SCM
Studies in Communication and Media

Gegründet im Auftrag der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft e. V. (DGPuK).

Established on behalf of Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft e. V. (DGPuK).

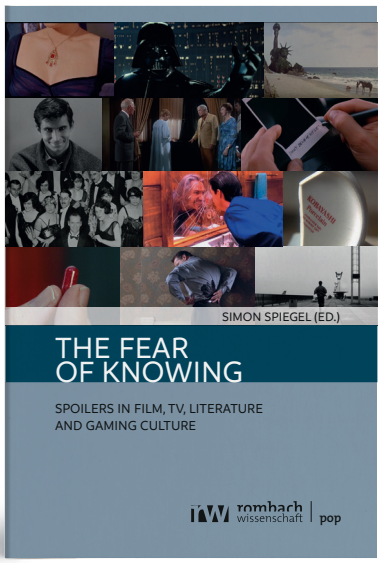
Aus dem Inhalt

4 2025
14. Jahrgang
Seite 471 – 623
ISSN 2192-4007

Nomos

Spoiler Alert

Transdisciplinary Views on Spoilers



OPEN ACCESS

Simon Spiegel [Ed.]

The Fear of Knowing

Spoilers in Film, TV, Literature and Gaming Culture

2025, 327 pp., pb., € 84.00
ISBN 978-3-98858-114-3
E-Book 978-3-98858-115-0
(Pop: Kultur | Medien | Ästhetik, vol. 3)

The fear of spoilers is now a global phenomenon, but it had been barely researched until recently. For the first time, this volume brings together scholars and practitioners from a variety of fields to explore the concept. Do spoilers really diminish our enjoyment of films, books or games? How do different fan communities deal with the issue? Has spoiling always been frowned upon, or were there times when it was handled differently? And how has the fear of spoilers affected the way media content is produced and marketed? The contributors to 'The Fear of Knowing' come from the fields of film, literature, game and fan studies, as well as empirical psychology, and their findings are—spoiler alert!—not always what you might expect.

Simon Spiegel is Privatdozent and senior researcher in film studies at the University of Zurich. He has published widely on science fiction and utopian films, and is the chief editor of the interdisciplinary journal 'Zeitschrift für Fantastikforschung'.

SC|M
Studies in Communication and Media

14. Jahrgang
Heft 4/2025
Seite 471 – 623

Inhalt

Impressum

Herausgeber/Publisher: Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft e. V. (DGPuK), vertreten durch Prof. Dr. Marko Bachl, Berlin | Prof. Dr. Christine Lohmeier, Salzburg | Prof. Dr. Constanze Rossmann, München | Prof. Dr. Helena Stehle, Münster

Redaktion/Editorial Office:
Ruth Kasdorf, Hochschule Wismar, Fakultät Gestaltung, Philipp-Müller-Straße 14, 23966 Wismar
e-mail: scm@nomos.de

Verlag/Publishing Company: Nomos Verlagsgesellschaft mbH & Co. KG,
Waldseestraße 3-5, 76530 Baden-Baden, Germany
Telefon/Phone: 0049 (0)7221/2104-0, Fax: 0049 (0)7221/2104-899

Internet: www.scm.nomos.de

Erscheinungsweise/Publication: viermal jährlich/four times a year

ISSN: 2192-4007

Hinweise für Autorinnen und Autoren finden Sie unter/Hints for authors: www.scm.nomos.de

SCM
Studies in Communication and Media

EDITORIAL

Ready or not, here I come. How synthetic media challenge epistemic institutions

Editorial to the Special Issue

Bereit oder nicht, hier komme ich. Wie synthetische Medien epistemische Institutionen herausfordern.

Editorial zum Sonderheft

Alexander Godulla & Christian Pieter Hoffmann

Alexander Godulla (Prof. Dr.), Leipzig University, Institute of Communication and Media Studies (IfKMW), Nikolaistraße 27–29, 04109 Leipzig, Germany. Contact: alexander.godulla@uni-leipzig.de. ORCID: https://orcid.org/0000-0002-1011-0639
Christian Pieter Hoffmann (Prof. Dr.), Leipzig University, Institute of Communication and Media Studies (IfKMW), Nikolaistraße 27–29, 04109 Leipzig, Germany. Contact: christian.hoffmann@uni-leipzig.de. ORCID: https://orcid.org/0000-0002-5282-6950

EDITORIAL

Ready or not, here I come. How synthetic media challenge epistemic institutions
Editorial to the Special Issue

Bereit oder nicht, hier komme ich. Wie synthetische Medien epistemische Institutionen herausfordern
Editorial zum Sonderheft

Alexander Godulla & Christian Pieter Hoffmann

Abstract: This editorial examines how synthetic media and deepfakes unsettle the epistemic foundations of contemporary public communication. We outline how rapidly advancing generative technologies erode long-standing assumptions about the authenticity of visual and audiovisual content and challenge the institutional capacities of journalism, science, politics, and the arts to maintain credibility and public trust. The contributions to this Special Issue demonstrate these dynamics across different national contexts and communicative domains, highlighting how synthetic media transform political campaigning, newsroom practices, audience cognition and strategies of verification. The resulting picture is one of accelerating technological complexity confronting comparatively slow-moving epistemic institutions. We therefore argue for a coordinated, interdisciplinary research agenda that addresses challenges in media reception and effects, political communication, journalism studies, visual communication, media education, media ethics, media law, and communication history. Such an agenda is essential for safeguarding the integrity of shared knowledge in an increasingly synthetic information environment.

Keywords: Synthetic media, deepfake, journalism, detection, truth, artificial intelligence, trust

Zusammenfassung: Dieser einführende Beitrag untersucht, wie synthetische Medien und Deepfakes die epistemischen Grundlagen zeitgenössischer öffentlicher Kommunikation destabilisieren. Wir zeigen, wie schnell voranschreitende generative Technologien etablierte Annahmen über die Authentizität visueller und audiovisueller Inhalte untergraben und die institutionellen Fähigkeiten von Journalismus, Wissenschaft, Politik und Kunst, Glaubwürdigkeit und Vertrauen herzustellen, herausfordern. Die Beiträge des Special Issues illustrieren diese Dynamiken in unterschiedlichen nationalen Kontexten und Kommunikationsdomänen und verdeutlichen, wie synthetische Medien politische Kampagnen, redaktionelle Arbeitsprozesse, kognitive Rezeptionsmuster und Verifikationsstrategien verändern. Insgesamt ergibt sich das Bild einer technologischen Beschleunigung, die auf epistemische Institutionen trifft, deren Anpassungsfähigkeit vergleichsweise langsam bleibt. Wir plädieren daher für ein koordiniertes, interdisziplinäres Forschungsprogramm, das zentrale Heraus-

forderungen in Medienwirkungsforschung, politischer Kommunikation, Journalismusforschung, visueller Kommunikation, Medienpädagogik, Medienethik, Medienrecht und Kommunikationsgeschichte adressiert. Ein solches Programm ist entscheidend, um die Integrität gemeinsamen Wissens in zunehmend synthetischen Informationsumgebungen zu sichern.

Schlagwörter: Synthetische Medien, Deepfake, Journalismus, Erkennung, Wahrheit, Künstliche Intelligenz, Vertrauen

1. Introduction

The term "deepfake" was first coined in 2017 by a Reddit user in a forum dedicated to discussing the creation of pornographic content (Somers, 2020). It was meant to denote the use of deep-learning technology to create fake depictions of real human beings (Citron & Chesney, 2019). Today, the less ominous term "synthetic media" is commonly applied to AI-generated visual, auditory or audiovisual media (Brady & Meyer-Resende, 2020). While often used interchangeably in public discourse, it could be argued that deepfakes constitute a subtype of synthetic media, as deepfakes depict real individuals in artificially generated contexts. That is what characterizes the potentially deceptive nature of deepfakes and what motivates their close association with "fake news" or disinformation (Altuncu et al., 2022; Dan et al., 2021; Hoffmann et al., 2025; Weikmann & Lecheler, 2023).

Instances of synthetic media that, instead, do not depict actual human beings are rarely considered problematic. Synthetic media can be used for utterly benign purposes, such as the arts and entertainment. In fact, even deepfakes, under specific circumstances, can be employed for constructive purposes, such as education, news, or in the creative industries (Bendahan Bitton et al., 2025). Yet, both in public discourse and in extant research on deepfakes, concerns about their deceptive potential dominate (Bendahan Bitton et al., 2025; Godulla et al., 2021).

In 2021, the authors published a systematic literature review in *Studies in Communication and Media,* highlighting that research on deepfakes, at the time, was (1) dominated by legal studies and computer science, and (2) overwhelmingly focused on risk mitigation (necessary amendments to legal frameworks and technological approaches to deepfake detection). In the social sciences, a range of studies explore user abilities to detect deepfakes and the impact of deepfake encounters on user attitudes (Bray et al., 2023; Lewis et al., 2023; Thaw et al., 2020). Numerous studies find that users struggle to accurately distinguish real from deepfake pictures and videos, even when supported by detection software (for a review, see Somoray et al., 2025; see also Holmes et al, 2025 and Vief et al., 2025, both in this issue).

The latter is a noteworthy finding given the recency of the deepfake or synthetic media technology and its rapid proliferation across society. Within less than a decade since its inception, the average human will no longer be capable of reliably distinguishing a real depiction of actual events from a computer-generated facsimile. We argue that the social and cultural impact of this development is still ill-understood. Most extant research focuses on individuals struggling to recognize specific instances of deepfakes. The wider implication of this failure, however,

affects the epistemic institutional order buttressing modern society. Since the invention of daguerreotype in 1839, humans have been conditioned to trust in the accuracy of photographic depictions of reality (Hoy, 2006). Journalism fundamentally relies on visual and audiovisual media to accurately, reliably and engagingly convey information (Noelle-Neumann, 2000).

Several studies, consequently, find that encounters with deepfakes induce a deep sense of uncertainty in audiences and shake trust in journalism – even bolstering media cynical attitudes (Dobber et al., 2020; Hameleers et al., 2024; Hoffmann et al., 2025; Lee et al., 2021; Vaccari & Chadwick, 2020). The term "liar's dividend" denotes a tactic of discounting unflattering or inconvenient visual and audiovisual depictions as AI-generated (Farid, 2025). In an environment of generalized epistemic uncertainty, any claim to reality can be challenged. Journalism struggles to implement technologies or processes to reliably verify visual and audiovisual digital content. While some studies examine the adoption of artificial intelligence in journalism (Arguedas & Simon, 2023; Graßl et al., 2022; Simon, 2024), and even potential journalistic applications of deepfake technology (Davis & Attard, 2025, in this issue; Raemy et al., 2025), few explore how the rapid proliferation of synthetic media and the ensuing epistemic shock challenge the institutional role of journalism in society.

Beyond journalism, recent examples of fraudulent uses of AI in academic publishing (Hong, 2025) indicate the challenge of generative AI to science. Countless journals now publish AI-generated slop (Naddaf, 2025). Synthetic media, specifically, render established research methods less reliable (Gu et al., 2022). It could even be argued that the epistemic function of the arts is challenged by synthetic media as AI dissolves any boundaries of realistic artistic expression. In short, epistemic institutions face a novel and profound challenge posed by synthetic media and deepfakes. Time plays a key role here, as the tremendous pace of proliferation of the technology is fundamentally at odds with the slow pace of institutional reform and adaptation. New norms of establishing and delineating truth in the absence of reliance on audiovisual representations will likely take decades to evolve.

In many ways, AI-based technologies such as synthetic media and deepfakes build on and contribute to trends that are associated with social media: Journalism no longer maintains its gatekeeping role (Godulla & Wolf, 2024) but rather has accurately been characterized as gatewatching (Bruns, 2009). Social media shakes trust in established institutions – by increasing transparency to a frequently uncomfortable degree, by giving voice to critics, challengers and outsiders, by providing a platform to those challenging authority (Donges et al., 2024; Gurri, 2018; Jungherr & Schroeder, 2021). Science is also subject to these challenges, as, for example, social media played a key role in questioning and undermining scientists' epistemic authority during the Covid-19 pandemic (cf., Park et al., 2022; Van Dijck & Alinejad, 2020).

Likely, those dissatisfied with the status quo and critical of established (epistemic) institutions will be especially drawn to using synthetic media to advance their interests (e.g., Geise et al., 2025, in this issue). Already, deepfakes are used to illustrate critiques that *feel* true to those involved, rather than literally being true (e.g., the deepfake of Democratic candidate Kamala Harris self-describing as a "diver-

sity hire" shared by Elon Musk on X during the 2024 US presidential election; Tenbarge, 2024). Previous studies have shown that misinformation is shared even when known to be untrue if it supports the sharer's worldview (Altay et al., 2022). Conversely, misleading deepfakes are perceived as more credible if they are deemed plausible (Barari et al., 2025; Hameleers et al., 2024), which depends on the content's congruence with viewers' preconceived notions.

Recent events, such as the wars in Ukraine and Gaza, have illustrated how partial, manipulated, decontextualized, or misattributed imagery is shared on social media to misleadingly advance political interests (Hameleers, 2025). Journalism struggles to keep up with and verify such content (Godulla, 2014). Synthetic media and deepfake technology will not just render the verification of visual and audiovisual content more difficult; they will also embed such conflicts of epistemic judgment and authority in a context of generalized uncertainty and mistrust towards media and other epistemic institutions. As noted above, new norms will have to emerge to adjust the epistemic institutional order to a techno-social environment shaped by social media *and* synthetic media or deepfakes (see Vogler et al., 2025, in this issue).

Grappling with the impact of synthetic media and deepfakes on society, thus, requires an inter- and transdisciplinary research effort. Legal studies, computer science, cultural studies, psychology, philosophy, history, sociology and political science, and, of course, communication and media studies need to apply their unique perspectives and methods, and need to collaborate across disciplinary boundaries to establish an understanding of the implications of the rapid proliferation of synthetic media for the epistemic institutional order of the future. The present Special Issue, therefore, had called for contributions from across the various subfields of communication and media studies grappling with the "age of synthetic media".

2. Contributions in the Special Issue

The contributions gathered in this Special Issue respond directly to this call for interdisciplinary engagement. They offer concrete empirical and conceptual insights into how synthetic media are reshaping the conditions under which communication, verification and truth discernment take place. By approaching the phenomenon from multiple angles, the articles illustrate the diversity of challenges that arise when established epistemic institutions encounter rapidly evolving generative technologies. The following sections briefly summarize and contextualize these studies and outline their contributions to understanding the societal implications of synthetic media.

In their full paper, *A new face of political advertising? Synthetic imagery in the 2025 German federal election campaigns on social media*, Stephanie Geise, Anna Ricarda Luther, Sabine Reich and Michael Linke (2025) examine how artificial intelligence is transforming political communication through the strategic use of AI-generated visuals. Based on a quantitative content analysis of more than 1,800 Instagram posts published by Germany's major political parties and their youth organizations during the 2025 federal election campaign, the study identifies 68

synthetic images, corresponding to roughly four percent of all posts. The findings reveal that the Alternative for Germany (AfD) employed such visuals far more frequently than any other party, primarily using photorealistic depictions designed to appear authentic. None of the analyzed images were labeled as artificially produced, raising significant ethical concerns regarding transparency and the potential manipulation of voter perception. The authors demonstrate that AI-generated imagery was used mainly for emotional and ideological framing, particularly through portrayals of "ordinary citizens" and symbolic metaphors that sought to evoke belonging, pride or resentment. Methodologically, the study highlights the limitations of automated AI-detection tools and underscores the superior consistency of structured manual coding. Theoretically, it situates these findings within the concepts of the disinformation order and Habermasian communication ethics, arguing that unlabeled generative visuals undermine the principles of truthfulness and informed deliberation.

The second article, *"The morass is just getting ... deeper and deeper and deeper": Synthetic media and news integrity* by Michael Davis and Monica Attard (2025), explores how Australian newsrooms are responding to the opportunities and challenges posed by generative AI and synthetic media. Drawing on a two-phase qualitative study with editors and product leads from a broad range of media organizations, the authors analyze how journalists perceive and implement AI tools in newsroom workflows, and how concerns over news integrity shape these practices. Their findings reveal an extremely cautious adoption of generative AI in Australian newsrooms, especially regarding the production of audience-facing synthetic media. Most experimentation remains confined to back-end applications such as transcription, summarization, and translation, with limited exploration of synthetic voice or image generation. Across all participating organizations, fears about audience trust, authenticity, and the erosion of editorial standards strongly constrain implementation. The study demonstrates that these apprehensions are grounded not only in professional ethics but also in a broader understanding of journalism's sociopolitical role as a democratic institution. Davis and Attard conclude that while Australian newsrooms recognize the transformative potential of AI, their restrained approach reflects a principled defense of journalistic integrity against both technological hype and the growing dominance of platform economies in shaping information environments.

In the third paper *Spotting fakes: How do non-experts approach deepfake video detection?*, Mary Holmes, Klaire Somoray, Jonathan D. Connor, Darcy W. Goodall, Lynsey Beaumont, Jordan Bugeja, Isabelle E. Eljed, Sarah Sai Wan Ng, Ryan Ede and Dan J. Miller (2025) investigate how individuals without technical expertise attempt to identify deepfake videos and which cognitive and perceptual strategies they employ. Drawing on two complementary studies, the authors examine both self-reported reasoning and eye-tracking data to better understand human behavior in deepfake detection. Study 1, an online experiment with 391 participants, tested whether providing a list of written detection tips could improve accuracy. Although detection rates remained modest, content analysis revealed that the intervention shifted participants' focus on visual cues such as skin texture and facial movement, while the control group relied more on intuition or body language.

Study 2, a laboratory eye-tracking experiment with 32 participants, found similar accuracy levels and revealed that participants primarily directed their gaze to the eyes and mouth, rather than the body, with no differences in gaze patterns between authentic and deepfake videos or between correct and incorrect classifications. The authors conclude that improving human detection may depend on redirecting visual attention from the eyes to more diagnostic cues, such as inconsistencies between face and body or irregularities at facial boundaries, offering valuable insights for future educational and training programs.

In the fourth contribution *Support for deepfake regulation: The role of third-person perception, trust, and risk*, Daniel Vogler, Adrian Rauchfleisch and Gabriele de Seta (2025) analyze how citizens' perceptions of deepfakes relate to their support for state or industry regulation of this emerging technology. Drawing on a pre-registered online survey of 1,361 participants in Switzerland – a country characterized by direct-democratic mechanisms such as referendums – the authors examine whether third-person perception, trust in institutions and risk awareness predict attitudes toward regulation. The study finds strong evidence of a perceptual third-person effect: Respondents believe that deepfakes influence others' opinions more than their own. This perceived influence on others serves as a weak but significant predictor of regulatory support, while the presumed effect on oneself does not. Contrary to expectations, the data reveal no general second-person effect, though exploratory analyses suggest that such a relationship may exist among women, who are disproportionately affected by non-consensual deepfake pornography. In addition, higher trust in political and journalistic institutions as well as heightened risk perception – particularly regarding media, the economy and individual privacy – are positively associated with stronger support for regulation. The authors conclude that public endorsement of deepfake regulation is rooted less in personal vulnerability than in broader concerns about societal risk and institutional trust, highlighting the democratic relevance of perception gaps in emerging technology governance.

In the fifth and final article *Synthetic disinformation detection among German information elites – Strategies in politics, administration, journalism, and business*, Nils Vief, Marcus Bösch, Saïd Unger, Johanna Klapproth, Svenja Boberg, Thorsten Quandt and Christian Stöcker (2025) investigate how professional actors with expertise in disinformation attempt to identify AI-generated content across text, visual and audio formats. Based on guided interviews with 41 elite actors from four sectors of German society – politics, administration, journalism and business – the authors explore which detection strategies these groups employ and which skills and resources they use in the authentication process. The study distinguishes between internal strategies based on intuition and prior knowledge and external strategies relying on verification through other sources. The findings reveal marked differences between the groups: Journalists consistently apply analytical, externally oriented methods, while actors in politics, administration and business mainly rely on intuition or describe no systematic strategy at all. Across all sectors, respondents perceive synthetic disinformation detection as a race between technological progress and human verification skills. Visual content evokes the highest concern, while audio-based disinformation remains largely overlooked. Journalists rely on con-

textual verification, reverse image search, and specialized software, but anticipate that AI will soon outpace human detection capabilities. The study concludes that external, context-based authentication strategies offer the most promising defense against synthetic disinformation yet are currently limited to the media sector.

To summarize, the Special Issue brings together empirical and conceptual work that, first, advances our understanding of how synthetic media reshape the epistemic foundations of contemporary societies. Across methodological approaches and empirical settings, the contributions illuminate how deepfakes and other forms of AI-generated content affect practices of political persuasion, journalistic verification, regulatory practices and elite strategies in information management. Together, the articles demonstrate that synthetic media not only introduce new modes of manipulation, but also challenge institutional norms of authenticity, credibility and public accountability.

Second, the issue spans a broad set of international contexts and thereby highlights that the implications of synthetic media unfold differently across media systems, political cultures and professional traditions. The studies examine the German federal election campaign, Australian newsrooms, Swiss regulatory preferences and the perspectives of German information elites, complemented by experimental research engaging participants from diverse backgrounds. This comparative breadth underscores that synthetic media constitute a global technological phenomenon whose societal effects are mediated by local institutional arrangements, political dynamics and communicative practices.

Third, the contributions approach synthetic media from distinct analytical perspectives, ranging from lay audiences and voters to journalists, political parties and elite actors in public administration, business and politics. They cover key areas of contemporary debate: Human detection capabilities, newsroom adoption and implementation, campaign communication strategies and public support for regulatory interventions. Across these domains, concerns about misinformation, epistemic uncertainty and declining trust recur as central themes. The combined insights of the articles point to a widening gap between the acceleration of synthetic media and the comparatively slow adaptation of epistemic institutions tasked with safeguarding the integrity of public communication. The following contributions address various aspects mentioned above.

3. Future research

Looking ahead, the rapid proliferation of synthetic media calls for a more systematic and programmatic research agenda that addresses the technological, psychological and institutional challenges outlined in this Special Issue. While the existing literature provides important early insights, the accelerating complexity and diffusion of generative models require a broader, more coordinated effort across the subfields of communication and media studies. Future research must therefore clarify how synthetic media reshape established practices of reception, persuasion, verification and representation, and identify which competencies, norms and regulatory frameworks will be necessary to safeguard the epistemic integrity of public communication in the years to come.

Research on *media reception and effects* will need to move beyond documenting losses of trust and instead specify the psychological mechanisms through which synthetic media alter the interpretation of audiovisual content. Future studies should examine how attention, involvement and entertainment value interact with credibility judgments, and which dispositional factors (such as prior knowledge, political attitudes or epistemic vigilance) structure these responses. In addition, robust experimental and field-based research is required to identify scalable interventions that effectively weaken the influence of deepfake misinformation without inducing generalized media cynicism. In *political communication*, a central task for future research is to determine how synthetic media reshape electoral persuasion, strategic messaging and the production and dissemination of political disinformation. While individual persuasion effects remain important, scholars must also investigate how political actors integrate synthetic visuals into campaign repertoires, conflict narratives and targeted mobilization efforts. Comparative and longitudinal designs will be essential to understanding how exposure to political deepfakes shapes voters' beliefs, emotional responses and democratic engagement across political systems and over time.

For *journalism studies*, future research should clarify how professional standards can be maintained in an environment in which the provenance of visual and audiovisual material becomes increasingly uncertain. Systematic work on labeling regimes, verification protocols and transparency practices is needed to determine how synthetic media may be incorporated without eroding the credibility of news products. At the same time, research must examine which technical, analytical and ethical skills journalists require to navigate deepfakes and how these competencies can be integrated into training and newsroom routines. Similarly, the field of *visual communication* faces the task of mapping how synthetic media alter the cultural and cognitive foundations of visual authenticity. Future studies should compare the persuasive power of audiovisual deepfakes with that of text-based or hybrid forms and specify which features, such as plausibility cues, contextual coherence, prior attitudes or psychological predispositions, amplify or weaken credibility. This line of research should also investigate how the very notion of authenticity evolves when the distinction between recorded and generated imagery becomes increasingly opaque.

Media education research must address how citizens can be equipped with the cognitive, technical and ethical competencies needed to critically evaluate synthetic media. Beyond traditional media literacy, future work should identify which specific skills help audiences detect manipulations, question the provenance of audiovisual content and maintain a healthy balance between skepticism and trust. Particular attention should be given to the protection of children and adolescents, who are highly exposed to algorithmically curated visual environments and especially vulnerable to harmful applications. Therefore, future research in *media ethics* must articulate normative boundaries for the creation and circulation of synthetic media, especially when real individuals are depicted in fabricated contexts. Scholars will need to clarify the conditions under which generated content may be used to represent real events, and which obligations arise for educators, journalists and strategic communicators who employ such material. Ethical analysis should also

consider the implications of resurrecting deceased individuals through synthetic media and the responsibilities inherent in shaping public memory through artificial means.

Legal research will need to develop regulatory models capable of preventing harmful uses of deepfake technology without unduly restricting creative expression, innovation or freedom of speech. This includes clarifying the scope of personality rights, privacy protections and liability in cases where synthetic media are used to mislead, defame or deceive. Future work should also address the legal status of synthetic depictions of the deceased and determine under what circumstances such uses may be permissible or require explicit safeguards. Furthermore, *communication history* offers an essential framework for situating synthetic media within a longer trajectory of manipulation, remediation and technological augmentation. Future research should compare contemporary deepfakes with historical practices such as photographic retouching, staged newsreels or digital image editing, and examine how earlier authenticity crises shaped audience expectations. By placing synthetic media within these lineages, scholars can illuminate how trust in audiovisual representation has been constructed, eroded and renegotiated across successive technological epochs.

References

Altay, S., De Araujo, E., & Mercier, H. (2022). "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, *10*(3), 373–394. https://doi.org/10.1080/21670811.2021.1941163

Altuncu, E., Franqueira, V. N., & Li, S. (2022). Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review. arXiv preprint. https://doi.org/10.48550/arXiv.2208.10913

Arguedas, A. R., & Simon, F. M. (2023). *Automating democracy: Generative AI, journalism, and the future of democracy.* Balliol Interdisciplinary Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:0965ad50-b55b-4591-8c3b-7be0c587d5e7

Barari, S., Lucas, C., & Munger, K. (2025). Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics*, *87*(2), 510–526. https://doi.org/10.1086/732990

Bendahan Bitton, D., Hoffmann, C. P., & Godulla, A. (2025). Deepfakes in the context of AI inequalities: Analysing disparities in knowledge and attitudes. *Information, Communication & Society*, *28*(2), 295–315. https://doi.org/10.1080/1369118X.2024.2420037

Brady, M., & Meyer-Resende, M. (2020). *Deepfakes: A new disinformation threat.* Democracy Reporting International.

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1), 1–18. https://doi.org/10.1093/cybsec/tyad011

Bruns, A. (2009). Vom Gatekeeping zum Gatewatching [From gatekeeping to gatewatching]. In C. Neuberger, C. Nuernbergk, & M. Rischke (Eds.), *Journalismus im Internet* (pp. 107–128). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91562-3_3

Citron, D. K., & Chesney, R. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753–1820. https://scholarship.law.bu.edu/faculty_scholarship/640

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

Davis, M., & Attard, M. (2025). "The morass is just getting ... deeper and deeper and deeper": Synthetic media and news integrity. *SCM Studies in Communication and Media*, *14*(4), 517–549. https://doi.org/10.5771/2192-4007-2025-4-517

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Donges, P., Hoffmann, C. P., & Pentzold, C. (2024). Modes of recognition and the persistence of center-periphery constellations in the digital public sphere. *International Journal of Communication*, *18*(2024), 4579–4602.

Farid, H. (2025). Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. *PNAS nexus*, *4*(7). https://doi.org/10.1093/pnasnexus/pgaf194

Geise, S., Luther, A. R., Reich, S., & Linke, M. (2025). A new face of political advertising? Synthetic imagery in the 2025 German federal election campaigns on social media. *SCM Studies in Communication and Media*, *14*(4), 485–516. https://doi.org/10.5771/2192-4007-2025-4-485

Godulla, A. (2014). Authentizität als Prämisse. Moralisch legitimiertes Handeln in der Pressephotografie [Authenticity as a premise: Morally legitimated action in press photography]. *Communicatio Socialis*, *47*(4), 402–410. https://doi.org/10.5771/0010-3497-2014-4-402

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

Godulla, A., & Wolf, C. (2024). Journalismus und Internet [Journalism and internet]. In M. Löffelholz, & L.Rothenberger (Eds.), *Handbuch Journalismustheorien*. (2nd ed., pp. 401–415). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-32153-6_27-1

Gu, J., Wang, X., Li, C., Zhao, J., Fu, W., Liang, G., & Qiu, J. (2022). AI-enabled image fraud in scientific publications. *Patterns*, 3(7). https://doi.org/10.1016/j.patter.2022.100511

Gurri M. (2018). *The revolt of the public and the crisis of authority in the new millennium* (2nd ed.). Stripe Press.

Graßl, M., Schützeneder, J., & Meier, K. (2022). Künstliche Intelligenz als Assistenz: Bestandsaufnahme zu KI im Journalismus aus Sicht von Wissenschaft und Praxis [Artificial intelligence as assistance: Insights into the use of artificial intelligence from a scientific and practical perspective]. *Journalistik: Zeitschrift für Journalismusforschung*, *5*(1), 3–27. https://doi.org/10.1453/2569-152X-12022-12021-de

Hameleers, M., Van Der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, *152*. https://doi.org/10.1016/j.chb.2023.108096

Hameleers, M. (2025). The visual nature of information warfare: The construction of partisan claims on truth and evidence in the context of wars in Ukraine and Israel/Palestine. *Journal of Communication*, *75*(2), 90–100. https://doi.org/10.1093/joc/jqae045

Hoffmann, C. P., Bendahan Bitton, D., & Godulla, A. (2025). The role of self-efficacy and intellectual humility in the relationship between perceived deepfake exposure and media cynicism. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/08944393251354977

Holmes, M., Somoray, K., Connor, J. D., Goodall, D. W., Beaumont, L., Bugeja, J., Eljed, I. E., Ng, S. S. W., Ede, R., & Miller, D. J. (2025). Spotting fakes: How do non-experts approach deepfake video detection? *SCM Studies in Communication and Media*, *14*(4), 550–569. https://doi.org/10.5771/2192-4007-2025-4-550

Hong, S. (2025, August 15). AI-based fake papers are a new threat to academic publishing. Times Higher Education. https://www.timeshighereducation.com/opinion/ai-based-fake-papers-are-new-threat-academic-publishing

Hoy, A. H. (2006). *Enzyklopädie der Fotografie. Die Geschichte – Die Technik – Die Kunst – Die Zukunft* [Encyclopedia of photography. The history – the technique – the art]. National Geographic Deutschland.

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). https://doi.org/10.1177/2056305121988928

Lee, Y., Huang, K. T., Blom, R., Schriner, R., & Ciccarelli, C. A. (2021). To believe or not to believe: Framing analysis of content and audience response of top 10 deepfake videos on YouTube. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 153–158. https://doi.org/10.1089/cyber.2020.0176

Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. *Royal Society Open Science*, *10*(11). https://doi.org/10.1098/rsos.231214

Noelle-Neumann, E. (2000). Wirkung der Massenmedien auf die Meinungsbildung [The effect of the mass media on opinion formation]. In E. Noelle-Neumann, W. Schulz, & J. Wilke (Eds.), *Fischer Lexikon Publizistik, Massenkommunikation* (pp. 518–571). Fischer.

Naddaf, M. (2025). Hundreds of suspicious journals flagged by AI screening tool. *Nature*, *645*, 294–295. https://doi.org/10.1038/d41586-025-02782-6

Park, Y. J., Chung, J. E., & Kim, J. N. (2022). Social media, misinformation, and cultivation of informational mistrust: Cultivating Covid-19 mistrust. *Journalism*, *23*(12), 2571–2590. https://doi.org/10.1177/14648849221085050

Raemy, P., Bendahan Bitton, D., Ötting, H., Hoffmann, C., Godulla, A., & Puppis, M. (2025). Deepfakes and journalism: Normative considerations and implications. *Journalism Studies*, *26*(14), 1724–1744. https://doi.org/10.1080/1461670X.2025.2547300

Simon, F.M. (2024). *Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena*. Tow Center for Digital Journalism Publications.

Somers, M. (2020). *Deepfakes explained*. MIT.

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, *2025*. https://doi.org/10.1155/hbe2/1833228

Tenbarge, K. (2024, August 1). Elon Musk made a Kamala Harris deepfake ad go viral, sparking a debate about parody and free speech. NBC News. https://www.nbcnews.com/tech/misinformation/kamala-harris-deepfake-shared-musk-sparks-free-speech-debate-rcna164119

Thaw, N. N., July, T., Wai, A. N., Goh, D. H., & Chua, A. Y. K. (2020). Is it real? A study on detecting deepfake videos. *Proceedings of the 83rd Annual Meeting of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.366

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

Van Dijck, J., & Alinejad, D. (2020). Social media and trust in scientific expertise: Debating the Covid-19 pandemic in the Netherlands. *Social Media+ Society*, *6*(4). https://doi.org/10.1177/2056305120981057

Vief, N., Bösch, M., Unger, S., Klapproth, J., Boberg, S., Quandt, T., & Stöcker, C. (2025). Synthetic disinformation detection among German information elites – Strategies in politics, administration, journalism, and business. *SCM Studies in Communication and Media*, *14*(4), 594–623. https://doi.org/10.5771/2192-4007-2025-4-594

Vogler, D., Rauchfleisch, A., & de Seta, G. (2025). Support for deepfake regulation: The role of third-person perception, trust, and risk. *SCM Studies in Communication and Media*, *14*(4), 570–593. https:doi.org/10.5771/2192-4007-2025-4-570

Weikmann, T., & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, *25*(12), 3696–3713. https://doi.org/10.1177/14614448221141648

SCM
Studies in Communication and Media

FULL PAPER

A new face of political advertising? Synthetic imagery in the 2025 German federal election campaigns on social media

Ein neues Gesicht politischer Werbung? Synthetische Bilder im Wahlkampf der Deutschen Bundestagswahl 2025 auf Social Media

Stephanie Geise, Anna Ricarda Luther, Sabine Reich & Michael Linke

Stephanie Geise (Prof. Dr.), University of Bremen, Centre for Media, Communication and Information Research (ZeMKI), Linzer Str. 4, Bremen, Germany. Contact: sgeise@uni-bremen.de. ORCID: https://orcid.org/0000-0003-4553-4762
Anna Ricarda Luther (M. Sc.), University of Bremen, Institute for Information Management (ifib), Am Fallturm 1, Bremen, Germany. Contact: aluther@ifib.de. ORCID: https://orcid.org/0000-0002-1169-9297
Sabine Reich (Dr.), University of Bremen, Centre for Media, Communication and Information Research (ZeMKI), Linzer Str. 4, Bremen, Germany. Contact: s.reich@uni-bremen.de. ORCID: https://orcid.org/0000-0002-9474-3635
Michael Linke (M. Sc.), University of Bremen, Centre for Media, Communication and Information Research (ZeMKI), Linzer Str. 4, Bremen, Germany. Contact: mlinke@uni-bremen.de.

FULL PAPER

A new face of political advertising? Synthetic imagery in the 2025 German federal election campaigns on social media

Ein neues Gesicht politischer Werbung? Synthetische Bilder im Wahlkampf der Deutschen Bundestagswahl 2025 auf Social Media

Stephanie Geise, Anna Ricarda Luther, Sabine Reich & Michael Linke

Abstract: The rise of AI-generated content represents a new frontier in political communication. As synthetic media become more sophisticated and accessible, their role in shaping voter perceptions and influencing public discourse warrants closer examination. This study examines the use of AI-generated images in the 2025 German federal election campaign, assessing their prevalence, strategic use, and transparency. We conducted a content analysis of Instagram posts from the major German political parties and their youth organizations in the six weeks leading up to the election. Our analysis focused on identifying AI-generated visuals, evaluating their labeling practices, and examining their communicative and ideological functions. We also compared differences in adoption and usage patterns across parties to assess potential implications for democratic processes. Our findings indicate that the far-right Alternative for Germany (AfD) uses synthetic visuals significantly more than other parties. These AI-generated images are predominantly photorealistic and often lack clear labeling, raising concerns about transparency and potential voter deception. The AfD primarily uses such visuals for emotional and ideological messaging, using AI-generated content to reinforce its political narratives and mobilize support. Our findings provide a structured assessment of AI-generated content in German political communication and highlight the potential risks associated with unregulated use of synthetic media in electoral campaigns. Our research also contributes to the broader discourse on the ethical implications of synthetic media in democratic societies.

Keywords: Synthetic images, generated images, generative AI, election campaigning, campaign strategies, German elections

Zusammenfassung: Die Zunahme von KI-generierten Inhalten stellt eine neue Herausforderung für die politische Kommunikation dar. Da synthetische Medien sich stetig weiterentwickeln und immer zugänglicher werden, muss ihre Rolle für die Meinungsbildung der Wähler*innen und für die öffentliche Debatte genauer untersucht werden. Die vorliegende Studie befasst sich mit der Verwendung KI-generierter Bilder im Wahlkampf zur Bundestagswahl 2025 und zeichnet deren Verbreitung, strategischen Einsatz und Transparenz nach. Anhand einer Inhaltsanalyse der Instagram-Beiträge der großen deutschen Parteien und ihrer Jugendorganisationen in den sechs Wochen vor der Wahl identifizieren wir KI-generierte Bilder, analysieren die Kennzeichnungspraktiken und untersuchen ihre kommunikativen und ideologischen Funktionen. Außerdem vergleichen wir die Unterschiede in der Akzeptanz und Nutzung der Bilder durch die verschiedenen Parteien, um mögliche Auswirkungen auf demokratische Prozesse zu bewerten. Unsere Ergebnisse zeigen, dass die

rechtsextreme Alternative für Deutschland (AfD) deutlich mehr synthetische Bilder verwendet als andere Parteien. Diese KI-generierten Bilder sind überwiegend fotorealistisch und oft nicht eindeutig gekennzeichnet, was Bedenken hinsichtlich der Transparenz und einer möglichen Täuschung der Wähler aufkommen lässt. Die AfD nutzt solche Bilder in erster Linie für emotionale und ideologische Botschaften und setzt KI-generierte Inhalte ein, um ihre politischen Narrative zu verstärken und Unterstützung zu mobilisieren. Unsere Ergebnisse liefern eine strukturierte Bewertung von KI-generierten Inhalten in der deutschen politischen Kommunikation, die die potenziellen Risiken hervorhebt, die mit der unkontrollierten Verwendung solcher Inhalte verbunden sind. Unsere Forschung dient auch einer breiteren Diskussion über die ethischen Implikationen synthetischer Medien in demokratischen Gesellschaften.

Schlagwörter: Synthetische Bilder, generierte Bilder, generative KI, Wahlkampf, Wahlkampfstrategien, deutsche Wahlen.

1. Introduction

The rise of AI-generated content represents a new frontier in political communication. Recent advances in artificial intelligence have made it easier, cheaper, and more effective to create synthetic images, deepfake videos, and other forms of digital content that are nearly indistinguishable from reality (Bray et al., 2023; Lu et al., 2023; Peng et al., 2025). AI's ability to generate synthetic images – defined as visual content that is entirely generated by artificial intelligence and has no photographic source or real-world reference – can blur the line between reality and fiction and raises concerns about misinformation and propaganda (Godulla et al., 2021; Momeni, 2025). In addition, AI-generated content often reflects biases embedded in the training data, resulting in distorted representations of political issues, events, or social groups (Laba, 2024). These biases can reinforce stereotypes, amplify existing power dynamics, and shape public perception in ways that privilege certain narratives over others (Hameleers & Marquart, 2023; Laba, 2024). In political communication, this is particularly problematic as it can distort the democratic debate, manipulate voter sentiment, and contribute to a more polarized information environment (Dobber et al., 2021; Hameleers et al., 2024; Vaccari & Chadwick, 2020). This corresponds to what Bennett and Livingston (2018) refer to as the "disinformation order," in which digital media environments facilitate affective and fragmented communication strategies that can be used to gain a political advantage. In line with these ideas, the increasing accessibility of generative AI tools raises questions about the authenticity of political communication, the ethical boundaries of campaign tactics, and the risks associated with disinformation and voter manipulation (Godulla et al., 2021; Momeni, 2025; Peng et al., 2025). These concerns also address fundamental principles of communication ethics (Habermas, 1983), which emphasize truthfulness, transparency, and the rationality of public discourse as these values are undermined when synthetic media is used without disclosure. Against this background, our study examines the role of synthetic images in campaign advertising, specifically their use on social media by German political parties in the 2025 federal election. These developments are embedded in a broader transformation of political com-

munication, which has undergone profound changes in recent decades, driven by the interplay of digitization, mediatization, and professionalization (Esser & Strömbäck, 2014). While traditional models of voter behavior have emphasized long-term party identification as a stable determinant of electoral choice (Campbell, 1960), research has also highlighted the increasing volatility of voter preferences. The erosion of party loyalty and the rise of undecided and swing voters (Dalton, 2018) have made voting decisions more susceptible to short-term influences, including media framing, campaign strategies, and emotional appeals. As a result, political actors are constantly adapting their communication strategies to take advantage of new technological opportunities to maximize voter engagement and persuasion, and social media platforms have become a central arena for contemporary political campaigns, allowing parties to engage with voters in a highly targeted and interactive manner.

Scholars have described these profound changes in political communication, especially in election campaigns, as the "fourth age" of political campaigning, characterized by the integration of digital technologies, data analytics, and artificial intelligence (e.g., Magin et al., 2017; Semetko & Tworzecki, 2017). As a result, political communication has become increasingly differentiated, with parties and politicians using digital platforms to engage with voters in increasingly precise and sometimes divisive ways (Vaccari & Chadwick, 2020), as newer technologies such as AI-driven predictive analytics allow political actors to dynamically refine their messages and ensure maximum resonance with target audiences (Semetko & Tworzecki, 2017). Such findings are consistent with the broader idea that traditional mass communication methods are increasingly being supplanted by strategies that prioritize direct voter engagement and real-time narrative adjustments.

However, the increased reliance on digital platforms also poses challenges at the societal level, particularly regarding polarization, disinformation, and foreign interference (Vaccari & Chadwick, 2020). For example, Gerbaudo (2018) has argued that the proliferation of social media facilitates the spread of emotional and polarizing content, contributing to the rise of populism in which emotional appeals can overshadow evidence-based discourse. Engesser et al. (2017) showed that such developments can amplify fringe perspectives, as evidenced by the growing popularity of populist parties among younger voters in Germany, who are attracted to their digital-first communication strategies. Some scholars suggest these innovations shape not only the strategies available to political campaigns but also voter perceptions and democratic norms (Perloff, 2021; Vaccari & Chadwick, 2020). In this changing environment, the use of AI-generated imagery in political advertising adds a new dimension to these challenges. First studies show that AI-generated content, particularly synthetic images and deepfakes, has the potential to reinforce political biases, fuel disinformation, shape public perceptions, and influence election outcomes (Dobber et al., 2021; Hameleers & Marquart, 2023; Hameleers et al, 2024). In addition to such micro-level effects, synthetic images that present biased or misleading narratives can also undermine public trust in the media, further complicating the information environment in

which voters make decisions (Hameleers & Marquart, 2023; Ternovski et al., 2022; Vaccari & Chadwick, 2020).

Despite these concerns, research on the role of AI-generated visuals in political communication is still in its infancy. While some scholars suggest that generative AI will become an increasingly integral part of political campaigns (Dobber et al., 2021; Hameleers & Marquart, 2023), empirical evidence on its actual use in elections remains scarce (De Vreese & Votta, 2023; Hameleers et al., 2024; Momeni, 2025). Election campaigns are a particularly relevant context for studying AI-generated content because they involve heightened political messaging, strategic communication, and voter persuasion. If political parties incorporate synthetic visuals into their campaign materials, it could have significant consequences for public opinion formation and the integrity of democratic discourse.

Against this backdrop, our study addresses an urgent empirical and conceptual gap. How are synthetic images currently being used in real-world election campaigns, and what strategic, visual, and ideological functions do they fulfill? Linking the use of AI-generated visuals to concerns about disinformation, emotionalization, and framing in political communication allows us to derive a set of research questions to guide our empirical analysis. This study, therefore, focuses on the prevalence and characteristics of AI-generated imagery in election campaigns based on a quantitative content analysis. While this examination does not address the potential media effects of generative imagery, it will lay the groundwork for future studies on the impact of AI-generated images on democratic processes.

2. Aim of the study

To address these conceptual and empirical challenges, our study focuses on synthetic content specifically in the context of political campaigning. For this study, we specifically focus on synthetic imagery defined as fully AI-generated images (AIGIs), content with no real-world reference. Unlike digitally manipulated visuals, which maintain a connection to reality, AI-generated, synthetic photographs create fictional, photo-realistic scenes from scratch. This definition is based on both conceptual and normative grounds. Conceptually, synthetic photographs represent a qualitative shift in political communication because they fabricate visual "realities" that have no basis in actual events, objects, or materials (Momeni, 2025; Peng et al., 2025). Normatively, synthetic images raise distinct ethical concerns as they exploit the persuasive power of realistic imagery while concealing their artificial origin (Bray et al., 2023; Hausken, 2025). We focus on this form of content because we believe it poses unique challenges to transparency, authenticity, and democratic discourse, especially in the emotionally charged, visually driven context of election campaigning.

Using the 2025 federal German election campaign, the study addresses six research questions: To what extent are synthetic images integrated into campaign ads (1), are AI-generated visuals explicitly labeled to inform the public of their artificial nature (2), and which formats (e.g., video, photography) and applied image types (e.g., portraits, symbolic representations) of AI-generated visuals are

used in political advertising (3)? We also explore how these visuals are linked to specific political issues and campaign strategies (4) and examine differences in their use across political parties (5). Furthermore, we explore which visual characteristics facilitate the identification of AI-generated images as synthetic within the context of political campaign communication (6).

To investigate these aspects, we conducted a quantitative content analysis of Instagram posts from the major German political parties and their youth organizations in the six weeks leading up to the 2025 federal election, measuring the prevalence, labeling and strategic use of AI-generated visuals as well as their characteristics, allowing us to compare differences between parties. Our study provides a structured assessment of AI-generated content in political communication, at least in the German context. The findings contribute to debates on the ethics of AI in elections, transparency in digital campaigns, and risks such as disinformation or voter manipulation (De Vreese & Votta, 2023; Vaccari & Chadwick, 2020). By raising awareness, we aim to inform policymakers, researchers and the public and promote the responsible use of AI in political advertising.

3. Theoretical framework

This study assesses the role of AI-generated images in political campaigning by drawing on four interrelated theoretical strands: The concept of a "disinformation order" (Bennett & Livingston, 2018); Habermas' (1983) ideas of political deliberation and communication ethics; visual and multimodal framing theories; and the mediatization of digital campaigning as a meta-trend in political communication. In the following section, we aim to integrate these strands into a coherent analytical framework that enables us to evaluate the strategic logic and normative implications of synthetic media in electoral communication.

The theoretical framework starts with the theory of *mediatization*, which emphasizes how political communication is increasingly influenced by the logic of digital media (Esser & Strömbäck, 2014). In contemporary campaigning, visibility, emotional resonance, and aesthetic optimization are paramount. Generative AI aligns seamlessly with this logic; it enables political actors to produce compelling and scalable visuals that can dominate social media feeds, bypass journalistic scrutiny, and maximize engagement. This transformation in campaign practice creates fertile ground for the diffusion of synthetic content, particularly among actors willing to experiment outside of conventional communicative norms (Corsi et al, 2024; Momeni, 2025).

Within this mediatized and digitized landscape, the concepts of *visual and multimodal framing* help us understand how AI-generated images and their textual companions (campaign slogans, claims, headlines) contribute to the creation of meaning in political contexts. While visual framing refers to the representational and stylistic choices within individual images that highlight certain aspects of reality while obscuring others (Geise & Baden, 2015; Messaris & Abraham, 2001), multimodal framing builds on this by emphasizing the interplay of visual, textual and other semiotic elements in the creation of meaning (Geise & Xu, 2024; Moernaut et al, 2020; Powell et al., 2019). Building on the work of Grabe and

Bucy (2009), Messaris and Abraham (2014), and Geise and Baden (2015), we conceptualize campaign posts as active rhetorical devices that strategically frame issues, evoke emotional responses, and construct ideological narratives rather than as neutral representations. Studies have shown that it is particularly the photo-realistic quality of the embedded images that amplifies their persuasive impact (Seo, 2020), allowing campaigns to simulate scenarios designed to elicit emotions such as fear, hope, pride, and outrage. Likewise, the photorealistic aesthetic of many AI-generated visuals strengthens this effect by presenting simulated political realities in ways that feel authentic and thus more convincing (Peng et al., 2025).

At the same time, the strategic use of such imagery must be considered in the context of the proposed *disinformation order*, described as a shift toward fragmented, emotionally driven, and often misleading political communication (Bennett & Livingston, 2018). Synthetic visuals embedded in political campaigns, especially when unlabeled, can function as tools of deception, reinforcing polarizing narratives or distorting public understanding (De Vreese & Votta, 2023). These dynamics are particularly salient in electoral contexts, where even subtle manipulations of perception can influence voter sentiment and undermine democratic deliberation. These developments raise urgent concerns about the *ethics of political communication* and campaigning. The idea of deliberation and ethically responsible communication is a well-theorized expectation in democratic societies, particularly within the Habermasian tradition, viewing the public sphere as a space for rational, critical debate based on mutual understanding (Habermas, 1983). According to this perspective, political communication is not merely a tool for persuasion, but rather a normative practice governed by principles such as truthfulness, transparency, and justification. It assumes that, even when strategic, political communication operates within a framework of communicative responsibility and accountability. These expectations are not merely abstract ideals but rather function as institutional guardrails that help sustain public trust and democratic legitimacy. The covert use of AI-generated imagery that mimics reality or conceals its synthetic origin obviously violates these core principles. When political actors disseminate photo-realistic yet fabricated visuals without disclosure, they exploit citizens' trust in visual evidence and circumvent the conditions necessary for making informed judgments. This practice calls into question the authenticity of political communication and undermines the deliberative foundations of democratic participation.

Building on these four strands – mediatization, multimodal framing, disinformation dynamics, and communication ethics – we propose an analytical framework that enables us to examine AI-generated campaign imagery along two axes: (1) its strategic communicative function within mediatized campaigning, and (2) its normative implications for democratic discourse.

This conceptual structure allows us to assess both how and why AI-generated images are used in campaign communication – and what their proliferation implies for the health and integrity of democratic processes. In the empirical sections that follow, we apply this framework to analyze the prevalence, function, and transparency of synthetic images in the 2025 German federal election campaign.

4. Method

4.1 Data collection

A comprehensive content analysis of the Instagram posts of the major German political parties (CDU/CSU, SPD, Bündnis 90/Die Grünen, FDP, AfD, BSW, & Die Linke) and their youth organizations (Junge Union, Jusos, Grüne Jugend, Junge Liberale, Linksjugend/Solid) was conducted during the six weeks before the 2025 federal election (January 12–February 23, 2025). This period was deliberately chosen as it represents the most intense phase of the election campaign, during which parties communicate strategically and rely heavily on multimodal social media content. This period is a well-established time frame to investigate electoral campaigning in Germany (Brettschneider et al., 2007; Wilke & Reinemann, 2003).

To gain a comprehensive understanding of political communication strategies on Instagram, we analyzed both official party channels and their youth organizations. Political parties act as central organizing entities in election campaigns, shaping overarching narratives, policy priorities, and strategic messaging (Farrell & Schmitt-Beck, 2002). While individual politicians may have their own communication styles, party-related content ensures a more consistent and institutionally embedded perspective on campaign strategies. In addition, party accounts often reach a broader audience and serve as the primary vehicle for mobilization and agenda setting on social media (Gibson & McAllister, 2015). By analyzing party communications rather than individual politicians, we aim to capture the structured, collective approach to digital campaigning rather than the personalized and sometimes idiosyncratic strategies of individual candidates.

Political youth organizations play a crucial role in digital campaigning as they often engage in more experimental, activist, and provocative communication styles compared to their parent parties (Ward, 2011; Weber, 2017). They also serve as an important link between parties and young voters, who are particularly active on digital and social media (Hooghe et al., 2004; Weber, 2017). By including both entities, we capture a broader range of campaign strategies, messaging techniques, and audiences, allowing for a more nuanced analysis of how political actors engage different demographics in the digital sphere.

For data collection, a systematic retrieval of all Instagram posts was conducted using *Instaloader* (Graf & Koch-Kramer, 2020), a Python-based tool for downloading social media content. Following the scraping, the Instagram data was checked for completeness by comparing it to the respective Instagram accounts. Collaborative posts (e.g., with individual politicians) were kept in the dataset. Each embedded image was analyzed separately, even if they were part of the same post.

No filtering of the dataset was necessary after scraping. This approach ensured a complete and unbiased dataset of the images and videos that German parties used in their political communication on Instagram. We collected 1,553 Instagram posts from the parties' channels and 315 posts from the corresponding youth organizations as the starting point for further analysis.

For this analysis, we used a sequential procedure, drawing on visual and multimodal framing research. First, we examined the images as independent visual frames. This step was particularly relevant given our focus on AI-generated imagery and our aim to identify distinctive visual characteristics, such as style and synthetic indicators (*see Section 4.3, Coding Categories*). Second, we analyzed the Instagram posts as multimodal ensembles, treating the combination of image and caption as a unified communicative act (Geise & Baden, 2015; Moernaut et al, 2020).

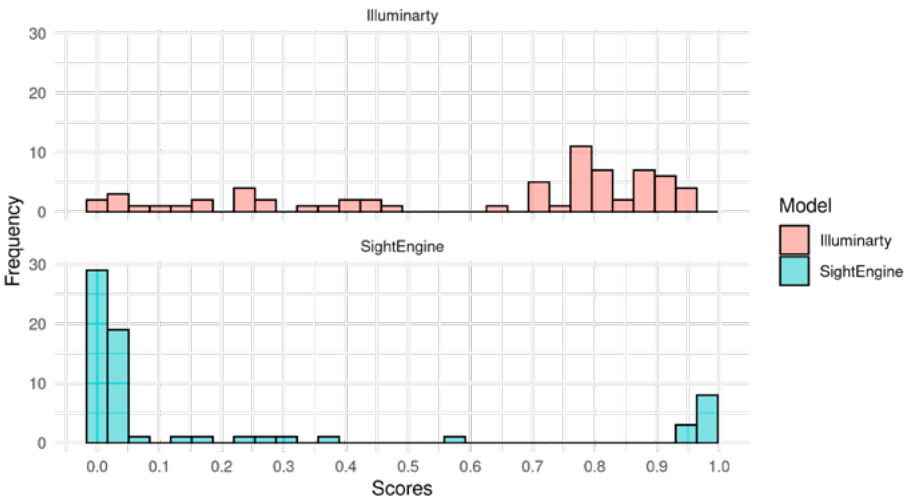
This approach reflects the understanding that communicative meaning, as manifested in the articulation of campaign issues, for example, and strategic framing – as reflected in the promotion of election campaign strategies – often emerges from the interplay of visual and textual elements (Coleman, 2010; Müller & Geise, 2015). Thus, we conceptualize AI-generated visuals as symbolic amplifiers and framing devices within political discourse, both in isolation and as integral components of broader multimodal communication strategies.

4.2 Two-step classification of generated images

After compiling the dataset, we categorized multimodal posts (containing text and images or videos) based on their generative nature, distinguishing between human-created visuals and potentially AI-generated images. To ensure optimal classification accuracy, a two-step validation process was implemented, combining human and automated coding. In the first step, four trained human coders systematically assessed whether an image appeared synthetic based on visual cues and contextual indicators (Mathys et al., 2024). These AI indicators were formalized within a codebook (*see Appendix in OSF*).

Given the potential for human judgment to be subjective, in a second step, images and videos suspected of being AI-generated were further validated using two established AI detection tools (sightengine.com and Illuminarty.ai). As these tools were expected to provide additional insight into whether an image or video has been artificially generated, this should additionally ensure that the classification is reliable. In prior research, SightEngine was shown to be able to achieve a high accuracy compared to other alternatives (Li et al., 2024). Illuminarty has also been tested as a detector of AI-generated images, showing mixed results (Gosselin, 2025).

Figure 1. Comparison of the classification scores of the models



Our analysis revealed significant discrepancies between the automated classification results and our manual coding, as well as inconsistencies between the two AI detection tools. Even in cases where AI generation was either highly likely or very unlikely, both models often produced unreliable or conflicting results. *Figure 1* shows histograms comparing the confidence scores assigned by the two tools, which range from 0–1, with higher values indicating greater confidence that an object was AI-generated.

Illuminarty's classification was slightly closer to manual coding, with a median score of 0.77, while *SightEngine* produced a median score of only 0.02, classifying most images as not AI-generated. While Illuminarty's performance is somewhat in line with previous research, it still deviates significantly from manual classification. SightEngine, on the other hand, performed unexpectedly poorly. One possible explanation could be the nature of the images analyzed, which often contain additional text and graphical elements that may affect the model's performance. However, even this does not explain the large divergence in scores for structurally similar images.

Overall, automated detection tools did not provide reliable validation of AI-generated content due to two key issues. First, there was a high degree of inconsistency – not only between manual and automated coding, but also between the AI models themselves. Second, these tools lack interpretability, as they do not explain why an image is classified as AI-generated or not. This "black box" nature makes the classification process opaque and, in many cases, seemingly erratic.

Although our dataset includes images for which we cannot be completely sure of the degree of AI generation or processing, the substantial discrepancies, especially in cases where classification should be straightforward, undermine the reliability of the automated approach. While human coding is not entirely free of subjectivity, our structured coding scheme and expert review provided greater

reliability and transparency. In contrast, the AI detection models struggled with robustness and generalizability, particularly when faced with images containing text overlays or graphic elements. Therefore, we concluded that automated classification would introduce more uncertainty rather than improve accuracy. As a result, we relied on manual coding, which, despite its limitations, provided a more consistent and interpretable method for evaluating AI-generated content.

4.3 Coding categories

Following the manual classification process, we subjected the identified synthetic posts to a *standardized content analysis*. The coding process was based on a predefined codebook encompassing categories designed to systematically capture patterns in how political actors use synthetic media and how this affects campaign narratives:

Addressing RQ1, we measured the *prevalence of AI-generated visuals* in campaign ads, compared to the number of social media posts in general. Regarding the transparency of AI-generated content, the category *labeling* assessed whether and how synthetic images are marked as AI-generated. Following recent suggestions of practitioners (Burrus et al., 2024; Epstein et al., 2023; Wittenberg et al., 2023), this includes four levels: Clear labeling, where the image is explicitly identified as AI-generated; indirect or hidden labeling, where disclosure is not immediately recognizable; no labeling, where no indication of artificial generation is provided; and deceptive representation, where synthetic images are deliberately presented as real. For the coding of labeling, we took the visual content of the post into account and closely inspected the accompanying text to assess whether any disclosure of AI generation was provided here. This categorization directly addresses RQ2, which investigates the extent to which political actors provide transparency when using AI-generated visuals.

To record the political messaging and political strategy in the election campaign post, corresponding categories were included in the codebook: First, we coded the central *political issue* of each post. Based on a predefined list of 17 categories (cf., Leidecker-Sandmann & Thomas, 2023; Wilke & Leidecker, 2013), this classification covers a broad range of topics, including domestic policy, foreign policy, internal security, social and labor policy, migration, economy, and climate change policy. The codebook also identifies various *election campaign strategies*, each of which can be used to frame political messages and influence public perception. In line with prior research (Klinger et al., 2023; Leidecker-Sandmann & Geise, 2020; Leidecker-Sandmann & Thomas, 2023; Wilke, & Leidecker, 2013), the respective coding category includes 15 commonly used campaign strategies, ranging from personalization, where candidates focus on their personal qualities, to negative campaigning, which targets political opponents, help shape the tone of the posts and thematic focus, which highlights specific issues like climate change or social justice, and emotionalization, which aims to evoke strong feelings. These strategies are coded based on their prominence within the post and can be linked to different political issues, as they may guide the use of AI-generated visuals and their connection to specific campaign objectives. This

enables an analysis of whether synthetic images are used strategically in relation to specific political narratives and whether their presence varies across different issue areas.

Further categories have been implemented in the codebook to better define the style of the post and the image content. We coded the *format of the post*, recording the basic presentation form of the post. The variable measures whether the post contains text, images or videos. The category *visual style of the content* distinguished between different visual styles such as photography, video, graphic illustrations, photomontages, cartoons, memes, and other experimental formats. This classification is essential for answering RQ3, as it allows us to examine whether synthetic images are more prevalent in specific visual styles, such as AI-generated illustrations or manipulated photographs.

We also coded the dominant *image type* used in the posts to examine the communicative strategy behind the visual content. Following the work of Grittmann (2007), this category captures the main theme of each post and includes different picture types, such as portraits of politicians, testimonial images featuring ordinary citizens, symbolic images representing abstract concepts, negative visual stereotypes used to reinforce political narratives, campaign slogans, protest images, and on-the-ground interactions between politicians and the public. Understanding the distribution of these image types is crucial to answering RQ1 and RQ3, as it will allow us to determine whether synthetic images more frequently use certain motifs and picture types, such as AI-generated portraits or visual metaphors, or whether they are used strategically in combination with specific political issues and campaign strategies.

In addition to visual style, content, political messaging, and campaign strategies, the codebook includes a category identifying visual characteristics that suggest an image may be AI-generated, as suggested by prior research (Geise & Yu, under review; Mathys et al., 2024). These *visual AI indicators* include (1) faulty textures or unrealistic surfaces (2) unrealistic facial features or expressions, (3) distorted or unusual body proportions, (4) incoherent combinations or implausible interactions (5) exaggerated colors or unnatural color balance, (6) unnatural lighting or shadowing, (7) irregularities in texts, symbols or numbers, (8) centered composition and symmetry, (9) high level of staging/hyperrealism and (10) visible image or representation errors. The category allowed coders to document up to four key visual markers that signal an AI origin. A more detailed description with example images for each category can be found in the codebook (*see Appendix in OSF*).

By systematically analyzing the visual features, frequency, and types of AI-generated images used across different political parties, our approach offers a thorough assessment of how synthetic images are strategically employed in digital political communication. This methodology contributes to a deeper understanding of the role AI plays in shaping public perception during election campaigns.

4.4 Coding process & intercoder reliability

Two independent coders jointly analyzed a total of 20 posts. After coding the first ten posts, a joint discussion was held to review and resolve any discrepancies and to ensure a common understanding of the coding scheme. Ten further posts were then double-coded to assess inter-coder reliability. The analysis showed satisfactory reliability for the variables examined. For the formal categories *post style* (agreement: 100%, Krippendorff's alpha: 1.00) and *style form of the visual* (agreement: 100%, α: 1.00), coders showed perfect agreement, indicating a clear and objective classification process. Similarly, *AI Labeling* (agreement: 95%, α: 0.89) showed high reliability, reflecting a strong consensus in identifying AI-generated content markers. The *central topic of the post* (agreement: 90%, α: 0.85) and image type (agreement: 90%, α: 0.86) also achieved substantial agreement, confirming that coders were largely in agreement when categorizing the thematic focus and visual format of the posts. For *campaign strategy* (agreement: 85%, α: 0.78) and *AI identifier* (agreement: 80%, α: 0.74), where multiple coding was allowed and coding was more complex and challenging, agreement was slightly lower. However, the values remained within an acceptable range, supporting the reliability of the classification process.

Overall, these results confirm that the coding framework provides a robust and reliable basis for analyzing the use of synthetic imagery in political advertising, with only minor variations in the more complex coding categories.

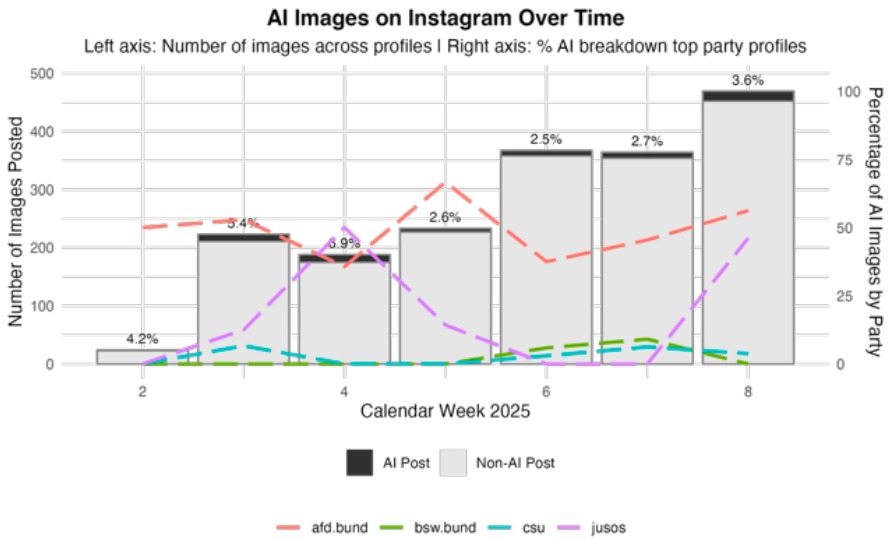
5. Results

With RQ1, we examine the extent to which German political parties use synthetic images in their campaign ads. We identified and downloaded a total of 1,553 images on the Instagram profiles of the parties and 315 images on the profiles of the youth organizations during the study period (January 12–February 23, 2025). Of these, we classified a total of 68 as AI-generated as part of the manual analysis, of which 53 fall on the accounts of the parties and 15 of the youth organizations. This corresponds to a share of 3.8% of AI-generated images in the total volume of posts published on Instagram during the study period. A week-by-week breakdown (*cf., Figure 2*) shows that the share of AI-generated images among all posted images throughout the campaign remained small. For the top posting party, AfD, AI-generated images kept a stable share of around 50% during the election campaign.

In RQ2, we asked to what extent synthetic images or AI-generated posts are explicitly labeled to inform the public of their artificial nature. The standardized content analysis of the 2025 campaign posts revealed that *not a single* political party or youth organization labeled their AI-generated images to inform the public of their artificial nature. This lack of transparency is concerning, as it raises questions about the ethical implications of using synthetic images in political messaging without clear disclosure. The lack of labeling suggests that voters were not made aware of the manipulated nature of the images they were exposed to, potentially leading to a distorted understanding of the candidates or issues being

presented. This could contribute to the manipulation of public perception, as AI-generated imagery often has subtle visual markers that may be recognizable to some but go unnoticed by others. The failure to disclose the use of AI undermines the integrity of political communication, making it more difficult for voters to critically assess the authenticity of campaign content and the motives behind its creation. This lack of transparency in AI use highlights a significant gap in ensuring fair and ethical digital campaigning and raises concerns about potential disinformation and voter manipulation.

Figure 2. AI content over time



Note. The bar chart relates to the left axis, indicating the number of images posted in that respective week. The black highlighted portion of this bar indicates the number of AI images from the entirety of the images posted in that week. The number above the bar displays the percentage of all AI images from all images posted this week across all parties. The dotted line chart relates to the right axis, showing the percentage of AI images per party, in relation to all images that each party posted in the respective week.

RQ3 sought to identify the types of AI-generated visuals, including video, photography, illustration, collage, photomontage and cartoon, used in campaign ads and to examine which specific image types (e.g., portraits of politicians, testimonial images of citizens, symbolic representations, negative visual stereotypes) were used. The analysis revealed that the dominant type of AI-generated image used across all parties was photography (73.5%), followed by collage (25%) and graphic illustration (1.5%). This strong reliance on photorealistic images suggests an intentional effort to create visuals that closely resemble real-life representations, likely enhancing their credibility and persuasive impact on voters. This effect is further intensified by the finding that no AI post is labeled. This is particularly problematic, as prior research has shown that audiences are more likely to percei-

ve AI-generated images as genuine when they resemble real photographs (e.g., Lu et al., 2023) and when depicting humans (e.g., Bray et al., 2023). Without clear labeling, voters may struggle to differentiate between authentic and AI-generated content, increasing the risk of misleading or manipulative campaign tactics.

We also analyzed the main image motifs or picture types to uncover key patterns in the visual strategies used by political parties. This allowed us to assess whether AI-generated images were mainly used for symbolic, emotional, or personalized appeals, and to understand how these choices aligned with broader campaign strategies.

Table 1. Prevalence of AI-generated image types in campaign ads

Rank	Topic label	*n*	Percent
1	Symbolic image/metaphor	32	47.1
2	Testimonial portrait (citizen solo)	21	30.9
3	Testimonial group portrait	5	7.4
4	Politician portrait (solo)	4	5.9
5	Negative visual stereotype	4	5.9
6	Image compilation (e.g., in video)	2	2.9

Our results indicate that AI-generated campaign visuals predominantly feature a narrow set of image types, with symbolic images and testimonial portraits being the most used (*see Table 1*). Symbolic images and visual metaphors (47.1%) serve as the dominant category, likely because they allow for "easy" abstract messaging and emotional engagement without explicitly referencing real-world events or individuals. Example images for the three most prominent image types of symbolic image/metaphor, testimonial portrait and testimonial group can be found in *Figure 3*.

Figure 3. Example images for the image types symbolic image/metaphor, testimonial portrait and testimonial group (from left to right)

Note. The translated text elements from left to right: "How our society looks like, when we invest one billion euros"; "Time for cheap energy – Time for Germany"; "Master plan to strengthen the Bundeswehr and Germany's defence – Swipe now"

Testimonial portraits – both individual (30.9%) and group-based (7.4%) – play a crucial role in personalizing campaign messages by showcasing "ordinary citizens", suggesting a strong strategic focus on portraying the party as "close to the people." In contrast, AI-generated portraits of politicians (5.9%) appear relatively infrequently, suggesting that synthetic visuals focus more on broader narratives than individual political figures. Negative visual stereotypes (5.9%) – while a small category – raise concerns as they could reinforce biases or serve divisive campaign tactics. Image compilations (2.9%), used primarily in video formats, are rare, possibly due to technical limitations or lower effectiveness in short-term campaign messaging.

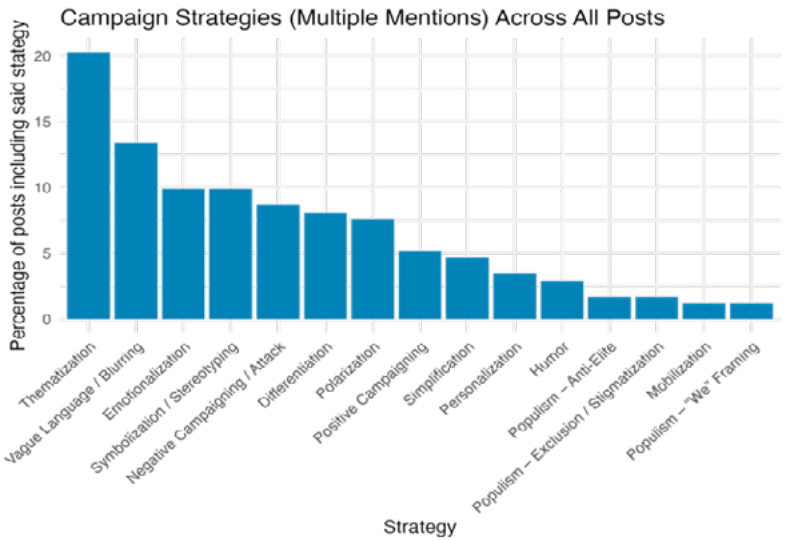
Overall, the findings highlight the selective and strategic use of AI-generated imagery in campaign communication, with an emphasis on abstraction, emotional engagement, and citizen testimonials. The limited variety of image types suggests that parties have not yet fully diversified their AI-generated visual strategies, possibly due to resource constraints or the novelty of these tools in the campaign context.

RQ4 examines how these visuals are associated with political issues and campaign strategies. When examining the *political issue*s communicated with AI synthetic visuals, our analysis revealed that "social policy and justice" emerged as the most frequently referenced subjects. Economy and trade-related issues, as well as asylum and migration policy, were also prominent themes. Each party exhibited a distinct emphasis on these issues in their respective AI-generated posts. The AfD placed significant emphasis on economic issues, migration, and domestic security, while the BSW employed AI visuals exclusively for social justice subjects. The CSU's AI-generated posts primarily addressed the economy and internal security, while the SPD's youth organization (Jusos) concentrated more on social issues compared to the larger parties.

The second part of the question relates to *campaign strategies*. Multiple coding was provided here; up to 3 strategies per contribution could be recorded. In general, analysis of AI-generated posts by political party revealed that the most prevalent strategy adopted was the utilization of thematization, employed in 20.3% of the posts. As a campaign strategy, thematization refers to the deliberate emphasis of specific issues, thereby influencing the salience of particular issues in public discourse. As thematization is a fundamental tool frequently used by parties and candidates to align their messages with voter concerns and media agendas (Perloff, 2021), the high prevalence in AI-generated campaign posts is not surprising. Unlike more specific strategies such as emotionalization or polarization, thematization serves as a basic function of political messaging. However, when combined with these more targeted strategies, it can contribute to a more populist style of communication.

In addition to thematization, AI-generated posts frequently used vague language and blurring (13.4%), as well as emotionalization (9.9%), symbolizing and stereotyping (9.9%). Other recurring strategies, each appearing in more than 5% of the posts, included negative campaigning, differentiation, positive campaigning, and polarization. The picture becomes clearer if only the AfD, which produced the most AI-generated posts, is considered (*see RQ5*).

Figure 4. The percentages of posts employing different campaign strategies



Since only six image types appear in the sample, showing little overall variance, further analysis of the association of specific image types with political strategies or issues is limited (see *Table 2* & *A2*). It should also be noted that up to three strategies could be coded per post. Taking these issues into account, the analysis shows that symbolic images, the most frequently used type of visuals, are mainly used to set themes. However, they often appear in combination with strategies such as stereotyping, differentiation, polarization and negative campaigning, suggesting that they are also employed to sharpen ideological divides and reinforce simplified narratives. This aligns with findings in populist communication research, where simplified, emotionally charged imagery is used to delegitimize political adversaries (Ernst et al., 2019; Schmuck & Matthes, 2017).

Table 2. AI-generated image types and their associated campaign strategies

Rank	Sujet (image type)	Strategy	Count
1	Symbolic image/metaphor	Thematization	19
2	Testimonial portrait (citizen solo)	Thematization	12
3	Testimonial portrait (citizen solo)	Vague language/blurring	12
4	Symbolic image/metaphor	Symbolization/stereotyping	10
5	Testimonial portrait (citizen solo)	Emotionalization	10
6	Symbolic image/metaphor	Differentiation	9
7	Symbolic image/metaphor	Negative campaigning/attack	8
8	Symbolic image/metaphor	Polarization	7
9	Symbolic image/metaphor	Simplification	7
10	Symbolic image/metaphor	Vague language/blurring	6

Testimonial portraits are often combined with vague language and emotionalization, aligning with their intended function: Testimonials are designed to signal proximity to voters and foster an emotional connection. By featuring (AI-generated) "ordinary citizens" and pairing them with emotionally charged yet ambiguous messaging, campaigns aim to create a sense of relatability and engagement while leaving room for broad identification. However, traditionally, such portraits build trust and emotional connections by featuring *real people* who support a party's message. This makes the AfD's reliance on AI-generated, entirely fictional individuals particularly paradoxical: While these images are meant to represent "citizens like you and me", they instead depict synthetic figures with no real agency. As a result, they become carefully controlled representations rather than authentic endorsements, raising critical concerns about credibility, transparency, and potential voter skepticism.

The analysis of the image type crossed with the central political themes is subject to similar limitations as the analysis of the association of image types and strategies, but here, only one central theme was coded per post. The image types, in combination with the central political themes, show interesting patterns. The most frequently used image type, symbolic image/metaphor, is particularly used in relation to the two political issues "social policy & justice" and "economy, trade & finance". Testimonial portraits are also frequently used, especially combined with "economy, trade & finance" and "culture & education", suggesting that personal connection and authenticity are emphasized in these areas. Negative visual stereotypes are used less frequently but are particularly associated with sensitive issues such as asylum and migration policy or social policy and justice, suggesting a strategic use of negative images to shape public perception. Portraits of politicians (alone) are more often associated with elections and election campaigns, illustrating the emphasis on individual political personalities in campaign imagery.

RQ5 asked how the use of synthetic imagery varies across political parties. Here, a key difference between the parties can be seen in the frequency of use of AI-generated images: The Alternative for Germany (AfD) has the highest frequency of AI use, with nearly half of all posts containing AI-generated images (n = 39). Other major parties had significantly lower usage rates, such as the Christian Social Union (CSU) with eight posts and the Sarah Wagenknecht Alliance (BSW) with three posts. Among the youth organizations, only the Jusos showed a significant level of AI usage with 13 AI-generated images, followed by the Linksjugend with two AI images. These results suggest that synthetic images were particularly prevalent in the AfD's digital communication strategy, while other parties, including the youth organizations, used AI to a much lesser extent. This suggests that AI-generated images may play a greater role in the campaign tactics of certain parties, particularly those that target specific voter groups, address specific campaign issues, or pursue specific campaign strategies.

A differentiated view by party also shows a clearer picture of the strategies used, especially as the AfD produced the most AI-generated posts. Analysis shows that the most common strategy applied by AfD is thematization (16 posts), closely followed by emotionalization (14 posts) and vague language/blurring (13 posts). The other parties posted significantly less AI-generated content, and no

clear strategic trends were observed in their posts. This suggests that the AfD's use of AI in its communication is more intentional and focused on specific strategies. These findings strongly suggest a populist style of communication (Ernst et al., 2019; Hameleers & Schmuck, 2017). Populist (visual) rhetoric aims to mobilize audiences through simple messages, visual stereotypes, and emotional appeals, while delegitimizing opponents and framing politics as a binary struggle, often at the expense of democratic norms (Bast, 2024; Ernst et al., 2019; Schmuck & Matthes, 2017). The incorporation of generative imagery into these strategies further amplifies their effectiveness, raising concerns about misinformation and manipulation. Given the increasing role of AI in political communication, understanding these dynamics seems critical to addressing the broader implications for democratic discourse and electoral integrity.

In addition, primarily the AfD relied heavily on testimonial portraits of individuals and groups (*see Appendix, Table A1 in OSF*). This suggests a strategic focus on citizen representations and emotionally charged narratives, potentially reinforcing populist messaging styles. Symbolic images and metaphors are widely used across parties, emphasizing the role of abstract visual messaging in AI-generated campaign communication. While the AfD leads in this category as well (n = 15), the CDU, CSU, and the Greens also employ this strategy. Another notable aspect is the use of negative visual stereotypes, almost exclusively found in AfD (n = 2), CSU (n = 1), and BSW (n = 1) posts. This highlights differences in visual communication strategies between parties, with some employing more polarizing imagery than others.

Table 3. Visual AI identifier represented in AI-generated election posts

Rank	Identifier	*n*	Percent
1	Faulty textures or unrealistic surfaces	40	58.8
2	Visible image or representation errors	25	36.8
3	Unrealistic facial features or expressions	24	35.3
4	Incoherent combinations or implausible interactions	23	33.8
5	High level of staging/hyperrealism	22	32.4
6	Unnatural lighting or shadowing	14	20.6
7	Irregularities in text, symbols, or numbers	8	11.8
8	Exaggerated colors or unnatural color balance	7	10.3
9	Distorted or unusual body proportions	6	8.8
10	Centered composition and symmetry	1	1.5

RQ 6 explores the visual characteristics that facilitate the identification of AI-generated images as synthetic within the context of political campaign communication. Our analysis revealed that key visual markers that facilitate the identification of AI-generated images are present in the context of political campaign communication (*see Table* 3). The most prevalent visual AI identifier was "faulty textures or unrealistic surfaces," which was observed in more than half of all images (58.8%). This finding suggests that a considerable proportion of AI-generated images are deficient in their depiction of realistic surface textures, a deficiency that can serve as a discernible indication of their synthetic origin. Inspecting sur-

faces such as clothing or hair, particularly when in motion, can aid in the identification of AI-generated visuals. *Figure 5* gives an example of such faulty textures, evident in the implausible movement of the clothing texture.

Figure 5. Example image for AI identifier "faulty textures or unrealistic surfaces". This was posted by the @afd.bund account on Instagram on 14.01.2025.

Note. Translated text elements: "Finally free in your own country – Time for Germany"

The second most prevalent characteristic, "visible image or representation errors," was identified in more than one-third of the images (36.8%), underscoring the prevalence of errors in the representation of objects or scenes. This identifier is likely most unambiguous due to its clear faultiness, such as six fingers or hovering objects. As these errors frequently occur in smaller details, they might require a more detailed inspection. *Figure 6* depicts an example image for this AI identifier from our dataset. Here, one visible image error is the change in material of the sitting bench from yellow plastic to brown wood.

Figure 6. Example image for the AI identifier "visible image or representation errors" and "unrealistic facial features or expressions". This was posted by the @ jusos account on Instagram on 24.01.2025

Note. Translated text element: "Punctual public transport everywhere"

Additionally, "unrealistic facial features or expressions" were observed in one third of the images (35.3%), suggesting that AI models face challenges in accurately replicating natural facial expressions, potentially resulting in unnatural or distorted depictions of individuals. *Figure 6* can also serve as an example for this AI identifier due to the distorted facial features of the depicted girl. A detailed inspection of the facial features, particularly eyes, ears and mouth, allows for the identification of this error. Furthermore, our analysis showed that "incoherent combinations or implausible interactions" were present in 33.8% of the images, suggesting that AI models frequently encounter difficulties in generating logical and coherent interactions between people, objects and scenes, resulting in images that may appear implausible.

Figure 7. Example image for the AI identifier "incoherent combinations or implausible interactions". This was posted by the @afd.bund account on Instagram on 06.02.2025.

Note. Translated text elements: "Exclusive analysis: This is how the CDU is financing the terror against themselves – Time for Germany"

Figure 7 illustrates this AI identifier. Illogical combinations of image parts from Friedrich Merz are evident due to the mixing of scenes from different sources that do not harmonize with each other. The interaction between the two depictions of Friedrich Merz is also implausible, not only regarding the scene itself but also due to the incorrect posture and relation between the two arms. To recognize this identifier, detailed attention to interaction points (e.g., the parts where a hand is grasping an object) as well as to the overarching scene (e.g., how the bodies are positioned to one another) is necessary.

The "high level of staging/hyperrealism" category, which appeared in 32.4% of the images, suggests that AI tends to generate highly idealized, almost surreal visuals, thereby creating a hyperrealistic atmosphere that may appear oversimplified and artificial. *Figure 8* illustrates this example, displaying an unnaturally polished appearance characterized by precise lighting and exaggeratedly composed poses. This hyperreal aesthetic, which lacks the subtle irregularities of authentic photography, can indicate synthetic image generation.

Figure 8. Example image for the AI identifier „high level of staging/hyperrealism". This was posted by the @afd.bund account on Instagram on 16.02.2025.

Note. Translated text elements: "Now it's our turn – Time for Germany"

Other less frequent but still significant visual AI-features included unnatural lighting or shadowing (20.6%) and irregularities in text, symbols, or numbers (11.8%). These errors often point to the AI's inability to replicate real-world complexities like correct typographic elements. The least prevalent features were exaggerated colors or unnatural color balance (10.3%), distorted or unusual body proportions (8.8%), and centered composition and symmetry (1.5%). These findings imply that, while AI has achieved substantial progress in generating visuals, it continues to grapple with the creation of entirely realistic and coherent representations of the physical world.

The identification of visual markers indicative of synthetic imagery, such as unrealistic textures, distorted facial features, inconsistent lighting, or unnatural proportions, is crucial for assessing the authenticity of political imagery. By identifying these markers, researchers and voters can be more informed about the origins of the visuals they encounter, which is crucial in an era where the lines between real and fake can easily be blurred. From the perspective of the user, these findings are of particular significance as they underscore the challenges encountered by AI-generated visuals in the context of political campaign communication. As synthetic imagery becomes more prevalent in political campaigns, these visual markers can serve as indicators for users to critically assess the authenticity of content. The identification of these characteristics empowers users to discern when an image may lack authenticity, thereby contributing to the maintenance of transparency and the mitigation of potential manipulation or misrepresentation. In a political context, the ability to identify AI-generated images is of particular importance, as

these visuals could be used to shape public opinion or influence voters by presenting idealized or fabricated representations of candidates, events, or policies.

6. Discussion

Our analysis of AI-generated imagery in the political campaign around the German federal election in 2025 reveals systematic patterns in how AI-generated images are applied in election campaigns. Building on the theoretical framework outlined above, this discussion interprets our empirical findings along two central analytical axes: (1) the strategic communicative function of AI-generated images within mediatized campaigning, and (2) their normative implications for democratic discourse. This dual perspective allows us to examine how generative visuals are used in practice and how they reflect broader transformations in mediatized political communication.

6.1 Strategic use of AI-generated imagery in mediatized election campaigning

Our findings show a significant difference in the frequency with which political parties use synthetic images. The AfD stands out as the most frequent and systematic user of AI-generated images. Other major parties, including the CSU and BSW, used AI-generated images only sporadically. This asymmetry suggests that the AfD has integrated AI tools as a core part of its campaign strategy, while other parties have remained more cautious or traditional in their approach. Closely linked to this strategic adoption is another pattern: The predominant use of photo-realistic images, and the relatively limited use of other image types (e.g., collages, graphic illustrations) highlights a preference for visuals that appear authentic. This photo-realistic visual style serves a dual purpose: From a multimodal framing perspective, this strategy serves to capture attention and enhance emotional appeal, reinforcing credibility through the illusion of authenticity. Both functions closely align with the logic of mediatized campaigning, which prioritizes emotional resonance and visibility over deliberative content – advantages that AI-generated imagery can help deliver more effectively for parties willing to innovate within this logic.

Many AI-generated visuals featured so-called "ordinary citizens", representing testimonials. While this strategy humanizes campaign messages and suggests proximity to the electorate, the use of fictitious, AI-generated individuals introduces a paradox: Employing entirely synthetic personas to promote party credibility undermines the very authenticity these visuals seek to convey, revealing the tension between strategic emotional appeal and the risk of credibility loss. Such communicative practices compromise the conditions necessary for open, rational, and informed public discourse, even within persuasive election campaigning.

It is noteworthy how little creative variation political actors display when using generative image AI. In our sample, the range of subjects and types of images is mostly limited to standard campaign imagery, such as pseudo-portraits of candidates or supporters and symbolic representations of issues. This limited use contrasts with the broader range of political imagery documented in previous

studies. For example, Grittmann's (2007) typology of political image types and Müller's (1997) historical analysis of visual strategies in U.S. presidential campaigns illustrate how political actors have long used diverse image motifs to construct identity, credibility, and emotional appeal. These findings suggest that political parties have not yet fully exploited the aesthetic and narrative potential of AI-generated visuals for election campaigns.

Most visuals conformed to familiar campaign tropes – symbolic images and portraits – suggesting that even innovative tools are subsumed under traditional visual campaign logic rather than used for novel messaging. The analysis also highlights that symbolic images are often combined with strategies such as stereotyping, differentiation, and polarization, suggesting that these images serve not only to engage voters emotionally but also to reinforce ideological divides and simplify political narratives. This again was particularly evident in AfD content, which used such visuals to create binary oppositions and reinforce ideological divisions. The use of negative visual stereotypes, while less common in our sample, is of particular concern in this context as it targets sensitive issues such as asylum and migration policy or social policy and justice, potentially using negative imagery to divide public opinion. In these examples, the visuals not only conveyed policy positions but served to delegitimize political opponents through affective framing. Such practices reflect the logic of the "disinformation order" (Bennett & Livingston, 2018), which – as outlined in our theoretical framework – emphasizes the erosion of rational discourse through emotionally charged media content. In such campaigns, AI-generated imagery can become a vehicle for further eroding democratic communication norms.

The specific policy issues addressed in AI-generated posts reveal clear patterns. Social policy and justice is the most common theme, followed by economy and trade, as well as asylum and migration policy. The parties vary in their foci, with the AfD emphasizing economic concerns, migration, and internal security, while other parties, such as the BSW, focus more on social justice issues. These themes are consistent with the broader visual strategies, with symbolic images and emotional appeals serving to shape the public's perception of these issues. These findings reflect a multimodal framing logic in which images are not merely illustrations but rather central devices for ideological positioning. The emotional framing of these issues through AI-generated imagery underlines how mediatization enables the amplification of affective and symbolic narratives, reinforcing party-specific ideological positions and voter mobilization strategies.

6.2 Normative implications for democratic discourse

Turning to the normative perspective, the common unlabeled use of photorealistic synthetic images challenges the principles of communicative responsibility that are essential to deliberative democracies, as suggested by Habermas' (1983) communication ethics, which emphasize sincerity, truthfulness, and rational justification as foundations of discourse. By disguising fabricated visuals as authentic representations, political actors undermine the public's ability to make informed judgments and violate core deliberative norms such as transparency, truthfulness,

and justification. Especially when combined with populist rhetorical strategies, (unlabeled) generative images can be used to fabricate misleading narratives, reinforce stereotypes, and influence election outcomes (Dobber et al., 2021; Hameleers et al, 2024).

The strategic amplification of polarizing and emotionally charged imagery also exacerbates the fragmentation of public debate. Using visually amplified, emotional, antagonistic, yet stereotypical and under-complex messages narrows the space for rational deliberation, mirroring the democratic risks associated with Bennett and Livingston's idea of disinformation order. That way, AI-generated images can contribute to the erosion of informed, rational political debate, further exacerbate societal polarization, and weaken democratic norms. Given AI's growing role in political communication, it is crucial to understand these dynamics early on to address the broader implications for democratic discourse and electoral integrity.

6.3 Potential avenues for regulation, resilience and research

Despite these challenges, our analysis identifies potential avenues for resilience as some AI-generated visuals still exhibit noticeable characteristics that can be identified by laypeople without technical expertise – particularly when prompted to scrutinize the image. A close examination of textures and lighting, as well as common inconsistencies in specific areas of the human body (e.g., eyes, hands, ears, hair), or the background can help voters recognize AI-generated images. While the detection of synthetic images remains challenging even for trained coders and automated tools, some of the AI indicators can still be identified by laypeople, given that they are aware of them and spent some time inspecting the image more closely. The presence of detectable artifacts in some synthetic images provides a tangible leverage point for media literacy interventions. Encouraging citizens to critically inspect visuals and recognize AI-generated cues could mitigate the risk of manipulation, fostering an electorate that is more informed and capable of navigating the media-saturated and AI-permeated information landscape. While these indicators likely evolve as AI technology advances rapidly, their current presence provides an opportunity to enhance public awareness and critical engagement with political visuals.

Additionally, our findings underscore the urgent need for regulatory measures, such as the mandatory labeling of AI-generated content, to ensure transparency and accountability. Alongside media literacy efforts, strengthening transparency regulations and labeling practices are crucial for countering the normative threats posed by synthetic campaign imagery and protecting democratic legitimacy.

This highlights a possible way for restoring deliberative integrity through institutional safeguards, such as labeling, as well as civic education and media literacy. These methods reinforce the normative conditions that underpin democratic communication, emphasized in communication ethics and our theoretical framework.

The strategic and normative analyses show that AI-generated visuals are a political instrument, not just a technical innovation. Their deployment reflects the two analytical axes introduced in our theoretical framework. Strategically, they

function as tools for mediated campaigning and affective and multimodal framing. Normatively, they raise significant concerns about the erosion of deliberative democratic principles. This dual role highlights the appeal and democratic risks of AI-generated imagery in political contexts. They serve distinct strategic functions within the logic of mediated campaigning while raising profound normative challenges to democratic discourse and electoral integrity. Applying our two-dimensional theoretical framework, which focuses on strategic function and normative implications, to our content analytical data allows us to better understand the appeal and risks of AI-generated campaign content.

Future research should explicitly address the impact of AI-generated imagery on voter perception, public opinion formation, and the broader democratic process. Although our study, conceptualized as content analysis, cannot empirically assess these effects, the use of photorealistic synthetic images, emotional framing, and polarizing visual strategies observed suggests the potential influence of AI-generated imagery on voter trust, the spread of disinformation, and social polarization. Investigating how audiences interpret and respond to such imagery is essential to comprehensively evaluating the societal consequences of AI-mediated political communication. This research could also better inform the development of effective regulatory and educational interventions to protect democratic discourse in an increasingly AI-saturated media environment.

7. Limitations

Our study investigates the use of synthetic images in campaign advertising, with a particular focus on their presence on social media during the 2025 German federal elections. By conducting a content analysis of Instagram posts of the major German political parties, we aimed to explore the extent to which synthetic images were integrated, whether AI-generated visuals were explicitly labeled, and the types of synthetic visuals used. We examined how these images were associated with specific political issues and campaign strategies, and how their use differed across political parties. This study is novel in the context of German political campaigns, as it is the first to assess the role of AI-generated images in this specific electoral setting. However, this novelty is reflected in the relatively small sample size, with only 68 posts identified. While this number allows for an initial understanding of the use of synthetic images, it is a limitation for a more in-depth analysis. This study provides valuable insights, but further research with a larger sample size would be beneficial to confirm and extend these findings.

Due to recurring access issues with Instaloader, the data collection process proved challenging and required continuous manual verification of the scraped content against the original Instagram posts. While this iterative comparison ensured the completeness and accuracy of the dataset, it significantly undermined the intended benefit of automation. As a result, the process became time-intensive and only partially scalable, highlighting a key limitation in relying on third-party scraping tools for systematic social media research.

Another limitation of our study is the exclusive focus on Instagram as the social media platform. While we hypothesize that other platforms may yield similar

results, this remains speculative and future research would need to include multiple platforms to fully assess the extent of synthetic image use in political campaign advertising. In addition, our study does not address the potential influence of synthetic imagery on voter perception or behavior, which may be an interesting avenue for future research.

The attempted automated classification also had a few critical limitations: A comparison between more than two classifiers would have been more insightful, but two was the only option within the given time frame. Decisions made by these models are not transparent due to their design as black boxes (in terms of the architecture and the data used to train them). Their exact performance cannot be calculated based on the given data, due to the human coders' own uncertainty. The limited access restricted our possibilities to perform extensive tests. For instance, the performance on partial images could not be tested, so it cannot be ruled out that the classification was influenced by subsequent edits, like inserted logos or text. Only images that had previously been manually coded as AI images were processed. It would have been interesting to see the full confusion matrix, which, however, would come with its own problems, since the dataset would have been highly imbalanced.

It is also important to note that the analysis was conducted in the specific context of the German federal elections, and the findings may not be readily transferable to other political contexts. The German political system, with its multi-party structure and the situational aspects of the election, such as the early dissolution of the government leading to a snap election, are factors that could influence the results. These contextual elements need to be considered when interpreting the results and applying them to other electoral settings or political systems.

Online appendix

Available at the OSF repository https://osf.io/y59um

References

Bast, J. (2024). Managing the image. The visual communication strategy of European right-wing populist politicians on Instagram. *Journal of Political Marketing*, *23*(1), 1–25. https://doi.org/10.1080/15377857.2021.1892901

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1), 1–18. https://doi.org/10.1093/cybsec/tyad011

Brettschneider, F., Niedermayer, O., & Weßels, B. (2007). Die Bundestagswahl 2005: Analysen des Wahlkampfes und der Wahlergebnisse [The German federal election 2005: Analyses of the election campaign and results]. In: F. Brettschneider, O. Niedermayer, B. Weßels (Eds.), *Die Bundestagswahl 2005* (pp. 9–18). VS https://doi.org/10.1007/978-3-531-90536-5_1

Burrus, O., Curtis, A., & Herman, L. (2024). Unmasking AI: Informing authenticity decisions by labeling AI-generated content. *Interactions*, *31*(4), 38–42. https://doi.org/10.1145/3665321

Campbell, A. (1960). Surge and decline: A study of electoral change. *Public Opinion Quarterly*, *24*(3), 397–418. https://psycnet.apa.org/doi/10.1086/266960

Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-140

Dalton, R. J. (2018). *Citizen politics: Public opinion and political parties in advanced industrial democracies*. CQ Press.

De Vreese, C. D., & Votta, F. (2023). AI and political communication. *Political Communication Report*, *2023*. http://doi.org/10.17169/refubium-39047

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

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society*, *20*(8), 1109–1126. https://doi.org/10.1080/1369118X.2016.1207697

Ernst, N., Blassnig, S., Engesser, S., Büchel, F., & Esser, F. (2019). Populists prefer social media over talk shows: An analysis of populist messages and stylistic elements across six countries. *Social Media+ Society*, *5*(1). https://doi.org/10.1177/2056305118823358

Epstein, Z., Arechar, A. A., & Rand, D. (2023). What label should be applied to content produced by generative AI? *PsyArxiv preprint*, 2023. https://doi.org/10.31234/osf.io/v4mfz

Esser, F., & Strömbäck, J. (2014). *Mediatization of politics: Understanding the transformation of Western democracies*. Springer. https://doi.org/10.1057/9781137275844

Farrell, D.M., & Schmitt-Beck, R. (Eds.). (2002). *Do political campaigns matter? Campaign Effects in Elections and Referendums* (1st ed.). Routledge. https://doi.org/10.4324/9780203166956

Geise, S., & Baden, C. (2015). Putting the image back into the frame: Modeling the linkage between visual communication and frame-processing theory. *Communication Theory, 25*(1), 46–69. https://doi.org/10.1111/comt.12048

Geise, S., & Xu, Y. (2024). Effects of visual framing in multimodal media environments: A systematic review of studies between 1979 and 2023. *Journalism & Mass Communication Quarterly*, *102*(3), 796–823. https://doi.org/10.1177/10776990241257586

Gerbaudo, P. (2018). Social media and populism: an elective affinity? *Media, Culture & Society*, *40*(5), 745–753. https://doi.org/10.1177/0163443718772192

Gibson, R. K., & McAllister, I. (2015). Normalising or equalising party competition? Assessing the impact of the web on election campaigning. *Political Studies*, *63*(3), 529–547. https://doi.org/10.1111/1467-9248.12107

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

Gosselin, R. D. (2025). AI detectors are poor western blot classifiers: A study of accuracy and predictive values. *PeerJ*, 13. https://doi.org/10.7717/peerj.18988

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195372076.001.0001

Graf & Koch-Kramer (2020) *Instaloader*. Retrieved January 10, 2025, from https://github.com/instaloader/instaloader.

Grittmann, E. (2007). *Das politische Bild: Fotojournalismus und Pressefotografie in Theorie und Empirie* [The political image: Photojournalism and press photography in theory and empirical research]. Herbert von Halem.

Habermas, J. (1983). *Moralbewußtsein und kommunikatives Handeln* [Moral consciousness and communicative action]. Suhrkamp.

Hameleers, M., & Marquart, F. (2023). It's nothing but a deepfake! The effects of misinformation and deepfake labels delegitimizing an authentic political speech. *International Journal of Communication*, *17*, 6291–6311.

Hameleers, M., & Schmuck, D. (2017). It's us against them: A comparative experiment on the effects of populist messages communicated via social media. *Information, Communication & Society*, *20*(9), 1425–1444. https://doi.org/10.1080/1369118X.2017.1328523

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, 152. https://doi.org/10.1016/j.chb.2023.108096

Hausken, L. (2024). Photorealism versus photography. AI-generated depiction in the age of visual disinformation. *Journal of Aesthetics & Culture*, *16*(1). https://doi.org/10.1080/20004214.2024.2340787

Hooghe, M., Stolle, D., & Stouthuysen, P. (2004). Head start in politics: The recruitment function of youth organizations of political parties in Belgium (Flanders). *Party Politics*, *10*(2), 193–212. https://doi.org/10.1177/1354068804040503

Klinger, U., Koc-Michalska, K., & Russmann, U. (2023). Are campaigns getting uglier, and who is to blame? Negativity, dramatization and populism on Facebook in the 2014 and 2019 EP election campaigns. *Political Communication*, *40*(3), 263–282. https://doi.org/10.1080/10584609.2022.2133198

Laba, N. (2024). Engine for the imagination? Visual generative media and the issue of representation. *Media, Culture & Society*, *46*(8), 1599–1620. https://doi.org/10.1177/01634437241259950

Leidecker-Sandmann, M., & Geise, S. (2020). Tradition statt Innovation. Die deutsche Presseberichterstattung über die Wahlkampfstrategien der Parteien zur Bundestagswahl 2017 [Tradition instead of innovation. The German press coverage of political parties' campaign strategies in the run-up to the 2017 parliamentary elections]. *SCM Studies in Communication and Media*, *9*(2), 264–307. https://doi.org/10.5771/2192-4007-2020-2-264

Leidecker-Sandmann, M., & Thomas, F. (2023). "Never was there more to do." Use of vaguely formulated statements in the 2021 German national election campaign and their potential effects. In C. Holtz-Bacha, (Ed.), *Die (Massen-)Medien im Wahlkampf: Die Bundestagswahl 2021* (pp. 43–66). Springer Fachmedien Wiesbaden.

Li, Y., Liu, Z., Zhao, J., Ren, L., Li, F., Luo, J., & Luo, B. (2024). The adversarial AI-art: Understanding, generation, detection, and benchmarking. In *European Symposium on Research in Computer Security* (pp. 311–331). Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2404.14581

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *Advances in Neural Information Processing Systems*, 36, 25435–25447. https://doi.org/10.48550/arXiv.2304.13023

Mathys, M., Willi, M., & Meier, R. (2024). Synthetic photography detection: A visual guidance for identifying synthetic images created by AI. *arXiv preprint arXiv:2408.06398*. https://doi.org/10.48550/arXiv.2408.06398

Magin, M., Podschuweit, N., Haßler, J., & Russmann, U. (2017). Campaigning in the fourth age of political communication. A multi-method study on the use of Facebook by German and Austrian parties in the 2013 national election campaigns. *Information, Communication & Society*, *20*(11), 1698–1719. https://doi.org/10.1080/1369118X.2016.1254269

Messaris, P., & Abraham, L. (2001). The role of images in framing news stories. In S. D. Reese, O. H. Gandy, Jr., & A. E. Grant (Eds.), *Framing public life: Perspectives on media and our understanding of the social world* (pp. 231–242). Routledge.

Moernaut, R., Mast, J., & Pauwels, L. (2020). Visual and multimodal framing analysis. In L. Pauwels, D. Mannay (Eds.), *The SAGE Handbook of Visual Research Methods* (pp. 484–499). SAGE Publications. https://doi.org/10.4135/9781526417015.n30

Momeni, M. (2025). Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *Journal of Creative Communications*, *20*(1), 41–56. https://doi.org/10.1177/09732586241277335

Müller, M. G. (1997). Visuelle Wahlkampfkommunikation: Eine Typologie der Bildstrategien im amerikanischen Präsidentschaftswahlkampf [Visual campaign communication: A typology of image strategies in the American presidential election campaign]. *Publizistik, 42*(2), 205–228. https://doi.org/10.1007/BF03654575

Perloff, R. M. (2021). *The dynamics of political communication: Media and politics in a digital age*. Routledge. https://doi.org/10.4324/9780429298851

Powell, T. E., Boomgaarden, H. G., De Swert, K., & de Vreese, C. H. (2019). Framing fast and slow: A dual processing account of multimodal framing effects. *Media Psychology*, *22*(4), 572–600. https://doi.org/10.1080/15213269.2018.1476891

Peng, Q., Lu, Y., Peng, Y., Qian, S., Liu, X., & Shen, C. (2025, April). Crafting synthetic realities: Examining visual realism and misinformation potential of photorealistic AI-generated images. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–12). https://doi.org/10.1145/3706599.3719834

Schmuck, D., & Matthes, J. (2017). Effects of economic and symbolic threat appeals in right-wing populist advertising on anti-immigrant attitudes: The impact of textual and visual appeals. *Political Communication*, *34*(4), 607–626. https://doi.org/10.1080/10584609.2017.1316807

Semetko, H. A., & Tworzecki, H. (2017). Campaign strategies, media, and voters: The fourth era of political communication. In J. Fisher, E. Fieldhouse, M.N. Franklin, R. Gibson, M. Cantijoch & C. Wlezien (Eds.), *The Routledge Handbook of Elections, Voting Behavior and Public Opinion* (pp. 293–304). Routledge.

Seo, K. (2020). Meta-analysis on visual persuasion–does adding images to texts influence persuasion? *Athens Journal of Mass Media and Communications*, *6*(3), 177–190. https://doi.org/10.30958/ajmmc.6-3-3

Ternovski, J., Kalla, J., & Aronow, P. (2022). The negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety*, *1*(2). https://doi.org/10.54501/jots.v1i2.28

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), 2056305120903408. https://doi.org/10.1177/2056305120903408

Ward, J. (2011). Reaching citizens online: How youth organizations are evolving their web presence. *Information, Communication & Society*, *14*(6), 917–936. https://doi.org/10.1080/1369118X.2011.572982

Weber, R. (2017). Political participation of young people in political parties. *Zeitschrift für Politikwissenschaft*, *27*, 379–396. https://doi.org/10.1007/s41358-017-0106-z

Wilke, J., & Leidecker, M. (2013). Regional – national – supranational. How the German press covers election campaigns on different levels of the political system. *Central European Journal of Communication*, *6*(1(10)), 122–143

Wilke, J., & Reinemann, C. (2003). Die Bundestagswahl 2002: Ein Sonderfall? [The German federal election 2002: A special case?]. In: C. Holtz-Bacha, C. (Eds.), *Die Massenmedien im Wahlkampf* (pp. 29–46). VS. https://doi.org/10.1007/978-3-322-80461-7_3

Wittenberg, C., Epstein, Z., Berkinsky, A.J., & Rand, D.G. (2023). Labeling AI-generated content: Promises, perils, and future directions. Topical Policy Brief, *MIT Schwarzman College of Computing*. https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf

SCM
Studies in Communication and Media

FULL PAPER

"The morass is just getting ... deeper and deeper and deeper": Synthetic media and news integrity

„Der Morast wird immer ... tiefer und tiefer und tiefer": Synthetische Medien und Nachrichtenintegrität

Michael Davis & Monica Attard

Michael Davis (Dr.), University of Technology Sydney, Centre for Media Transition, 15 Broadway, Ultimo NSW 2007, Sydney, Australia. Contact: Michael.Davis@uts.edu.au. ORCID: https://orcid.org/0000-0002-1113-3664
Monica Attard (Prof.), University of Technology Sydney, Centre for Media Transition, 15 Broadway, Ultimo NSW 2007, Sydney, Australia. Contact: Monica.Attard@uts.edu.au. ORCID: https://orcid.org/0000-0003-0798-4340

FULL PAPER

"The morass is just getting ... deeper and deeper and deeper": Synthetic media and news integrity

„Der Morast wird immer ... tiefer und tiefer und tiefer": Synthetische Medien und Nachrichtenintegrität

Michael Davis & Monica Attard

Abstract: With the arrival of generative AI (genAI) in 2022, waves of hype and handwringing struck the news industry. These initial responses have proved overblown, if not without foundation. The challenges and opportunities of synthetic media for news are real, if more humdrum than the hype would suggest. In this paper, we draw from a two-phase qualitative study to explore how these challenges and opportunities have manifested in Australian newsrooms. We focus on: 1) How are newsrooms implementing genAI in the production of synthetic media? 2) How do newsrooms perceive the potential impacts of synthetic media on news integrity? 3) How are perceived impacts on news integrity mediating the implementation of genAI, particularly for synthetic media? Industry surveys have shown that uptake of genAI in Australian newsrooms is low relative to comparable markets. In phase 1, we found almost no use of genAI to produce synthetic media for publication. This reflected apprehension over the limitations of genAI and acute consciousness of threats to trust and news integrity. Phase 2 found some moderation of concern as low-risk opportunities had emerged, though applications in audience-facing content were still limited. Participants continued to express strong concerns about news integrity and audience trust. We apply both a technological process lens and a normative lens focused on the concept of news integrity to interpret participant insights. We conclude that the limited uptake of genAI in Australian newsrooms is driven by concerns about news integrity in a broad sense, going beyond journalistic standards to encompass the sociopolitical functions of journalism as well as concerns about continued platformisation of the media economy and an increasingly degraded information environment.

Keywords: Journalism, generative AI, synthetic media, news integrity, trust

Zusammenfassung: Seit dem Aufkommen generativer KI (genKI) im Jahr 2022 erlebt die Nachrichtenindustrie Wellen der Begeisterung wie auch Besorgnis. Wenn auch die Herausforderungen und Möglichkeiten synthetischer Medien für Nachrichten real sind, erweisen sich diese als weniger aufregend als erwartet. Im Beitrag wird eine qualitative Studie in zwei Phasen vorgestellt, die untersucht, wie sich diese Herausforderungen und Möglichkeiten in australischen Nachrichtenreaktionen entfalten. Dabei fokussieren wir uns auf: 1) Wie wenden Redaktionen genKI in der Produktion synthetischer Medien an? 2) Wie nehmen Redaktionen potenzielle Auswirkungen synthetischer Medien auf die Nachrichtenintegrität wahr? 3) Wie beeinflussen diese Auswirkungen die Anwendung von genKI, insbe-

sondere für synthetische Medien? Erhebungen innerhalb der Industrie haben gezeigt, dass der Einsatz von genKI in australischen Nachrichtenredaktionen gering ist. In Phase 1 konnten wir feststellen, dass kaum genKI zur Produktion synthetischer Medien verwendet wurde, was auf Bedenken hinsichtlich technischer Begrenzungen wie auch Bewusstsein über die Gefährdung von Vertrauen und Nachrichtenintegrität hindeutet. Phase 2 deutet zwar auf das Aufkommen weniger risikoreicher Möglichkeiten hin, wenn auch die Anwendung in publikumsorientierten Inhalten weiterhin begrenzt blieb und Teilnehmende nach wie vor Besorgnisse hinsichtlich der Nachrichtenintegrität und des Vertrauens des Publikums äußerten. Zur Interpretation der Aussagen der Teilnehmenden wenden wir sowohl eine technologische Prozessperspektive als auch eine normative Perspektive an, die auf das Konzept der Nachrichtenintegrität fokussiert ist. Wir schließen daraus, dass die begrenzte Anwendung von genKI in australischen Redaktionen auf Bedenken hinsichtlich der Nachrichtenintegrität zurückzuführen ist. Diese Bedenken gehen über journalistische Standards hinaus und umfassen sowohl die sozio-politischen Funktionen des Journalismus als auch Sorgen über die anhaltende Plattformisierung der Medienökonomie und ein zunehmend degradiertes Informationsumfeld.

Schlagwörter: Journalismus, generative KI, synthetische Medien, Nachrichtenintegrität, Vertrauen

1. Introduction

With the arrival of generative AI (genAI) in 2022, waves of hype and handwringing struck the news industry. On the one hand, the technology heralded a new era of automation that would escalate production without increasing costs and deliver novel formats that would rejuvenate declining audiences. On the other, it would threaten jobs and undermine news quality. Meanwhile, increasingly sophisticated deepfakes would degrade political discourse, damage electoral integrity and accelerate the decline in public trust (Beckett & Yaseen, 2023; Ternovski et al., 2022). These polarised "utopian and dystopian portrayals" (Cools & Diakopoulos, 2024, p. 1) have proved overblown (Simon et al., 2023), if not without foundation. The challenges and opportunities of synthetic media for news are real, if both more humdrum and more profound than the hype suggests.

In this paper, we draw from a two-phase empirical study into the impact of genAI on public-interest journalism in Australia to explore how these challenges and opportunities are being negotiated in Australian newsrooms.

AI has been making its way into news output for nearly two decades, mostly through automated reporting from structured data sources (Bäck et al., 2019, p. 11). In many cases, "the technology has slowly moved into news production and distribution, often without readers (or journalists) really noticing" (Simon & Isaza-Ibarra, 2023, p. 8). Discourse was polarised even in these early days, with one side championing the potential transformation of news production through technological innovation, and the other focused on industry disruption, particularly threats to journalists' jobs, as in discussions of "robot journalism" (Lindén & Dierickx, 2019).

In part, the polarised discourse surrounding AI must be understood in the "larger context of the digitization of media and public life", which has transformed journalism, "undercutting business models, upending work routines, and

unleashing a flood of information alternatives to news" (Lewis, in Broussard et al., 2019). Questions about ethical practice, news values and journalistic purpose are also never far from mind. As Moran and Shaikh (2022, p. 1757) attest, debates within newsrooms about technology are embedded in broader conversations about the role and efficacy of journalism, and about where its boundaries lie. These centre on the question of how technologies "advance or hinder a particular normative vision for journalism".

With the emergence of genAI, both the technological and the normative questions have been amplified. The potential transformation of production is seen as more significant than earlier AI technologies, but so is the potential impact on news as an industry and sociopolitical institution. On one hand, AI represents "the next level" of technical sophistication. On the other hand, AI is "fraught with myths, political connotations and emotional responses that stand in the way of an informed debate on AI, within and outside newsrooms" (Helberger et al., 2022, p. 1606).

In our study, we find deep engagement within the news industry with both the technological and normative questions. We find that the implementation of genAI in newsrooms is mediated largely by concerns about ethical practice and the sociopolitical functions of journalism, though resource limitations also play a role. We apply both a technological process lens and a normative lens to investigate the implications of AI-generated synthetic media for the integrity of news. GenAI's technical capabilities and limitations are inseparable from normative questions about the desirability of its application in news. In examining both, we can build a fuller picture than by, e.g., applying a classical technology acceptance model (Bagozzi, 2007), or an ethical analysis divorced from the economic and labour imperatives driving technological adoption.

While the normative lens considers common journalistic standards such as accuracy and fairness, our interviews reveal journalists are thinking about AI-generated synthetic media more broadly by framing it in terms of the sociopolitical functions of public-interest journalism and its critical importance in an increasingly degraded information environment. We employ the concept of news integrity to capture this broader lens.

Given the novelty of genAI, research into newsroom implementation is only emerging. In the Australian context, studies are limited. A report on the first phase of our research at the UTS Centre for Media Transition was the first comprehensive study of newsroom implementation in Australia (Attard et al., 2023). Thomson et al. (2024) observe the impact of genAI on visual journalism in seven countries, including Australia, while a report from RMIT provides insights into audience as well as journalist perceptions of AI (Thomson et al., 2025).

The scholarly contribution of this study is not limited to Australia, however. Despite national differences in approaches to implementation and in industry and sociopolitical context, newsrooms worldwide face the same issues of news integrity as those revealed in our research.

2. Literature review

2.1 Synthetic media

Whittaker et al. (2020, p. 91) define synthetic media as "all automatically and artificially generated or manipulated media," including but not limited to synthesised audio, virtual reality, and advanced digital-image creation. Squicciarini et al. (2024, pp. 15–16) use synthetic content to cover a similar range, defined as "digital output generated or modified by algorithms, typically AI techniques, such as machine learning," including text, audio, imagery or multimedia. They use synthetic media to refer to a subset of synthetic content intended for or available to audiences, with deepfakes a further subset of synthetic media. Martin and Newell (2024, p. 448) refer to synthetic media as "synthetic outputs ... that are often (though not always) produced by generative AI systems and intended for people to consume," with AI slop referring to low-quality synthetic media. Synthetic media encompasses but is not exhausted by the new wave of genAI technologies, including GPTs (He & Fang, 2024, pp. 40–43), though they are the focus of our study.

Harris (2024, p. 131) observes that the distinction between synthetic and non-synthetic or human-produced media is not a "clear binary." Rather, genAI output could be thought to exist on a spectrum from lightly modified to fully synthesised. Barnes and Barraclough (2020, p. 214) note that most types and uses of synthetic media are benign. However, deepfakes, by their mere existence, "cast a shadow on the veracity of any given audiovisual record."

The terms synthetic content and synthetic media arise from, and remain primarily associated with, visual and auditory media rather than journalism (Feher, 2024, p. 353; Schell, 2024, p. 19). We found them very infrequently used amongst our interview subjects, who preferred AI-generated news or content. In scholarship, automated or robot journalism is common. There is some use in industry-wide guidelines, such as the Paris Charter on AI and Journalism (2023) and the Partnership on AI's Responsible Practices for Synthetic Media (2023, p. 3), which goes beyond journalism to include other synthetic media.

For newsrooms, the distinction between the use of genAI to create audience-facing synthetic content (or synthetic media, cf. Martin & Newell, 2024; Squicciarini et al., 2024) and internal-only uses is critical. Synthetic content is an umbrella term that includes audience-facing content. Accordingly, throughout this paper, we specify whether the use, content or media is audience-facing/front-end or internal/back-end where it is not. For newsrooms, there is also an important distinction between synthetic media or content produced internally (whether for back-end or audience-facing uses) and externally sourced synthetic media used in a news story, such as a video of a breaking news event circulating on social media.

2.2 AI adoption in newsrooms

Surveys conducted since the arrival of ChatGPT reveal accelerating AI implementation in newsrooms globally. A 2023 survey found that almost half of newsrooms were actively working with genAI, though use was infrequent and confined to a small number of users (Roper et al., 2023, pp. 5–6). By 2025, a Thomson Reuters survey found 49% of journalists used AI daily (Radcliffe, 2025, p. 17). Usage in Australia is markedly lower. A 2025 report found 63% of journalists had not used genAI in their work during the previous year (Medianet, 2025, p. 48). 88% reported concern about the potential effects of genAI on the overall integrity or quality of journalism. Our study explores newsroom perspectives to identify the drivers of this limited adoption.

Globally, experimentation has mostly aimed at making newsroom workflows more efficient and scalable (Cools & Diakopoulos, 2024, p. 12). This includes automating routine tasks – often those made necessary by other forms of technology, such as metatagging (Diakopoulos et al., 2024, p. 16) – or by augmenting human capabilities, e.g., in large-scale data and document analysis (Radcliffe, 2025, p. 14). It also includes both internal and audience-facing content creation and other editorial tasks. The Thomson Reuters survey found 30% of journalists used genAI for text creation and 21% for multimedia creation (Radcliffe, 2025, p. 18). A study by Møller and others (2025, p. 14) found that content creation has the lowest perceived potential for journalistic applications of genAI, with the highest in information analysis and content optimisation (e.g., SEO).

There have been some notable attempts at full article generation using genAI, with what might generously be called mixed results (Farhi, 2023; Mahadevan, 2025). Some newsrooms have developed AI-generated newsreaders, complete with social-media profiles (Samosir, 2023). But many outlets are using genAI mainly for internal content-manipulation tasks like summarisation, transcription or information synthesis (Diakopoulos et al., 2024, p. 11). Others are experimenting with limited audience-facing content creation, subject to editorial scrutiny before publication. This includes headlines, social-media posts, article summaries, translations, data visualisations, and synthetic voice (Borchardt et al., 2024). Limitations arise from the complexity of the newsgathering, production and distribution processes, which are "messy and unpredictable" rather than "an assembly line of neatly defined components" that can be easily or fully automated (Simon, 2024, p. 20). For this reason, it can be difficult to blend automation tools into existing workflows (Gutierrez Lopez et al., 2023, p. 485).

To mitigate some of the weaknesses of consumer AI tools – including inaccuracy, hallucination, bias, and generic output, as well as legal and intellectual property concerns – well-resourced newsrooms have moved to develop customised, in-house AI models (Simon & Isaza-Ibarra, 2023, p. 10). One third of respondents to a global March 2025 survey reported their organisations were using AI tools trained on their own content (Center for News, Technology & Innovation, 2025, p. 38). These include archival search tools or proofreading tools trained on internal style guides (Borchardt et al., 2024, p. 74). Several newsrooms have in-

corporated audience-facing chatbots into their websites (Oliver, 2024; WashPost-PR, 2024).

A long-running narrative accompanying moves towards automation is that it will free up journalists to do high-value work such as lengthy investigations and analysis (Meir, 2015; Tran, 2006). Widespread experimentation with genAI has so far yielded a relatively narrow range of beneficial uses, mostly in back-end, rather than audience-facing, production tasks (Cools & Diakopoulos, 2024, p. 11; Simon, 2024, p. 18). This is partly due to information-integrity problems such as inaccuracy and bias, but also a lack of news value in much AI output, including oversimplification, failure to highlight newsworthy information, or homogenisation of news content (Cools & Diakopoulos, 2024, p. 12). The work required to produce quality, newsworthy output – or to compile and edit low-quality output – may outweigh any potential time saving (Diakopoulos et al., 2024, p. 20; Radcliffe, 2025, p. 22; Simon, 2024, pp. 18–19; Simon & Isaza-Ibarra, 2023, p. 10). This is particularly the case with off-the-shelf products. But developing AI in house is very resource intensive, potentially for only modest productivity gains (Simon & Isaza-Ibarra, 2023, p. 10).

Some question whether genAI heralds a new era of innovation or is just another in a long line of hyped technologies accessible only to well-resourced newsrooms, leaving local outfits and many in the global south at a disadvantage (Ferrucci & Perreault, 2021; Min & Fink, 2021). While the accessibility of consumer AI has democratised the technology in newsrooms (Cools & Diakopoulos, 2024, p. 1), it is less useful than custom products, the cost of which may put them out of reach of many. Amid ongoing pressure to produce more content to satisfy the digital market, another key question is whether AI will merely facilitate a rise in low-quality content or "churnalism" rather than freeing up capacity for high-quality journalism (Golding & Murdock, 2022, p. 40; Montaña-Niño, 2024, pp. 30–31; Simon, 2024, p. 19).

2.3 News integrity

Concepts of integrity (e.g., news, journalistic, editorial, media & information integrity) are increasingly invoked in both industry and scholarly discourse amidst rising concern about the degradation of the online information ecosystem and the impact of digital platforms on news. Despite their popularity and broad application beyond these recent concerns, concepts of news integrity remain undertheorised. Here we tease out some essential elements of these concepts before looking in the next section at their relevance to genAI and synthetic media.

Integrity is often invoked in discussion of journalistic practice as a commitment to shared ideals and to the structures and practices that have evolved to promote them. As Borden and Tew (2007, p. 302) observe, "When journalists present news in a way that distorts the truth, their performance is at odds with the commitment to truthfulness that their role substantively requires." This normative commitment is what most clearly distinguishes journalism from other activities in the media marketplace (Borden & Tew, 2007, p. 303). Thus, for Kieran (1998, p. 23) to accuse a journalist of bias "is to impugn his journalistic integrity

in the deepest possible sense" – to claim that he is, "intentionally or otherwise, not adhering to the truth-respecting methods required for him to achieve the proper goal of journalism: Arriving at the truth of the matter."

Both tie journalistic integrity to the sociopolitical functions of journalism: For Kieran (1998, p. 23), the proper goal of truth arises from the democratic function of the media as an "unofficial fourth estate." For Borden and Tew (2007, p. 303), journalistic standards of reliability, truthfulness, and independence, pursued through "a discipline of verification," provide an "epistemologically defensible" framework for creating and communicating knowledge that, ultimately, helps citizens participate meaningfully in the public sphere. The SPJ Code of Ethics in the US founds the concepts of independence and integrity on the "highest and primary obligation of ethical journalism," which is to "serve the public" (Society of Professional Journalists, 2014). Many newsrooms' editorial policies explicitly reference integrity and its relation to serving the public interest and preserving trust (Riordan, 2014).

While there is variety in how the public-service or public-interest value of news is articulated, it typically includes what Hall (2025, p. 101) calls the three core democratic functions of news: Informing the public about public-interest issues, holding power to account, and providing a forum for public debate. Public trust in the news depends on the perception that these functions have not been undermined, e.g., by poor practice or by commercial or political pressures. The Peace Institute's Media Integrity Matters report (Petković, 2014, pp. 21–22) conceives of media integrity as encompassing the policies, structures and practices which "enable the media to serve the public interest and democratic processes," by providing "accurate and reliable information to citizens" and ensuring that citizens "have access to and are able to express a wide range of views and opinions without being exposed to bias and propaganda." Where these structures and practices are weak, the public can no longer trust that the news they read is accurate, reliable and free from bias or the influence of vested interests.

Integrity relates not only to journalistic practice but also to the news itself. News produced with integrity has integrity in turn. Public trust extends not only to particular outlets but to the news they produce. Adherence to professional standards is vital to "the confidence consumers have in the integrity of news material being reported to them" (Australian Press Council, 2023, p. 3). That is, the purpose of journalistic integrity as an integrity of process is to ensure news integrity as an integrity of product.

As a feature of the product as well as the process, integrity can be undermined at any stage of news production and distribution, including after the news has been published and is no longer under the newsroom's control. This underlies longstanding concerns about the distribution of news on digital platforms, where the integrity and continuity of a publisher's broader coverage can be lost in the torrent of atomised content (Wilding et al., 2018, p. 37). The structures and practices that maintain news integrity can also be weakened by market forces, e.g., the loss of advertising revenues to digital platforms and diminishing consumer demand for traditional news, and consequent reductions in the journalistic workforce or in news coverage (Australian Competition and Consumer Commission,

2019, pp. 309–322). These industry effects bring into play the concept of information integrity (Elbeyi et al., 2025, pp. 7–10) and the ability of the news industry to deliver public-interest news within a broader information environment in which it is playing a weakened role and over which it has limited control.

2.4 AI and news integrity

Emerging empirical studies of journalists' perspectives on genAI (Cools & Diakopoulos, 2024; Thomson et al., 2024), of metajournalistic discourse (Ananny & Karr, 2025) and emerging codes and guidelines on AI use (Becker et al., 2025; de-Lima-Santos et al., 2024) show that genAI adoption is attended by strong concerns about its potential impact on news integrity and public trust. Our participants regularly raised concerns about news integrity, suggesting that these concerns are mediating and constraining uptake in Australian newsrooms. Even where integrity is not explicitly invoked, we contend that integrity concepts provide a useful framework for understanding how genAI technologies are being adopted as well as the attitudes of industry members towards them, encompassing journalistic practice, audience trust, and the sociopolitical functions of news.

2.4.1 Editorial standards

One leading concern about genAI in news is the potential for newsrooms to unintentionally propagate the inaccuracies, hallucinations and bias that are notorious features of much genAI output (Jones et al., 2023, p. 4). This threatens the integrity of news as a product, requiring an integrity of process to mitigate it. Zier and Diakopoulos (2024, p. 1) argue that careful editorial oversight is required to preserve journalistic integrity and the integrity of news or information. While Cools and Diakopoulos (2024, p. 5) focus on the importance of ethical principles, they frame these in terms of integrity, arguing that ethical principles can serve as "a compass for preserving the integrity of journalistic practices" as AI is implemented in news workflows.

GenAI is accompanied by concerns about loss of editorial control within the newsroom, e.g., that eagerness to experiment might override ethical practice, particularly when driven by management (Gutiérrez-Caneda et al., 2024, p. 4; Møller et al., 2025, p. 16). There are also worries about the robustness of oversight measures, given the opacity of AI systems (Cools & Koliska, 2024, p. 666; Jones et al., 2022, p. 1736), a lack of AI literacy (Cools & Diakopoulos, 2024, p. 13; Jones et al., 2023, p. 4) and the pressures of the digital news cycle, which had already strained traditional verification processes before the advent of genAI (Hermida, 2015, pp. 39–41).

2.4.2 Editorial control in the digital information ecosystem

While the onus is on news publishers to ensure the integrity of the news they publish, once the news moves into the broader information ecosystem, they can no longer do so and must instead rely on third parties that make use of that news to

maintain its integrity. This includes AI systems that use the news as data for training or grounding generative models. Examples abound of genAI tools misrepresenting, misattributing or even hallucinating news stories (C. Moran, 2023). News Integrity in the Age of AI (European Broadcasting Union, 2025), a joint initiative of the European Broadcasting Union (EBU) and WAN-IFRA, responds to these issues, proposing five principles to "counter the misinformation crisis and protect the value of trusted news." These include requiring authorisation, attribution and accuracy for news content in genAI models; fair recognition of the value of up-to-date, high-quality news; and ensuring AI harnesses the diversity or plurality of the news media, presumably by not limiting deals to powerful media organisations. Principle 10 of the Paris Charter on AI and Journalism offers a similar prescription, requiring access to journalistic content to be "conditional on respect for the integrity of the information and the fundamental principles of journalistic ethics."

2.4.3 Authenticity and trust

Issues of authenticity and trust arise even where oversight processes are robust. As Mike Ananny states, once synthetic content is incorporated into news, "we can't necessarily know if the news that we're reading was made by humans or made by machine learning models or made by some mixture of those two things" (Avishai, 2023). Moran and Shaikh (2022, pp. 1766–1767) suggest concerns about authenticity rely on unquestioned assumptions about what "real journalism" is. But audiences value authenticity (Jones et al., 2023, pp. 4, 8; Wintterlin et al., 2020, p. 230), and audience expectations concerning authenticity and journalistic integrity are strongly linked to human creation of news (Jones et al., 2023, p.4). Studies have also found people view AI-generated text and chatbots as more objective and credible than humans (Lin & Lewis, 2022, p. 1635; Salas et al., 2023), while others have found transparency over AI use can decrease trust (Toff & Simon, 2024). That is, AI-generated content is credible, but paradoxically, journalists producing it are not.

The proliferation of synthetic media on digital platforms, including deepfakes and AI slop, raises concerns about the capability and capacity of newsrooms to verify externally sourced material, particularly images, video and audio (Thomson et al., 2024, pp. 11–12), threatening information integrity and public trust (Cazzamatta & Sarısakaloğlu, 2025, p. 3) and causing collateral damage to news in the form of a liar's dividend (Chesney & Citron, 2018, p. 1758).

2.4.4 Economic impacts on news integrity

The potential impact of genAI on news integrity must be understood against the backdrop of the broader media economy and the shift in journalism's place within it. Several emerging studies of genAI in newsrooms explore industry concerns over broader political and economic factors or interpret these through a political-economic lens. Borchardt et al. (2024, pp. 23–24) highlight fears that as more users access news through chatbots, AI will exacerbate the problem of news visibility in atomised platform environments and further threaten revenue (Dodds et

al., 2025, p. 6). Others see increasing dependence on technology companies for news production and distribution as an ongoing process of infrastructure capture that undermines journalistic autonomy (Simon, 2022, p. 1833; Sjøvaag, 2024, p. 247), especially as it is in many cases the same digital platform companies that are playing an outsized role in AI (Dodds et al., 2025, p. 6). Discourses of efficiency and "freeing up" journalists have been interpreted as supplanting labour by stealth (Matich et al., 2025, pp. 10–11). And some argue that casting genAI systems as tools for creativity or even as autonomous undermines the moral rights of journalists and other creators on whose work the systems have been trained (Montaña-Niño, 2024, p. 31).

Concerns about news integrity go hand in hand with impacts on journalistic labour. Automation may increase efficiency but decrease the role of human judgement (Cools & Koliska, 2024, p. 664). Journalists' concerns about authenticity, objectivity and voice have been interpreted as a form of boundary work to preserve independence and editorial control (Ananny and Karr, 2025, p. 13). At the organisational level, this can manifest in discussions about preserving brand integrity. But journalists – at least those who see themselves as observers or watchdogs rather than as mobilisers or entertainers (Møller et al., 2025, p. 15) – take the ethical implications of AI seriously because they take the sociopolitical functions of journalism seriously.

These considerations suggest that a broad view of news integrity – encompassing ethical journalistic practice, the sociopolitical functions of news, the impacts of the media economy and the relations between news and the broader information environment – is necessary for a comprehensive understanding of how newsrooms are implementing and responding to genAI. Taking these considerations into account, our research questions are as follows:

RQ1) How are newsrooms implementing genAI in the production of synthetic content?

RQ2) How do newsrooms perceive the potential impacts of genAI on news integrity?

RQ3) How are perceived impacts on news integrity mediating the implementation of genAI?

3. Methodology

In the study's first phase (July–October 2023), we interviewed 11 newsroom editors and one product lead from eight Australian media organisations. In the second phase (August–November 2024), we interviewed 13 news editors and six product leads from 14 news organisations, including the majority from phase one. In November 2024, we held a day-long workshop attended by many of the interviewees and additional participants (cf. Table A in Appendix).

The study population was defined using criterion-based expert sampling (Etikan, 2016, p. 2), based on expertise in newsroom editorial management or prod-

uct development, and involvement in AI implementation or policy development. In phase one, we focused exclusively on newsrooms producing public-interest journalism or "hard news" in different markets, models and media types to achieve sample variation. Phase two was broadened to test whether implementation differed at the margins of public-interest journalism, such as in factual lifestyle content. We also sought to include at least one editor and one product lead from each newsroom, as these roles represent different imperatives within a newsroom's implementation process. Participants were approached directly or via the researchers' networks and selected based on willingness to participate. Further participants were identified using snowball sampling.

Interviews were semi-structured. A set of general questions was posed to all participants, based on a literature review in mid-2023 and updated over time. Others were aimed at specific newsrooms based on their characteristics. Further questions arose from participant responses. Questions covered uses of AI; implementation processes; practical limitations; risks to news integrity, journalistic ethics, and audience trust; legal risks; and risks for the industry and the broader information environment.

Sixteen participants attended the workshop, which was conducted under the Chatham House rule to encourage discussion. The workshop was split into three sessions, focusing on: (1) use cases and implementation, particularly relating to synthetic content generation; (2) principles and guideline development; and (3) cross-industry issues, including the integrity of the broader information environment, and closer collaboration between newsrooms and AI companies, particularly for the purposes of mitigating risk. Session 1 was led by colleagues researching audience perceptions of AI in journalism, while the authors led sessions 2 and 3. Sessions 1 and 2 were attended by news editors, content editors and product managers. In session 3, these were joined by two representatives from AI companies and two industry consultants.

The workshop was also semi-structured in approach. To facilitate discussion, the first session included a slide presentation on genAI use cases and audience perceptions of AI use compiled from our colleagues' prior research. Participants discussed whether they had implemented or considered any of these uses, and where they perceived risk. We also shared general themes from our interviews. Before the second session, participants were provided with a handout of example AI guidelines and principles drawn from guidelines by news organisations and industry bodies in Australia, the UK and Europe. These were sorted into categories: Journalistic principles (accuracy, impartiality, etc.); transparency; human oversight and accountability; use restrictions; evaluation and testing; and organisational and legal issues such as privacy, licensing and distributing risk, responsibility and liability. The handout also included discussion questions. The third session was informed by the discussion in the two prior sessions.

The workshops enabled multidimensional knowledge transfer, with the researchers sharing findings on implementation, audience perceptions and guideline development, and participants sharing with the researchers and each other their practical experiences and perceptions.

The interviews and the workshop sessions were recorded, transcribed and coded thematically in NVivo. Broad themes were based on our research questions, including AI implementation and use; constraints on implementation; and perceptions of risk, particularly to news integrity. Finer-grain codes were inferred inductively from the interview and workshop data. Coded data was then analysed based on a four-way classification of our participants (see Table A in Appendix) across three organisational variables: Market (national, metropolitan, or regional); medium (television, radio, hardcopy newspaper or online); and model (public, commercial, or non-profit); as well as a single personal variable: Professional role (news editor, factual content editor or product manager). For broadcasting and print, market generally reflects size, with national organisations the largest and regional the smallest, though subsidiary relationships complicate this. Online-only outlets are generally small but have national reach, and in two cases are backed by international organisations. The views of the AI company representatives and industry consultants attending the third workshop session have been excluded from the present sample.

The research has undergone ethics approval at the University of Technology Sydney (ETH21-5787-24-2) and conforms with all relevant requirements and guidelines. Participants were provided with information about the purposes and conduct of the study and about data retention and use. Written consent was obtained, and participants and organisations have been de-identified.

This research forms part of an ongoing study, and only a subset of our findings is reported here. These have been selected solely based on their relevance to this special issue. An industry-targeted research report on phase one has previously been published, and some of those findings are included here (Attard et al., 2023). Some quotes have been edited for clarity.

4. Findings

In phase one (July–Oct 2023), participants were cautiously optimistic about the opportunities brought by genAI. There was trepidation over how rapidly the next wave of disruption was approaching. While all participants thought genAI would have a momentous impact on the news industry and journalism, there was uncertainty over precisely what it would be. A mantra of "no genAI in published content" served as the default short-term safeguard, reflecting apprehension over the limitations of genAI, a reluctance to undermine journalistic output, and an acute consciousness of the threats to trust and brand integrity. There was significant concern about navigating the proliferation of online synthetic media, where increasing technical sophistication and a degraded information ecosystem amplify the need for robust verification processes and undermine the ability to undertake them.

Our second-phase investigations (Aug–Nov 2024) found moderation of concern as experimentation had identified opportunities to enhance workflow. Despite this, implementation remained limited, and experimentation was carefully controlled, with most organisations focused almost exclusively on back-end productivity and efficiency gains. Few had experimented with audience-facing syn-

thetic content, confined to a narrow range of low-risk applications. There was notable variation across the three organisational variables of market, medium and model. Larger, national organisations, particularly the public broadcasters and commercial radio networks, had progressed much further in experimentation and implementation than smaller, regional organisations.

Two main constraints on genAI implementation and experimentation emerged across the study: (1) A perception that the utility of genAI was limited, with cost often outweighing benefits; and (2) an overriding, principled focus on the integrity of news. Both constraints were clearly apparent across all organisations. While news integrity was a universal concern, the cost–benefit calculus yielded different results across markets, media and models in parallel with differing levels of implementation.

4.1 Emerging uses of synthetic media in newsrooms

In both phases, we found that most news organisations see the biggest opportunities for genAI in back-end functionality, particularly news gathering and production, reflecting other studies (Cools & Diakopoulos, 2024; Diakopoulos et al., 2024; Møller et al., 2025; Radcliffe, 2025). Even in phase one, participants were contemplating deeper investigation of the opportunities of genAI, including in front-end output, and were gathering the resources to begin experimenting. Many, especially larger national and metropolitan outfits, had formed working groups comprising editorial, product development and legal personnel to manage implementation and develop AI policy.

> *We have a huge technology, product and digital team here. ... We are really trying to understand how the tech works, what we might build in house, what we might use, what we might license. (P1-09)*

By phase two, all participants had established such groups, though their formality, size and progress differed according to the size of the organisation, suggesting that resources are an important factor in newsrooms' ability to manage implementation. Across the board, implementation remained mostly experimental, focusing on low-risk opportunities with potential for good returns on investment, such as increased efficiencies or audience expansion.

4.1.1 Audience-facing content

In phase one, no participant organisations had experimented with audience-facing synthetic content, and many, seeing only downside risk, had ruled it out in the near term. However, we found differences between and within organisations according to purpose, market and brand. While they still had an eye on potential opportunities, print outlets and public broadcasters were very cautious. "The general policy is we don't want journalists using ChatGPT for their journalism" (P1-05).

In phase two, many of these organisations were still very wary of using AI for audience-facing content. A regional newspaper (P2-06) was not contemplating

genAI to create synthetic content at all, even in areas such as data visualisation. Some had begun experimenting, chiefly in digital content rather than news, and always with human oversight. For most, the scope of application was still limited to short texts and ideation.

> *So there is news, which is the pointy end, and a very ... strict approach, ... whereas in content we accept that there the audience expectation is a little bit different. (P2-15)*

The online lifestyle publisher was experimenting heavily with a wide range of efficiency-focused back-end use cases, but was as reticent as our other participants about using AI for audience-facing content (P2-04).

Several organisations were interested in exploring chatbots and other content delivery and personalisation uses. Given resource limitations, it was generally a lower priority than newsgathering and production. One public broadcaster had experimented with older types of AI in 2015 to develop a chatbot to deliver news and other information, but the project had stalled (P2-13). The organisation is now testing a genAI chatbot, confined to research and back-end tasks.

Despite the relatively limited implementation of genAI to produce synthetic content across all organisations, we found increased experimentation in several distinct areas. In audience-facing content, these were largely limited to synthetic voice, image generation, short-text generation like headlines or alt text, and some translation. Much more common were back-end newsgathering and production tasks, including transcription, summarisation, and idea generation.

4.1.2 Synthetic voice

Synthetic voice has emerged as a significant opportunity across different use cases. For the public broadcasters, improvements to accessibility and representation are a particular focus, as is connecting synthetic voice with the translation capability of genAI to serve Indigenous communities and migrant language groups.

In phase one, one of Australia's largest commercial radio networks was investigating how synthetic voice could be deployed for simple information services like short weather reports that otherwise require significant time for a journalist to produce.

> *We are not talking about a developing situation like a cyclone coming into Cairns. It's 26 degrees and sunny, so a very short sentence. But ... there are actually quite a lot of touchpoints. Whereas if you could automate that process, and you've got 99 radio stations, you could be saving a good couple of hours of someone's time. (P1-09)*

In phase two, this network had implemented audience-facing synthetic voice in the lower-risk areas of hyper-local weather reports and fuel-price updates (P2-16). This resulted in substantial time savings – especially important for Australian commercial radio broadcasters, which are required by law to provide a certain amount of local content per day (Australian Communications and Media Authority, n.d.).

Another commercial radio network had developed a multi-faceted internal tool that can source content from around the world, draft scripts for short news bulletins in the distinct house styles of the network's various stations (reflecting market demographics) and synthesise those bulletins using cloned voices of their own journalists. The tool was still in testing and had not been used to publish audience-facing news content (P2-03).

4.1.3 Image generation

In phase one, some were contemplating synthetic image generation, though with little official testing. In phase two, more organisations had experimented in this area. Data visualisation was an opportunity in both internal analysis and audience-facing uses. Still, all were cautious about full-scale image generation for audiences.

> *We've done some internal experiments with illustration for articles, seeing that as low risk. We haven't put that in front of audiences. (W-01)*

Some organisations were more liberal with non-news uses of image generation and image modification, animation or extension rather than full generation. One editor at a public broadcaster was clear that even image extension could undermine audience trust (W-05).

4.1.4 Headlines, short text generation and ideation

In phase one, short text generation was mostly a perceived opportunity rather than a subject of testing. In phase two, there had been much more experimentation, though application was still limited in audience-facing uses. This was the case across different media types and markets. Using AI to analyse a large set of images and to generate alt text was a common use case. Many were using it for headlines, but there was reluctance to push too far.

Many organisations had also found a use for genAI in ideation. It was perceived by all as an assistive technology, not a substitute for human creativity.

4.1.5 Transcription, summarisation and translation

In phase one, many organisations saw a potential application for genAI in transcription and translation. By phase two, many had implemented AI transcription tools, mostly in internal use, and had seen real efficiency gains. Public broadcasters and other organisations which produce content across different media types see strong value in automated transcription.

One public broadcaster had developed a customised large language model, principally for transcriptions, as off-the-shelf tools were inadequate.

> *The in-house one was ... trained on our own content, and it performed a lot better when it came to nouns, Australian place names, Indigenous lan-*

> *guage, etc.; whereas, you know, an off-the-shelf [tool] that's built on a global language just doesn't perform quite as well. (P2-14)*

Outside our interview cohort, one Australian broadcaster has implemented a tool that repurposes human-authored TV news scripts into online news stories (9News staff, 2024).

Demonstrated time savings also saw summarisation used across all participant organisations, principally for research. Use of genAI for translation was partly dependent on market and audience. Public broadcasters were experimenting with translation and synthetic voice in languages other than English. For commercial media, translation was still mostly viewed as a future opportunity.

4.2 Constraints on implementation

The limited scope of AI implementation even in phase two points to strong constraints based on: (1) A lack of perceived utility and value in AI tools, particularly for those not sufficiently well-resourced to develop in-house products; and (2) concerns for audience trust and news integrity. These are not unrelated: Most participants saw the limitations of genAI as directly threatening the integrity of news, and audience trust as hinging on perceptions of authenticity. We found that the implementation of genAI is mediated largely by concern for news integrity and trust, underpinned by broader cognisance of the sociopolitical role of journalism. Labour concerns were raised, but these were also often cast as a risk to news integrity, and most editors thought them misplaced in the short term.

4.2.1 Lack of utility and value

Many newsrooms have so far found limited use cases for AI. Few saw value in using genAI to produce synthetic content, even in back-end tasks, as the need for human oversight might outweigh efficiency gains. This was particularly pointed out for smaller teams, including those that sit within larger organisations.

> *Is that really where we're going to put our time into using those tools? ... If it's basically going to mean somebody's got to go back into it, go through it, check another source, make it two times the length of time that you're looking at for that? (W-12)*

> *Accuracy is the key point there, and I think ... in fast-moving newsrooms or small teams where you're really conscious you don't have a lot of ... resources to go back and check things beyond the rigorous fact checks you're already doing on stories, ... then that starts to impact trust for all of us. (P2-11)*

Some observed that humans provide much more value in content creation. This is connected with ideas about the value of originality and authorial voice and the sense that while AI is good at stringing words or pixels together, its output has a tendency to be bland and homogeneous.

> *I feel like, in terms of it generating content, and especially content that we would use, we're so far away from that just because we are the experts in that field. (W-08)*

While product teams saw potential in a larger range of use cases and could meet resistance from editorial staff, they were acutely aware of what journalists need from AI tools to maintain editorial standards.

> *In journalism, things need to be in certain places, and word order matters. It's far more precise than people give it credit for, when you're dealing with high-quality journalism. If you don't have those standards, you can get away with stuff. But if you do, it's going to be more work to get it to that standard. (P2-13)*

Many observed a particular lack of value in consumer-level AI due to its poor accuracy, bias and other limitations. But developing internal tools is expensive and slow, even for larger organisations.

> *We don't have $100 million spare ... to run around just trying all this stuff out. It takes heaps of time to do it properly and to have some faith in the integrity of the process. (W-05)*

Newsrooms with more resources can invest in better models that reduce risk, allowing them to find more utility in AI. As well as larger budgets, larger national organisations have large product teams and massive news archives, and tend to operate across different media. These factors incentivise and permit greater application of resources to experimentation and implementation. Still, even the larger organisations in Australia lag behind large news organisations in Europe or the US in their level of experimentation and implementation, particularly in audience-facing uses.

4.2.2 Concerns about news integrity

In both phases, the constraints on experimentation and almost complete preclusion of audience-facing synthetic media content reflected deep concern over the potential for AI to undermine news integrity and audience trust, which mediated how participant organisations were thinking about potential uses and the safeguards needed to govern implementation. This was true across all participant organisations.

> *For us, the key thing that will be top of mind every step of the way is safeguarding and retaining trust. And then that being at the centre of every decision we take in relation to genAI, but that not being at the cost of potential efficiencies or things that could actually help the audience. (P1-05)*

> *Integrity is so important to the journalism that we do. So, I struggle to see, as far as our storytelling goes, that we will be doing much with it for a little while, just because we're not ready. Integrity is very, very important. However, I think it would be very unwise to ignore it, as well. (P1-11)*

Concepts of integrity and trust were deeply rooted in organisational culture, reflected in the observations of both product leads and editors. The lifestyle publisher, who had ruled out audience-facing content despite heavy experimentation in back-end uses, also reached for news integrity to explain this choice, noting they were members of the Australian Press Council and abided by its code of practice and consumer complaints scheme (P2-04).

Amongst all participants, there was greater willingness to experiment and adopt where trust was not threatened, as in low-risk, back-end applications.

> *The biggest takeaway for me is how audience trust is embedded in everything we do. People see the availability of AI tools that will build efficiencies in their workflow, but we also need to consider if this would have an impact on audiences. Ultimately, we want to use AI in a way that enhances our services and maintains editorial standards so audiences know they can trust all our content. (P2-18)*

Most felt that, while risks need to be considered, fundamental journalistic standards – accuracy, impartiality, fairness and independence – remain unchanged. Instead, what is needed is guidance on how they apply to AI. While guidelines help in making decisions about use, what came through as most important was clarity about deeper principles and purposes.

> *When you parse the various principles and guidance that're being published by others, I think when you are deep into them and you can see the code that people are speaking, what sort of external references they're acknowledging, what kind of imperatives they're acknowledging, it's more useful. (W-05)*

4.2.2.1 Authenticity and transparency

Often, the conversation turned to questions of authenticity, including the risk that AI use would blur the boundaries between reality and representation. This was a concern even in non-news content. One participant (P2-12) noted they would never use AI to expand an image, "because if the photograph is documentary in nature, then you don't know what was beyond here, and you can't pretend to the audience that you did."

Most participants insisted on the importance of transparency in maintaining trust. News integrity was at the heart of these concerns.

> *I think it's inevitable that more AI tools are going to be adopted in journalism, but it's absolutely essential that we are upfront about our use of that, and that we communicate with our audiences about that. And because of the importance of trust in news organisations ... the audience has to know. They have to have evidence and faith in the fact that if a news organisation is using Gen AI, ... they will tell you the ways in which they are using it and still guarantee the quality of the journalism. So I think, you know, slip-*

> *ping it in under the radar is not where the news industry should be going. (P2-01)*

For regional newsrooms, notions of authenticity and transparency tied back in to the value of journalists on the ground gathering news in local communities.

> *If we're going to sort of get through this journey with genAI and still continue to provide what I think is an essential service to democracy and to anyone who lives in a democracy, then we've got to bring people with us, and they need to be able to distinguish between professional news outlets providing independent, reliable information that's been fact checked by, you know, professional journalists living in their community and understanding how things work, and a bot spitting something out based on an algorithm. (P2-01)*

Given the "black box" nature of AI systems, some felt that it would not always be possible to explain to audiences how newsrooms were using AI, and thus to maintain trust in the integrity of their product.

> *I can't tell you how those tools are actually working. So the explainability thing to me is a really big one, if or when we go down that path of audiences interacting with AI products and being able to explain to them really in ways that they would understand. I think that's going to be a real hurdle. (W-05)*

4.2.2.2 External information pollution

In both phases, news integrity, authenticity and trust also arose as acute concerns in discussions about the effects of AI on the integrity of the broader information ecosystem, as an area that largely lies outside newsroom control. This was often tied to the potential for AI to pollute the information environment through spurious "pink slime" or misinformation and to generate a liar's dividend.

> *I am more concerned about the dangers it poses for the news ecosystem as a whole. That's my major concern ... (P1-01)*

> *If there's a whole lot of bullshit out there generated by AI, then that shifts the entire landscape into bullshit. (P2-01)*

The flipside of this was the potential for quality news to become increasingly valuable in a degraded information ecosystem. Others felt that even if audiences seek out trusted news, it will be increasingly difficult to find in an atomised media environment.

> *Outlets like [ours] have the ability to stand out in coming years. ... You know you can trust every single thing we say because we've done the legwork to establish and confirm what's occurred. (P1-05)*

> *As information sources splinter, the morass people will have to wade through to try to find reliable information is just getting deeper and deeper and deeper. (P1-01)*

4.2.2.3 Verification

Nearly all participants were concerned about the pressure that external synthetic content would put on newsroom verification processes. Unlike many other concerns that accompanied the arrival of genAI, this had not moderated as we moved to phase two. All agreed that newsrooms need to be extra-vigilant, but many were concerned that verification processes, no matter how rigorously applied, are not always effective, particularly given the spectrum of synthetic content ranges from cheap fakes to sophisticated deep fakes and subtly altered content.

> *The possibilities of AI hoodwinking the media are now limitless and the fakes are coming at us all the time. (P1-10)*

With prominent local cases of poor processes leading to the publication of inaccurate and offensive content (Dunstan & Ortolan, 2024), others were concerned about standards slipping due to the competitive pressure of the news environment.

For those working in regional newsrooms, verification was less of a problem, as their original news stories mostly feature people familiar to the journalists or present in their archives.

4.2.2.4 Workforce risks

The final areas of concern focused on market risks to news integrity, including job losses and fair use of news content. Some reported substantial concern about job losses from junior journalists; however, at the senior editorial level, all insisted that threats to news integrity and trust ruled out replacing journalists or fundamental reporting tasks. Rather than replacement, participants were thinking about augmentation. This was true even in radio, where the success of synthetic voice heightens the perceived threat to jobs. On the lifestyle end of the news spectrum, where there is potentially more leeway to explore AI, we found strong commitment to improving workflows rather than reducing staff costs (P2-04).

> *There's a lot of anxiety. But then when you get people using it, and they realise that it's got limitations and that it doesn't necessarily replace them, but can help them. Then you get eyes lighting up. (P2-13)*

Regional outlets foresaw that AI might lead to some replacement of human work on time-consuming but low-value tasks, such as churning out stories from wire services or press releases in metropolitan newsrooms, but believed it could never replace the value of on-the-ground reporting (P2-03). Some had heard concerning views from management about the potential to reduce headcount and needed to insist on the importance of maintaining journalist numbers to cover public-interest news. This touches again on the potential for revenue pressures to lead to more automation and a relaxing of editorial oversight.

4.2.2.5 Platform power

Many participants felt that the biggest threat AI poses to news integrity is not hallucinations or bias; though these are certainly of concern, they are mostly within newsrooms' editorial control. The biggest threat lies in the potential undermining of traffic as tech platforms increasingly sequester users inside "walled gardens" built on information scraped from news sites.

This concern was apparent already in phase one, but had grown by phase two. Some product leads were eager for deals or other forms of collaboration to gain access to high-powered custom AI tools, as OpenAI was reportedly doing with newsrooms internationally (P2-13). A few felt striking deals with AI companies offered an opportunity for news media to monetise their quality content, but a common view was that, in the long term, deals would only undermine the news business (P2-10). Many newsrooms were blocking AI scrapers, although there was a pervasive feeling that the horse has already bolted. For small newsrooms in particular, the power of tech companies means there is a significant bargaining imbalance, and there is concern that market developments would favour the larger news companies.

> *A lot of these things sound great in theory, but actually in practice, they're really, really difficult for small and medium-sized publishers, you just don't get in the door. (P2-04)*

Some argued that while there seemed to be a great deal of public concern about whether news media would use AI responsibly, there needs to be greater discussion about the responsibilities of tech companies.

> *I think our industry needs to behave responsibly with respect to AI, but it's also a challenge across the tech titan ecosystem, and I think we're a ways away from that. (W-09)*

5. Discussion

Returning to RQ1, our findings show profound caution in Australian newsrooms, reflecting a recent industry survey that revealed relatively low AI adoption rates amongst Australian journalists (Medianet, 2025, p. 48). The vociferous concern we saw in the first phase of our research had moderated 12 months later, and experimentation had increased. The scope of implementation remained tightly governed and relatively narrow, focused on increasing the efficiency of back-end tasks such as transcription and summarisation, reflecting the findings of other recent studies (Cools & Diakopoulos, 2024, p. 12; Diakopoulos et al., 2024, p. 16). There had been very little experimentation with AI-augmented personalisation or delivery, such as chatbots or even article summaries. Even within this narrow scope, there were questions about whether efficiency gains were outweighed by the need for continual verification and oversight. Very few organisations were experimenting with audience-facing synthetic content.

While implementation was relatively constrained across all our participant organisations, there were notable differences across the variables of market (national, metropolitan and regional), medium (print, TV, radio, online) and model (public, commercial, or non-profit). In print and online news outlets, use of genAI for audience-facing content is virtually non-existent, even amongst metropolitan and national outlets. Experimentation was also least advanced, particularly amongst those serving regional markets. Implementation is more advanced in radio, with synthetic voice emerging as the most likely audience-facing use case in Australia in the near term. This reflects a lengthy history of synthetic voice experimentation in radio (Furtáková & Janáčková, 2023, p. 95).

The large commercial radio networks were well ahead of other participants in their willingness to test audience-facing uses. One had implemented synthetic voice for service information such as regional weather and fuel-price updates, and one had developed an end-to-end tool to search the web, script and synthesise news bulletins using synthetic voice, though it had yet to roll it out. There is still significant wariness about uses that might impact jobs or audience trust.

The public broadcasters, with a national reach across television and radio and large product teams, had also engaged in significant experimentation, with the prospect of some audience-facing uses on the horizon. These were focused on factual content rather than news, which had stricter parameters, and on serving linguistically diverse communities – a reflection of their public-service obligations.

These significant differences in implementation – within nationally low uptake rates – reflect variable resourcing and distinct organisational purposes. Smaller, regional print outlets have more constrained finances and very small product teams. They saw few beneficial front-end use cases, highlighting the expertise of journalists in newsgathering and reporting, particularly on local issues. The low rate of adoption means that regional news organisations risk falling behind industry developments, potentially exacerbating sustainability concerns as audiences increasingly move online (Eder & Sjøvaag, 2025).

National and metropolitan publications, while better resourced than regional outlets, also have relatively small product teams and constrained finances, with a stronger focus on national and international coverage and investigative reporting. These outlets saw opportunities to optimise a variety of back-end tasks, including summarisation, transcription, and data analysis, but little opportunity for front-end production outside of data visualisation.

For regional commercial radio, AI was seen as an opportunity to deliver on its regulatory obligations to broadcast local content while minimising labour costs. And for metropolitan radio, AI was seen as an easy opportunity to synthesise press releases for broadcast, though none had yet put this into practice. We also observed a distinction between these and publicly funded organisations with legislated public-service mandates and commercial outlets, with the latter experimenting widely across back-end tasks but little in audience-facing content.

Looking at RQ2 and RQ3, we found very little variation across our participant organisations. AI adoption was mediated in all newsrooms by concerns over the potential impact on brand integrity and audience trust, should journalistic processes break down. Journalistic standards were thus seen as critical to counteract

the perils of genAI (Cools & Diakopoulos, 2024). But the need for continual and robust oversight to ensure the integrity of the news product was often perceived as a drain on resources with little benefit, reducing the perceived utility of AI and constraining implementation. The lack of control over the functioning of AI tools exacerbated these concerns. Organisations with more advanced roll out of genAI were as concerned about news integrity as others, but better resourcing or the nature of the market or medium had opened a greater range of cases which were perceived to be low risk, even in some audience-facing areas such as synthetic voice.

Participants' reflections on the importance of ethical practice were also couched in an awareness of broader and deeper threats. They were sensitive to the reckless disruption of the media economy and information ecosystem by powerful AI companies, and the tension between securing deals, maintaining independence, and managing declining revenues, which might increase pressure for automation in editorial workflows (Borchardt et al., 2024, pp. 23–24; Simon, 2022, p. 1833; Sjøvaag, 2024, p. 247). Editors saw a critical need for rigorous, original journalism, particularly in an environment polluted by misinformation and fraught political discourse, to counteract the blurring of the boundaries between reality and representation (Chesney & Citron, 2018; Montaña-Niño, 2024). While senior staff echo the long-running discourse around freeing journalists from the mundanity of digital workflows (Matich et al., 2025; Meir, 2015; Tran, 2006), they recognise that revenue and management pressures could weaken the safeguards against job losses, increasing risks to brand and news integrity. Perhaps the strongest of our participants' concerns was the scraping of freely available but costly news to service the training and retrieval needs of AI platforms, which seem increasingly likely to undercut the public market for the news they have taken without compensation or attribution. Adding to this, the inherent flaws of genAI tools undermine the values of accuracy and reliability that underpin public trust in the news and sustain the industry that produces it.

Our study validates the utility of a broad conception of news integrity that encompasses both internal journalistic processes and adherence to editorial standards – what we have called process integrity – and the integrity of news as a product once it has been published into the information ecosystem – or what we have called product integrity. This twofold conception of news integrity recognises that the ability of news to fulfil its democratic functions depends not only on journalistic process and editorial standards but also on external factors largely outside a newsroom's control. Even where news is produced to the highest standards, its integrity may be threatened as it is ingested as data for AI training and grounding, and synthesised into generative output.

Despite these concerns, many of our participants were optimistic about the assistive opportunities of AI, augmenting workflows, facilitating time-consuming tasks and opening new possibilities of analysis, ideation and even content creation – suitably constrained, of course, by editorial safeguards. The larger and better-resourced, in particular, are certainly experimenting and alive to AI's transformative potential, reflecting that "AI-infused journalism will be better and

worse simultaneously, and in ways that only vaguely come into view as we see generative AI's early sprouts" (Dodds et al., 2025, p. 5).

6. Conclusion

In this study, we reported on two phases of ongoing research into the implementation of genAI in Australian newsrooms. We found that deep concerns over the integrity of news are driving relatively low adoption rates, constraining experimentation and potentially limiting the uptake of opportunities observed in overseas organisations. These concerns were apparent across all our participant organisations, suggesting that it is a significant constraint on implementation, in the Australian context at least. Variations in implementation rates were explained thus not by greater or lesser concern for the integrity of news, but largely by differences in market and resourcing, as well as the demands and opportunities brought by different media types and business models. All organisations were sensitive to the need to maintain audience trust and not undercut their own sustainability. Equally, they were concerned about the potential for genAI to threaten the integrity of news in areas outside their control. News companies perceive technological adoption as an additional strain on already-limited resources. But faced with the potential for that very technology to undermine the industry's sustainability by pulling audiences away from news – even as it uses news to sustain itself – their concern does not seem misplaced.

No newsroom, on our count, is about to lay waste to the integrity of their product through reckless adoption of genAI. That is not to say that the threats are not there. There will no doubt be occasional acts of error and negligence. But we should not let that distract us from the ongoing undermining of the news and information ecosystem that may see us all end up in the same deepening morass.

Acknowledgements

Thanks to our capable research assistants on this project, Miguel D'Souza and Tamara Markus, and to the anonymous reviewers for their very helpful suggestions.

Funding

This research was funded by the Minderoo Foundation.

References

9News staff. (2024, March 7). 9ExPress. 9News. https://www.9news.com.au/technology/9express/16480c33-636a-461f-9c4f-d0e2522c722a

Ananny, M., & Karr, J. (2025). How media unions stabilize technological hype: Tracing organized journalism's discursive constructions of generative artificial intelligence. *Digital Journalism*. https://doi.org/10.1080/21670811.2025.2454516

Attard, M., Davis, M., & Main, L. (2023). *Gen AI and journalism*. UTS Centre for Media Transition. https://doi.org/10.6084/m9.figshare.24751881.v3

Australian Communications and Media Authority. (n.d.). Local content on regional commercial radio. Retrieved April 24, 2025, from https://www.acma.gov.au/local-content-regional-commercial-radio

Australian Competition and Consumer Commission. (2019). Digital platforms inquiry final report. https://www.accc.gov.au/about-us/publications/digital-platforms-inquiry-final-report

Australian Press Council. (2023, August). Submission to the Department of Industry, Transport, Regional Development and Communications on the Exposure Draft of the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 (sub. E3250). https://www.infrastructure.gov.au/sites/default/files/documents/acma2023-e3250-australian-press-council.pdf

Avishai, T. (Director). (2023, December 4). Synthetic media: AI and journalism (No. 6) [Broadcast]. In Knowing Machines. https://engelberg-center-live.simplecast.com/episodes/synthetic-media-ai-and-journalism

Bäck, A., Diakopoulos, N., Granroth-Wilding, M., Haapanen, L., Leppänen, L. J., Melin, M., Moring, T. A., Munezero, M. D., Siren-Heikel, S. J., Södergård, C., & Toivonen, H. (2019). News automation: The rewards, risks and realities of "machine journalism." World Association of Newspapers and News Publishers, WAN-IFRA.

Bagozzi, R. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244–254. https://doi.org/10.17705/1jais.00122

Barnes, C., & Barraclough, T. (2020). Deepfakes and synthetic media. In R. Steff, J. Burton, & S. R. Soare (Eds.), *Emerging Technologies and International Security: Machines, the State, and War* (pp. 206–222). Routledge. https://doi.org/10.4324/9780367808846

Becker, K. B., Simon, F. M., & Crum, C. (2025). Policies in parallel? A comparative study of journalistic AI policies in 52 global news organisations. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2024.2431519

Beckett, C., & Yaseen, M. (2023). *Generating change: The journalism AI report*. Polis, London School of Economics and Political Science. https://www.journalismai.info/s/Generating-Change-_-The-Journalism-AI-report-_-English.pdf

Borchardt, A., Simon, F., Zachrison, O., Bremme, K., Kurczabinska, J., Mulhall, E., & Johanny, Y. (2024). *Trusted journalism in the age of generative AI*. European Broadcasting Union. https://ora.ox.ac.uk/objects/uuid:8c874e2e-34de-4813-ba23-84e6300af110

Borden, S. L., & Tew, C. (2007). The role of journalist and the performance of journalism: Ethical lessons from "fake" news (seriously). *Journal of Mass Media Ethics*, *22*(4), 300–314. https://doi.org/10.1080/08900520701583586

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, *96*(3), 673–695. https://doi.org/10.1177/1077699019859901

Cazzamatta, R., & Sarısakaloğlu, A. (2025). Mapping global emerging scholarly research and practices of AI-supported fact-checking tools in journalism. *Journalism Practice*, *19*(10), 2422–2444. https://doi.org/10.1080/17512786.2025.2463470

Center for News, Technology & Innovation. (2025). *What it means to do journalism in the age of AI: Journalist views on safety, technology and government*. https://innovating.news/2024-journalist-survey/

Chesney, R., & Citron, D. K. (2018). Deep fakes: A looming challenge for privacy, democracy, and national security. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3213954

Cools, H., & Diakopoulos, N. (2024). Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*, 1–19. https://doi.org/10.1080/17512786.2024.2394558

Cools, H., & Koliska, M. (2024). News automation and algorithmic transparency in the newsroom: The case of the Washington Post. *Journalism Studies*, *25*(6), 662–680. https://doi.org/10.1080/1461670X.2024.2326636

de-Lima-Santos, M.-F., Yeung, W. N., & Dodds, T. (2024). Guiding the way: A comprehensive examination of AI guidelines in global media. *AI & Society*, *40*, 2585–2603. https://doi.org/10.1007/s00146-024-01973-5

Diakopoulos, N., Cools, H., Li, C., Helberger, N., Kung, E., Rinehart, A., & Gibbs, L. (2024). Generative AI in journalism: The evolution of newswork and ethics in a generative information ecosystem. https://doi.org/10.13140/RG.2.2.31540.05765

Dodds, T., Zamith, R., & Lewis, S. C. (2025). The AI turn in journalism: Disruption, adaptation, and democratic futures. *Journalism*. Advance online publication https://doi.org/10.1177/14648849251343518

Dunstan, J., & Ortolan, M. (2024, January 31). An AI-generated image of a Victorian MP raises wider questions on digital ethics. ABC News. https://www.abc.net.au/news/2024-02-01/georgie-purcell-ai-image-nine-news-apology-digital-ethics/103408440

Eder, M., & Sjøvaag, H. (2025). Falling behind the adoption curve: Local journalism's struggle for innovation in the AI transformation. *Journal of Media Business Studies*, *22*(4), 325–343. https://doi.org/10.1080/16522354.2025.2473301

Elbeyi, E., Bruhn Jensen, K., Aronczyk, M., Asuka, J., Ceylan, G., Cook, J., Erdelyi, G., Ford, H., Milani, C., Mustafaraj, E., Ogenga, F., Yadin, S., Howard, P. N., Valenzuela, S., Brulle, R., Jacquet, J., Lewandowsky, S., & Roberts, T. (2025). Information integrity about climate science: A systematic review. International Panel on the Information Environment (IPIE). https://doi.org/10.61452/BTZP3426

Etikan, I. (2016). Comparison of Convenience Sampling and Purposive Sampling. American *Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

European Broadcasting Union. (2025, May 5). Media outlets worldwide join call for AI companies to help protect news integrity. https://www.ebu.ch/news/2025/05/media-outlets-worldwide-join-call-for-ai-companies-to-help-protect-news-integrity

Farhi, P. (2023, January 17). CNET used AI to write articles. It was a journalistic disaster. The Washington Post. https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/

Feher, K. (2024). Exploring AI media. Definitions, conceptual model, research agenda. *Journal of Media Business Studies*, *21*(4), 340–363. https://doi.org/10.1080/16522354.2024.2340419

Ferrucci, P., & Perreault, G. (2021). The liability of newness: Journalism, innovation and the issue of core competencies. *Journalism Studies*, *22*(11), 1436–1449. https://doi.org/10.1080/1461670X.2021.1916777

Furtáková, L., & Janáčková, Ľ. (2023). AI in radio: The game changer you did not hear coming. In M. Prostináková Hossová, M. Martovič, & M. Solík (Eds.), Marketing identity: AI – The future of today. Proceedings from the International Scientific Conference. University of Ss. Cyril and Methodius. https://mmidentity.fmk.sk/wp-content/uploads/2024/10/MM_2023_eng.pdf

Golding, P., & Murdock, G. (2022). The political economy of contemporary journalism and the crisis of public knowledge. In S. Allan (Ed.), *The Routledge Companion to News and Journalism* (2nd ed., pp. 36–45). Routledge. https://doi.org/10.4324/9781003174790-5

Gutierrez Lopez, M., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., & Missaoui, S. (2023). A question of design: Strategies for embedding AI-driven tools into journalistic work routines. *Digital Journalism*, *11*(3), 484–503. https://doi.org/10.1080/21670811.2022.2043759

Gutiérrez-Caneda, B., Lindén, C.-G., & Vázquez-Herrero, J. (2024). Ethics and journalistic challenges in the age of artificial intelligence: Talking with professionals and experts. *Frontiers in Communication*, *9*. https://doi.org/10.3389/fcomm.2024.1465178

Hall, C. J. (2025). Platform journalism on YouTube: A democratic functions approach to analysing journalism on digital platforms. *Australian Journalism Review*, *47*(1), 97–115. https://doi.org/10.1386/ajr_00178_7

Harris, K. R. (2024). Synthetic media detection, the wheel, and the burden of proof. *Philosophy & Technology*, *37*(131). https://doi.org/10.1007/s13347-024-00821-0

He, X., & Fang, L. (2024). Regulatory challenges in synthetic media governance: Policy frameworks for AI-generated content across image, video, and social platforms. *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization*, *9*(12), 36–54. https://helexscience.com/index.php/JRPAAIW/article/view/2024-12-13

Helberger, N., van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism*, *10*(10), 1605–1626. https://doi.org/10.1080/21670811.2022.2152195

Hermida, A. (2015). Nothing but the truth: Redrafting the journalistic boundary of verification. In M. Carlson & S. C. Lewis (Eds.), *Boundaries of Journalism* (pp. 37–50). Routledge.

Jones, B., Jones, R., & Luger, E. (2022). AI 'everywhere and nowhere': Addressing the AI intelligibility problem in public service journalism. *Digital Journalism*, *10*(10), 1731–1755. https://doi.org/10.1080/21670811.2022.2145328

Jones, B., Jones, R., & Luger, E. (2023). Generative AI & journalism: A rapid risk-based review. University of Edinburgh. https://www.research.ed.ac.uk/en/publications/generative-ai-amp-journalism-a-rapid-risk-based-review

Kieran, M. (1998). Objectivity, impartiality and good journalism. In M. Kieran (Ed.), *Media Ethics* (1st ed., pp. 23–36). Routledge.

Lin, B., & Lewis, S. C. (2022). The one thing journalistic AI just might do for democracy. *Digital Journalism*, *10*(10), 1627–1649. https://doi.org/10.1080/21670811.2022.2084131

Lindén, T. C.-G., & Dierickx, L. (2019). Robot journalism: The damage done by a metaphor. *Unmediated*, *2*, 152–155.

Mahadevan, A. (2025, March 20). An Italian newspaper launched a generative AI experiment. It's not going well. Poynter. https://www.poynter.org/tech-tools/2025/il-foglio-newspaper-generated-artificial-intelligence/

Martin, A., & Newell, B. (2024). Synthetic data, synthetic media, and surveillance. *Surveillance & Society*, *22*(4), 448–452. https://doi.org/10.24908/ss.v22i4.18334

Matich, P., Thomson, T. J., & Thomas, R. J. (2025). Old threats, new name? Generative AI and visual journalism. *Journalism Practice*, *19*(10), 2402–2421. https://doi.org/10.1080/17512786.2025.2451677

Medianet. (2025). *2025 Australian media landscape report*. https://engage.medianet.com.au/2025-media-landscape-report

Meir, N. (2015, June 15). Automated earnings stories multiply. The Associated Press. https://www.ap.org/the-definitive-source/announcements/automated-earnings-stories-multiply/

Min, S. J., & Fink, K. (2021). Keeping up with the technologies: Distressed journalistic labor in the pursuit of "shiny" technologies. *Journalism Studies*, *22*(14), 1987–2004. https://doi.org/10.1080/1461670X.2021.1979425

Møller, L. A., Cools, H., & Skovsgaard, M. (2025). One size fits some: How journalistic roles shape the adoption of generative AI. *Journalism Practice*, 1–22. https://doi.org/10.1080/17512786.2025.2484622

Montaña-Niño, S. (2024). Automated journalistic assemblages. A conceptual approach to the normative and ethical debates on AI implementation in newsrooms. *Problemi Dell'informazione*, *1*, 17–40. https://doi.org/10.1445/113227

Moran, C. (2023, April 6). ChatGPT is making up fake Guardian articles. Here's how we're responding. The Guardian. https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article

Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism*, *10*(10), 1756–1774. https://doi.org/10.1080/21670811.2022.2085129

Oliver, L. (2024, November 1). This chatbot helps tell the story of how women are affected by drug trafficking in Paraguay. Reuters Institute News. https://reutersinstitute.politics.ox.ac.uk/news/chatbot-helps-tell-story-how-women-are-affected-drug-trafficking-paraguay

Paris Charter on AI and Journalism. (2023, November 10). https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20Charter%20on%20AI%20and%20Journalism.pdf

Partnership on AI. (2023, February 27). PAI's responsible practices for synthetic media. https://partnershiponai.org/

Petković, B. (2014). *Media integrity matters: Reclaiming public service values in media and journalism* (1st ed). Peace Institute, Institute for Contemporary Social and Political Studies.

Radcliffe, D. (2025). Journalism in the AI era (TRF Insights). Thomson Reuters Foundation. https://www.trust.org/wp-content/uploads/2025/01/TRF-Insights-Journalism-in-the-AI-Era.pdf

Riordan, K. (2014). Accuracy, independence, and impartiality: How legacy media and digital natives approach standards in the digital age. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/our-research/accuracy-independence-and-impartiality-how-legacy-media-and-digital-natives-approach

Roper, D., Henriksson, T., Hälbich, K., & Martin, O. (2023). Gauging generative AI's impact on newsrooms. World Association of News Publishers (WAN-IFRA). https://wan-ifra.org/insight/gauging-generative-ais-impact-in-newsrooms/

Salas, A., Rivero-Calle, I., & Martinón-Torres, F. (2023). Chatting with ChatGPT to learn about safety of COVID-19 vaccines – A perspective. *Human Vaccines & Immunotherapeutics*, *19*(2). https://doi.org/10.1080/21645515.2023.2235200

Samosir, H. (2023, July 12). More countries across Asia are debuting digital artificial intelligence news readers. Could Australia follow suit? ABC News. https://www.abc.net.au/news/2023-07-13/artificial-intelligence-news-readers-becoming-common-in-asia/102591790

Schell, K. (2024). AI transparency in journalism: Labels for a hybrid era. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-01/RISJ%20Fellows%20Paper_Katja%20Schell_MT24_Final.pdf

Simon, F. M. (2022). Uneasy bedfellows: AI in the news, platform companies and the issue of journalistic autonomy. *Digital Journalism*, *10*(10), 1832–1854. https://doi.org/10.1080/21670811.2022.2063150

Simon, F. M. (2024). Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena. Tow Center for Digital Journalism. https://journalism.columbia.edu/news/tow-report-artificial-intelligence-news-and-how-ai-reshapes-journalism-and-public-arena

Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-127

Simon, F. M., & Isaza-Ibarra, L. F. (2023). AI in the news: Reshaping the information ecosystem? Oxford Internet Institute. https://www.oii.ox.ac.uk/wp-content/uploads/2023/08/Minderoo_Report_Simon_Ibarra.pdf

Sjøvaag, H. (2024). The business of news in the AI economy. *AI Magazine*, *45*(2), 246–255. https://doi.org/10.1002/aaai.12172

Society of Professional Journalists. (2014). SPJ code of ethics. https://www.spj.org/spj-code-of-ethics/

Squicciarini, M., Valdez Genao, J., & Sarmiento, C. (2024). Synthetic content and AI policy: A primer. UNESCO. https://policycommons.net/artifacts/17958669/synthetic-content-and-its-implications-for-ai-policy/18857919/

Ternovski, J., Kalla, J., & Aronow, P. M. (2022). The negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety, 1*(2). https://doi.org/10.54501/jots.v1i2.28

Thomson, T. J., Thomas, R. J., & Matich, P. (2024). Generative visual AI in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*, 1–22. https://doi.org/10.1080/21670811.2024.2331769

Thomson, T. J., Thomas, R., Riedlinger, M., & Matich, P. (2025). Generative AI & journalism. RMIT University. https://doi.org/10.6084/m9.figshare.28068008

Toff, B., & Simon, F. M. (2024). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, *30*(4), 881–903. https://doi.org/10.1177/19401612241308697

Tran, M. (2006, August 18). Robots write the news. The Guardian. https://www.theguardian.com/news/blog/2006/aug/18/robotswriteth

WashPostPR. (2024, November 7). The Washington Post launches "Ask the Post AI," a new search experience. The Washington Post. https://www.washingtonpost.com/pr/2024/11/07/washington-post-launches-ask-post-ai-new-search-experience/

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of AI-generated media. *IT Professional*, *22*(5), 90–99. https://doi.org/10.1109/MITP.2020.2985492

Wilding, D., Fray, P., Molitorisz, S., & McKewon, E. (2018). The impact of digital platforms on news and journalistic content. UTS Centre for Media Transition. http://hdl.handle.net/10453/159124

Wintterlin, F., Engelke, K. M., & Hase, V. (2020). Can transparency preserve journalism's trustworthiness? Recipients' views on transparency about source origin and verification regarding user-generated content in the news. *SCM Studies in Communication and Media*, *9*(2), 218–240. https://doi.org/10.5771/2192-4007-2020-2-218

Zier, J., & Diakopoulos, N. (2024, October 26). Labeling AI-generated news content: Matching journalist intentions with audience expectations. Proceedings of the Computation and Journalism Symposium 2024. https://cplusj2024.github.io/

Appendix

Table A. Overview of participants

Identifier*	Professional role	Medium	Type	Market
P1-01	News editor	Online	Non-profit	National
P1-02	News editor	Hardcopy + online	Commercial	Metropolitan
P1-03	Factual content editor	Hardcopy + online	Commercial	Metropolitan/national
P1-04	Factual content editor	Hardcopy + online	Commercial	Metropolitan/national
P1-05	News editor	TV + radio + online	Public	National TV/radio + regional radio
P1-06	News and product manager	TV + radio + online	Public	National TV/radio + regional radio
P1-07	Factual content editor	TV + radio + online	Public	National TV/radio + regional radio
P1-08	News editor	TV + radio + online	Public	National
P1-09	News editor	Radio	Commercial	Metropolitan + regional
P1-10	News editor	Online	Commercial	National
P1-11	News editor	Hardcopy + online	Commercial	Regional
P1-12	News editor	Hardcopy + online	Commercial	Regional
P2-01	News editor	Hardcopy + online	Commercial	Regional
P2-02	News editor	Hardcopy + online	Commercial	National
P2-03	News editor	Radio	Commercial	Metropolitan + regional
P2-04	News editor (lifestyle)	Online	Commercial	National
P2-05	News editor	TV + online	Commercial	National
P2-06	News editor	Hardcopy + online	Commercial	Regional
P2-07	Product manager	Hardcopy + online	Commercial	Regional
P2-08	Product manager	Radio	Commercial	Metropolitan + regional
P2-09	Product manager	Hardcopy + online	Commercial	National

P2-10	News editor	Online	Commercial	National
P2-11	News editor	TV + radio + online	Public	National
P2-12	News editor	TV + radio + online	Public	National TV/radio + regional radio
P2-13	Product manager	TV + radio + online	Public	National TV/radio + regional radio
P2-14	News editor and product manager	TV + radio + online	Public	National TV/radio + regional radio
P2-15	Factual content editor	TV + radio + online	Public	National TV/radio + regional radio
P2-16	News editor	Radio	Commercial	Metropolitan + regional
P2-17	News editor	Radio	Commercial	Metropolitan + regional
P2-18	Product manager	TV + radio + online	Public	National
W-01	Product manager	TV + radio + online	Public	National TV/radio + regional radio
W-02	News editor	Hardcopy + online	Commercial	Regional
W-03	Product manager	TV + radio + online	Public	National TV/radio + regional radio
W-04	News editor	Radio	Commercial	Metropolitan + regional
W-05	News editor	TV + radio + online	Public	National TV/radio + regional radio
W-06	News editor	Online	Commercial	National
W-07	Product manager	Hardcopy + online	Commercial	Regional
W-08	Factual content editor	Hardcopy + online	Commercial	National
W-09	News editor	Hardcopy + online	Commercial	Metropolitan/national
W-10	Product manager	Radio	Commercial	Metropolitan + regional
W-11	News editor	Hardcopy + online	Commercial	Regional
W-12	News editor	TV + radio + online	Public	National

Note. *P1 = phase 1 interview; P2 = phase 2 interview; W = workshop

SCM
Studies in Communication and Media

FULL PAPER

Spotting fakes: How do non-experts approach deepfake video detection?

Fälschungen feststellen: Wie können Nicht-Experten Deepfake-Videos erkennen?

Mary Holmes, Klaire Somoray, Jonathan D. Connor, Darcy W. Goodall, Lynsey Beaumont, Jordan Bugeja, Isabelle E. Eljed, Sarah Sai Wan Ng, Ryan Ede & Dan J. Miller

Mary Holmes, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0001-0764-8024
Klaire Somoray (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology & Margaret Roderick Centre for Mental Health Research, 4814, Townsville, Australia. Contact: klaire.somoray@jcu.edu.au. ORCID: https://orcid.org/0000-0001-7521-1425
Jonathan D. Connor (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. Contact: jonathan.connor@jcu.edu.au. ORCID: https://orcid.org/0000-0003-3246-8858
Darcy W. Goodall, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.
Lynsey Beaumont, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.
Jordan Bugeja, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.
Isabelle E. Eljed, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0005-6597-0440
Sarah Sai Wan Ng, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia. ORCID: https://orcid.org/0009-0005-5154-8249
Ryan Ede, James Cook University, College of Healthcare Sciences, Department of Psychology, 4814, Townsville, Australia.
Dan J. Miller (PhD), James Cook University, College of Healthcare Sciences, Department of Psychology & Margaret Roderick Centre for Mental Health Research, 4814, Townsville, Australia. Contact: daniel.miller1@jcu.edu.au. ORCID: https://orcid.org/0000-0002-3230-2631

FULL PAPER

Spotting fakes: How do non-experts approach deepfake video detection?

Fälschungen feststellen: Wie können Nicht-Experten Deepfake-Videos erkennen?

Mary Holmes, Klaire Somoray, Jonathan D. Connor, Darcy W. Goodall, Lynsey Beaumont, Jordan Bugeja, Isabelle E. Eljed, Sarah Sai Wan Ng, Ryan Ede & Dan J. Miller

Abstract: Intervening to bolster human detection of deepfakes has proven difficult. Little is known about the behavioural strategies people employ when attempting to detect deepfakes. This paper reports two studies in which non-experts completed a deepfake detection task. As part of the task, participants were presented with a series of short videos – half of which were deepfakes – and asked to categorise each video as either deepfake or authentic. In Study 1 (N = 391), an online study, participants were randomly assigned to a control or intervention group (in which they received a list of detection strategies before the detection task). After the detection task, participants elaborated on the approach they employed during the task. In Study 2 (N = 32), a laboratory-based study, participants' gaze behaviour (fixations and saccades) was recorded during the detection task. No detection strategies were provided to Study 2 participants. Consistent with prior research, Study 1 participants showed modest detection accuracy (M = .61, SD = .14) – only somewhat above chance levels (.50) – with no difference between the intervention and control groups. However, content analysis of participants' self-reports revealed that the intervention successfully shifted participants' attention toward cues such as skin texture and facial movements, while the control group more frequently reported relying on intuition (gut feeling) and features such as body language. Study 2 found similar levels of detection accuracy (M = .65, SD = .20). Participants focused their gaze primarily on the eyes and mouth rather than the body, showing a slight preference for the eyes over the mouth. No differences in gaze were found between authentic and deepfake videos or between correctly and incorrectly categorised videos. The findings suggest interventions can modify detection behaviours (even without improving accuracy). Future interventions may benefit from directing attention from the eyes toward more diagnostic features, such as face–body inconsistencies and the face boundary.

Keywords: Deepfakes, AI-generated media, synthetic media, detection, self-report, eye-tracking

Zusammenfassung: Es hat sich als schwierig erwiesen, Maßnahmen zur Verbesserung der menschlichen Erkennung von Deepfakes zu ergreifen. Über die Verhaltensstrategien, die Menschen bei der Erkennung von Deepfakes anwenden, ist nur wenig bekannt. Dieser Ar-

tikel präsentiert zwei Studien, in denen Nicht-Experten eine Deepfake-Erkennungsaufgabe absolvierten. Im Rahmen dieser Aufgabe wurde den Teilnehmern eine Reihe von kurzen Videos gezeigt – von denen die Hälfte Deepfakes waren – und sie wurden gebeten, jedes Video entweder als Deepfake oder als authentisch zu kategorisieren. In Studie 1 (N = 391), einer Online-Studie, wurden die Teilnehmer nach dem Zufallsprinzip einer Kontroll- oder Interventionsgruppe zugewiesen (in der sie vor der Erkennungsaufgabe eine Liste mit Erkennungsstrategien erhielten). Nach der Erkennungsaufgabe erläuterten die Teilnehmer den Ansatz, den sie während der Aufgabe verwendet hatten. In Studie 2 (N = 32), einer Laborstudie, wurde das Blickverhalten (Fixationen und Sakkaden) der Teilnehmer während der Erkennungsaufgabe aufgezeichnet. Den Teilnehmern von Studie 2 wurden keine Erkennungsstrategien zur Verfügung gestellt. In Übereinstimmung mit früheren Untersuchungen zeigten die Teilnehmer der Studie 1 eine mäßige Erkennungsgenauigkeit (M = 0,61, SD = 0,14) – nur geringfügig über dem Zufallsniveau (0,50) – ohne Unterschied zwischen der Interventions- und der Kontrollgruppe. Die Inhaltsanalyse der Selbstauskünfte der Teilnehmer ergab jedoch, dass die Interventionsgruppe ihre Aufmerksamkeit erfolgreich auf Hinweise wie Hautstruktur und Gesichtsbewegungen lenkte, während die Kontrollgruppe häufiger angab, sich auf ihre Intuition (Bauchgefühl) und Merkmale wie Körpersprache zu verlassen. Studie 2 ergab eine ähnliche Erkennungsgenauigkeit (M = 0,65, SD = 0,20). Die Teilnehmer richteten ihren Blick hauptsächlich auf die Augen und den Mund und weniger auf den Körper, wobei sie eine leichte Präferenz für die Augen gegenüber dem Mund zeigten. Es wurden keine Unterschiede im Blickverhalten zwischen authentischen und Deepfake-Videos oder zwischen korrekt und falsch kategorisierten Videos festgestellt. Die Ergebnisse deuten darauf hin, dass Interventionen das Erkennungsverhalten verändern können (ohne die Genauigkeit zu verbessern). Zukünftige Interventionen könnten davon profitieren, die Aufmerksamkeit von den Augen auf diagnostischere Merkmale wie Inkonsistenzen zwischen Gesicht und Körper und die Gesichtskonturen zu lenken.

Schlagwörter: Deepfakes, KI-generierte Medien, synthetische Medien, Erkennung, Selbstauskunft, Eye-Tracking

1. Introduction

Deepfakes are a form of AI-manipulated media in which an existing person's likeness is inserted into an extant piece of media (be it a static image, piece of audio, or a video). They can be highly realistic. The most common type of deepfakes are "face replacement" deepfakes (Silva et al., 2022). Deepfakes can be used to make it appear as if someone has done or said something they have never done or said. As such, deepfake technology, when used maliciously, can cause serious harms. These harms can occur at the individual and societal level. Examples of individual-level harms include scams (Miller et al., 2025) and the use of non-consensual digitally altered sexual imagery for harassment and extortion (Flynn et al., 2022). Potential societal-level harms include the spread of disinformation and misinformation, manipulation of political campaigns and public opinion, the erosion of trust in democratic institutions and legitimate media reporting (Godulla et al., 2021), and military deception (Smith & Mansted, 2020).

Automated deepfake detection tools have advanced significantly (Abbas & Taeihagh, 2024). However, these technologies are still generally inaccessible to

the public. Furthermore, political climate can influence the implementation of these technologies, as demonstrated by Meta's recent decision to discontinue third-party fact-checking on Facebook, Threads, and Instagram (McMahon et al., 2025). Thus, the public is typically still left to their own devices to verify the digital content they consume. There is a serious need to develop behavioural interventions to mitigate the adverse impacts of AI-created content, such as deepfakes (eSafety Australia, 2022; World Economic Forum, 2024).

To date, the development of effective deepfake detection interventions has been hampered by our lack of knowledge of the strategies and processes people employ when attempting deepfake detection. Very little research has examined the specific approaches – conscious or unconscious – that individuals adopt during deepfake detection. Without this foundational knowledge, interventions may be poorly aligned with natural detection behaviours.

The current research sought to address this gap by collecting self-report and eye-tracking data while participants knowingly engaged in a video-based deepfake detection task. This approach provides a more ecologically valid representation of how members of the public process potentially manipulated content when actively searching for deception. To this end, we conducted two complementary studies. Study 1 replicated and extended Somoray and Miller (2023) – discussed below – using an alternative recruitment method. It aimed to further evaluate the efficacy of Somoray and Miller's (2023) passive, visual-anomaly-focused intervention and to examine participants' self-reported strategies for detecting deepfakes. Study 2 investigated implicit detection processes by using eye-tracking methods to directly measure participants' gaze during a deepfake detection task.

2. Literature review

Meta-analytic evidence indicates that the general public typically performs at chance levels on deepfake detection tasks across media modalities, including video (Diel, Lalgi, et al., 2024). Various detection interventions have been developed and tested to improve the public's ability to discern deepfakes (for an overview, see Somoray et al., 2025). Interventions can vary in both focus (e.g., identifying common visual and/or auditory anomalies [also called "artifacts"], increasing motivation to perform well, or assessing the plausibility of message content) and level of interactivity (passive interventions vs. more active interventions in which feedback on performance is provided).

Attempts to increase detection by bolstering motivation have generally proven ineffective (Somoray et al., 2025). For instance, Köbis et al. (2021) found that raising awareness about the dangers of deepfakes or offering cash incentives for correct detections did not enhance detection accuracy. This suggests that the inability to detect deepfakes reflects a skill deficit, rather than a lack of motivation to perform well.

Active interventions have shown some promise. Feedback-based interventions have been found to improve detection for static images (Diel, Teufel, & Bäuerle, 2024; Robertson et al., 2018). However, other studies have failed to replicate these findings when using higher-quality stimuli (Kramer et al., 2019). Tailored

media literacy lectures have also been shown to impact perceptions of deepfake video credibility (El Mokadem, 2023), and one-on-one "walk-through" examples have been employed successfully to enhance video detection accuracy (Tahir et al., 2023).

By comparison, passive intervention approaches appear to be less effective. For example, Somoray and Miller (2023) tested a written, visual-anomaly-focused intervention adapted from detection advice provided by the MIT Media Lab. They found the proportion of videos correctly identified on a detection task to be nearly identical in their control (60%) and intervention group (61%). However, these null findings may partly reflect their recruitment method: Via a post on Reddit. If participants happened to share detection strategies in the post thread, this would have "washed out" the effect of their intervention. Bray et al. (2023) and Kramer et al. (2019, Study 2) similarly found that providing anomaly-based detection advice (either once or repeatedly) did not improve detection for static images.

Passive, anomaly-based interventions are simple and scalable, making them attractive options for use in public safety campaigns. However, they currently lack demonstrated efficacy. Refining such interventions requires a clearer understanding of the behaviours people engage in during deepfake detection. Eye-tracking studies have the potential to elucidate this issue. Yet, existing eye-tracking studies in this domain have methodological limitations that may constrain the insights they offer into video detection behaviours. Many have relied on still image stimuli (Caporusso et al., 2020; Cartella et al., 2024) or video stimuli viewed by participants naïve to the study's purpose (Gupta et al., 2020; Wöhler et al., 2021). Tahir et al. (2021) did incorporate videos in a detection task, but tracked gaze only in relation to static screenshots, not during dynamic viewing. Study 2 in the present research sought to address these issues by recording eye-tracking data during dynamic viewing of video stimuli.

3. Study 1

3.1 Method

3.1.1 Design

Study 1 employed an online between-subjects experimental design in which participants were randomly assigned to either an intervention (receiving a list of written detection strategies) or control condition, before completing a deepfake detection task. The Human Research Ethics Committee of James Cook University granted ethical approval to conduct the study. The study was preregistered on OSF (https://osf.io/vutb8) on April 21, 2023, before data collection.

3.1.2 Stimulus materials and measures

The same written detection strategies were used as in Somoray and Miller (2023). These strategies were sourced from the MIT Media Lab (https://www.media.mit.edu/projects/detect-fakes/overview/). These strategies are provided in the Supplementary Material (Table S1 in OSF file).

The detection task involved the presentation of 20 stimulus videos. The same videos were used as in Somoray and Miller (2023). They were originally sourced from the Deepfake Detection Challenge (DFDC) Dataset (Dolhansky et al., 2020). All videos were 10 seconds in length and featured regular people rather than public figures. Stimulus videos depicted an equal number of male and female models and included models of various skin tones.

Each participant saw exactly 10 deepfake and 10 authentic videos. Before the detection task, participants were informed as to how many videos they would be presented with and what proportion would be deepfakes. Two sets of videos were created. That is, Set A contained the authentic version of Video 1, whereas Set B contained the deepfaked version, et cetera. Participants were randomly assigned to receive either Set A or B. The order of the presentation of videos within sets was randomised to mitigate order effects.

Participants responded to each video with one of two binary options: *This video is a deepfake* or *this video is real*. Detection accuracy was calculated as the number of videos correctly categorised divided by the total number of videos categorised (e.g., correctly categorising 13 out of 20 videos would give a detection accuracy score of .65). After the detection task, participants were presented with an open-ended question asking what strategies they employed during the task. The wording of this question differed between conditions: Control condition = "What strategy/s did you use when doing the detection activity?"; intervention condition = "Which, if any, of the strategies provided at the beginning of the experiment helped you the most during the detection activity? Additionally, what other strategy/s, if any, did you use during the detection activity?" Participants were also asked about their perceptions of their susceptibility to deepfake-based scams and misinformation. These findings are reported elsewhere (Dornbusch et al., 2025).

3.1.3 Procedure

Following Somoray and Miller (2023), participants were randomly assigned to either the intervention (provided with a list of written detection tips) or control condition. Participants were then given information about the detection task and presented with two comprehensive check questions, which they were required to answer correctly before they could start the detection task. These questions concerned the definition of deepfakes and the proportion of deepfaked videos in the detection task video set (50%). They were also informed that, at the end of the study, they would receive a score reflecting the number of videos they correctly categorised. Participants then completed the detection task before being asked to provide demographic information. Participants were able to watch each video as

many times as they wished. After the detection task, participants were debriefed and provided with their detection accuracy score.

3.1.4 Recruitment and Participants

Participants were recruited via a student participation scheme at the authors' institution and by sharing the study via the authors' professional networks and snowball recruitment. Student participants were provided with course credit in exchange for their participation. Recruitment occurred from April 2023 to February 2024.

The study was accessed by 474 people. Participant data were removed if participants: 1) did not provide consent, 2) did not attempt the detection task, 3) spent on average under 10 seconds watching each stimulus video, or 4) indicated that this was not their first time participating in the study. This left a final sample of 391 participants. Demographic characteristics of the sample are reported in Table 1.

Table 1. Participant demographics for Study 1 (N = 391) and Study 2 (N = 32)

Variable	Study 1	Study 2
	M (*SD*)	
Age	25.80 (10.24)	26.32 (7.95)
	n (%)	
Gender		
Male	116 (29.7)	11 (34.4%)
Female	271 (69.3%)	21 (65.6%)
Non-binary	3 (0.8%)	-
Country of residence		
Australia	213 (54.5%)	32 (100.0%)
Singapore	159 (40.7%)	-
China	8 (2.0%)	-
Other countries	11 (2.8%)	-
Highest level of education		
High school graduate	166 (42.5%)	12 (37.5%)
TAFE/other vocational studies	43 (11.0%)	4 (12.5%)
Undergraduate degree	137 (35.0%)	6 (18.8%)
Some postgraduate study or a postgraduate degree	45 (11.5%)	10 (31.3%)

3.1.5 Codebook development

Quantitative content analysis was used to analyse responses to the open-ended question. A codebook was developed to facilitate this process. Initially, three investigators independently coded 10% of responses while blinded to the experi-

mental condition, generating potential coding categories (e.g., *voice*, *blur*, *gut feeling*) and organising these into putative groupings (e.g., *visual artefacts*, *feeling*). The investigators then met to develop a pilot codebook containing groupings, codes, definitions, and examples. To test the codebook's reliability, two authors independently coded an additional 10% of responses. The coders had a 75% agreement in categorising these responses, demonstrating moderate intercoder reliability (Burla et al., 2008). Following this, the raters met to make necessary modifications to the pilot codebook. For example, a code was added (e.g., *skin – general – any mention of wrinkles, blemishes, smoothness or agedness of the skin, without specification as to whether this is on the face or body*). The finalised codebook is provided in the Supplementary Material (Table S2 in OSF file). The remainder of responses were coded by one investigator. To prevent rater drift, coding was completed in blocks with regular codebook review.

3.2 Results

3.2.1 Detection accuracy

In the overall sample, mean detection accuracy was .61 (*SD* = .14), suggesting that participants, on average, correctly identified 12 out of the 20 videos. This is above the degree of accuracy that would be expected by chance alone (.50). The poorest performers correctly categorised 4 out of 20 videos (.20), while the best performers correctly categorised 19 out of 20 videos (.95). The control (*M* = .61, *SD* = .14) and intervention groups (*M* = .60, *SD* = .14) did not differ on detection accuracy, $t(389) = 0.46$, $p = .646$, Cohen's $d = 0.05$.

3.2.2 Content analysis of self-reported detection strategies

Of the 392 participants, 47 did not respond to the open-ended question and were therefore removed from the content analysis. Analysis of participant responses indicated that most participants reported employing more than one strategy. A total of 640 detection strategies were reported across the 345 participants who responded to the open-ended question. Table 2 provides the percentage of participants who reported each type of strategy for the whole sample and broken down by experimental condition. Colour gradient heat-mapping (green for higher values, white for lower values) is used to visualise which strategies were more commonly reported. Across the entire sample, the most frequently reported detection strategy was to look for *visual attributes* (this coding category was defined as "Any mention of shadows, lighting, textures, colours or resolution. This does NOT include glitches or blurring"; for definitions for all codes see Table 2) with just over a third of participants giving a response which could be categorised under this code. Other popular strategies (reported by > 10% of the overall sample) included: *Body movement*; *face movement – eyes*; and *facial features* – eyes.

Table 2. Frequency of self-report strategies in overall Study 1 sample and by condition

Grouping	Code	Codebook description	Overall sample #	Overall sample %	Condition	
				(N = 344)	Control % (n = 178)	Intervention % (n = 166)
Audio artefacts	Audio attribute	Any mention of sound quality that is NOT about background noise or voice.	8	2.3%	3.9%	0.6%
	Background noise	Any mention of background noise or environmental noise.	4	1.2%	1.1%	1.2%
	Voice	Any mention of tone/modulation of voice or accents.	20	5.8%	10.7%	0.6%
Body	Body feature	Any mention of a body feature other than the face and does NOT mention skin. This includes neck, collarbone or hands, "composition" or body.	7	2.0%	1.1%	3.0%
	Body movement	Any mention of abnormality regarding body language, demeanour or posture, body movement or hand gestures.	54	15.7%	24.7%	6.0%
Face	Face expression	Any mention of facial expression. This includes mood, expression, and emotions.	48	14.0%	16.3%	11.4%
	Facial feature - eyes	Any mention of the eye area that does NOT mention movement (e.g., blinking).	48	14.0%	8.4%	19.9%
	Facial feature - general	Any mention of a holistic evaluation of the face that is NOT to do with the expression of emotion.	36	10.5%	10.1%	10.8%
	Facial feature - hair	Any mention of facial hair or hair line. Eyebrows/lashes don't count as hair.	15	4.4%	1.7%	7.2%
	Facial feature - mouth	Any mention of movement around the mouth, including lips and teeth.	14	4.1%	3.4%	4.8%
	Facial movement - eyes	Any mention of the movement of the eyes, including blinking or eyebrow movement.	52	15.1%	9.6%	21.1%
	Facial movement - general	Any mention of the overall movement of the face such as angle that does NOT specify a particular facial feature.	32	9.3%	9.0%	9.6%
	Facial movement - mouth	Any mention of the movement of the mouth, including teeth, lips.	9	2.6%	3.4%	1.8%
Feeling	Gut feeling	Anything relating to participant's emotional reaction or feelings. This includes "uncanny valley" "the vibe being off" etc.	19	5.5%	10.1%	0.6%
Skin	Discrepancy of the skin on face and body	Any mention of discrepancy between the skin on face and body. For example, any mention of the agedness of the skin not matching across face and body.	1	0.3%	0.0%	0.6%
	Skin - face	Any mention of wrinkles, blemishes (e.g., moles), smoothness, folding or agedness of the person in the facial area. This does NOT include other parts of the body (see below). The term "complexion" can be considered to be referring to the face.	18	5.2%	1.7%	9.0%
	Skin - general	Any mention of wrinkles, blemishes, smoothness or agedness of the skin, without specification as to whether this is on the face or body.	28	8.1%	1.1%	15.7%
Sync	Sync between voice and mouth movement	Any mention of the discrepancy between the synchronisation of the voice/speech and mouth movement. For example, any mention of match or mismatch of voice/speech and mouth movement.	17	4.9%	7.3%	2.4%
Visual artefacts	Blur	Any mention of blurring or smoothness, including the face.	15	4.4%	3.9%	4.8%
	Glitch	Any mention of glitching, including the face.	38	11.0%	14.0%	7.8%
	Visual attribute	Any mention of shadows, lighting, textures, colours or resolution. This does NOT include glitches or blurring.	123	35.8%	28.7%	43.4%
Other	None/unsure	Code for people who said no, unsure, or N/A.	16	4.7%	1.7%	7.8%
	Not covered by above codes	Participant gives a response not covered by an existing coding unit.	9	2.6%	3.4%	1.8%
	Unclear response	Response is unclear or ambiguous.	8	2.3%	4.5%	0.0%

As seen in Table 2, differences between the control and intervention group were observed for some codes. Compared to participants in the control condition, participants in the intervention group more frequently reported engaging in strategies falling under the following codes: *Visual attribute* (control = 28.7%, intervention = 43.4%); *skin – general* (control = 1.1%, intervention = 15.7%); *facial features – eyes* (control = 8.4%, intervention = 19.9%); *facial movement – eyes* (control = 9.6%, intervention = 21.1%); and *skin – face* (control = 1.7%, intervention = 9.0%). In contrast, the control group more frequently reported strategies falling under codes such as *body movement* (control = 24.7%, intervention = 6.0%); *voice* (control = 10.7%, intervention = 0.6%); and *gut feeling* (control = 10.1%, intervention = 0.6%).

3.3 Discussion

In Somoray and Miller (2023) the intervention group – who received a list of strategies they could apply to aid themselves in the detection task – did not outperform the control group. There are a number of possible reasons for this lack of an effect, including 1) recruitment via social media platforms undermining the validity of the experimental manipulation (e.g., if information was shared to the control group in discussion threads), 2) the detection guidance provided to participants being ineffective (e.g., incorrect or difficult to apply), or 3) intervention-group participants choosing not to apply the strategies outlined in the provided detection guidance.

The overall samples' detection accuracy in the current study was virtually identical to that observed in Somoray and Miller (2023) – Study 1: M = .61, SD = .14; Somoray and Miller (2023): M = .61, SD = .13. Consistent with Somoray and Miller (2023), the Study 1 intervention group performed almost identically to the control group. This suggests that the lack of an experimental effect observed in Somoray and Miller (2023) is not solely attributable to the authors' recruitment approach.

Content analysis of participants' self-reports does suggest that the intervention influenced participants' behaviours. The intervention group appeared to focus on areas reflective of those highlighted in the detection tips they were provided with. For instance, the intervention group were more likely to self-report examining the skin on the models' faces for anomalies, reflecting one of the detection strategies ("Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? Deepfakes are often incongruent on some dimensions."). In contrast, participants in the control group were more likely to self-report relying on their "gut feeling" or irrelevant features such as the model's body language (a likely ineffective strategy, given that deepfakes are typically face manipulations). This suggests that the non-significant results observed in Study 1 and in Somoray and Miller (2023) were not due to participants in the intervention group simply ignoring the detection strategies provided to them. This casts doubt on whether these strategies are fit for purpose.

4. Study 2

4.1 Method

4.1.1 Design

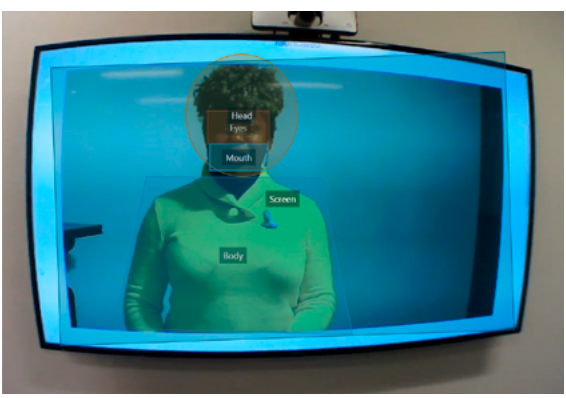
Study 2 was an in-person laboratory study in which participants completed a detection task while their gaze behaviours were recorded. Unlike in Study 1, a detection intervention was not introduced. The Human Research Ethics Committee of James Cook University granted ethical approval to conduct the study.

4.1.2 Materials, measures, and apparatus

The detection task was similar to that used in Study 1. This time, however, five practice trial videos were presented prior to the presentation of ten detection task videos. As in Study 1, stimulus videos were sourced from the DFDC dataset, although the specific videos used differed. Following the Study 1 procedure, two sets of videos were created, and the order in which videos were presented within sets was randomised. The videos depicted models of various genders and skin tones. All models were non-public figures.

The same detection accuracy index was used as in Study 1. Three gaze variables were analysed as part of this study: Average fixation duration (the average duration of participants' fixations, measured in milliseconds), fixation count (the frequency with which participants fixated their gaze), and saccade count (the frequency with which participants made saccades, i.e., shifted their gaze between fixations). These variables were recorded in relation to five areas of interest (AOIs): The screen, the stimulus model's body, the stimulus model's head, the stimulus model's eye area, and the stimulus model's mouth area. However, we report results only for the eyes, mouth, and body AOIs (as the eyes and mouth AOIs are situated within the head AOI and all other AOIs are situated within the screen AOI). These AOIs are depicted in Figure 1.

Figure 1. An example of the areas of interest (eyes, mouth, body, head, and screen) created during data processing. Image representative of participants' field of view.



Eye-tracking information was recorded using Pupil Labs' *Pupil Invisible* model glasses. These glasses fit like normal prescription eyeglasses, allowing for naturalistic movement. They record the movement of each eye. The specifications of this equipment are provided in the Supplementary Material hosted on OSF. Participants completed the study sitting in a chair approximately 57cm away from a 70cm HD TV screen. The stimulus videos took up most of the screen. Figure S1 (in OSF file) depicts the experimental setup.

4.1.3 Procedure

At the start of the study, participants were informed that exactly half of the detection task videos were deepfakes and that they would receive a detection accuracy score at the study's conclusion. Participants then completed the same comprehension check questions as in Study 1. The eye-tracking glasses were then calibrated to the participant, and the five practice trial videos were presented. Following each practice trial, participants received feedback indicating whether their detection decision was correct or incorrect. Participants then completed the detection task. Unlike in Study 1, participants were not permitted to watch stimulus videos more than once. Following the detection task, participants completed demographic questions and received a debriefing that included their detection accuracy score. Investigators read from a pre-established script when explaining the study procedure to participants.

4.1.4 Recruitment and participants

As in Study 1, participants were recruited via undergraduate student recruitment channels (in exchange for course credit) and snowball recruitment within the researchers' personal and professional networks. Recruitment occurred from December 2023 to August 2024. Student participants were offered course credit for their participation, and non-student participants were entered into a prize draw for a gift card. Those who require eyeglasses for up-close work were excluded (unless wearing contact lenses), as the eye-tracking glasses do not fit comfortably over regular eyeglasses. Demographic characteristics of the sample are reported in Table 1.

4.1.5 Data processing

Eye-tracking data was processed using Pupil Labs' iMotions 10 software. This involved manually creating AOIs and moving these to match the movement of the stimulus video model (e.g., moving the eyes AOI to the left as the stimulus video model moved their head to the left of screen). This was done for all AOIs, for all 10 detection task videos, for each participant. Practice trial videos were excluded from this process, as this data was not included in the analysis. Further technical details of the data processing are provided in the Supplementary Material in OSF.

4.2 Results

Mean detection accuracy in the overall sample was .65 (*SD* = .20), indicating that participants, on average, correctly identified 6.5 out of 10 videos. This is above the degree of accuracy that would be expected by chance alone (.50). The best performer correctly categorised all 10 videos (1.00), while the worst performer correctly categorised 2 videos only (.20).

Table 3 presents descriptive statistics for average fixation duration, fixation count, and saccade count broken down by AOI (body, eyes, mouth), video authenticity (deepfake or authentic), and decision (correct or incorrect categorisation of video) for the overall sample. The table suggests that participants' visual attention was directed predominantly towards the eyes and mouth, rather than towards the body.

Table 3. Descriptive statistics for gaze variables across Study 2 sample (*N* = 32) by video authenticity, decision, and area of interest (AOI)

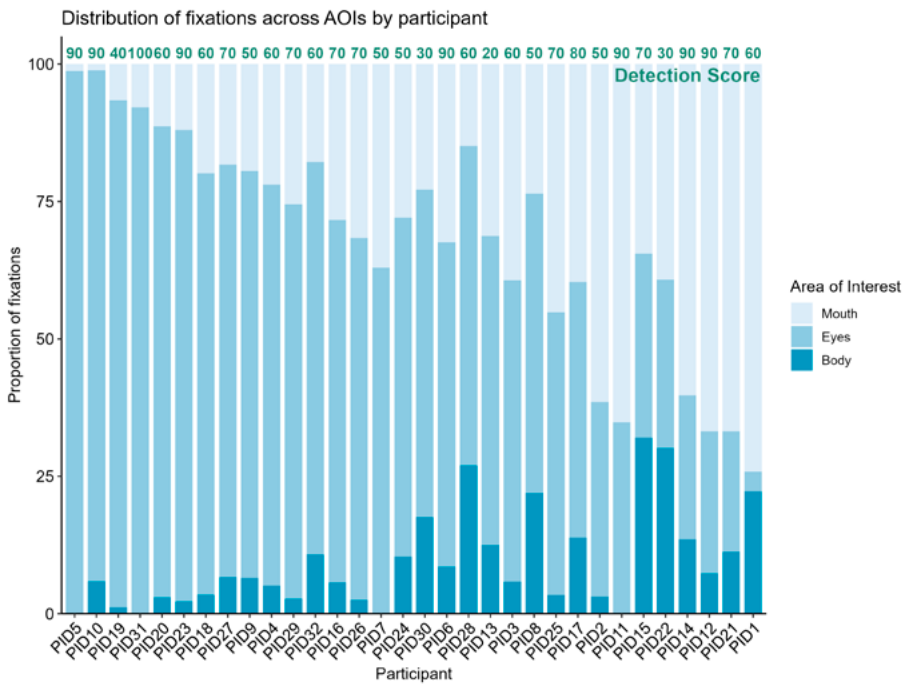
AOI	All Videos	Video authenticity		Video decision	
		Deepfake	Authentic	Correct	Incorrect
Average fixation duration (ms)					
Body	85.70 (119.19)	84.84 (118.97)	86.57 (119.78)	82.30 (120.72)	92.08 (116.55)
Eyes	395.91 (232.78)	415.29 (232.87)	376.41 (231.77)	387.39 (231.06)	411.86 (236.18)
Mouth	352.14 (263.12)	353.87 (276.02)	350.39 (250.33)	349.95 (256.02)	356.24 (277.07)
Fixation count					
Body	1.61 (2.54)	1.55 (2.63)	1.66 (2.46)	1.45 (2.41)	1.89 (2.76)
Eyes	10.64 (6.73)	11.12 (6.84)	10.15 (6.60)	10.85 (7.07)	10.24 (6.04)
Mouth	5.77 (5.54)	5.35 (5.44)	6.18 (5.62)	6.29 (5.97)	4.78 (4.48)
Saccade count					
Body	1.93 (3.15)	1.86 (3.11)	2.01 (3.19)	1.76 (3.01)	2.27 (3.38)
Eyes	12.53 (10.58)	12.90 (10.42)	12.15 (10.76)	12.78 (11.00)	12.05 (9.78)
Mouth	6.70 (7.59)	6.30 (7.67)	7.10 (7.51)	7.20 (7.78)	5.76 (7.16)

A 3 (AOI) × 2 (video authenticity) × 2 (video decision) repeated measures ANOVA was conducted for each outcome variable (average fixation duration, fixation count, saccade count). To account for potential interactions, each ANOVA included four interaction terms: AOI × authenticity; AOI × decision; authenticity × decision; and AOI × authenticity × decision. The details of these analyses are provided in the Supplementary Material (Tables S3–S11 in OSF file). These analyses indicated that participants had significantly longer fixations when looking at the eyes and mouth, relative to the body ($p < .001$ in both cases). The difference in average fixation length between the eyes and mouth AOIs was non-significant ($p > .999$). Further, participants made significantly more fixations on the eyes

than the mouth (p = .033) or body (p < .001). They also fixated more frequently on the mouth than the body (p < .001). Similarly, participants engaged in more saccades in the eyes AOI compared to the body AOI (p < .001). Saccades were more frequent in the mouth AOI compared to the body AOI (p = .006) but not the eyes AOI (p = .122). All reported p-values have been Bonferroni corrected. Average fixation duration, fixation count, and saccade count did not differ between correctly or incorrectly categorised videos or between deepfake and authentic videos.

While, on average, the eyes tended to attract the most visual attention, there did appear to be individual differences around this. In Figure 2, it can be seen that some participants focused almost exclusively on the eyes, some focused almost exclusively on the mouth, and others spent a roughly equal amount of time on each AOI. Detection accuracy was unrelated to proportion of time spent looking at the eyes, $r(30)$ = .10, p = .603, mouth, $r(30)$ = .04, p = .826, or body, $r(30)$ = -.34, p = .055. In the latter case, results are bordering on significance, which could suggest that those who spent more time looking at the body tended to perform worse on the detection task.

Figure 2. Distributions of fixations across areas of interest, along with detection accuracy scores for all participants



4.3 Discussion

Study 2 suggests that, when trying to determine the authenticity of videos, participants show a strong preference for looking at eyes of stimulus models rather than the body, and a moderate preference for the eyes rather than the mouth. Participants' apparent focus on the eyes is somewhat inconsistent with past studies, which have found that attention is often directed away from the eyes and towards other regions of the face when participants view high-quality deepfakes (Wöhler et al., 2021). This said, visualisation of the data (Figure 2) suggests that there was a subset of participants who adopted a "mouth-focused" approach. Detection accuracy was unrelated to proportion of time spent looking at the eyes and proportion of time spent looking at the mouth. Spending a greater proportion of time looking at the body may be associated with poorer detection performance.

Participants exhibited similar gaze patterns regardless of whether they correctly or incorrectly categorised videos, as evidenced by the lack of main effects for video decision. Participants may have employed a consistent visual search strategy across all videos – such as rapidly scanning the eye area for anomalies – with variable success depending on the presence and detectability of visual cues (i.e., some videos may contain obvious anomalies that others do not).

Participants exhibited similar gaze patterns when viewing authentic versus deepfake videos, as evidenced by the absence of main effects for video authenticity. This indicates that participants did not subconsciously modify their visual behaviour in response to deepfake content, at least not in ways captured by our gaze measurements. These findings contrast with previous research (Gupta et al., 2020; Wöhler et al., 2021), which documented distinct gaze patterns when participants unknowingly viewed deepfake videos. A critical methodological difference may explain this discrepancy: Unlike previous studies, participants in our experiment were explicitly aware they were in a deepfake detection task. This awareness may have resulted in participants adopting a more deliberate visual search strategy, which they actively applied to all videos.

5. Overall discussion

These studies sought to investigate laypeople's behaviour when faced with the problem of trying to identify deepfake videos. This was done through the analysis of participants' self-reports of the strategies they employed on a deepfake detection task (Study 1) and gaze data collected during a detection task (Study 2). Many of the findings are relevant to those seeking to design better deepfake detection training modules.

First, both studies corroborate prior research indicating that deepfake detection is difficult for most individuals (Diel, Lalgi, et al., 2024; Köbis et al., 2021; Somoray & Miller, 2023), with participants performing only marginally better than chance. Importantly, this poor performance occurred despite participants being explicitly warned that they would encounter deepfakes. For this reason, we

should expect "real-world" detection rates to be even lower than those observed in Studies 1 and 2.

Second, Study 1 suggests that the provision of written detection tips is not enough to meaningfully bolster detection rates (as also found in Somoray & Miller, 2023). These null findings are consistent with the results of other studies into the efficacy of passive, anomaly-based detection interventions (Bray et al., 2023; Kramer et al., 2019). However, the findings do indicate that people will shift their behaviour on detection tasks in response to detection instructions. That is, the detection approaches self-reported by participants in the intervention condition showed a greater alignment with the detection instructions than those of the control group. This highlights the possibility of improving the public's detection abilities through passive detection interventions (even if the specific advice tested in Study 1 was itself ineffective).

Third, a consistent finding across both studies is that many people gravitate towards the face region, particularly the eyes, when trying to ascertain video veracity. This may reflect Western cultural norms around eye contact (Senju et al., 2013). Most deepfakes involve face replacement (Silva et al., 2022) – imposing the face of a target person onto a model, while leaving the model's body unadjusted. Thus, a focus on the face is advisable during deepfake detection. However, for this same reason, assessing for discrepancies between the face and body may also be informative (e.g., looking for discrepancies in the agedness of the skin on the face and hands). Future training modules may benefit from overtly directing participants towards this strategy, while, ideally, also providing visual examples. It is also worth noting that the eyes may be less diagnostic than other face regions. Areas such as the boundary of the face (which may show visual peculiarities, particularly during head movement) or the lips (which may reveal errors in audio-mouth synchronisation) could provide more reliable cues. Future studies should investigate whether explicitly directing participants' attention away from the eyes and towards other regions of the face improves detection accuracy.

Future research should also investigate the gaze behaviour of deepfake detection experts. Across most domains, experts demonstrate more efficient and selective visual scanning than novices, strategically directing attention to task-relevant areas and maintaining longer fixations on critical information (Brams et al., 2019). The domain of medicine represents a notable exception, where experts exhibit more extensive visual spans. The gaze patterns of superior detectors could reveal optimal visual strategies for deepfake identification. To facilitate the identification of individuals with exceptional detection abilities, population norms should be established by administering standardised video sets to large representative samples.

In interpreting the study's findings, it is important to consider the choice of stimulus videos, which all depicted non-public figures discussing mundane topics. This is both a strength and limitation of the study. It is a strength in that it minimises the influence of prior knowledge or contextual biases, forcing participants to rely on visual and auditory cues. By controlling for these factors, the study provides a clearer picture of deepfake detection behaviour "in a vacuum" and

offers insight into how people identify manipulated content when contextual information is limited.

At the same time, the use of non-public figures discussing mundane topics limits ecological validity. In real-world scenarios, videos often feature public figures or emotionally salient messages, where context cues and prior knowledge and attitudes play an important role. For example, when assessing videos of public figures such as politicians, detectors can draw on visual and auditory cues while *also* evaluating whether the message content aligns with what they know of the figure's beliefs ("Would this person ever say something like this?"). Familiarity with the deepfaked subject may even enhance ability to pick up on visual anomalies (Thaw et al., 2020). These factors would likely increase detection performance. Conversely, the use of known figures discussing charged topics may, in some instances, undermine performance. For example, detectors are less likely to correctly identify deepfakes when message content aligns with their existing personal beliefs (Sütterlin et al., 2023). Holding strong prior attitudes towards the deepfaked subject may also influence detection decisions (Ng, 2023).

Several other limitations also warrant consideration. First, the study did not account for individual differences in perceptual expertise that may influence deepfake detection ability. Future research should explore whether factors such as experience with digital media production moderate gaze behaviour and detection accuracy. Second, while this study examined overall gaze patterns, it did not differentiate between deepfakes of varying sophistication. Research suggests that deepfake quality impacts detection performance (see Somoray et al., 2025) and that individuals unconsciously adjust their visual behaviour based on deepfake quality (Wöhler et al., 2021), warranting further investigation of this factor. Finally, in both studies, individuals with a particular interest in deepfakes may have been more inclined to participate, introducing the possibility of sampling bias. If greater familiarity with, or interest in, deepfakes is linked to enhanced detection performance, the sample's performance may have been greater than that of the general population. However, this concern is somewhat mitigated by the recruitment of student participants, who were likely motivated to participate by external factors (e.g., course credit) rather than a specific interest in deepfakes.

GenAI declaration

Generative AI (Claude 4.0 and ChatGPT-5) was used for basic copy-editing.

Supplementary material

A supplementary material file can by found on OSF: https://osf.io/tzpd7/files/osfstorage/6922940d5f3279069d76fc29. All other materials associated with the study can be found on the OSF page for Study 1: https://osf.io/tzpd7.

References

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, *252*. https://doi.org/10.1016/j.eswa.2024.124260

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M., & Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin*, *145*(10), 980–1027. https://doi.org/10.1037/bul0000207

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

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research, 57*(2), 113–117. https://doi.org/10.1097/01.NNR.0000313482.33917.7d

Caporusso, N., Zhang, K., & Carlson, G. (2020). Using eye-tracking to study the authenticity of images produced by generative adversarial networks. *2020 International Conference on Electrical, Communication, and Computer Engineering* (ICECCE), 1–6. https://doi.org/10.1109/ICECCE49384.2020.9179472

Cartella, G., Cuculo, V., Cornia, M., & Cucchiara, R. (2024). Unveiling the truth: Exploring human gaze patterns in fake images. *IEEE Signal Processing Letters*, 1–5. https://doi.org/10.1109/LSP.2024.3375288

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*. https://doi.org/10.1016/j.chbr.2024.100538

Diel, A., Teufel, M., & Bäuerle, A. (2024). *Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy*. OSF. https://doi.org/10.31219/osf.io/muwnj

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Cristian Canton Ferrer. (2020). *The DeepFake Detection Challenge (DFDC) Dataset*. arXiv. https://doi.org/10.48550/arxiv.2006.07397

Dornbusch, A., Tye, T., Somoray, K., & Miller, D. J. (2025). *Third person effects and the base-rate fallacy: Cognitive biases in deepfake detection* [Manuscript in preparation].

El Mokadem, S. S. (2023). The effect of media literacy on misinformation and deep fake video detection. *Arab Media & Society*, *35*, 53–78. https://www.arabmediasociety.com/

Flynn, A., Powell, A., Scott, A. J., & Cama, E. (2022). Deepfakes and digitally altered imagery abuse: A cross-country exploration of an emerging form of image-based sexual abuse. *The British Journal of Criminology*, *62*(6), 1341–1358. https://doi.org/10.1093/bjc/azab111

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

Gupta, P., Chugh, K., Dhall, A., & Subramanian, R. (2020). The eyes know it: FakeET- An eye-tracking database to understand deepfake perception. *Proceedings of the 2020 International Conference on Multimodal Interaction, 5*19–527. https://doi.org/10.1145/3382507.3418857

Köbis, N, C., Doležalová, B., & Soraperra., I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11). https://doi.org/10.2139/ssrn.3832978

Kramer, R. S., Mireku, M. O., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, *4*(1). https://doi.org/10.1186/s41235-019-0181-4

McMahon, L., Kleinman, Z., & Subramanian, C. (2025, January 8). Facebook and Instagram get rid of fact checkers. *BBC News*. https://www.bbc.com/news/articles/cly74mpy8klo

Miller, D. J., Somoray, K., & Stevens, H. (2025). *A shallow history of deepfakes*. SSRN. http://dx.doi.org/10.2139/ssrn.5130379

Ng, Y. L. (2023). An error management approach to perceived fakeness of deepfakes: The moderating role of perceived deepfake targeted politicians' personality characteristics. *Current Psychology*, *42*, 25658–25669. https://doi.org/10.1007/s12144-022-03621-x

Robertson, D. J., Mungall, A., Watson, D. G., Wade, K. A., Nightingale, S. J., & Butler, S. (2018). Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research: Principles and Implications*, *3*. https://doi.org/10.1186/s41235-018-0113-8

Senju, A., Vernetti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., & Johnson, M. H. (2013). Cultural background modulates how we look at other persons' gaze. *International Journal of Behavioral Development*, *37*(2), 131–136. https://doi.org/10.1177/0165025412465360

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, *4*. https://doi.org/10.1016/j.fsisyn.2022.100217

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*. https://doi.org/10.1016/j.chb.2023.107917

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, *2025*. https://doi.org/10.1155/hbe2/1833228

Smith, H., & Mansted, K. (2020). *Weaponised deep fakes: National security and democracy* [Policy brief]. Australian Strategic Policy Institute. https://www.aspi.org.au/report/weaponised-deep-fakes

Sütterlin, S., Ask, T. F., Mägerle, S., Glöckler, S., Wolf, L., Schray, J., Chandi, A., Bursac, T., Khodabakhsh, A., Know, B. J., Canham, M., & Lugo, R. G. (2023). Individual deep fake recognition skills are affected by viewer's political orientation, agreement with content and device used. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition: 17th International Conference, Held as Part of the 25th HCI International Conference, Copenhagen, Denmark, Proceedings: Vol. 14019 (pp. 269–284). Springer, Cham.* https://doi.org/10.1007/978-3-031-35017-7_18

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445699

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2020). Is it real? A study on detecting deepfake videos. *Proceedings of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.366

Wöhler, L., Zembaty, M., Castillo, S., & Magnor, M. (2021). Towards understanding perceptual differences between genuine and face-swapped videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445627

World Economic Forum (2024). *The Global Risks Report 2024*. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

SCM
Studies in Communication and Media

RESEARCH IN BRIEF

Support for deepfake regulation: The role of third-person perception, trust, and risk

Unterstützung für Deepfake-Regulierung: Die Rolle von Third-Person-Perception, Vertrauen und Risiko

Daniel Vogler, Adrian Rauchfleisch & Gabriele de Seta

Daniel Vogler (Dr.), University of Zurich, Research Center for the Public Sphere and Society, Andreasstrasse 15, Zurich, Switzerland. Contact:daniel.vogler@foeg.uzh.ch. ORCID: https://orcid.org/0000-0002-0211-7574
Adrian Rauchfleisch (Prof. Dr.), National Taiwan University, Graduate Institute of Journalism, Taipei, 10617, Taiwan. Contact: adrian.rauchfleisch@gmail.com. ORCID: https://orcid.org/0000-0003-1232-083X
Gabriele de Seta (PhD), University of Bergen, Department of Linguistic, Literary and Aesthetic Studies, Langes Gate 1–3, Bergen, Norway. Contact: gabriele.seta@uib.no. ORCID: https://orcid.org/0000-0003-0497-2811

RESEARCH IN BRIEF

Support for deepfake regulation: The role of third-person perception, trust, and risk

Unterstützung für Deepfake-Regulierung: Die Rolle von Third-Person-Perception, Vertrauen und Risiko

Daniel Vogler, Adrian Rauchfleisch & Gabriele de Seta

Abstract: Like other emerging technologies, deepfakes present both risks and benefits to society. Due to harmful applications such as disinformation and non-consensual pornography, calls for their regulation have increased recently. However, little is known about public support for deepfake regulation and the factors related to it. This study addresses this gap through a pre-registered online survey (n = 1,361) conducted in Switzerland, where citizens can influence political regulation through direct democratic instruments, such as referendums. Our findings reveal a strong third-person perception, as people believe that deepfakes affect others more than themselves (*Cohen's d* = 0.77). This presumed effect on others is a weak but significant predictor of support for regulation (β = 0.07). However, we do not find evidence for the second-person effect – the idea that individuals who perceive deepfakes as highly influential on both themselves and others are more likely to support regulation. However, an exploratory analysis indicates a potential second-person effect among females, who are specifically affected by deepfakes; a result which must be further explored and replicated. Additionally, we find that higher perceived risk and greater trust in institutions are positively associated with support for deepfake regulation.

Keywords: Deepfake technology, regulation, third-person effect, second-person effect, risk perception, trust

Zusammenfassung: Wie andere aufkommende Technologien bringen Deepfakes sowohl Risiken als auch Vorteile für die Gesellschaft mit sich. Aufgrund schädlicher Anwendungen wie Desinformation und nicht einvernehmlicher Pornografie sind die Forderungen nach einer Regulierung von Deepfake-Technologie jüngst gestiegen. Allerdings ist wenig darüber bekannt, inwieweit die Öffentlichkeit eine Regulierung von Deepfakes unterstützt und welche Faktoren dabei eine Rolle spielen. Diese Studie adressiert diese Forschungslücke mit einer präregistrierten Online-Befragung (n = 1.361) in der Schweiz, einem Land, in dem Bürgerinnen und Bürger durch direktdemokratische Instrumente wie Referenden Einfluss auf die politische Regulierung nehmen können. Unsere Ergebnisse bestätigen die Third-Person-Perception: Menschen glauben, dass Deepfakes andere stärker beeinflussen als sich selbst (*Cohen's d* = 0,77). Dieser vermutete Effekt auf andere ist ein schwacher, aber signifikanter Prädiktor für die Unterstützung einer Regulierung (β = 0,07). Allerdings finden wir keine Hinweise auf den Second-Person-Effekt–die Annahme, dass Personen, die Deepfakes sowohl bei anderen als auch bei sich selbst als besonders einflussreich wah-

rnehmen, eine stärkere Unterstützung für Regulierungsmaßnahmen zeigen. Eine explorative Analyse weist allerdings auf einen potenziellen Second-Person-Effekt bei Frauen hin, die besonders von Deepfakes betroffen sind; dieses Ergebnis muss weiter untersucht und repliziert werden. Darüber hinaus stellen wir fest, dass eine höhere Risikowahrnehmung sowie ein größeres Vertrauen in Institutionen positiv mit der Unterstützung für eine Regulierung von Deepfakes zusammenhängen.

Schlagwörter: Deepfake-Technologie, Regulierung, Third-Person-Effekt, Second-Person-Effekt, Risikowahrnehmung, Vertrauen

1. Introduction

Emerging technologies usually come with benefits and risks for society. How and if a technology can establish itself in society depends on how individuals perceive its risks and benefits (Gardner & Gould, 1989; Lima et al., 2005; Slovic et al., 1982). A common approach to coping with the risks of technology is regulation by the state or self-regulation by technology providers. Calls for regulation are often articulated in the public by citizens, journalists, politicians, or non-governmental organizations when the risk of a technology is perceived as outweighing its benefits (Nguyen, 2023). In the field of communication technology, regulatory initiatives have targeted the internet, social media platforms, and AI – often in response to concerns about problematic content, such as disinformation, pornography, or potential negative effects on users, including privacy issues, well-being, violence, and addiction (de Ruiter, 2021; Kim, 2025; Paradise & Sullivan, 2012; Yu et al., 2023).

While deepfake technology has beneficial applications in certain industries and for personal recreation (Bendahan Bitton et al., 2024; Rauchfleisch et al., 2025), it also poses significant risks, particularly in relation to disinformation (Godulla et al., 2021; Hameleers et al., 2022; Vaccari & Chadwick, 2020) and pornography (de Ruiter, 2021). To mitigate these risks, technological detection methods, (digital) media literacy initiatives, as well as regulation by the state or the industry itself, are currently being discussed (Birrer & Just, 2024). However, regulating deepfakes is legally complex, may create economic disadvantages, and is often perceived as a restriction on freedom of speech (Godulla et al., 2021).

In democracies, public acceptance of regulations is crucial, particularly in Switzerland, where referendums can be held on proposed regulations. However, little is known about citizens' support for regulating deepfake technology and the factors related to such support. From studies on disinformation, we know that the perceived negative effects of disinformation are positively related to support for the regulation of content and platforms (Jungherr & Rauchfleisch, 2024). The literature also shows third-person effects related to regulation of technology, as the perception of others' high vulnerability to disinformation or other harmful content is positively associated with support for regulation (Chen et al., 2023; Chung & Wihbey, 2024; Kim, 2025; Riedl et al., 2022).

Our pre-registered online study conducted in Switzerland addresses this gap by drawing on third-person effect literature (Baek et al., 2019; Davison, 1983; Gunther & Storey, 2003).[1] The study shows that people believe deepfakes have a greater influence on others than on

1 Preregistration and full list of hypotheses available at https://aspredicted.org/s2gt-7rwr.pdf

themselves (perceptual third-person effect) and that the perceived effect on others is positively related to support for deepfake regulation. An additional exploratory analysis indicates that gender plays a role. While the presumed effect on others explains support for regulation among male citizens, we observed a potential second-person effect for female citizens, as those who perceive deepfakes as influential on themselves and others show even stronger regulatory support. Furthermore, the study indicates a positive association between support for deepfake regulation and both trust in institutions and perceived risks associated with deepfakes.

2. Conceptual framework

One way to mitigate the risks posed by technology is through regulation. Deepfakes, often associated with disinformation, pornography, and criminal activity in public discourse in Switzerland (Rauchfleisch et al., 2025) and other countries (Gosse & Burkell, 2020; Yadlin-Segal & Oppenheim, 2021), have prompted calls for state-led regulation or self-regulation by platforms. Although few specific laws targeting deepfakes currently exist, they are often addressed within broader regulatory frameworks concerning AI, disinformation, and privacy. In Europe, for instance, providers and moderators of deepfake technology are subject to the Digital Services Act (DSA) and the AI Act. The AI Act requires systems that generate and manipulate images to meet minimal transparency standards (Karaboga et al., 2024). Switzerland recently rejected a specific regulation regarding deepfakes (Swissinfo, 2025), but existing laws, such as criminal law and privacy rights, can still apply to cases involving deepfakes (Thouvenin et al., 2023). This indicates that regulating technologies like deepfakes is a continuum that encompasses multiple frameworks.

2.1 Third-person effect and behavioral second-person effect

The extent to which emerging technologies are regulated depends mainly on the risks and benefits associated with them (Slovic et al., 1982). In the case of deepfakes, their potential impact on public opinion, particularly as a tool for disinformation, is a central concern. Research on the perceived negative effects of communication technology like deepfakes, social media platforms, or games suggests a third-person effect (Davison, 1983), where individuals tend to view the harms as greater for unknown others than for themselves (Ahmed, 2023; Chen et al., 2023; Paradise & Sullivan, 2012; Riedl et al., 2022; Yu et al., 2023), with further notable differences between close and distant others (Altay & Acerbi, 2024; Corbu et al., 2020). Initially developed as a primarily perceptual phenomenon by Davison (1983), the concept was later expanded to include a behavioral dimension. Such extensions suggest the existence of an "influence of presumed influence" (Gunther & Storey, 2003, p. 199), which leads individuals to adjust their behavior based on the belief that others are influenced by the media (Baek et al., 2019).

To date, few studies have analyzed third-person perceptions of the influence of deepfakes. A noteworthy exception is the study by Ahmed (2023), which is based on the third-person perception framework and demonstrates that individuals in the US and Singapore perceive deepfakes as influencing others more than themselves. Many studies have demonstrated the

third-person effect for disinformation: Individuals perceive themselves as more capable of detecting disinformation (Corbu et al., 2020) and less vulnerable to it (Jang & Kim, 2018; Kim, 2025; Liu & Huang, 2020) than others.

The third-person effect is positively related to higher support for regulating communication technologies. Chung and Wihbey (2024) show that presumed media effects on others are related to support for governmental platform regulation as well as self-regulation (i.e., content moderation) in the US, the UK, and South Korea. Thereby, the perceived ability of others to spot misinformation acts as an antecedent of the third-person effect. Similarly, Kim (2025) showed a positive relation between third-person perception of COVID-related disinformation and support for regulating such content. Riedl et al. (2022) identified the third-person perception for perceived effects of social media content on others and platform moderation. However, not all studies lend support to this relationship (Chen et al., 2023). Interestingly, Jang and Kim (2018) demonstrate in their US-based study that the third-person perception of disinformation is positively related to support for media literacy interventions, but not for regulatory approaches by the state or platforms.

In the context of fake news and platform regulation, prior research in some cases supports a second-person effect instead of a third-person effect for the behavioral hypothesis. For example, Riedl et al. (2022) observe a behavioral second-person effect, meaning that people who perceive effects of social media content as high on both themselves and others support extended content moderation but not stronger platform regulation through the state. Similarly, Baek et al. (2019) also identify a second-person effect for the presumed effect of fake news and support for regulation. In our study, we first assume, as a perceptual third-person hypothesis, a difference between the presumed effect of deepfakes on self and others:

> *H1: Individuals will presume a greater deepfake effect on "others" than on the "self".*

The presumed effect on others alone might explain support for regulation. Here, we follow the literature on the "influence of presumed influence" (Gunther & Storey, 2003, p. 199). The following hypothesis can also serve as an alternative explanation if we do not find support for a second-person effect (H3) where the association between the presumed effect on others and support for regulation is moderated by the presumed effect on oneself (Baek et al., 2019).

> *H2: Individuals' presumed deepfake effect on "others" is positively related to their support for the regulation of deepfakes.*

Prior research in the context of online communication (Riedl et al., 2022) and disinformation (Baek et al., 2019) indicated a second-person effect. We also assume, as a behavioral hypothesis, a second-person effect in the context of deepfakes, which would be supported by a significant interaction effect between the presumed effect on others and the self. In contrast, the third-person effect suggests that the issue is perceived primarily as a problem affecting others, rather than oneself. If the interaction is not significant, a significant positive estimate for presumed effect on others and a negative presumed effect on self would support a strict third-person effect. Only a significant negative estimate for presumed effect on others would support

the less strict influence of presumed influence as stated in H2:

> *H3: Individuals with both high presumed deepfake effects on "others" and "self" will show stronger support for the regulation of deepfakes.*

2.2 Trust in institutions

In democracies, regulation is often at least partially delegated to the state. Together with technology providers and experts, state regulators develop frameworks for technology regulation. The delegation of power and responsibility for regulation to a third party requires trust (Six, 2013; Verhoest et al., 2025). However, "in regulatory regimes, the provision of third-party trust is only useful as long as citizens trust the third party" (Verhoest et al., 2025, p. 365). In his theory of justified public trust in regulation, Wolf (2021) highlights that to be trustworthy a regulatory regime must "fairly and effectively manage risk, must be 'science based' in the relevant sense, and must in addition be truthful, transparent, and responsive to public input" (p. 29). We argue that two central institutions ensuring such trustworthy regulatory frameworks are politics and journalism. Politics is the primary actor in drafting, developing, and implementing state-led regulatory frameworks. In an experimental study by PytlikZillig et al. (2017), the participants' trust in water regulatory institutions was positively related to their general trust in government. In a study encompassing 33 European countries, Marien and Hooghe (2011) demonstrate that low trust in the institutions of the political system is associated with a higher acceptance of illegal behavior, such as tax fraud, indicating that individuals are less likely to follow governmental regulations. Journalism, in its role as a watchdog, critically observes the regulatory process and detects weaknesses and undesirable developments (Kalogeropoulos et al., 2022). Therefore, we expect a positive relation between trust in institutions and support for deepfake regulation:

> *H4: Individuals with higher trust in institutions will show stronger support for the regulation of deepfakes.*

2.3 Risk perceptions

New technology always comes with potential risks and benefits for society. The implementation of technology, and how it can be utilized, depends on how these risks and opportunities are perceived by members of a society (Gardner & Gould, 1989; Lima et al., 2005). Calls for state-led regulation of technology usually emerge when individuals or groups perceive the risks as outweighing the benefits of a technology. The perception of risks also depends on the field of application of a technology, as possible benefits may occur in one field and risks might be identified in another. Regarding deepfakes, the risks to politics might be perceived as more severe than those related to the economy, making support for regulation more likely when the risks to politics are regarded as high. Research on disinformation has shown that higher problem perception increases support for regulating online environments (Jungherr & Rauchfleisch, 2024). Considering differences in application fields, we therefore hypothesize that higher risk perceptions will be associated with stronger support for regulating deepfakes.

H5: Individuals with higher risk perception of deepfakes for a) politics, b) the media, c) the economy, and d) the "self" will show stronger support for the regulation of deepfakes.

3. Methods

Our pre-registered study was approved by the ethics committee of the Faculty of Arts and Sciences of the University of Zurich. We used an online panel (Respondi-Bilendi) for our survey, which was conducted in September 2023 (N = 1,361 participants). Participants are individuals residing in Switzerland who are 16 years of age or older. The sample includes participants from both the French and German language regions. Before we began the survey, we ensured that we had sufficient power for our statistical tests. For a sample of 1,200, we had a power of more than 0.9 for all our statistical tests (see Appendix C for more details). The surveys were programmed and administered in both languages using Unipark software. Because the natural fallout in our sample resulted in some age groups having a disproportionate number of female respondents, we computed survey weights based on Swiss population data. In the main paper, we present the model using weighted data (see Appendix D.2.1 for the model with unweighted data).

3.1 Measures

The dependent variable, *support for regulation of deepfakes*, was measured with four items covering support for (1) a general ban of deepfakes, (2) a regulatory framework for prohibiting deepfakes, (3) state-led regulation and (4) self-regulation of deepfakes by platforms (M = 5.10; SD = 1.45; α = .77). We used the items from Baek et al.'s (2019) study and adapted them to the context of our study (overview of the main measures is provided in Appendix B.1).

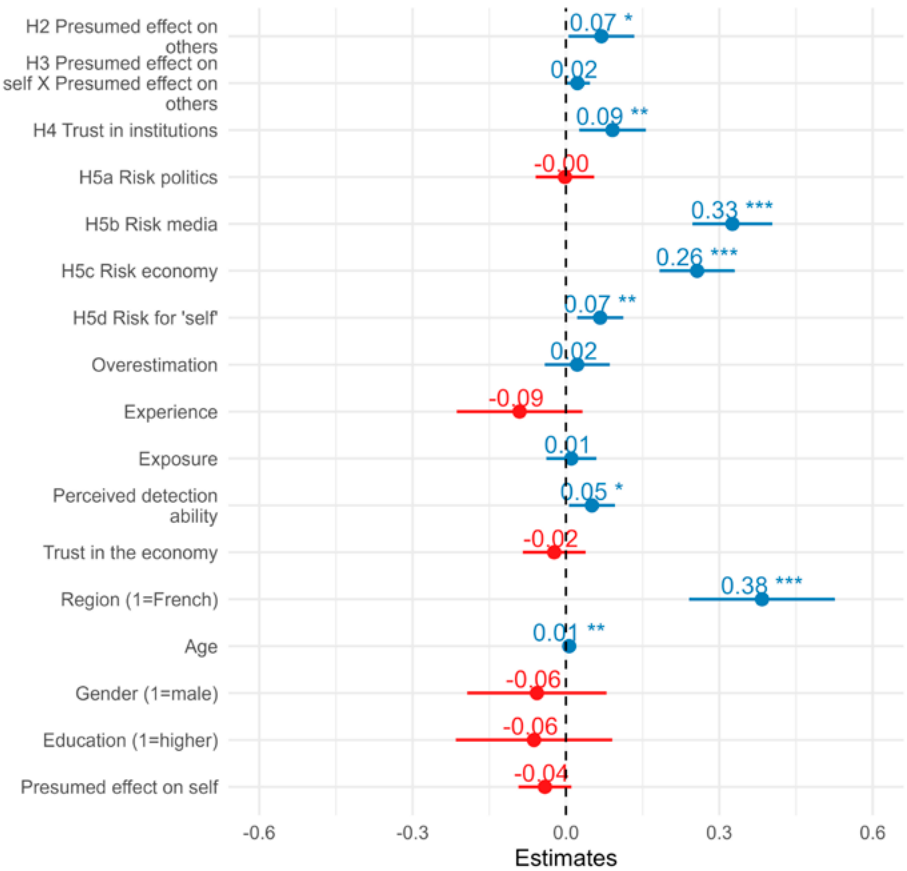
Presumed effects of deepfakes on self and others were measured with two single items by asking participants to estimate how deepfakes influence their own opinions [M = 3.53; SD = 1.70] and the opinions of the Swiss population [M = 4.70; SD = 1.45]. *Trust in institutions* was measured using two items that covered trust in political institutions and journalism (M = 3.71; SD = 1.31; α = .74). We assessed risk perceptions for the different application fields using two items each. We included risks for politics (M = 4.98; SD = 1.64; α = .89), journalism (M = 5.81; SD = 1.20; α = .70), the economy (M = 4.88; SD = 1.46; α = .81) as well as individual risks, for instance, privacy-related concerns (M = 4.10; SD = 1.83; α = 0.73).

As pre-registered we also included variables for *overestimation of deepfakes*, *prior experience with deepfakes*, *prior exposure to deepfakes*, the *perceived ability to detect deepfakes*, *trust in the economy*, *gender*, *age*, and *educational attainment* (for a complete overview of measures, see Appendix B.1). As an analytical strategy, we follow Baek et al.'s (2019) recommendation and test the presumed effect on self and others as an interaction term in the regression model. This approach allows us to clearly identify a first-person effect, a second-person effect (H3: significant interaction term), a strict third-person effect (significant positive presumed effect on others and negative presumed effect on self), and the less strict presumed effects on others (H2: significant positive presumed effect on others; Gunther & Storey, 2003). Presumed effects on self and others were both mean-centered before estimating the model.

4. Results

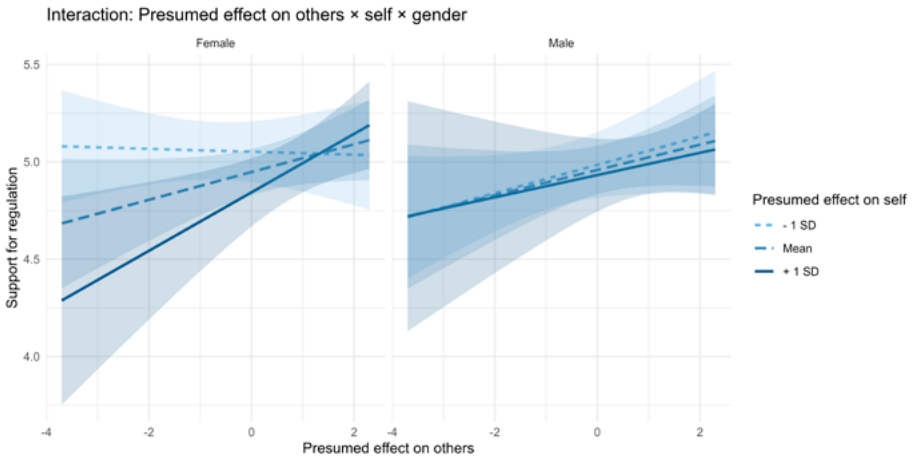
Our data support the perceptual hypothesis (H1), as people perceive deepfakes to have a stronger effect on others than on themselves. A paired-samples t-test ($t(1360) = -28.54$, $p < .001$; *Cohen's d* = 0.77) indicated a significant difference between the two variables, with the presumed effect on self ($M = 3.53$, $SD = 1.70$) being over one scale point lower than the presumed effect on others ($M = 4.70$, $SD = 1.45$). We also find support for H2 as a higher presumed effect on others is positively associated with stronger support for regulation of deepfakes ($b = 0.07$, 95% CI [0.01,0.13], $p = .035$, $\beta = 0.07$; see Figure 1 for all estimates and Appendix D.1.1 for the complete model). However, we do not find support for H3. While the interaction effect is positive, which would be an indicator for a second-person effect, the estimate is not significant ($b = 0.02$, 95% CI [-0.00,0.05], $p = .074$). We find support for H4 as higher trust in institutions is positively

Figure 1. All estimates from the regression model with 95%-CIs



Note. Estimates are shown with significance level: *$p < .05$, **$p < .01$, ***$p < .001$.

Figure 2. Interaction effect between presumed effect on others, presumed effect on self, and gender.



related to higher support for regulation (b = 0.09, 95% CI [0.03,0.16], p = .006, β = 0.07). Also H5 is mostly supported as people with higher risk perception for the media (b = 0.33, 95% CI [0.25,0.40], p < .001, β = 0.27), economy (b = 0.26, 95% CI [0.18,0.33], p < .001, β = 0.25), and self (b = 0.07, 95% CI [0.02,0.11], p = .004; β = 0.08) have stronger support for regulation. However, for politics, H5 could not be supported (b = -0.00, 95% CI [-0.06,0.06], p = .946).

4.1 Additional exploratory analysis with gender

In contrast to our analysis using weighted data, the model based on unweighted data indicates a second-person effect (see Appendix D.2.1). Therefore, we decided to conduct an additional exploratory analysis with a three-way interaction term involving gender, as the imbalance of gender in the sample appears to influence the outcome of the analysis. The reasoning behind this approach is that gender potentially plays a role with regard to a second-person effect in the context of deepfakes. Indeed, when adding gender as a three-way interaction term (see Appendix D.1.2 for the complete model), we identified a significant difference (b = -0.05, 95% CI [-0.10,-0.00], p = .043, β = 0.09). For male respondents, we find primarily a difference in presumed effects on others but no substantial difference in the presumed effect on self (see Figure 2). In contrast, for females, we observe a potential second-person effect in our data, as the presumed effect on self moderates the relationship of the presumed effect on others. Thus, females with a high presumed effect on others and themselves show the strongest support for regulation. However, the overall pattern remains less clear-cut, as female participants with low values on both variables also indicate relatively high support.

5. Discussion and conclusion

Our study is one of the first to examine the relationship between the perception of deepfake technology and support for its

regulation. The results support the existing literature on third-person perception (Corbu et al., 2020; Davison, 1983). When asked about the influence of deepfakes, the perceived effects on one's own opinions are significantly lower than those perceived on the opinions of others. The analysis also sheds light on the behavioral dimension of such third-person perceptions (Gunther & Storey, 2003). The perceived influence of the effects of deepfakes on others is positively related to the support for deepfake regulation. A similar relationship has been found between perceptions of disinformation and regulations of platforms and their content (Chen et al., 2023; Kim, 2025; Riedl et al., 2022). However, our main analysis was unable to identify a second-person effect. Given the small effect size and limited statistical significance in our study, future research should further examine the third-person perception and second-person effects in the context of deepfake regulation.

As we identified differences between the models using weighted and unweighted data, we also focused on gender as part of an exploratory analysis, which was not pre-registered. Our data indicate that a potential second-person effect may apply to female participants but not to male ones. For women, the perceived impact of deepfakes on their own opinions is positively associated with support for deepfake regulation. This might be linked to perceived threats related to deepfake pornography, which predominantly targets women (de Ruiter, 2021; Jungherr & Rauchfleisch, 2025; Rauchfleisch et al., 2025; Wang & Kim, 2022). Although we asked about the effects of deepfakes on opinions, such threats may resonate more strongly with women, leading to a greater inclination to support regulation. This argument is also supported by a significant difference ($t\text{-}Welch(970.57) = 3.28, p = .001$) between males ($M = 3.88$, $SD = 1.88$) and females ($M = 4.22$, $SD = 1.79$) in terms of risk perception for the self. For the other risk perception domains, we do not find such gender differences. For males, the presumed effect of deepfakes on others is positively related to support for regulation, whereas the perceived effect on oneself is not. This noteworthy difference between female and male participants warrants replication and further exploration in future studies, especially given the statistical uncertainty for the estimate of this interaction and the not fully consistent pattern (see Figure 2).

Our study reveals that trust in institutions is positively associated with support for regulating deepfake technology. This finding has practical implications: When trust in institutions is strong, people are more willing to delegate power and responsibility for deepfake regulation. Our measure of institutional trust included politics and journalism as key institutions. In the model following the pre-registration (see Figure 1), we also examined trust in the economy as a predictor, which did not yield any significant association with support for regulation. Further studies could compare the relationship between support for regulation and trust in different kinds of institutions.

The results further confirm that the perceived risks of a technology are positively associated with support for regulation (Gardner & Gould, 1989; Lima et al., 2005). This relationship holds across various application fields. However, contrary to expectations, perceived risks in the political domain do not correlate with support for regulation. This is noteworthy, as previous literature has emphasized the political risks associated with deepfake technology, including its impact on elections and votes (Godulla et al., 2021; Hameleers et al., 2022; Vaccari & Chadwick, 2020). A possible explanation is that the agency for

regulation is most likely seen as a political responsibility. As a result, while people may recognize the high risks associated with deepfakes in politics, they may not believe that these risks can be effectively addressed through state-led regulation.

Our study comes with some limitations. First, we use the case of Switzerland, which, due to its direct-democratic instruments (referendums), is a particularly suitable example of a country where public opinion might be relevant when it comes to regulations. However, the generalizability of the findings remains limited, although we cautiously suggest some degree of applicability to other Western European countries. Future studies could compare the link between perceptions of communication technology and support for its regulation in different countries. Furthermore, we also inquired about general aspects of regulation, specifically restrictions on the use of deepfake technology, and did not differentiate between state-led approaches and self-regulation, which studies have shown to be relevant for regulating social media platforms (Chung & Wihbey, 2024; Riedl et al., 2022). Therefore, further studies could investigate different approaches for regulating deepfake technology, considering state-led or self-regulation. Our collected data showed some imbalance regarding gender, which we could address through weighting. While this imbalance affected the result of the assumed second-person effect, other results, such as trust in institutions and risk perceptions, remained stable. Despite the limitations, our study sheds light on the relationship between individual perceptions of deepfake technology and support for its regulation, an issue that is increasingly raised in the public and addressed by politics.

Funding statement

Daniel Vogler's research and the data collection for the project were funded by the Swiss Foundation for Technology Assessment (TA-SWISS). Adrian Rauchfleisch's work was supported by the National Science and Technology Council, Taiwan (R.O.C.) (Grant No. 114-2628-H-002-007-) and by the Taiwan Social Resilience Research Center (Grant No.114L9003) from the Higher Education Sprout Project by the Ministry of Education in Taiwan. Gabriele de Seta's work was supported by a Trond Mohn Foundation Starting Grant (TMS2024STG03).

References

Ahmed, S. (2023). Examining public perception and cognitive biases in the presumed influence of deepfakes threat: Empirical evidence of third person perception from three studies. *Asian Journal of Communication*, *33*(3), 308–331. https://doi.org/10.1080/01292986.2023.2194886

Altay, S., & Acerbi, A. (2024). People believe misinformation is a threat because they assume others are gullible. *New Media & Society*, *26*(11), 6440–6461. https://doi.org/10.1177/14614448231153379

Baek, Y. M., Kang, H., & Kim, S. (2019). Fake news should be regulated because it influences both "others" and "me": How and why the influence of presumed influence model should be extended. *Mass Communication and Society*, *22*(3), 301–323. https://doi.org/10.1080/15205436.2018.1562076

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241253138

Bendahan Bitton, D. B., Hoffmann, C. P., & Godulla, A. (2024). Deepfakes in the

context of AI inequalities: Analysing disparities in knowledge and attitudes. *Information, Communication & Society*, 295–315. https://doi.org/10.1080/1369118X.2024.2420037

Chen, M., Yu, W., & Liu, K. (2023). A meta-analysis of third-person perception related to distorted information: Synthesizing the effect, antecedents, and consequences. *Information Processing & Management*, *60*(5). https://doi.org/10.1016/j.ipm.2023.103425

Chung, M., & Wihbey, J. (2024). Social media regulation, third-person effect, and public views: A comparative study of the United States, the United Kingdom, South Korea, and Mexico. *New Media & Society*, *26*(8), 4534–4553. https://doi.org/10.1177/14614448221122996

Corbu, N., Oprea, D.-A., Negrea-Busuioc, E., & Radu, L. (2020). 'They can't fool me, but they can fool the others!' Third person effect and fake news detection. *European Journal of Communication*, *35*(2), 165–180. https://doi.org/10.1177/0267323120903686

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, *47*(1), 1–15. https://doi.org/10.1086/268763

de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

Gardner, G. T., & Gould, L. C. (1989). Public perceptions of the risks and benefits of technology. *Risk Analysis*, *9*(2), 225–242. https://doi.org/10.1111/j.1539-6924.1989.tb01243.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

Gosse, C., & Burkell, J. (2020). Politics and porn: How news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497–511. https://doi.org/10.1080/15295036.2020.1832697

Gunther, A. C., & Storey, J. D. (2003). The influence of presumed influence. *Journal of Communication*, *53*(2), 199–215. https://doi.org/10.1111/j.1460-2466.2003.tb02586.x

Hameleers, M., Van Der Meer, T. G. L. A., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221116346

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior*, *80*, 295–302. https://doi.org/10.1016/j.chb.2017.11.034

Jungherr, A., & Rauchfleisch, A. (2024). Negative downstream effects of alarmist disinformation discourse: Evidence from the United States. *Political Behavior*, *46*(4), 2123–2143. https://doi.org/10.1007/s11109-024-09911-3

Jungherr, A., & Rauchfleisch, A. (in press). Public Opinion on the Politics of AI Alignment: Cross-National Evidence on Expectations for AI Moderation from Germany and the United States. *Social Media + Society*.

Kalogeropoulos, A., Toff, B., & Fletcher, R. (2022). The watchdog press in the doghouse: A comparative study of attitudes about accountability journalism, trust in news, and news avoidance. *The International Journal of Press/Politics*, *29*(2), 485–506. https://doi.org/10.1177/19401612221112572

Karaboga, M., Frei, N., Puppis, M., Vogler, D., Raemy, P., Ebbers, F., Runge, G., Rauchfleisch, A., de Seta, G., Gurr, G., Friedewald, M., & Rovelli, S. (2024). Deepfakes und manipulierte Realitäten: Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz [Deepfakes and manipulated realities: Technology impact assessment and policy recommendations for Switzerland]. vdf Hochschulverlag AG.

Kim, M. (2025). A direct and indirect effect of third-person perception of COVID-19 fake news on support for fake news regulations on social media: Investigating the role of negative emotions and political views. *Mass Communication and Society*, *28*(2), 229–252. https://doi.org/10.1080/15205436.2023.2227601

Lima, M. L., Barnett, J., & Vala, J. (2005). Risk perception and technological development at a societal level. *Risk Analysis*, *25*(5), 1229–1239. https://doi.org/10.1111/j.1539-6924.2005.00664.x

Liu, P. L., & Huang, L. V. (2020). Digital disinformation about COVID-19 and the third-person effect: Examining the channel differences and negative emotional outcomes. *Cyberpsychology, Behavior, and Social Networking*, *23*(11), 789–793. https://doi.org/10.1089/cyber.2020.0363

Marien, S., & Hooghe, M. (2011). Does political trust matter? An empirical investigation into the relation between political trust and support for law compliance: does political trust matter? *European Journal of Political Research*, *50*(2), 267–291. https://doi.org/10.1111/j.1475-6765.2010.01930.x

Nguyen, D. (2023). How news media frame data risks in their coverage of big data and AI. *Internet Policy Review*, *12*(2). https://policyreview.info/articles/analysis/how-news-media-frame-data-risks-big-data-and-ai

Paradise, A., & Sullivan, M. (2012). (In)visible threats? The third-person effect in perceptions of the influence of Facebook. Cyberpsychology, Behavior, and Social Networking, *15*(1), 55–60. https://doi.org/10.1089/cyber.2011.0054

PytlikZillig, L. M., Kimbrough, C. D., Shockley, E., Neal, T. M. S., Herian, M. N., Hamm, J. A., Bornstein, B. H., & Tomkins, A. J. (2017). A longitudinal and experimental study of the impact of knowledge on the bases of institutional trust. *PLOS ONE*, *12*(4). https://doi.org/10.1371/journal.pone.0175387

Rauchfleisch, A., Vogler, D., & de Seta, G. (2025). Deepfakes or synthetic media? The effect of euphemisms for labeling technology on risk and benefit perceptions. *Social Media + Society*. https://doi.org/10.1177/20563051251350975

Riedl, M. J., Whipple, K. N., & Wallace, R. (2022). Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others. *Information, Communication & Society*, *25*(11), 1632–1649. https://doi.org/10.1080/1369118X.2021.1874040

Six, F. (2013). Trust in regulatory relations: How new insights from trust research improve regulation theory. *Public Management Review*, *15*(2), 163–185. https://doi.org/10.1080/14719037.2012.727461

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Why study risk perception? *Risk Analysis*, *2*(2), 83–93. https://doi.org/10.1111/j.1539-6924.1982.tb01369.x

Swissinfo.ch (2025, May 9). Switzerland rejects deepfake regulation. Retrieved from https://www.swissinfo.ch/eng/ai-governance/switzerland-rejects-deepfake-regulation/89277391

Thouvenin, F.; Eisenegger, M.; Volz, S.; Vogler, D.; Jaffé, M., (2023). *Governance von Desinformation in digitalisierten Öffentlichkeiten. Bericht für das Bundesamt für Kommunikation (BAKOM)* [Governance of disinformation in digitalized publics. Report for the Federal Office of Communication]. Retrieved from: https://www.bakom.admin.ch/bakom/de/home/elektronische-medien/studien/einzelstudien.html

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

Verhoest, K., Redert, B., Maggetti, M., Levi-Faur, D., & Jordana, J. (2025). Trust and regulation. In F. Six, J. A. Hamm, D. Latusek, E. V. Zimmeren, & K. Verhoest

(Eds.), *Handbook on Trust in Public Governance* (pp. 360–380). Edward Elgar Publishing. https://doi.org/10.4337/9781802201406.00030

Wang, S., & Kim, S. (2022). Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, *134*. https://doi.org/10.1016/j.chb.2022.107305

Wolf, C. (2021). Public trust and biotech innovation: A theory of trustworthy regulation of (scary!) technology. *Social Philosophy and Policy*, *38*(2), 29–49. https://doi.org/10.1017/S0265052522000036

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence: The International Journal of Research into New Media Technologies*, *27*(1), 36–51. https://doi.org/10.1177/1354856520923963

Yu, E., Song, H., Jung, J., & Kim, Y. J. (2023). Perception and attitude toward the regulation of online video streaming (in South Korea). *Online Media and Global Communication*, *2*(4), 651–679. https://doi.org/10.1515/omgc-2023-0059

Appendix

A. Pre-registration

The pre-registration can be accessed on AsPredicted (https://aspredicted.org/s2gt-7rwr.pdf). In the main paper, we discuss in detail only hypotheses H1–H3, H8, and H10. We use a different numbering system in the main paper, labeling them H1–H5. In the appendix, we report the complete analysis with all hypotheses. H11 remained in the pre-registration by oversight, as it was part of an earlier draft and was not carried forward into our final analysis. Here is a list of all pre-registered hypotheses:

H1: Individuals will presume a greater deepfake effect on "others" than on the "self".

H2: Individuals' presumed deepfake effect on "others" is positively related to their support for the regulation of deepfakes.

H3: Individuals with both high presumed deepfake effects on "others" and "self" will show stronger support for the regulation of deepfakes.

H4: Individuals overestimating deepfakes will show stronger support for the regulation of deepfakes.

H5: Individuals with prior experience with deepfakes will show stronger support for the regulation of deepfakes.

H6: Individuals with prior exposure to deepfakes will show stronger support for the regulation of deepfakes.

H7: Individuals with higher perceived deepfake detection ability will show weaker support for the regulation of deepfakes.

H8: Individuals with higher trust in institutions will show stronger support for the regulation of deepfakes.

H9: Individuals with higher trust in the economy will show lower support for the regulation of deepfakes.

H10: Individuals with higher risk perception of deepfakes for a) politics, b) the media, c) the economy, and d) the "self" will show stronger support for the regulation of deepfakes.

H11: Presumed deepfake effect on others strengthens the positive effect of risk perception of deepfakes on support of regulation of deepfakes.

We also pre-registered an analysis with risk perception of deepfakes as outcome variable. This analysis is completely missing in the main paper due

to space constraints. These models are reported in Section D.3. Here are the pre-registered risk perception hypotheses:

H12: Individuals with both high presumed deepfake effects on "others" and "self" will show stronger risk perception of deepfakes.

H13: Individuals overestimating deepfakes will show stronger risk perception of deepfakes.

H14: Individuals with prior experience with deepfakes will show stronger risk perception of deepfakes.

H15: Individuals with prior exposure to deepfakes will show stronger risk perception of deepfakes.

H16: Individuals with higher perceived deepfake detection ability will show weaker risk perception of deepfakes.

H17: Individuals with higher trust in institutions will show weaker risk perception of deepfakes.

B. Measures

B.1. Complete descriptive tables with all variables and items

Table 1. First part of descriptive statistics for all relevant variables and items

Variable	Question/operationalization	M	(*SD*)	n
H1/H3 Presumed effect of deepfakes on self	Deepfakes influence my own opinion.	3.53	(1.70)	1361
H1–H3 Presumed effect of deepfakes on others	Deepfakes influence the opinion of the Swiss population in general.	4.70	(1.45)	1361
(H4) Overestimating deepfakes (3 items, α = 0.73)	(1 = "do not agree at all", 7 = "totally agree")	4.58	(1.20)	1361
	Deepfakes can be produced for little money.	4.92	(1.47)	1361
	You can create deepfakes yourself with little prior knowledge.	4.36	(1.59)	1361
	Deepfakes are widespread.	4.46	(1.40)	1361
(H5) Prior experience with deepfakes (sum index)	1.11 (0.56)	1361		
	I had already heard about deepfakes before this study	57.02%		776
	I have already seen deepfakes	49.16%		669
	I have already shared or disseminated deepfakes	2.28%		31
	I have already made deepfakes myself	2.65%		36

(H6) Prior exposure to deepfakes (3 items, α = 0.8)	How often do you encounter deepfakes on the following channels? (1 = "Never", 7 = "Often")	3.81	(1.54)	1361
	on social media	4.12	(1.81)	1361
	in messenger apps such as Whatsapp or Telegram	3.31	(1.86)	1361
	on video platforms such as YouTube or Vimeo	4.00	(1.81)	1361
	I am able to distinguish deepfakes from real media content (1 = "do not agree at all", 7 = "totally agree")	3.39	(1.60)	1361
(H7) Perceived deepfake detection ability	(1 = "No trust at all", 7 = "Fully trust")	3.71	(1.31)	1361
H4 (H8) Trust in institutions (2 items, α = 0.74, *Spearman-Brown* = 0.74)	politics	3.62	(1.51)	1361
	media	3.80	(1.43)	1361

Note. Hypothesis numbers in parentheses indicate the pre-registered hypothesis number of a variable.

Table 2. Second part of descriptive statistics for all relevant variables and items

Variable	Question/operationalization	M	(*SD*)	n
(H9) Trust in the economy	(1 = "No trust at all", 7 = "Fully trust")	4.18	(1.37)	1361
H5a (H10a) Risks for politics (2 items, α = 0.89, *Spearman-Brown* = 0.89)	(1 = "do not agree at all", 7 = "totally agree")	4.98	(1.64)	1361
	Deepfakes can be used to manipulate the results of elections in Switzerland.	5.02	(1.71)	1361
	Deepfakes can be used to manipulate the results of referendum votes in Switzerland.	4.95	(1.74)	1361
H5b (H10b) Risks for media (2 items, α = 0.70, *Spearman-Brown* = 0.71)	(1 = "do not agree at all", 7 = "totally agree")	5.81	(1.20)	1361
	Deepfakes can be used to create fake news.	5.48	(1.48)	1361
	Deepfakes can undermine trust in Swiss media.	6.15	(1.24)	1361
H5c (H10c) Risks for economy (2 items, α = 0.81, *Spearman-Brown* = 0.81)	(1 = "do not agree at all", 7 = "totally agree")	4.88	(1.46)	1361
	Deepfake technology developed abroad threatens the Swiss economy.	4.66	(1.61)	1361
	Deepfakes can undermine trust in the Swiss economy.	5.09	(1.57)	1361
H5d (H10d) Risks for the "self" (2 items, α = 0.73, *Spearman-Brown* = 0.73)	(1 = "do not agree at all", 7 = "totally agree")	4.10	(1.83)	1361
	Deepfakes are a problem for my privacy.	4.11	(2.01)	1361
	I'm afraid that someone will create deepfakes with videos of me.	4.08	(2.10)	1361

H2–H5 (H2–H10) Support for deepfake regulation (4 items, α = 0.77)	(1 = "do not agree at all", 7 = "totally agree")	5.10	(1.45)	1361
	Deepfakes should be banned.	5.28	(1.81)	1361
	I support legislation to ban deepfakes.	5.22	(1.78)	1361
	Deepfakes should be regulated by internet companies like Google and Facebook.	4.85	(2.06)	1361
	Deepfakes should be regulated by the government.	5.04	(1.87)	1361
University degree		27.63%		1361
Gender male		36.00%		1361
Region (French)		33.28%		1361
Age		43.24	(16.28)	1361

Note. Hypothesis numbers in parentheses indicate the pre-registered hypothesis number of a variable.

C. Power analysis

We ran power analyses for the smallest expected effects. For a paired t-test (H1–two-sided) with *Cohen's d* = 0.2, we have a power of 0.9 with $n = 265$ (calculated with the pwr package in R*)*. We have a power of 0.9 for the regression models with 14 predictors and an effect size of $f2 = 0.02$ with $n = 1148$ (calculated with the pwr package in *R*). For the interaction term (H3), we have a power of 0.93 with a sample size of 1,200, with an effect size of $f2 = 0.01$ (power simulation in *R*, $p < 0.05$, *sigma* = 1, *intercept* = 1, *b self* = -0.1, *b others* = 0.1, *b interaction* = -0.1, 1,000 runs).

D. Model results

D.1 Complete models reported in the main paper

This section shows the complete model reported in the main paper. We first compared the gender and age distribution of our sample with the population data of Switzerland at the end of 2023. Although some groups are overrepresented (see the table below), we could generally get observations for each individual group (age and gender). Thus, models with survey weights are used for our analysis. We calculated weights for each single age year between 16 and 79, interlocked with gender (male and female/other).

Table 3. Sample and population data matching the distribution of the Swiss population with our sample

Age group	Gender	Sample count	Population count	Pop. proportion (%)	**Sample proportion (%)**
16–24	Female	157	300,266	5.87%	11.50%
16–24	Male	37	312,723	6.12%	2.72%
25–34	Female	217	382,557	7.48%	15.90%
25–34	Male	72	386,382	7.56%	5.29%
35–44	Female	182	393,599	7.70%	13.40%

35–44	Male	96	385,713	7.54%	7.05%
45–54	Female	124	427,435	8.36%	9.11%
45–54	Male	94	407,681	7.97%	6.91%
55–64	Female	110	503,961	9.86%	8.08%
55–64	Male	87	476,144	9.31%	6.39%
65+	Female	81	606,353	11.90%	5.95%
65+	Male	104	529,525	10.40%	7.64%

D.1.1 Support for regulation weighted data

Table 4. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	0.84	0.32	1.35	0.001
H2 Presumed effect on others	0.07	0.01	0.13	0.035
H3 Presumed effect on self X Presumed effect on others	0.02	-0.00	0.05	0.074
H4 Trust in institutions	0.09	0.03	0.16	0.006
H5a Risk politics	-0.00	-0.06	0.06	0.946
H5b Risk media	0.33	0.25	0.40	<0.001
H5c Risk economy	0.26	0.18	0.33	<0.001
H5d Risk for "self"	0.07	0.02	0.11	0.004
Overestimation	0.02	-0.04	0.09	0.497
Experience	-0.09	-0.21	0.03	0.148
Exposure	0.01	-0.04	0.06	0.676
Perceived detection ability	0.05	0.01	0.10	0.027
Trust in the economy	-0.02	-0.08	0.04	0.461
Region (1 = French)	0.38	0.24	0.53	<0.001
Age	0.01	0.00	0.01	0.003
Gender (1 = male)	-0.06	-0.19	0.08	0.414
Education (1 = higher)	-0.06	-0.22	0.09	0.422
Presumed effect on self	-0.04	-0.09	0.01	0.119
Observations R2/R2 adjusted	1361 0.326/0.317			

Note. The outcome variable is support for regulation.

D.1.2 Support for regulation weighted data with gender interaction

Table 5. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	0.78	0.26	1.30	0.003
Presumed effect on self X others X Gender	-0.05	-0.10	-0.00	0.043
Presumed effect on self X others	0.05	0.01	0.08	0.008
Presumed effect on self	-0.06	-0.13	0.01	0.078
Presumed effect on others	0.07	-0.01	0.16	0.071
Gender (1 = male)	0.01	-0.14	0.17	0.860
Overestimation	0.03	-0.04	0.09	0.417
Experience	-0.09	-0.21	0.03	0.157
Exposure	0.01	-0.04	0.06	0.737
Perceived detection ability	0.05	0.01	0.10	0.023
Trust in institutions	0.09	0.03	0.16	0.006
Trust in the economy	-0.02	-0.08	0.04	0.477
Risk economy	0.25	0.18	0.33	<0.001
Risk for "self"	0.07	0.02	0.11	0.004
Risk politics	-0.00	-0.06	0.05	0.912
Risk media	0.33	0.25	0.41	<0.001
Region (1 = French)	0.39	0.24	0.53	<0.001
Age	0.01	0.00	0.01	0.003
Education (1 = higher)	-0.06	-0.22	0.09	0.422
Presumed effect on self X Gender (1 = male)	0.05	-0.05	0.15	0.363
Presumed effect on others X Gender (1=male)	-0.01	-0.12	0.10	0.867
Observations R2/R2 adjusted	1361 0.328/0.318			

Note. The outcome variable is support for regulation.

D.2 Model with unweighted data

In this section, we report the model with the unweighted data. The main difference in the model with the weighted data is the observed second-person effect that vanishes when the weighted data are used to represent the age and gender distribution of the Swiss population.

D.2.1 Support for regulation unweighted data

Table 6. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	1.06	0.52	1.60	<0.001
Presumed effect on self	-0.06	-0.11	-0.00	0.031
H2 Presumed effect on others	0.07	0.00	0.13	0.038
H3 Presumed effect on self X Presumed effect on others	0.03	0.00	0.05	0.040
Overestimation	-0.02	-0.08	0.05	0.610
Experience	-0.02	-0.14	0.11	0.775
Exposure	0.03	-0.02	0.08	0.275
Perceived detection ability	0.05	0.01	0.10	0.028
H4 Trust in institutions	0.07	0.01	0.14	0.025
Trust in the economy	-0.02	-0.08	0.04	0.493
H5a Risk politics	0.02	-0.04	0.08	0.542
H5b Risk media	0.29	0.21	0.37	<0.001
H5c Risk economy	0.27	0.20	0.35	<0.001
H5d Risk for "self"	0.06	0.01	0.10	0.010
Region (1 = French)	0.30	0.16	0.45	<0.001
Age	0.01	0.00	0.01	0.012
Gender (1 = male)	-0.03	-0.18	0.11	0.648
Education (1=higher)	-0.11	-0.26	0.04	0.161
Observations R2/R2 adjusted	1361 0.297/0.288			

Note. The outcome variable is support for regulation.

D.3 Additional analyses from pre-registration with risk perception as dependent variable

In this section, we report the models with risk perceptions as outcome variables. These analyses were also pre-registered but would go beyond the scope of the current paper. Thus, we report them in the appendix. We also use the weighted data for these models.

D.3.1 Risk for politics

Table 7. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	4.70	4.20	5.20	<0.001
Presumed effect on self	-0.02	-0.08	0.05	0.633
Presumed effect on others	0.43	0.36	0.50	<0.001
Overestimation	0.04	-0.03	0.12	0.269
Experience	0.05	-0.10	0.20	0.480
Exposure	0.05	-0.01	0.11	0.111
Perceived detection ability	0.00	-0.05	0.06	0.944
Trust in institutions	0.00	-0.06	0.06	0.979
Region (1 = French)	-0.41	-0.58	-0.24	<0.001
Age	-0.00	-0.01	0.00	0.296
Education (1 = higher)	-0.07	-0.26	0.11	0.439
Gender (1 = male)	0.11	-0.05	0.28	0.178
Presumed effect on self X Presumed effect on others	0.02	-0.01	0.05	0.143
Observations R2/R2 adjusted	1361 0.182/0.175			

Note. The outcome variable is the perceived risk of deepfakes for politics.

D.3.2 Risk for media

Table 8. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	4.84	4.47	5.21	<0.001
Presumed effect on self	-0.05	-0.10	-0.01	0.020
Presumed effect on others	0.31	0.25	0.36	<0.001
Overestimation	0.15	0.09	0.21	<0.001
Experience	0.12	0.00	0.23	0.043
Exposure	0.01	-0.04	0.05	0.689
Perceived detection ability	-0.04	-0.08	0.00	0.079
Trust in institutions	0.07	0.02	0.12	0.004
Region (1 = French)	-0.22	-0.35	-0.10	0.001
Age	0.00	-0.00	0.00	0.581
Education (1 = higher)	0.04	-0.10	0.18	0.552
Gender (1 = male)	-0.09	-0.21	0.03	0.154

Presumed effect on self X Presumed effect on others	0.01	-0.01	0.04	0.205
Observations R2/R2 adjusted	1361 0.185/0.178			

Note. The outcome variable is the perceived risk of deepfakes for the media.

D.3.3 Risk for the economy

Table 9. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	4.07	3.63	4.51	<0.001
Presumed effect on self	0.05	-0.01	0.10	0.078
Presumed effect on others	0.37	0.30	0.43	<0.001
Overestimation	0.05	-0.02	0.12	0.170
Experience	0.01	-0.13	0.14	0.927
Exposure	0.02	-0.03	0.07	0.486
Perceived detection ability	-0.00	-0.05	0.05	0.895
Trust in institutions	0.05	-0.01	0.11	0.081
Region (1 = French)	-0.05	-0.20	0.10	0.513
Age	0.01	0.00	0.01	<0.001
Education (1 = higher)	-0.17	-0.34	-0.01	0.039
Gender (1 = male)	-0.02	-0.17	0.12	0.758
Presumed effect on self X Presumed effect on others	0.02	-0.01	0.04	0.210
Observations R2/R2 adjusted	1361 0.189/0.181			

Note. The outcome variable is the perceived risk of deepfakes for the economy.

D.3.4 Risk for the 'self'

Table 10. Linear regression model with 95%-CIs shown as LL and UL

Predictors	Estimate	LL	UL	p
Intercept	4.35	3.79	4.92	<0.001
Presumed effect on self	0.21	0.14	0.28	<0.001
Presumed effect on others	0.16	0.08	0.24	<0.001
Overestimation	-0.00	-0.09	0.08	0.928
Experience	-0.07	-0.24	0.10	0.430
Exposure	0.08	0.01	0.15	0.023

Perceived detection ability	0.04	-0.02	0.10	0.237
Trust in institutions	0.00	-0.07	0.07	0.957
Region (1 = French)	-0.04	-0.23	0.16	0.703
Age	-0.01	-0.02	-0.00	<0.001
Education (1 = higher)	-0.40	-0.61	-0.19	<0.001
Gender (1 = male)	-0.23	-0.42	-0.04	0.016
Presumed effect on self X Presumed effect on others	0.03	-0.00	0.07	0.079
Observations	1361			
R2/R2 adjusted	0.130/0.122			

Note. **The outcome variable is the perceived risk for the 'self'.**

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

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.

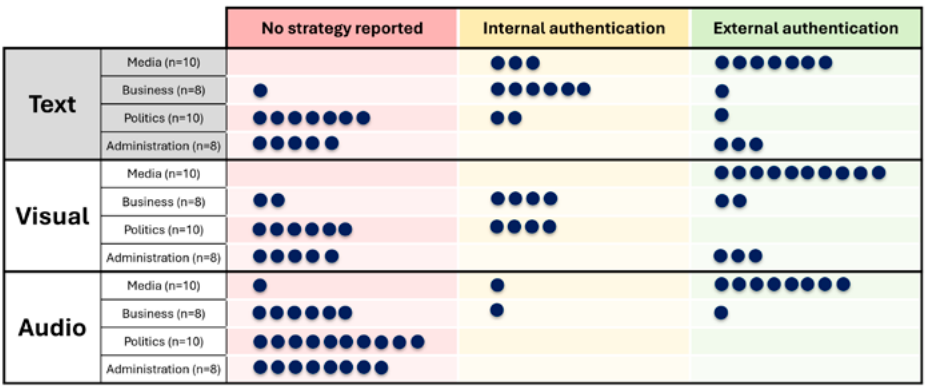
Based on the "audience's acts of authentication" framework by Tandoc et al., we classified the detection strategies that the respondents reported for the different types of synthetic disinformation (text, visual, audio) into one of the following three categories:

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

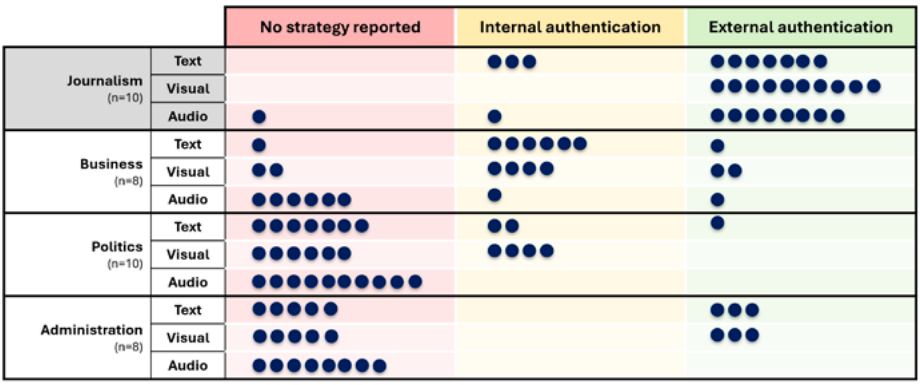
Wasner, B. (2013). *Eliten in Europa: Einführung in Theorien, Konzepte und Befunde* [Elites in Europe: Introduction to theories, concepts and findings]. Springer-Verlag.

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