication, service providers can at least be obliged to report on the development of their technical functions.

It would be conceivable to propose such a **legal innovation at EU level.** This could be justified - as for the regulations of the DSA - in principle by the EU's internal market competence under Art. 114 TFEU, which may require further harmonisation of the regulations for messenger services in the internal market.

4.2.5 Summarised Assessment

The suitability of narrow-spectrum and broad-spectrum measures is evaluated differently across various disciplines.

From a psychological perspective, prebunking is generally considered an effective approach. However, its impact depends on the method of application and on individual factors of the recipients. Prebunking measures can enhance resistance to disinformation and reduce the credibility of false information. Nevertheless, further research is needed to examine potential negative effects, such as the influence on trust in true information or the impact on the sharing of false information. The relative effectiveness of narrow- and broad-spectrum methods also requires additional investigation, since individual studies and theoretical considerations suggest that narrow-spectrum prebunking may be more effective overall.

Our interviews with communication science experts show that prebunking measures are generally regarded as connectable, although doubts about the identification and accessibility of the target groups, the short-term effectiveness and the willingness of messenger services to cooperate exist.

From a computer science perspective, broad spectrum measures are technically more straightforward to implement. Conversely, narrow spectrum measures entail the challenge of constructing databases and developing methods that allow the short-term and representative extraction of messages. Additionally, the measurability of the spread of disinformation poses problems.

From a legal perspective, a legal obligation to implement broad-spectrum prebunking would constitute a minimally invasive measure that protects fundamental rights.

Private communication in messenger services, which is not and should not be accessible for surveillance measures due to the protection of private communications, an obligation for service providers to implement prebunking measures is therefore a relatively lenient measure compared to alternatives (e.g. deleting, filtering or labelling chat messages).

If prebunking measures were mandated by law or voluntarily implemented by service providers, both empirical findings and legal requirements should be taken into account. Narrow- and broad-spectrum methods should, however, be further investigated to develop prebunking measures that are as effective and compliant with fundamental rights as possible. Two key points should be considered in future research: First, while most perspectives tend to favor broad-spectrum prebunking solutions, narrow-spectrum prebunking might yield significantly better results. Second, in the context of broad-spectrum measures, ethical discussions must address the extent to which increasing media literacy places significant responsibility on individual users, who may need further training. A balance must therefore be found between technical possibilities, fundamental rights-related considerations and the protection of the public, which should ideally reduce the cognitive burden on users (e.g., minimizing the need for skill acquisition).

5 Conclusion

Counter-publics that are state-sceptical or anti-constitutional are forming on messenger services by spreading disinformation via channels and groups. Counter-measures are needed that must be adapted on a legal, technical and civil society level.

In this policy paper, we have shown that certain legal measures that attempt to curb the spread of disinformation are already in place. However, the regulations are not sufficiently designed for the hybrid structure of messenger services, meaning that more concrete measures are needed. In addition, the private communication features (one-to-one, few-to-few) remain largely unregulated. Therefore, from legal, communication science, psychological and technical perspectives, existing measures must be supplemented.

We conclude that the emphasis should not only be placed on repressive measures, but above all on **combating and preventing the causes** of the spread of disinformation on messenger services. A better understanding of the underlying motives is required to prevent the emergence and staging of state-sceptical counter-publics, which does not primarily pluralise public discourse but endangers it. This can only be done by continuing the dialogue with those who have considerable doubts about established media and political institutions, who are turning their backs on democratic systems, and who are increasingly, and in some cases exclusively, obtaining information via messenger services. To this end, the communicative **advantages of the interpersonal level** must be intergrated **into the social dialogue**. To better understand the emergence, persistence, and dynamics of state-sceptical counter-publics that network through messenger services, it is important to engage in dialogue with each other on equal footing, rather than merely talking 'about' those within counter-publics.

Furthermore, journalists and politicians must **avoid following the argumentative strategies of disinformation actors**, not only but especially when they themselves use messenger services for their own communication. In our information ecosystem, which includes journalism, politics, platforms and individual users, political and journalistic actors are particularly important in spreading and combating disinformation.

Politicians bear a dual responsibility: to establish the **prerequisites to combat disinformation** and ensure that they themselves **do not spread disinformation**. During election campaigns in particular, information could potentially be used in a targeted manner to promote political goals. This can only be prevented by **rethinking and establishing new narratives** in **journalism and politics** to exclude populist strategies – such as those used by disinformation actors – from democratic discourse. Achieving this with effective and **more systemic solutions** that can be implemented in the short term is particularly necessary in crisis situations and before (upcoming) elections.

With the proposed prebunking measures, we therefore recommend a preventive set of measures for policymakers, researchers and those involved in combating disinformation. Given that not all segments of civil society (or only parts) can be reached, broad-based prebunking measures will usefully expand the promotion of media literacy. In this context, **media literacy by design** means that measures reach actors both within and outside the state-sceptical counter-publics, thus promoting constructive discourse and social cohesion to effectively counter the further spread of disinformation, even in fluid forms of communication.

To assess the feasibility and effectiveness of the scientifically researched recommendations for action and to refine them where necessary, **further scientific research** needs to be funded. This research should investigate whether broad-spectrum or narrow-spectrum prebunking is more effective against the spread of disinformation in messenger services. Empirical suitability, technical feasibility and fundamental rights considerations must all be considered in equal measure and a sensible division of responsibilities between service providers, civil society actors and individuals must be found.

DYNAMO

A joint project by

Fraunhofer
SIT

HOCHSCHULE
DER MEDIEN

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

UNIKASSEL
VERSITÄT

SPONSORED BY THE

Federal Ministry
of Education
and Research

DOI: 10.17185/duepublico/83259
URN: urn:nbn:de:hbz:465-20250328-103032-3