SCM
Studies in Communication and Media

FULL PAPER

Synthetic disinformation detection among German information elites – Strategies in politics, administration, journalism, and business

Erkennung synthetischer Desinformation unter deutschen Informationseliten – Strategien in Politik, Verwaltung, Journalismus und Wirtschaft

Nils Vief, Marcus Bösch, Saïd Unger, Johanna Klapproth, Svenja Boberg, Thorsten Quandt, & Christian Stöcker

 Studies in Communication and Media, 14. Jg., 4/2025, S. 594–623, DOI: 10.5771/2192-4007-2025-4-594

Nils Vief (M. A.), HAW Hamburg, Department Information, Media and Communication, Finkenau 35, Hamburg, Germany. Contact: nils.vief@haw-hamburg.de.
Marcus Bösch (M. A.), HAW Hamburg, Department Information, Media and Communication, Finkenau 35, Hamburg, Germany. Contact: marcus.boesch@haw-hamburg.de.
Saïd Unger (M. A.), University of Münster, Department of Communication, Bispinghof 9–14, Münster, Germany. Contact: said.unger@uni-muenster.de. ORCID: https://orcid.org/0000-0003-1266-2055
Johanna Klapproth (M. A.), University of Münster, Department of Communication, Bispinghof 9–14, Münster, Germany. Contact: johanna.klapproth@uni-muenster.de.
Svenja Boberg (M. A.), University of Münster, Department of Communication, Bispinghof 9–14, Münster, Germany. Contact: svenja.boberg@uni-muenster.de.
Thorsten Quandt (Prof. Dr.), University of Münster, Department of Communication, Bispinghof 9–14, Münster, Germany. Contact: thorsten.quandt@uni-muenster.de. ORCID: https://orcid.org/0000-0003-1937-0282
Christian Stöcker (Prof. Dr.), HAW Hamburg, Department Information, Media and Communication, Finkenau 35, Hamburg, Germany. ORCID: https://orcid.org/0000-0002-7182-167X

FULL PAPER

Synthetic disinformation detection among German information elites – Strategies in politics, administration, journalism, and business

Erkennung synthetischer Desinformation unter deutschen Informationseliten – Strategien in Politik, Verwaltung, Journalismus und Wirtschaft

Nils Vief, Marcus Bösch, Saïd Unger, Johanna Klapproth, Svenja Boberg, Thorsten Quandt, & Christian Stöcker

Abstract: Since the technology for generating synthetic media content became available to a wider audience in 2022, the social and communication sciences face the urgent question of how these technologies can be used to spread disinformation and how well recipients are equipped to deal with this risk. Research so far has focused primarily on the phenomenon of deepfakes, which mostly refers to visual media generated or modified by artificial intelligence. Most studies aim to test how well recipients can detect such deepfakes, and they generally conclude that recipients are rather poor at detecting them. In contrast, this analysis focuses on the broader concept of synthetic disinformation, which includes all forms of AI-generated content for the purpose of deception. We investigate the process of how actors with professional expertise in the field of disinformation try to detect AI-generated disinformation in text, visual and audio content and which strategies and resources they employ. To gauge an upper bound for societal preparedness, we conducted guided interviews with 41 actors in elite positions from four sectors of German society (politics, corporations, media and administration) and asked them about their strategies for detecting synthetic disinformation in text, visual and audio content. The respondents apply different detection strategies for the three media formats. The data shows substantial differences between the four groups when it comes to detection strategies. Only the media professionals consistently describe analytical, rather than simply intuitive, methods for verification.

Keywords: Synthetic disinformation, deepfakes, disinformation literacy, digital media literacy, generative AI, elite actors

Zusammenfassung: Seit die Technologie zur Generierung synthetischer Medieninhalte im Jahr 2022 einem breiteren Publikum zugänglich wurde, sehen sich die Sozial- und Kommunikationswissenschaften mit der dringlichen Frage konfrontiert, inwiefern diese Technologie zur Verbreitung von Desinformation genutzt werden kann und wie gut Rezipienten gerüstet sind, um mit diesem Risiko umzugehen. Die bisherige Forschung konzentriert sich primär auf das Phänomen der Deepfakes, welche sich zumeist auf visuelle Medieninhalte

beziehen, die durch Künstliche Intelligenz (KI) generiert oder modifiziert wurden. Die meisten Studien testen, wie gut Rezipienten darin sind, Deepfakes zu erkennen, und kommen zu dem Ergebnis, dass sie Deepfakes in den meisten Fällen von authentischen Medieninhalten nicht unterscheiden können. Im Gegensatz dazu stützt diese Analyse sich auf das breitere Konzept der synthetischen Desinformation, welches alle Formen von KI-generierten Medieninhalten zum Zweck der absichtlichen Falschinformation umfasst. Wir untersuchen die Strategien und Ressourcen, die Akteure mit professioneller Expertise im Bereich Desinformation einsetzen, um KI-generierte Desinformation in Text-, Bild- und Audioinhalten zu erkennen, um so ein tieferes Verständnis für den Prozess der Identifizierung von synthetischer Desinformation und die dafür benötigten Praktiken und Kompetenzen zu erlangen. Hierfür haben wir leitfadengestützte Interviews mit 41 Akteuren in Elitepositionen aus vier Sektoren der deutschen Gesellschaft (Politik, Wirtschaft, Journalismus und Verwaltung) durchgeführt und befragten sie zu ihren Strategien zur Detektion synthetischer Desinformation in Text-, Bild- und Audioinhalten. Die Befragten wenden für die drei Medienformate unterschiedliche Erkennungsstrategien an. Zusätzlich zeigen die Daten substanzielle Unterschiede zwischen den vier befragten Gruppen, wobei die Befragten aus dem Mediensektor am häufigsten analytische Erkennungsstrategien beschrieben, die sich nicht ausschließlich auf eigenes Wissen und Intuition verlassen, sondern externe Quellen zur Überprüfung heranziehen.

Schlagwörter: Synthetische Desinformation, Deepfakes, Desinformationskompetenz, digitale Medienkompetenz, generative KI, Eliten

1. Introduction

Artificial intelligence (AI) has been described as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17). Over the past years, AI or, more precisely, machine learning has become a transformative technology that is revolutionizing various aspects of our lives (Williamson & Prybutok, 2024), while also generating new kinds of problems. One of them is synthetically generated disinformation. One significant milestone for synthetic text generation was the release of the free version of a chatbot called GPT-3.5 by its maker, the company OpenAI, in November 2022. Just two months later, the application reached 100 million monthly users, making it the fastest-growing consumer application in history (Hu, 2023). In parallel, machine learning based systems for generating increasingly realistic images were released, e.g., DALL-E 2, also by OpenAI in September 2022 and Midjourney 5 in March 2023 by Midjourney, Inc. or the open-source text-to-image model Stable Diffusion by Stability AI. Further technology releases allowed the generation of realistic audio and video content by instant voice cloning (ElevenLabs, April 2023) and video voice cloning and lip-syncing (HeyGen Labs, September 2023). All these types of systems are often referred to as "generative AI" (Wu et al., 2023).

There is increasing concern about whether and how synthetic media created with generative AI is used to produce and spread disinformation and whether people are able to recognize such content (Goldstein et al., 2023).

Previous research suggests that recipients have some difficulty detecting AI-generated media content (especially for synthetic images), while overestimating

their own ability to do so (Bray et al., 2023). This is compounded by the fact that algorithmically curated platforms for serving media content to users are, because of their design and optimization goals, an ideal ecosystem for spreading disinformation content (Aïmeur et al., 2023; Stöcker, 2020).

The advent of synthetic disinformation content in the digital public also damages the trust of recipients in authentic news media (Godulla et al., 2021, p. 90). There is a growing body of research on the (negative) implications of these disruptive changes for media recipients and for democratic societies and the digital public sphere in general (Gambín et al., 2024; Roe et al., 2024). For example, an experiment by Dobber et al. (2021) shows that synthetic disinformation videos of politicians can severely impact the public's perception of them. Meanwhile, Russia's invasion of Ukraine provides the first real-life examples of synthetic disinformation being used in conjunction with warfare, with several incidents involving synthetic videos of Russian and Ukrainian government officials being used for disinformation and entertainment (Twomey et al., 2023). Research from the social and communication sciences has focused on the consequences for recipients, specifically on the topic of media literacy. Most of these studies address a specific question: Can people distinguish synthetic visual media from real images and videos, and if so, how well are they performing (Godulla et al., 2021; Rana et al., 2022; Stroebel et al., 2023)?

How people attempt to check content is an under-researched area. When do they decide to verify information? Which detection strategies do they use? What are the skills and resources that they rely on, and which aspects and design features of the content are reviewed during the authentication process? We see a strong focus on the concept of deepfakes in current research, which primarily refers to visual media. To our knowledge, the ability to detect fakes generated by generative AI systems has so far mostly been tested for images and videos. We argue that two other media formats play an important role in the spread of disinformation online that have received little attention in literacy research: Audio and text (Bösch & Divon, 2024; Calvo et al., 2020; Maros et al., 2021; Shao et al., 2018). We intend to fill this research gap and therefore use the term "synthetic disinformation" instead of "deepfakes" to capture the whole phenomenon of intentionally shared false information generated or modified by AI, including text and audio content.

Building on the concept of "acts of authentication" by Tandoc et al. (2018), we assume that internalized knowledge and skills, as well as the skillful use of external verification sources, are crucial for detecting synthetic disinformation content. For this reason, we surveyed individuals who we believe have expertise on the topic due to their prominent professional positions. We conducted guided interviews with 58 elite actors from four sectors of German society (politics, corporations, media and administration), who are either responsible for dealing with disinformation for their respective institutions or have special expertise on the topic. We conducted two rounds of interviews. The initial interviews took place in the fall of 2022, and 41 follow-up interviews in the fall of 2023.

During these interviews, we asked the respondents to elaborate on their strategies to detect disinformation content online for three different media formats:

Text, Video/Image and audio. Because the first wave of interviews took place before the release of critical technologies like Chat-GPT drew public attention to the topic of synthetic media, this analysis draws on the 41 follow-up interviews conducted in 2023. Respondents' awareness and concern regarding the emergence of synthetic disinformation had increased dramatically from 2022 to 2023.

We aim to get a better understanding of how disinformation experts in German politics, administration, media and corporations are affected by the emergence of synthetic disinformation and how well they are prepared to deal with it. Our rationale behind this is: Synthetic disinformation is poised to increase the well-described and researched disinformation problem that democratic societies already face. We tried to identify and interview groups of professionals best placed to deal with this emerging problem to gauge how these information elites deal with it. Since the rest of society is probably less well-equipped to deal with it than these professionals, our results mark a tentative upper bound for societal preparedness for the emerging problem of synthetic disinformation.

RQ: Which detection strategies do German disinformation elites use to identify different kinds of synthetic disinformation in textual, visual and audio content, and which aspects and design features of the content are reviewed during the authentication process?

2. Theoretical framework

2.1 Definition: Synthetic disinformation

Combining established definitions, we define synthetic disinformation as a special type of disinformation partly or fully generated/modified by AI and containing false information that is knowingly shared to cause harm (Millière, 2022; Wardle & Derakhshan, 2017, p. 5). The concept of synthetic disinformation differs from the concept of deepfakes in two respects: It is narrower in terms of the purpose of its distribution (intentional distribution with the intention of causing harm) and broader in terms of the included media formats (text-based, visual, and audio content).

Most research on AI-generated misinformation focuses on deepfakes, a term coined in 2017 by a Reddit user who circulated AI-generated pornographic videos with celebrity faces (Cole, 2017). The term combines "deep learning" and "fake", referring to the neural network-based tools used to create the fabricated content. In 2019, Deeptrace found that nearly 96% of 15,000 identified deepfake videos online were pornographic, indicating its primary use at the time (Simonite, 2019). Most deepfake research concentrates on visual media, with definitions like the UK Government's Centre for Data Ethics and Innovation (2019) describing deepfakes as "artificial intelligence-based image synthesis technique that involves creating fake but highly realistic video content", through which it is possible to "change how a person, object or environment is presented" (CDEI, 2019). Only some authors like Gambín et al. (2024, p. 64) include audio and text in their deepfake conceptions. To describe the broad spectrum of all types of artificially

generated or modified media content (text, images, video, audio), the term synthetic media was introduced (Millière, 2022).

We combine the concept of synthetic media with the concept of information disorder by Wardle and Derakhshan (2017), who distinguish three types of problematic messages around the concepts of falseness and harm. By this definition, "disinformation" is information that is false and deliberately created to harm, in contrast to "misinformation", which is false but not created or spread with harmful intention, and "malinformation", which is based on reality, but used in a way designed to inflict harm on a person, organization or country, e.g., by leaving out important context. To avoid confusion, we use the term "synthetic disinformation", which encompasses all forms of AI-generated and intentionally disseminated false information.

2.2 Synthetic media literacy

Media literacy is understood as the human potential to acquire knowledge about media, operate media skillfully, critically evaluate them, and create media content. It also serves as a pedagogical goal to foster these abilities and transmit relevant knowledge in both formal and non-formal educational settings (Hugger, 2022). Rohs and Seufert argue that media literacy in a professional context also includes the ability to consider relevant, legal, ethical, and economic frameworks in the use and production of media (Rohs & Seufert, 2020).

AI and synthetic media present significant challenges for the concept of information and media literacy, particularly the issue of "explainability" in AI systems. Unlike classical systems, modern AI systems make decisions based on complex parameters that are not easily understood by humans, making it difficult for users to ascertain how information was obtained or why a particular output was generated. Users unaware of these limitations may struggle to validate AI-generated outputs and recognize misinformation (Tiernan et al., 2023). Over the last few years, various concepts of digital media competence have developed. However, there is yet no coherent literacy concept related to the detection of synthetic media content and, in particular, synthetic disinformation.

Martinez-Bravo et al. (2022) identify six key dimensions of competence that are central for digital media literacy: The ability to adopt a responsible and ethical approach to using technology and evaluating information (critical dimension), high-level thinking skills such as problem-solving, logical reasoning, and creativity in digital environments (cognitive dimension), the ability to engage socially and collaboratively in digital environments (social dimension), the instrumental and technical skills for using digital tools and understanding their underlying principles (operative dimension), the capacity of managing personal emotions and behaviors, building healthy relationships, and protecting one's well-being in digital spaces (emotional dimension). The sixth dimension addresses the ability to anticipate and innovate within dynamic digital environments, using foresight and technological understanding for problem-solving and scenario building (projective dimension) (Cho et al., 2024; Martínez-Bravo et al., 2022).

Lintner (2024) argues that three core competencies are essential when it comes to "AI-literacy": A technical understanding of AI that goes beyond just general awareness and implies a basic comprehension of the underlying principles and mechanisms of AI technologies, a critical understanding of how AI influences society in various sectors, such as economics, employment, privacy, and social structures and the awareness and understanding of the ethical considerations surrounding AI development and deployment. Other authors of educational sciences like Ng et al (2021) and Kong (2021) emphasize a fourth important competence: The ability to apply AI concepts in practical, real-world scenarios and even develop AI technologies.

However, it is not yet clear what specific skills are required to detect synthetic media that are intentionally used and disseminated to deceive. There is, so far, no clearly defined concept of synthetic disinformation literacy.

When it comes to the authentication of synthetic disinformation, several core questions can be raised: How do people attempt to verify the authenticity of content on the internet in general? And what are the strategies that they use to identify synthetic disinformation content and distinguish it from authentic information?

Tandoc et al. (2018, p. 2753) argue that people use a two-step authentication process. They examined the authentication strategies that 2501 people in Singapore used to authenticate news items they encountered through social media. On this basis, they established a conceptual framework called "audience's acts of authentication (3 As)." They argue that people first use internal and then external acts of authentication to determine the validity of an item.

The first step is the Internal act of authentication. It refers to an individual's initial encounter with news on social media. In this initial encounter, individuals rely on three main authentication framings: (1) the self, (2) the source, and (3) the message. First, at the most basic level, people rely on their own sense of judgment. They use their tacit stock of knowledge to examine whether a particular item is believable. For example, both respondents from Tandoc and from this survey answered that they detect misleading information based on "their gut feeling" (Tandoc et al. 2018, 2754) or that they will "just naturally notice" (S1) based on their common sense. Beyond their own stock of knowledge, individual users also consider the characteristics of the message itself and of the source. When the individual is satisfied with the authenticity of the information in this initial stage, the process ends there, and the information is accepted as authentic. However, if after this reading the individual remains unconvinced of the information's authenticity, then he or she proceeds to the next step, which includes external acts of authentication.

External acts of authentication, according to Tandoc et al., can be either intentional or incidental, by relying on interpersonal and institutional resources. Individuals can deliberately seek out ways to verify news items either through personal contacts or by seeking authentication in formalized sources (Tandoc et al., 2018, p. 2754).

Some people might opt not to try verifying the authenticity of digital content. The framework of Tandoc et al. is consistent with models from the field of cogni-

tive psychology, such as the dual-process model of information processing under uncertainty presented by Tversky and Kahneman (1974). "Internal acts of authentication" can be likened to what Tversky and Kahneman would call system 1 processing: Fast, intuitive, effortless, associative, implicit, based on experience but prone to heuristics that are a common source of cognitive distortions and biases. "External acts" of authentication would be more like system 2 processing, i.e., controlled, slower, effortful processing that is less prone to heuristics and thus to biases.

All three steps of the authentication process, according to Tandoc et al., have one thing in common: They rely on trust. First, whether the content is reviewed at all depends primarily on the person's trust in the source and their own abilities. Also, during internal authentication the individual will first look for markers of credibility within the content (message, source, style) and within themselves (internalized prior knowledge and instinctive reaction). Only when this internal trust is deemed insufficient to label a given piece of content as authentic does the individual move beyond the news item and beyond their own experiences to look for external markers of credibility. This suggests a strong social element to what content people will review at all and how they will do it (Frischlich 2019; Tandoc et al. 2018, 2758).

3. Literature review

3.1 Synthetic disinformation: Implications and literacy

The majority of research on the topic of synthetic disinformation is driven by computer science and law. It uses the concept of deepfakes and focuses on synthetic images and videos. Most studies from the field of computer science follow an experimental approach and concentrate on developing and testing technical systems for detecting AI-generated pictures and videos and/or tracing the source of the synthetic disinformation. For these studies, the research interest lies in judging the authenticity of the content and not in its political function and implications. The central goal is to determine whether a piece of content is fake or not and whether it was created using AI (Rana et al., 2022; Stroebel et al., 2023). In the field of law, most authors discuss the legal implications and regulations of synthetic media. In addition to the dissemination of synthetically generated disinformation, the legal perspective primarily addresses the legal issues surrounding the pornographic use of AI-generated content (Godulla et al., 2021, p. 86).

Since this study aims at identifying specific strategies that recipients use to detect synthetic disinformation, we will primarily discuss studies that examine the effect of synthetic disinformation on recipients or their ability to detect it. The proportion of research that investigates these aspects is significantly smaller and predominantly from the social and communication sciences (Godulla et al., 2021). Almost all these studies operate with the concept of deepfakes, not synthetic disinformation, and therefore have a slightly different focus regarding the media formats and the political function of the (false) content they examine.

To date, there have been few studies examining the effects of synthetic disinformation on recipients. These initial findings suggest that AI-generated visual content can further amplify the negative effects of disinformation by increasing its credibility, strengthening the intention to share, and damaging political attitudes and trust in politicians and the media. An experiment by Hwang et al. (2021) tested whether an AI-generated video would enhance the negative impact of a specific disinformation message on 316 Korean adults. The researchers measured how recipients rated the vividness, persuasiveness, and credibility of a disinformation message about Facebook CEO Mark Zuckerberg, as well as their intention to share the message. They showed two groups the same message, with one of the messages supplemented by a synthetic video. The results show a positive effect for the synthetic video: Respondents rated the liveliness, persuasiveness, and credibility of the synthetic version higher and expressed a greater intention to share the message. The authors suggest that this is where a key mechanism of synthetic disinformation comes into play. By supplementing false content with appropriate imagery, synthetic disinformation increases its credibility and dissemination. They also tested different types of media literacy education treatments: Deepfake-specific literacy education, general media literacy education and no literacy treatment at all. Their results show that literacy education helps reduce the effects of the disinformation message. Interestingly, for this study, "general disinformation literacy" reduced the effects just as well, sometimes even better, than specific "deepfake literacy" (Hwang et al., 2021).

Another study by Dobber et al (2021) argues that microtargeting techniques can amplify the effects of synthetic disinformation by enabling malicious political actors to tailor deepfakes to the susceptibilities of the receiver. In their online experimental study (N = 278), the researchers constructed a synthetic video by modifying an authentic video of a politician and examined its effects on political attitudes. They found that attitudes toward the depicted politician were significantly lower after viewing the artificially modified version, while attitudes toward the politician's party remained similar to the control condition. Only 12 of the 144 Participants from the treatment group identified the synthetic video as such. The authors also tested the effects for a microtargeted group and observed that both attitudes toward the politician and attitudes toward his party scored significantly lower than the control condition. This suggests that microtargeting techniques can indeed amplify the effects of synthetic disinformation content (Dobber et al., 2021).

Other early studies follow a broader approach and address the societal implications of synthetic disinformation. Twomey et al. (2023) conducted a thematic analysis of tweets that discussed deepfakes in relation to the Russian invasion of Ukraine. By analyzing public discourse on social media, they aimed to understand how people perceive and react to synthetic videos during a real-world conflict. The authors conclude that synthetic videos, especially in a high-stakes context like a military conflict, do contribute to undermining epistemic trust by fostering doubt and making it harder for individuals to rely on shared information. It highlights the real-world implications of synthetic disinformation beyond individual perception, impacting collective trust in knowledge (Twomey et al., 2023).

Another study by Vaccari and Chadwick (2020) found that individuals are more likely to experience a feeling of uncertainty after viewing synthetic disinformation videos, rather than being directly misled by them. This resulting uncertainty, in turn, reduces trust in news on social media. They conducted an experiment with a representative sample from the UK (n = 2005) using various AI-modified versions of a popular video of former US President Barack Obama and the US comedian Jordan Peele. Two of the versions were misleading, one disclosed the AI modification. The authors conclude that deepfakes may contribute to generalized indeterminacy and cynicism, further intensifying recent challenges to online civic culture in democratic societies (Vaccari & Chadwick, 2020).

The overwhelming majority of research that investigates recipients of synthetic mis- and disinformation concerns empirically testing if people can distinguish synthetic images and videos from authentic content (Bray et al., 2023).

The research suggests that recipients' ability to detect synthetic images is rather underdeveloped, sometimes not even better than chance. A study by Liu et al. found a labelling accuracy between 63.9 and 79.13%, depending on the dataset (various deepfake generators were tested). This was a mass processing task with a small sample, since 20 users had to classify 1,000 images. It took them an average of 5.14 seconds to do so (Liu et al., 2020). Two other studies by Nightingale and Farid (2022) and Shen et al (2021) tested the classification of images that showed faces and found accuracies of 48.2 and 49.1%, on par with a coin toss. The former study also found that the trustworthiness of AI images was rated higher than that of real images and that a second treatment group that received a "literacy tutorial" before the experiment reached an accuracy of just 59%. Other authors have criticized the experiments for a variety of methodological reasons (Bray et al., 2023, p. 5). Shen et al. also investigated whether the participants used other aspects of the images besides the faces for classification, so they repeated the experiment with a black background. The results were almost the same: 49.7% accuracy (black background) vs. 49.1% (Shen et al., 2021).

Bray et al conducted a study that tested three different kinds of intervention with a sample of 280 participants. One group was shown examples of synthetic images for familiarization, the second group was shown a list of 10 'tell-tale features' that synthetic images of this kind commonly contain, and the third group saw the same list of features and was reminded of these features below each image they had to classify. This study found accuracies above chance of around 60%. However, the interventions did not help. They slightly increased the detection accuracy for synthetic images, but at the same time reduced the accuracy for real images, leading to false positives. Also, participants tended to be overly confident in their ability to differentiate real and synthetic images (Bray et al., 2023).

Unlike with images, the results for video authentication varied considerably between 23 and 87% labelling accuracy for synthetic video detection. The participants performed much better when asked to recognize real video stimuli compared to AI-generated videos. In all studies that were examined in a literature review by Bray et al (2023, pp. 5–6), subjects labeled real videos correctly between 75 and 88% of the time. But while they rarely think that real videos are fake, they don't recognize fake videos as such. The authors criticize most studies on

synthetic video literacy extensively, pointing to mostly small samples and some test generators developed by the respective researchers themselves (sometimes closed source). A study with a larger sample was conducted by Groh et al, who investigated 304 paid participants and another 15,578 who took an online test for synthetic video classification. The mean accuracy was 66% (Groh et al., 2022). Another study, by Köbis et al. (2021), investigated video stimuli with two treatment groups. One received a monetary incentive, and the other read a text addressing the potential harm of AI-generated videos. They did not find measurable differences between the groups. The accuracy was significantly above chance at 57.6%. But they found that the participants' confidence in their classification decision was much higher than the actual detection accuracy (73.7–82.5% compared to 57.6%).

The current state of research suggests that synthetic disinformation (mostly studied in the form of synthetic images and videos) has considerable potential for damage to democratic societies. First, people are already rather bad at recognizing synthetic visual media (especially for synthetic images), while it can be assumed that the techniques for generating synthetic content will continue to improve dramatically over the coming years. Several studies suggest that the recipients overestimate their ability to detect synthetic disinformation. The appearance of synthetic media in the digital public sphere also damages the trust of recipients in authentic news media and can amplify the negative impact of online disinformation.

We see two gaps in the current body of research regarding synthetic media and online disinformation. First, while research has already produced numerous insights into the performance of synthetic disinformation literacy and especially synthetic image and video literacy among recipients, little is known about the process by which people attempt to recognize synthetic disinformation. We are not aware of any study that surveys participants who have specific expertise and/or influence on the handling of synthetic disinformation at a societally relevant level. Previous research on synthetic disinformation has focused almost exclusively on visual media content. However, initial research suggests that two other media formats play an important role in the spread of disinformation online that have so far received little attention in literacy research: Audio and text (Bösch & Divon, 2024; Calvo et al., 2020; Maros et al., 2021; Shao et al., 2018). This study aims to address these two research gaps.

Although previous research on synthetic media literacy suggests that for the majority of recipients, visual synthetic media content is not distinguishable from authentic content anymore, the experimental designs of these studies significantly limited participants' recognition strategies by not providing any external sources or context for the content under review. In most experiments, the participants had no other sources than the image or video itself and their own knowledge to verify it. Only internal acts of authentication were tested. However, if the synthetic content itself can hardly be distinguished from real content, the context becomes the decisive marker for the verification of the checked content.

For this analysis of German information elites' detection strategies, we therefore assume that strategies that rely on external acts of authentication are the

most promising to build robust resilience against synthetic disinformation. This is especially true when it comes to new forms of disinformation that the interviewees have no prior internalized knowledge about, since the reliability of internal detection strategies relies on internalized knowledge and skills. Since our interviewees have professional expertise on the topic of disinformation, it can be assumed that they also have an above-average repertoire of internalized knowledge that they can apply.

Most of the studies discussed so far attempt to compile samples that are representative of the respective population or user group under study. We are interested in the application of external acts of authentication in the detection of synthetic disinformation, which relies heavily on internalized knowledge and skills. We assume that these skills are most likely to develop through regular (and professional) exposure to synthetic disinformation. Therefore, we specifically surveyed "elite actors" (defined below) who we assume to have particularly extensive experience in dealing with synthetic disinformation. Our research question is therefore:

> *RQ: Which detection strategies do German disinformation elites use to identify different kinds of synthetic disinformation in textual, visual and audio content, and which aspects and design features of the content are reviewed during the authentication process?*

4. Methods

4.1 Synthetic disinformation elites

We follow a positional approach to the concept of "elite" actors (Wasner, 2013), meaning that they have to hold elite positions. "Elite" is defined as having the power and resources to enact decisions or to be able to influence political decisions and public opinion (Higley, 2018; Hoffmann-Lange, 2018; Wasner, 2013). We selected individuals in positions that grant them this elite status. Then we identified the societal sectors of politics, administration & government, media and private business as especially important as they are in a doubly relevant position when it comes to disinformation: On the one hand, they are, at least theoretically, in control of the means to tackle disinformation. On the other hand, they are also potentially high-value targets for disinformation.

Political and administrative elites establish policies, enact laws, and allocate funds for countermeasures, including funding research and education and involve security agencies and other administrative tools for detection and prosecution of criminal disinformation (Filipovic & Schülke, 2023; Pawelec & Sievi, 2023). Media elites are crucial due to their fact-checking expertise and role in building public trust (Graves & Amazeen, 2019), and accountable due to their role in holding other sectors. Private business elites, while less public, aim to protect their image and narratives, potentially lobbying for measures or being impacted by regulation (Guilbeault, 2018).

Second, as the public and potential disinformation actors and spreaders are aware of the status of societal elites, they are also affected as potential targets of synthetic disinformation. Politicians and high-ranking government officials are frequently central to conspiracy theories that fuel populist and anti-elite narratives, seeking to destabilize political systems (Koistinen et al., 2022). Journalists play a crucial role as information providers in the struggle against widespread online disinformation, acting as both adversaries and targets (Kalsnes et al., 2021). Beyond that, disinformation is an increasing concern for the private sector. While cybersecurity has long been a focus for businesses to combat hacking and espionage, the discussion of disinformation as a potential threat to companies and the markets they operate in is only just beginning (Akhtar et al., 2023; Petratos, 2021).

4.2 Sample

We conducted two waves of guided interviews with 58 (n1) key actors from four sectors of German society (politics, corporations, media and administration). The first wave took place September–December 2022, mostly in face-to-face interviews. Follow-up interviews were conducted one year later, September 2023–January 2024, with 41 (n2) participants from the first wave. For this analysis, only the interviews of the second wave were included, since the interest in and awareness of the topic of synthetic disinformation increased drastically in the second wave.

The interview partners were recruited in a multi-stage systematic procedure from the four sectors of German society that are professionally involved with the topic of disinformation. As we follow a positional approach to the identification of elites, we selected representatives of the sectors based on their position (Hoffmann-Lange, 2018). For each organization we contacted, we asked to get in touch with the person either responsible for dealing with the topic of disinformation or with the most expertise in that area.

1) Politics ($n1$ = 16, $n2$ = 10): We contacted politicians from all democratic parties represented in the German parliament in descending order of their position within the party's organizational hierarchy, ending up with 16 interviewees from the Christian Democrats (CDU), the Social Democrats (SPD), the Green Party (Die Grünen) and the Left Party (Die Linke). However, we could not recruit members of the Liberal Party (FDP), and we deliberately excluded the party Alternative für Deutschland (AfD) as using disinformation and disinformation campaigns has already become a part of the AfDs political strategy (Bennett & Livingston, 2023; Darius & Stephany, 2022; Leschzyk, 2021). Among the interviewed politicians are administrative heads of the parties, ministers and former ministers, treasurers and MPs leading parliament committees. 13 of these politicians also participated in the follow-up interviews. For this analysis, three interviewees had to be excluded from the sample, since they did not have the time to answer the questions about their detection strategies.
2) Administration ($n1$ = 17, $n2$ = 8): We used the ministerial structures of the German government to contact members of all ministries. Our sample covers

a broad range of representatives, e.g., from the interior and exterior ministry or the ministry of defense, as well as security agencies and adjacent institutions. We gathered 17 interview partners ranging from press secretaries to state secretaries and individual members in leadership roles at security or defense agencies. Ten of them participated in the follow-up interviews, but two had to be excluded from the sample, since there was not enough time to talk about their detection strategies.

3) Media (*n1* = 15, *n2* = 10): We interviewed journalists from private and publicly funded nationwide media outlets, as well as freelance journalists from newspapers, public and private broadcasters and research collectives specializing in fact-checking with editorial lines ranging from conservative to liberal. Within their respective organization, they mostly occupy roles of department heads, editors, or specialize in the field of social media in journalism. Ten of them also participated in the follow-up interviews.
4) Business (*n1* = 10, *n2* = 8): We recruited spokespeople of large private businesses listed on the German stock market, social media platforms and specifically businesses involved in critical infrastructure like banking, mobility or medical supplies. We were able to recruit ten interviewees from the business sector, working mainly as heads of communication and heads of security. Eight of them participated in the follow-up interviews.
5) We intentionally did not specify which professional positions the respondents should have within their organizations (e.g., only spokespeople) to be open to potentially very different professional approaches to the topic of disinformation and synthetic media within the organizations. These different approaches are reflected, for example, in the fact that some companies referred us to their heads of security, while others forwarded our request to their heads of communication. We did not explicitly ask for expertise in synthetic media or AI during the recruitment process, but for experience with disinformation in general. The focus on the topic of synthetically generated disinformation emerged during the interviews, particularly in the follow-up interviews, and reflects the focus and concerns of the interviewees for this specific period (autumn 2022–winter 2023/24). A more detailed overview of interview partners, their sector and position can be found in Figure 3 in the Appendix.

4.3 Data collection

The interviews followed a semi-structured guide evaluated in a pretest. During the initial interviews in autumn of 2022, five interviewers asked the interviewees about (a) their general experience and definition of disinformation, (b) their strategies for detecting disinformation for different types of media (text, image/video, audio and memes) and (c) their assessment of future developments with respect to the spread of disinformation and the efforts to combat it.

The follow-up interviews followed the same procedure but focused on the time since the last interview (autumn 2022–autumn 2023). We specifically asked for changes and new experiences since the last conversation. The most important change that preoccupied and worried many of the respondents during this period

was the perceived boom in synthetically generated disinformation after the release of Chat-GPT 3.5 and other tools for synthetic content creation.

Most interviews were conducted at the respondents' workplaces. Where this wasn't possible, we used video calls via Zoom or Microsoft Teams. The interviews generally lasted between 40 and 60 minutes. Audio recordings of the interviews were transcribed and pseudonymized according to the extended simple rules of Dresing and Pehl (2013). Using qualitative content analysis according to Mayring (2010), we deductively determined pre-defined categories and inductively developed categories during coding. The initial coding scheme was developed between the five interviewers, with disagreements being solved via discussion and consensus. After a first round of coding, the inductive code development was carried out by two coders with multiple rounds of coding conferences to ensure reliability.

5. Results

5.1 Detection strategies

We asked our interviewees about the exact procedure that they apply to authenticate online media content, and about which features or characteristics they use to identify disinformation content. Given that textual, visual and audio content all function differently in online media and have different effects on the audience (Dan et al., 2021; Hameleers et al., 2020; Powell et al., 2015; Vaccari & Chadwick, 2020), we asked for each of these media types individually.

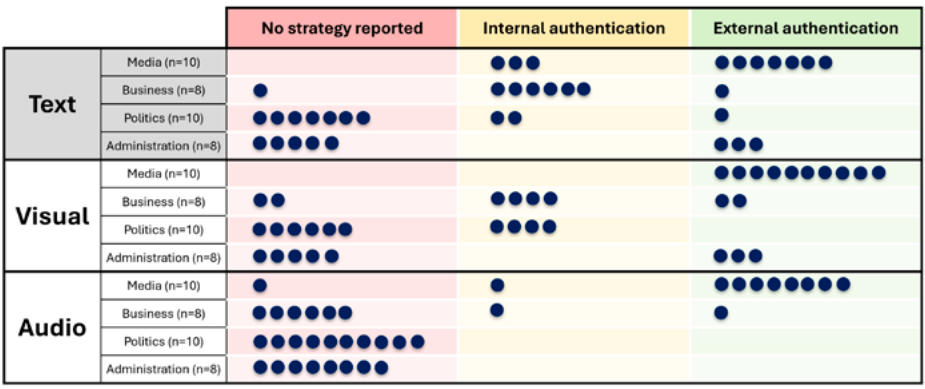
Based on the "audience's acts of authentication" framework by Tandoc et al., we classified the detection strategies that the respondents reported for the different types of synthetic disinformation (text, visual, audio) into one of the following three categories:

1) No strategy: This category was coded when the interviewees did not describe any authentication strategy at all.
2) Internal authentication: This category includes all strategies that are internal acts of authentication. Respondents "go with their gut" and only check their own (instinctive) knowledge and features of the source and the message itself that are immediately apparent to them without referring to any external sources of credibility.
3) External authentication: This category was coded when interviewees described more detailed and complex authentication strategies that go beyond an intuitive and quick comparison with their own experience and instantly apparent features and instead check other (external) sources for credibility. Such strategies correspond roughly to the everyday understanding of what most people would call "fact-checking".

Figure 1 shows an overview of the categories to which respondents from the four professional fields (media, business, politics and administration) were assigned for the three media formats surveyed: Text, visual and audio. The blue dots represent the individual respondents and indicate what type of recognition strategies they described.

Two findings are immediately apparent: First, when comparing the four social groups, journalists (labelled here as "media") distinguish themselves from the others, as they are the only ones who predominantly rely on external sources for content verification. The other three groups trust their internalized knowledge and gut feeling, or they do not describe any recognition strategies at all. Second, the results reveal a particular knowledge gap in audio verification. Apart from media workers, respondents do not appear to have any tools to detect synthetic audio disinformation content.

Figure 1. Detection strategies for different media formats by sector of society



Note. 3 Question asked: "Disinformation and campaigns use different media types like text, images/videos and audio. What characteristics do you use to identify disinformation in online media in the area of [text/visual/audio]? Can you describe concrete examples here?" Each dot represents one participant's responses pertinent to the respective subcategory (n = 36).

5.2 Text content detection

When asked about their methods for identifying disinformation in texts, participants' responses differ between the societal sectors. Figure 1 shows that most of the journalists reported sophisticated strategies that rely on external sources and require a detailed examination of the content, while most respondents from business rely on internal strategies, and most participants from politics and administrations described no strategy at all. This corresponds to the different work practices that the respondents described to us, which seem to result in different levels of engagement with online information in general. While for many journalists intensive scrutiny of the veracity of online texts is part of their daily routine and they primarily deal with news content, respondents from the business sector deal with a wide range of different text content. User reviews and comments on digital platforms play a greater role here, for example. These are primarily evaluated in terms of their harms to the companies. The respondents usually judge the accuracy based on their existing knowledge of the specialist area of their company.

How do the respondents approach verifying text content? Some of the respondents did not describe any detection strategy, because text verification simply is not part of their work.

The most common internal strategies (n = 11) are checking for three aspects of the text. The first marker for falsification is the immediate (formal) appearance of the Text. Spelling errors, lots of emojis or exclamation marks and the like are perceived as reasons to mistrust the information. The same applies to content-related features such as emotional and dramatic language or translation errors. Sloppy translation is understood as an indication of the use of AI, which in turn is almost always equated with an intention to deceive. The third set of features are keywords and "dog whistles". These are trigger words that refer to a narrative that the participant in question already believes to be false. The same procedure is applied to certain authors and sources whom the respondents generally distrust.

Another set of internal recognition strategies relies on the directly accessible knowledge of the respondents. They "go with their gut" and rely on "common sense" and their professional expertise. Or as one respondent from the field of administration put it: *"If someone like me is politically active, they will naturally quickly notice: This is, I think, a certain kind of feeling for language and content that is present"* (S1).

Twelve people also described detection strategies that rely on external sources, most of them working in the field of journalism. The most common of these is a cross-check of the sources mentioned in the message itself, as well as the author of the message. Most journalists also check for further evidence to support the message. Another important external source of credibility is institutionalized verification, especially on social media, as one respondent explains: *"Platform X is making it so difficult for us now since there are no longer any blue checkmarks where you can at least relatively easily know that the sender is OK"* (S1).

5.3 Visual content detection

Visual media is the category for which our respondents were most concerned with the problem of examining synthetic disinformation. They mostly subsumed this under the term "deepfakes" or just "AI". Once again, we see clear differences between the professional groups. While all journalists described elaborate strategies that involve external sources in the verification process, most interviewees from politics and administration told us that they also worry about deepfakes but believe that it is not possible to identify them anymore. For respondents from the corporate sector, the problem is somewhat different. They are more optimistic, since *"usually it's images showing our products that are changed. And we know what our products look like"* (W9).

However, all groups agree that synthetic disinformation technologies are improving rapidly and that distinguishing them from real content will sooner or later become impossible. They only differ in their assessment of the current stage of the technical development of synthetic media technology compared to their own detection skills. Several respondents from the fields of politics and administration said things along the lines of this answer: *"A year ago I would have said*

they were poorly edited images and videos. But I can't say that anymore, because unfortunately, they've gotten really good with this whole AI thing" (P5). The journalists that we talked to see the same problem, but their assessment is different: One put it like this: *"At the moment, you're still learning to pay attention to certain characteristics as a fact checker. And that's how you recognize that this is actually an AI-generated photo. These are often areas like the background or the hair, the hairline. The ears or eyes are sometimes different. But that's just a snapshot.*" Most journalists share this conclusion. For the moment, they are still confident to have sufficient means to recognize synthetic disinformation as such, but *"this will only be temporary, because in two years the AI will no longer be able to use five or six fingers"* (J9). In a nutshell, visual content authentication is perceived as a race between technology and synthetic media literacy, which all respondents expect to lose sooner or later.

How are the participants approaching the authentication of visual content? 13 respondents, primarily from politics and administration, did not describe any detection strategies. Most agree that authenticating synthetic visual disinformation content is impossible. The eight respondents who depicted internal strategies followed a similar approach to the one reported for text content. They either trusted their own knowledge or inspected the immediately apparent appearance of the message for an "unprofessional" or "alternative media aesthetic" and for dramatic and emotional presentations. These features were rated as indicators of inauthenticity.

15 Interviewees (all journalists, 2 from business, 3 from the field of administration) described strategies that relied on different external sources for credibility.

The most frequent way of doing this was a context check. The most frequently described case was not the synthetic generation of images, but the use of real images moved to a different context.

> *And then, we rarely see fake images. Neither through AI nor in any way that someone has done something with Photoshop. Instead, we actually see things being taken out of context. [...] The camera somehow points down a street. And while this live feed is running, two relatively tall buildings are razed to the ground by Israeli rocket attacks. And that actually happened. But it was two or three years old, I think. So, it's being shown again and again in connection with the current war. And that's what we see a lot in photography and video. A real photo, actually taken for some occasion, but it's presented in a completely false context. And it's claimed to be a recent photo. And it would show this and that. But in fact, some of it is years old. And we see that again and again. (J6)*

> *There's a photo of an Airbus A380. Inside this Airbus A380 are large water tanks, each containing 200 liters. And there was a photo that was published, and the water tanks are being used. The water is being pumped around to change the load in the aircraft, how it moves. This image was taken by so-called chemtrails conspiracy theorists to prove that these are containers containing chemical liquids that are then spread during the*

> *flight. It's like I have an image-text mismatch. The text doesn't fit the image, or the text is made to fit the image and isn't reflected in the image. (W6)*

For these cases of decontextualization of visual content, respondents told us *"that's where counter-research really helps"* (J2) and *"you can always do this reverse image search" (J15)*. To verify the context of visual content, digital platforms play an important role, since *"The easiest and fastest way is of course via Google, or other social platforms that are stricter with the awarding of blue checkmarks [to mark verified accounts], for example"* (J2).

The second type of external authentication strategy relies on a complex review of the image material itself. This applies to both artificially generated images and manipulated original content. Our interviewees mostly rely on additional software to do so: *"When something is manipulated, the image noise is often different at some point. With the right tool, you can visualize this"*(J10). The other way they check for image manipulation or generation is, again, context, as this example illustrates: During the German federal election campaign, an AI-fake photo of the Green Party's party conference received a lot of attention. It allegedly showed the event room after the party conference, which was littered with mountains of rubbish, especially large quantities of pizza boxes.

> *That was an AI-generated image, and you could see it. These are the kinds of things that you can still pay attention to now, when people suddenly have five fingers on their hand plus a thumb, or even just two fingers. Or when fashion accessories somehow don't match, clothes look a bit weird. When there are strange characters on the pizza boxes that look like Arabic characters. But really, the pizza delivery service or the restaurant should have some kind of meaningful print on them. So, we look at the content to see if the images are somehow not quite consistent. We pay attention to writing, we pay particular attention to, as stupid as it sounds, people's fingers. (J9)*

This form of authentication examines content-related features of the images and compares them with verifiable features of the (allegedly) depicted objects.

Since these strategies all have in common that they are time-consuming and laborious, many respondents from the field of journalism resort to a third strategy that is faster: *"If in doubt, ask your own followers a question. If the audience is large enough, you'll find many who have probably already considered the same question before"* (J2).

5.4 Audio content detection

Compared to textual and visual media, our respondents express a lower awareness of audio disinformation. Respondents from politics and administration did not describe any strategies for authenticating audio content, as did all but two business representatives. The question of why so many respondents did not describe any strategies here can only be answered inadequately based on this samp-

le. One possible explanation would be that the people concerned had not yet had contact with audio-based disinformation and synthetic audio content in their professional context. Only the journalists seemed to have encountered this problem so far. Those concerned with synthetic audio content and audio disinformation in general nearly always use analytical authentication strategies that rely on external sources. Only two of them reported internal authentication, by trusting to "have a feeling" for the sound and the language of it: *"If they are professionals, you don't notice. But if they are not professionals who are sending messages on answering machines or whatever, then you notice pretty quickly"* (W6).

Respondents who described external authentication strategies for audio content are mostly concerned with fake telephone calls to scam people and audio messages on messenger apps that spread disinformation. In most cases, they use specialized software for authentication. One journalist who worked with a specialized research institute to authenticate audio files told us:

> *We are now essentially dependent on experts or on software that experts create. And this software, especially when it comes to deepfake audio, isn't that widespread yet. So, that's another advantage. You actually have direct contact with the experts who actually create this software. (J6)*

The second external source of credibility is once again swarm intelligence on social media:

> *I've often found this tendency to engage in swarm fact-checking to be surprisingly strong. And I think it's often led me to think, when I wasn't sure what to think about things, that I might actually be inclined to say, 'Okay, maybe that's not true.' Or, 'Okay, maybe that's true, it could be'. (J7)*

6. Discussion and conclusion

We asked which detection strategies German (dis-)information elites use to identify different kinds of synthetic disinformation in textual, visual and audio content and what skills and sources they rely on during the process of authentication. The analysis shows that the "acts of authentication" model by Tandoc et al. (2018) provides a useful basis for understanding and classifying the different detection strategies for synthetic disinformation. We see potential for future research to further investigate the synthetic disinformation detection process.

Our results show that synthetic disinformation detection is perceived as a constant race between technology and harmful actors on one side and improving literacy and countermeasures on the other. For synthetic media content, the effectiveness of internal strategies is perceived to be declining and expected to continue to decline, since all forms of synthetic media, textual, visual or audio content will sooner or later reach a stage where they can no longer be distinguished from authentic content. This seems to be consistent with other research showing a decline in synthetic media detection accuracy (Bray et al., 2023; Groh et al., 2022; Köbis et al., 2021; Liu et al., 2020; Nightingale & Farid, 2022; Shen et al., 2021). An important aspect for future research on synthetic disinformation detection accu-

racy is the consideration of different recognition or other mitigation strategies when empirically testing them.

Our respondents do describe promising external strategies to verify deceptive synthetic content, the most important being context. The more a piece of online content cannot be verified by itself, the more important the context of the information it contains becomes. This applies to all three media formats we examined. In other words, the central question is not whether a medium is genuine or fabricated, but whether the information contained in the message is correct. Following Tandoc's "acts of authentication" framework, the most promising detection strategies are those that rely on external sources and check the context of the information (Tandoc et al., 2018). We see a great need for research here. Previous studies on the detection of synthetic disinformation are structured in such a way that they merely test whether respondents can distinguish authentic from synthetic content. The experimental designs do not allow respondents to verify the context of the stimuli using external sources; instead, their authenticity must be assessed exclusively based on the media content itself. Therefore, only detection strategies based on "internal acts of authentication" can be applied here (Bray et al., 2023; Dobber et al., 2021; Groh et al., 2022; Hwang et al., 2021; Köbis et al., 2021; Liu et al., 2020; Nightingale & Farid, 2022; Shen et al., 2021; Vaccari & Chadwick, 2020). According to our results, the key to synthetic disinformation detection is verifying the context and external sources. Since no representative sample was surveyed for this study, we cannot make any statements about which strategies are used by the general population and which groups are particularly vulnerable to synthetic disinformation. Furthermore, we were unable to empirically test the effectiveness of the described detection strategies in our survey. Future studies might address the question of strategies employed to detect synthetic disinformation with an experimental approach with larger samples and standardized stimulus material and methods, such as self-reporting, while making decisions about such material to get a more precise idea of how, and how successful, various strategies are employed in real-world situations.

The group of journalists can serve as a best practice example for synthetic disinformation detection strategies. They are the only group for which we can reasonably assume that they occupy an elite status regarding their synthetic disinformation literacy and clearly distinguish themselves from the average media recipients. They predominantly describe detection strategies that rely on external sources. Professional training in the authentication of media content, as is common among journalists, is doubtlessly helpful here. Journalists in our sample also use some "elite" detection strategies that aren't readily available to other recipients. For example, complex software tools were often used for audio verification. Also, some journalists rely on their professional networks and large numbers of social media followers to implement the "ask the crowd" strategy to verify online content. Journalists are more concerned about the phenomenon of synthetic disinformation than the other groups and express the most pessimistic outlook. This could also be interpreted as a sign that the other groups still underestimate the scope of the problem. Respondents from politics and administration, who are usually not trained in the verification of media content and whose daily work

rarely involves this activity, may be more vulnerable to synthetic disinformation because they cannot describe adequate methods to detect it. This also applies to the group from the field of business, which relied mainly on internalized knowledge and the resulting gut feeling when making decisions.

Previous research suggests that recipients' trust in digital content itself appears to be declining (Twomey et al., 2023; Vaccari & Chadwick, 2020). It is becoming even more important for the public to be able to rely on trustworthy sources (like democratic institutions and professional media outlets) that do not use synthetic media and do not misinform their audience, but provide context and sources for news and information.

Regarding the three media formats we looked at, our results show different detection approaches to text, visual and audio content.

For text-based disinformation content, respondents more often rely on internal strategies that only check obvious features and rely on what they deem "common sense." Synthetic text generation is described almost exclusively for one use case: The translation of fake news texts in the context of foreign influence operations with the intent to deceive. The most described external detection strategy focuses on comparing information with other sources and gathering further evidence.

In the area of visual disinformation content, our respondents are particularly concerned about synthetic disinformation. Here, the respondents' perceptions align with the focus of previous research. The strategies described primarily aim to verify the authenticity of visual media. The reported internal strategies mostly rely on their own "gut feeling" and expertise and look for obvious AI errors, while external strategies rely on technical tools to detect synthetic media. The most frequently described form of deception is not the fabrication of new content, but rather the alteration of real content to change its meaning or context. Therefore, the most important use case for further literacy research appears to be not the detection and testing of fully generated images and videos, but the detection of manipulation and decontextualization of authentic content.

Deceptive audio content as a category of disinformation is the least well-known to the interviewees. The interviewees from administration and politics, as well as all but two business representatives, described no detection strategies for this or believe that verification is impossible. Those who deal with the detection of audio disinformation (almost exclusively journalists) primarily use technical tools, for which they sometimes rely on additional external expertise. Here we see an urgent need for further research as well. Initial studies indicate that audio-based disinformation does exist, and its influence is growing (Bösch & Divon, 2024). When it comes to resilience, this study suggests that the greatest threat stems from those forms of disinformation that respondents are not yet aware of. The prerequisite for establishing robust detection strategies is problem awareness. One central finding of this study is that most respondents are primarily concerned with detecting deceptive content rather than synthetic content. Our respondents do not treat AI-generated disinformation as an isolated problem, but as another aspect of disinformation and information disorder. When it comes to the content they review, their primary concern is, reasonably enough, whether they are being lied to, not whether the content was synthetically created. Accordingly, many of

the strategies described are not primarily aimed at identifying traces of synthetic disinformation, but rather at assessing the credibility of the message as a whole. However, when synthetic content is identified, it is usually equated with an intention to deceive and viewed as a sign of unreliability. Disinformation as a societal problem is most definitely on the mind of every single person we interviewed.

To sum up, our results show that the information elites in Germany describe detection strategies that usually do not go beyond an internal gut feeling check and are not suitable for detecting new forms of synthetic disinformation. Audio is the biggest blind spot: Synthetic audio disinformation is the least understood and detected, posing a significant future threat. Even the participants themselves view this as a problem when considering the rapid pace of improvement in synthetic, AI-generated media. This does not bode well for the preparedness of society in general when it comes to dealing with this relatively new threat in the larger arena of disinformation.

The most promising detection strategies rely on external sources and, crucially, evaluating the context of the information, rather than just the authenticity of the media content itself. Journalists, due to their training and reliance on external verification, are better equipped to detect synthetic disinformation. Other elite groups (politics, administration, business) often lack adequate methods and may underestimate the problem.

The results also suggest some promising avenues for mitigation: Professional training and methods in verification and analysis seem to be helpful, judging from the answers we recorded in the group of journalists. Problem awareness in all groups is high, which points to a potential willingness to learn the necessary skills. Considering context and consulting external sources for verification and analysis seem to be deemed most useful by those participants who report their strategies most clearly. Future research should focus on these strategies in more detail, since that was beyond the scope of our interview for this study. Future research might then also address how these and other tools can be used and taught – not just to elite actors, since synthetic disinformation is poised to be a major problem for society.

References

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01028-5

Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., Khan, Z., & Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research*, *327*(2), 633–657. https://doi.org/10.1007/s10479-022-05015-5

Bennett, W. L., & Livingston, S. (2023). A brief history of the disinformation age: Information wars and the decline of institutional authority. In S. Salgado & S. Papathanassopoulos (Eds.), *Streamlining Political Communication Concepts* (pp. 43–73). Springer International Publishing. https://doi.org/10.1007/978-3-031-45335-9_4

Bösch, M., & Divon, T. (2024). The sound of disinformation: TikTok, computational propaganda, and the invasion of Ukraine. *New Media & Society*, *26*(9), 5081–5106. https://doi.org/10.1177/14614448241251804

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1). https://doi.org/10.1093/cybsec/tyad011

Calvo, D., Cano-Orón, L., & Abengozar, A. E. (2020). Materials and assessment of literacy level for the recognition of social bots in political misinformation contexts. *ICONO 14, Revista de Comunicación y Tecnologías Emergentes*, *18*(2), 111–136.

CDEI. (2019, September 12). *Snapshot paper – Deepfakes and audiovisual disinformation*. Centre for Data Ethics and Innovation. https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation

Cho, H., Cannon, J., Lopez, R., & Li, W. (2024). Social media literacy: A conceptual framework. *New Media & Society*, *26*(2), 941–960. https://doi.org/10.1177/14614448211068530

Cole, S. (2017, December 11). AI-assisted fake porn is here and we're all fucked. *VICE*. https://www.vice.com/en/article/gal-gadot-fake-ai-porn/

Dan, V., Paris, B., Donovan, J., Hameleers, M., & Roozenbeek, J. (2021). Visual mis- and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 641–664. https://doi.org/10.1177/10776990211035395

Darius, P., & Stephany, F. (2022). How the Far-Right polarises Twitter: 'Hashjacking' as a disinformation strategy in times of COVID-19. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex Networks & Their Applications X* (Vol. 1073, pp. 100–111). Springer International Publishing. https://doi.org/10.1007/978-3-030-93413-2_9

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (micro-targeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Dresing, T., & Pehl, T. (2013). *Praxisbuch Interview, Transkription & Analyse* [Practical guide interview, transcription & analysis] (5th ed.).

Filipovic, A., & Schülke, A. (2023). Desinformation und Desinformationsresilienz [Disinformation and disinformation resilience]. *Ethik Und Militär: Kontroversen in Militärethik & Sicherheitspolitik*, *1*, 34–41.

Frischlich, L. (2019, May 2). Kritische Medienkompetenz als Säule demokratischer Resilienz in Zeiten von "Fake News" und Online-Desinformation [Critical media literacy as a pillar for democratic resilience in times of "fake news" and online disinformation]. *Bundeszentrale* für *politische Bildung*. https://www.bpb.de/themen/medien-journalismus/digitale-desinformation/290527/kritische-medienkompetenz-als-saeule-demokratischer-resilienz-in-zeiten-von-fake-news-und-online-desinformation/

Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, *57*(3). https://doi.org/10.1007/s10462-023-10679-x

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes – An interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, *10*(1), 72–96. https://doi.org/10.5771/2192-4007-2021-1-72

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats*

and potential mitigations (arXiv:2301.04246). arXiv. https://doi.org/10.48550/arXiv.2301.04246

Graves, L., & Amazeen, M. (2019). *Fact-checking as idea and practice in journalism*. Oxford University Press. https://ora.ox.ac.uk/objects/uuid:a7450b2f-f5a7-4207-90e2-254ec5de14e2

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1). https://doi.org/10.1073/pnas.2110013119

Guilbeault, D. (2018). Digital marketing in the disinformation age. *Journal of International Affairs*, *71*(1.5), 33–42.

Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

Higley, J. (2018). Continuities and discontinuities in elite theory. In H. Best & J. Higley (Eds.), *The Palgrave Handbook of Political Elites* (pp. 25–39). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51904-7_4

Hoffmann-Lange, U. (2018). Methods of elite identification. In H. Best & J. Higley (Eds.), *The Palgrave Handbook of Political Elites* (pp. 79–92). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51904-7_8

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base – Analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hugger, K.-U. (2022). Medienkompetenz [Media competence]. In U. Sander, F. von Gross, & K.-U. Hugger (Eds.), *Handbuch Medienpädagogik* (pp. 67–80). Springer Fachmedien. https://doi.org/10.1007/978-3-658-23578-9_9

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Kalsnes, B., Falasca, K., & Kammer, A. (2021). *Scandinavian political journalism in a time of fake news and disinformation* (pp. 283–304). Nordicom, University of Gothenburg. https://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-40895

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11). https://doi.org/10.1016/j.isci.2021.103364

Kong, S.-C., Man-Yin Cheung, W., & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100026

Leschzyk, D. K. (2021). Infodemic in Germany and Brazil: How the AfD and Jair Bolsonaro are sowing distrust during the Corona pandemic. *Zeitschrift* Für *Literaturwissenschaft Und Linguistik*, *51*(3), 477–503. https://doi.org/10.1007/s41244-021-00210-6

Lintner, T. (2024). A systematic review of AI literacy scales. *Npj Science of Learning*, *9*(1). https://doi.org/10.1038/s41539-024-00264-4

Liu, Z., Qi, X., & Torr, P. H. S. (2020). *Global texture enhancement for fake face detection in the wild*. 8060–8069. https://openaccess.thecvf.com/content_CVPR_2020/html/Liu_Global_Texture_Enhancement_for_Fake_Face_Detection_in_the_Wild_CVPR_2020_paper.html

Maros, A., Almeida, J. M., & Vasconcelos, M. (2021). A study of misinformation in audio messages shared in whatsapp groups. In J. Bright, A. Giachanou, V. Spaiser, F. Spezzano, A. George, & A. Pavliuc (Eds.), *Disinformation in Open Online Media* (pp. 85–100). Springer International Publishing. https://doi.org/10.1007/978-3-030-87031-7_6

Martínez-Bravo, M. C., Sádaba Chalezquer, C., & Serrano-Puche, J. (2022). Dimensions of digital literacy in the 21st century competency frameworks. *Sustainability*, *14*(3). https://doi.org/10.3390/su14031867

Mayring, P. (2010). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis. Basics and techniques]. Beltz.

Millière, R. (2022). Deep learning and synthetic media. *Synthese*, *200*(3). https://doi.org/10.1007/s11229-022-03739-2

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100041

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8). https://doi.org/10.1073/pnas.2120481119

Pawelec, M., & Sievi, L. (2023). Falschinformationen in den sozialen Medien als Herausforderung für deutsche Sicherheitsbehörden und -organisationen [Disinformation on social media as a challenge for German security authorities and organizations]. *Kriminologie – Das Online-Journal | Criminology – The Online Journal*, *5*(5). https://doi.org/10.18716/ojs/krimoj/2023.4.7

Koistinen, P., Alaraatikka, M., Sederholm, T., Savolainen, D., Huhtinen, A.-M., & Kaarkoski, M. (2022). Public authorities as a target of disinformation. *European Conference on Cyber Warfare and Security*, *21*(1), 123–129. https://doi.org/10.34190/eccws.21.1.371

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, *64*(6), 763–774. https://doi.org/10.1016/j.bushor.2021.07.012

Powell, T. E., Boomgaarden, H. G., De Swert, K., & de Vreese, C. H. (2015). A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication*, *65*(6), 997–1017. https://doi.org/10/f3s2sj

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494–25513. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3154404

Roe, J., Perkins, M., & Furze, L. (2024). Deepfakes and higher education: A research agenda and scoping review of synthetic media. *Journal of University Teaching and Learning Practice*, *21*(10). https://doi.org/10.53761/2y2np178

Rohs, M., & Seufert, S. (2020). Berufliche Medienkompetenz [Professional media literacy]. In R. Arnold, A. Lipsmeier, & M. Rohs (Eds.), *Handbuch Berufsbildung* (pp. 339–363). Springer Fachmedien. https://doi.org/10.1007/978-3-658-19312-6_29

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06930-7

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). A study of the human perception of synthetic faces. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8. https://doi.org/10.1109/FG52635.2021.9667066

Simonite, T. (2019, October 7). Most deepfakes are porn, and they're multiplying fast. *Wired*. https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/

Stöcker, C. (2020). How Facebook and Google accidentally created a perfect ecosystem for targeted disinformation. In C. Grimme, M. Preuss, F. W. Takes, & A. Waldherr (Eds.), *Disinformation in Open Online Media* (Vol. 12021, pp. 129–149). Springer International Publishing. https://doi.org/10.1007/978-3-030-39627-5_11

Stroebel, L., Llewellyn, M., Hartley, T., Shan Ip, T., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, *7*(2), 83–113. https://doi.org/10.1080/23742917.2023.2192888

Tandoc, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2018). Audiences' acts of authentication in the age of fake news: A conceptual framework. *New Media and Society*, *20*(8), 2745–2763. https://doi.org/10/gc2fmd

Tiernan, P., Costello, E., Donlon, E., Parysz, M., & Scriney, M. (2023). Information and media literacy in the age of AI: Options for the future. *Education Sciences*, *13*(9). https://doi.org/10.3390/educsci13090906

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE*, *18*(10). https://doi.org/10.1371/journal.pone.0291668

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report DGI (2017)09*. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

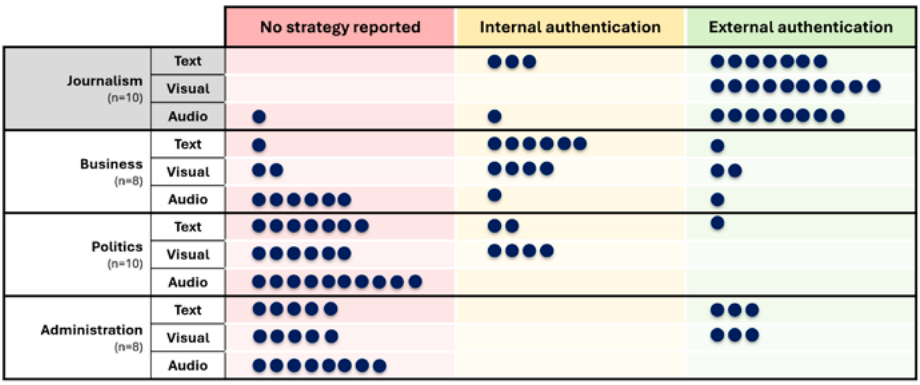
Wasner, B. (2013). *Eliten in Europa: Einführung in Theorien, Konzepte und Befunde* [Elites in Europe: Introduction to theories, concepts and findings]. Springer-Verlag.

Williamson, S. M., & Prybutok, V. (2024). The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*, *15*(6). https://doi.org/10.3390/info15060299

Wu, J., Gan, W., Chen, Z., Wan, S., & Lin, H. (2023). *AI-generated content (AIGC): A survey* (arXiv:2304.06632). arXiv. https://doi.org/10.48550/arXiv.2304.06632

Appendix

Figure 2. Detection strategies for different sectors of society by media format



Note. Question asked: "Disinformation and campaigns use different media types like text, images/videos and audio. What characteristics do you use to identify disinformation in online media in the area of [text/visual/*audio*]? Can you describe concrete examples here?" Each dot represents one participant's responses pertinent to the respective subcategory. (n = 36)

Figure 3. Respondent details

		Type of described detection strategy			
Sector/Subsector/ Party	Position	Text	Visual	Audio	Code
Journalism					
Magazine	department head	internal	external	external	J1
Public broadcaster	editor	external	external	external	J10
Research collective	project lead	external	external	external	J11
Public broadcaster	editor/journalist	external	external	no strategy	J13
Public broadcaster	freelancer	external	external	external	J15
Public broadcaster	multiple roles	external	external	internal	J2
Private broadcaster	department head	internal	external	external	J4
Private broadcaster	department head	internal	external	external	J6
Newspaper	editor	external	external	external	J7
Public broadcaster	staff	external	external	external	J9
Business					
Business association	department head	internal	internal	no strategy	W1
Energy	department head	internal	internal	no strategy	W11
Energy	department head	internal	external	no strategy	W12
Heavy industry	department head	internal	no strategy	no strategy	W2
Mobility	department head	no strategy	no strategy	no strategy	W4
Aerospace engineering	department head	external	external	internal	W6
Energy	department head	internal	internal	external	W7

		Type of described detection strategy			
Sector/Subsector/ Party	**Position**	**Text**	**Visual**	**Audio**	**Code**
Pharmaceuticals	staff	internal	internal	no strategy	W9
Politics					
Die Linke	leadership member	internal	internal	no strategy	P1
Die Grünen	leadership member	no strategy	no strategy	no strategy	P10
CDU	MP	no strategy	no strategy	no strategy	P13
Die Grünen	MP	no strategy	internal	no strategy	P14
Die Grünen	MP Staff	no strategy	internal	no strategy	P15
Die Grünen	leadership member	external	no strategy	no strategy	P3
CDU	department head	internal	no strategy	no strategy	P4
Die Grünen	MP	no strategy	no strategy	no strategy	P5
SPD	leadership member	no strategy	internal	no strategy	P8
CDU	leadership member	no strategy	no strategy	no strategy	P9
Administration					
Federal ministry	staff	external	no strategy	no strategy	S1
Federal government Agency	interim department head	no strategy	external	no strategy	S10
State security agency	staff	no strategy	no strategy	no strategy	S12
Federal ministry	staff	external	no strategy	no strategy	S18
Federal government Agency	vice department head	external	no strategy	no strategy	S3
Federal ministry	department head	no strategy	no strategy	no strategy	S4
Federal ministry	department head	no strategy	external	no strategy	S5
Federal ministry	staff	no strategy	external	no strategy	S8

Media Border Phenomena

OPEN ACCESS

Nora Benterbusch [Ed.]

Spotlights on Media Borders / Perspektiven auf Mediengrenzen

2025, 401 pp., pb., € 104.00
ISBN 978-3-7560-3018-7
E-Book 978-3-7489-6238-0
(Medienkomparatistik: Vergleichende Medien- und Kulturforschung | Comparative Media and Culture Studies, vol. 1)
In Englisch and German

Media border phenomena are fundamental to communicative practice. They allow for reflection on both the limiting properties of media and media constellations as well as the nature of these boundaries. This heterogeneous field—located in both artistic and everyday, historical and contemporary forms of communication—attracts broad disciplinary interest. However, terminological and analytical ambiguities often preclude communication between these perspectives. This book makes an important contribution to interdisciplinary discourse by bringing together diverse theoretical and methodological approaches and case studies, which also provide valuable insights into their respective fields.

With contributions by

Marco Agnetta | Lisa Bauer | Nora Benterbusch | Lars Elleström | Kathrin Engelskircher | Stefan Meier | Thomas Metten | Ana Peraica | Jasmin Pfeiffer | Sebastian R. Richter | Laura Rosengarten | Andrea Rostásy | Patrick Rupert-Kruse | Tobias Sievers | Manuel Van der Veen

Translator

Michael Windgassen