Combating Information Manipulation in the Digital Sphere:

Amplifying Artificial Intelligence to Counter Disinformation

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von
M.A. Isabel Sophia Bezzaoui

Tag der mündlichen Prüfung: 04.12.2025

Referent: Prof. Dr. Christof Weinhardt
Korreferentin: Prof. Dr.Olga Slivko

Karlsruhe, 2025

Abstract

Disinformation is no longer a marginal concern in the digital age; it is a central force in reshaping public discourse, eroding institutional trust, and threatening democratic resilience. This dissertation examines how artificial intelligence, and specifically explainable AI (XAI), can be designed and deployed to counter the scale, speed, and sophistication of online information manipulation. While digital platforms have enabled unprecedented access to information, they also incentivize emotionally charged, divisive, and often deceptive content. In this environment, disinformation thrives not only because of technological affordances but also due to human cognitive vulnerabilities and platform-level incentives. Adopting a socio-technical perspective grounded in Information Systems (IS), this research addresses three key questions: (1) How can disinformation be systematically conceptualized and classified? (2) How can XAI be designed to enhance transparency and user trust in detection systems? (3) What practical tools and strategies can be implemented to reduce the spread and influence of manipulated content? The dissertation contributes theoretically by proposing a comprehensive taxonomy and annotation framework for disinformation and methodologically through the development and evaluation of XAI detection prototypes. It further explores the forensic requirements for identifying synthetic media such as deepfakes and outlines a model for critical digital literacy that integrates technical, cognitive, and social dimensions. By combining AI innovation with human-centered design and interdisciplinary insight, this work offers both conceptual clarity and practical tools to support democratic resilience. It positions IS research as a vital contributor to the development of transparent, trustworthy, and ethically grounded responses to one of the most urgent challenges of the digital era.

Acknowledgments

In a world spinning faster than any algorithm can track, and amidst the twists and turns of political, technological, and societal change, this dissertation grew from many minds and many moments, and I am grateful to all who shaped its path. First and foremost, my heartfelt thanks go to Prof. Christof Weinhardt, who consistently opened new perspectives and challenged me to think more deeply. His guidance was both an anchor and a compass. I am equally indebted to Dr. Jonas Fegert, whose knowledge and encouragement kept this project moving forward when I felt stuck at times. His mentorship has been a steady light throughout this journey, and his support has not only shaped this work but also influenced my development as a researcher.
I am also deeply grateful to the colleagues who surrounded me during these years, both at WIN and at the FZI. The mix of rigorous debate, spontaneous brainstorming sessions, fun lunch breaks, and a shared sense of purpose made our academic world feel vibrant and alive. Equally memorable were the moments of laughter between tasks and the insights that emerged while revising drafts, which kept me energized and inspired throughout this journey.

Beyond the academic sphere, my friends have been invaluable. Thank you for being the wide-open windows when the walls of academia felt too close, reminding me that the world is larger than any chapter or citation list. To my sisters in spirit, Carla, Paulina, and Nana, thank you for filling my heart, my days, and occasionally my inbox with everything from joy to nonsense. Life would be much emptier without you.
I am grateful to my parents, for the many forms of support that helped guide me to this point, and to my love and partner in life, Florian – thank you for being my steady rock over these years, for carrying me through countless long days, for making sure that life exists beyond drafts and deadlines, and for believing in me even when I sometimes struggled to believe in myself. Your support, patience, and care have made all the difference, and I could not have completed this journey without you.
My dog and the office's happiness manager, Juri, also deserves a special place here. Between walks, cuddles, and his endless talent for stealing my attention, he made the long hours at the desk lighter, the workdays brighter, and reminded me that even in the midst of deadlines, there's always time to pause and enjoy the little things.

Finally, I dedicate this dissertation to my grandmother, Gretel Schober. Although she never had the chance to pursue formal education, she possessed a depth of kindness, resilience, and humility that no course could teach. She faced life's hardships with remarkable strength and a tireless generosity of spirit. Her example has shaped my outlook on life and remains my greatest inspiration. This work is for her.

Table of Contents

Part I

Foundation

1 Introduction[1]

"*The internet was supposed to set us free.*"

This sentiment, echoed by early digital pioneers, reflects the once-utopian vision of cyberspace as a realm of boundless information and democratic empowerment (Carr, 2020; Rushkoff, 2016; Vaidhyanathan, 2018). Initially envisioned as a transformative forum for free expression and inclusive participation, this realm was heralded as a refuge for marginalized voices and a platform for deliberative discourse (Schäfer, 2015). However, the digital sphere has evolved into a more contested and ambivalent environment – one that has increasingly become a breeding ground for polarization, manipulation, and disinformation (Bennett & Pfetsch, 2018; Bezzaoui et al., 2023). Tim Berners-Lee, one of the inventors of the World Wide Web, famously warned that his creation was "being weaponized" against its original ideals (Solon, 2018). Instead of fostering open discourse, the digital sphere has increasingly become a battleground, where disinformation spreads faster than facts, and division is incentivized over deliberation. What was once a hopeful vision of democratized information now finds itself increasingly undermined by the darker forces of exploitation and manipulation.

Rather than functioning solely as a marketplace of ideas, much of the internet has become a marketplace for attention – where content that elicits strong emotional reactions is amplified, often at the expense of rational and reasoned dialogue (Nelson-Field et al., 2013; Weinhardt et al., 2024). Platforms are structurally incentivized to promote divisive or sensational material, as user engagement correlates directly with advertising revenue (Munn, 2020). As a result, emotional intensity increasingly supersedes factual accuracy in shaping online discourse. The very mechanics that were designed to connect people and foster inclusive conversation have instead been manipulated to prioritize sensationalism, playing into the hands of those seeking to capitalize on emotional and divisive content.

Far-right populist movements have been particularly adept at exploiting the affordances of digital platforms to self-organize and mobilize supporters. Events such as the assault

[1] This chapter comprises excerpts of two articles that were published by Isabel Bezzaoui, Jonas Fegert and Christof Weinhardt in the following outlet with the following title: Distinguishing Between Truth and Fake: Using Explainable AI to Understand and Combat Online Disinformation. In The 16th International Conference on Digital Society, 2022, and Isabel Bezzaoui, Nevena Nikolajevic and Jonas Fegert in the following outlet with the following title: Demokratiegefährdende Plattform-Mechanismen – Erkennen, Verstehen, Bekämpfen. In KI, Konflikte, Konventionen – PolKomm, 2023. Formatting and reference style were adapted and references were updated.

on the German Bundestag in 2020, the storming of the U.S. Capitol in 2021, and the attack on Brazil's seat of government in 2023 illustrate how digital technologies have amplified political extremism and disrupted democratic norms (Bezzaoui et al., 2023). These incidents have demonstrated the transformative yet perilous power of the internet in shaping political discourse and action.

At the same time, digitalization has also driven positive democratic change by weakening authoritarian information control and enabling grassroots activism (Jackson & Kreiss, 2023). Global interconnectedness and rapid information flows have made it harder for oppressive regimes to suppress dissent. However, the same infrastructure has also facilitated the proliferation of hate speech, conspiracy theories, and disinformation – phenomena that place significant strain on democratic societies (Aïmeur et al., 2023; Bennett & Pfetsch, 2018). The internet's dual role as both a force for empowerment and a breeding ground for malign influence presents a paradox that is at the core of current debates about digital platforms and their role in democracy.

Hence, as Habermas (2022) argues, the digital public sphere has not fulfilled its normative potential as an egalitarian platform for reasoned discourse. Instead, it often exacerbates fragmentation and undermines collective deliberation. The widespread dissemination of disinformation has become a corrosive force in public opinion formation, threatening the integrity of democratic processes (McQuail, 1993; Strömbäck, 2005). Disinformation no longer merely exists in the margins of the digital landscape; it has become a central and influential force in shaping the ways people perceive and engage with reality, eroding trust in institutions, the media, and even in one another (Frischlich & Humprecht, 2021).

This challenge was vividly demonstrated during the COVID-19 pandemic, which triggered a surge of health-related disinformation and misinformation. The crisis underscored the critical importance of distinguishing trustworthy from misleading information (Sharma et al., 2021; Shu et al., 2020a). Disinformation campaigns about the virus' origins, transmission, and treatment contributed to confusion and, in some cases, directly undermined public health efforts. The ongoing war in Ukraine illustrates how digital technologies are leveraged not only as tools of influence but also as potent weapons in a global hybrid warfare arena (Bachmann et al., 2023). In this context, the Russo-Ukrainian war has brought renewed attention to state-driven disinformation campaigns aimed at influencing elections, deepening societal divides, and even encouraging radicalized or terrorist behavior.

Online Social Networks (OSN) such as Facebook and messenger services like Telegram have played pivotal roles in this ecosystem of manipulated information. In theory, these platforms could adopt stronger preventive mechanisms, but their commercial incentives privileging emotionally engaging content often conflict with democratic safeguards (Bezzaoui et al., 2022a; Walker et al., 2019). As a result, these platforms are both enablers and amplifiers of the very forces that destabilize democratic discourse.

More recently, rapid developments in artificial intelligence (AI), generative AI specifically, have added a new layer of complexity. These tools allow for the effortless creation of text, images, videos, and audio that can mimic authentic media. Generative Adversarial Networks (GANs), in particular, enable the production of highly realistic deepfakes, further complicating efforts to verify content authenticity (Akhtar, 2023; Hussain et al., 2021). As the sophistication of these technologies increases, so too does their potential to harm. Their ability to create hyper-realistic but entirely fabricated content makes it increasingly difficult to discern truth from fabrication.

Governments and supranational organizations have begun responding through regulatory frameworks such as the European Union's AI Act and the General Data Protection Regulation (GDPR), both of which carry significant implications for global tech platforms (Schmitt et al., 2024). Nevertheless, policy responses are uneven, and the risk remains that authoritarian regimes may exploit such regulations to curtail civil liberties (World Economic Forum, 2024). The *Global Risks Report 2024*, based on a survey of nearly 1500 experts worldwide, identifies disinformation as one of the most pressing global threats in the following ten years. The report warns that the strategic use of manipulated information by state and non-state actors will likely deepen political polarization, erode institutional trust, and fuel social unrest (French et al., 2024; Qureshi et al., 2021; World Economic Forum, 2024). For democratic governments, the challenge lies in crafting policies that address the very real threats posed by disinformation without impeding the free flow of information or enabling state control over digital platforms.

While media literacy campaigns, journalistic fact-checking, and content moderation remain vital components in the fight against disinformation, these measures alone are often insufficient to counter the scale, velocity, and technological sophistication of contemporary information manipulation (Bradshaw & Howard, 2019; Guess et al., 2020; Pennycook & Rand, 2021). Their effectiveness, although well-documented in targeted contexts, is frequently constrained by resource limitations and the reactive nature of such interventions (Matasick et al., 2020; Wardle & Derakhshan, 2017). The accelerative pace of generative AI and algorithmically amplified content further complicates these efforts, demanding more adaptive and anticipatory approaches (Buchanan et al., 2021; Goldstein et al., 2023).

In this regard, the field of Information Systems (IS) offers a uniquely valuable perspective, as it inherently bridges technological innovation with organizational and societal processes. IS researchers, particularly when engaged in interdisciplinary collaboration, are well-equipped to design and implement socio-technical systems that foster democratic resilience and informed public discourse (Weinhardt et al., 2024). As the digital environment becomes increasingly complex and globally interconnected, the scope of IS has evolved beyond traditional business applications to encompass broader societal concerns,

positioning the discipline as a key contributor to the development of an "Information Society" (Dolensky et al., 2015; Vom Brocke et al., 2015). Moreover, IS research emphasizes the integration of data-driven technologies with context-sensitive understanding (Pfeiffer et al., 2024) – making it ideally suited to harness AI-powered tools in ways that are not only technically robust but also ethically grounded and responsive to social dynamics. If responsibly designed, such technologies can empower users to critically engage with digital content, recognize disinformation, and participate more meaningfully in democratic processes (Mahyoob et al., 2020; Yu & Lo, 2020). However, realizing this potential requires a deliberate synthesis of computational capabilities with human-centered design principles and long-term societal foresight – an area where IS research, with its methodological pluralism and systemic orientation (Stieglitz et al., 2018), can make a profound and lasting contribution.

Considering these complex realities, this dissertation addresses critical questions at the intersection of IS research and democratic resilience. It aims to contribute in three key ways: (1) by deepening our conceptual understanding of digital disinformation and refining methods for its detection by both humans and machines; (2) by developing and evaluating Explainable AI (XAI) tools to improve the transparency and user trust of detection systems; and (3) by proposing practical strategies to mitigate the spread and impact of manipulated content. This work bridges theoretical insights with real-world applications, offering actionable contributions for researchers, platform designers, and policymakers working to safeguard democratic discourse in the digital age.

To achieve these objectives, the dissertation is guided by the following main research questions:

- *RQ1: How can online disinformation be characterized and differentiated based on conceptually grounded characteristics?*
- *RQ2: How does an XAI component for disinformation detection have to be designed to help users trust the algorithm's assessment?*
- *RQ3: How can the key challenges in detecting information manipulation be effectively addressed through practical tools and strategies?*

This dissertation is structured into five key sections (Figure 1). These sections collectively introduce the topic of digital disinformation, outline the methodological approaches employed in the studies, present the studies themselves, and discuss their findings. The final section provides a critical assessment of strategies for combating information manipulation in the digital sphere, with a particular emphasis on the role of artificial intelligence in enhancing these efforts.

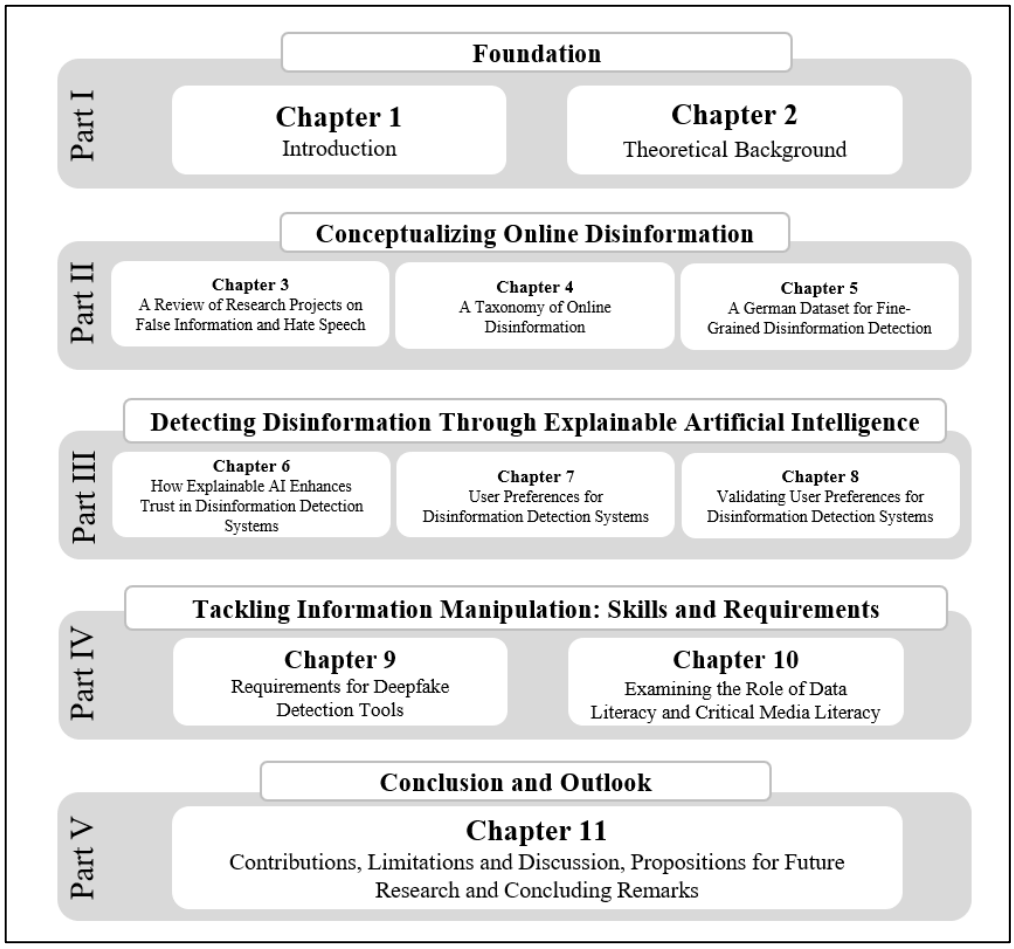

Figure 1. Structure of this dissertation.

Before effective countermeasures against disinformation can be developed, it is essential to establish a comprehensive understanding of its origins, mechanisms, and impact. **Part I** lays this foundation by exploring the conceptual basis of digital disinformation, tracing its historical evolution, and identifying the key actors and the role of platforms shaping the contemporary information landscape. Chapter 1 defines the broader scope of this dissertation, emphasizing the significance of disinformation as a critical socio-political issue and positioning it within the wider discourse on information disorder. Chapter 2 expands on this by constructing a theoretical framework, beginning with a critical analysis of disinformation as a concept, followed by an examination of the digital platforms and algorithmic systems that enable its spread. This chapter also explores the psycho-cognitive factors that make individuals susceptible to manipulated narratives and discusses the consequences of disinformation at both individual and societal levels. The final section introduces intervention strategies, focusing on the role of AI in combating disinformation, as well as the emerging potential of XAI in improving detection systems.

Part II (*RQ1*) is dedicated to conceptualizing digital disinformation as a foundation for developing effective detection and mitigation strategies. Chapter 3 presents a comprehensive review of existing research on false information and hate speech, offering insights into the current state of academic and practical efforts in this field while identifying key gaps and challenges. Building on this, Chapter 4 introduces a taxonomy of online disinformation, providing a structured classification of different disinformation types and key indicators for detection. This taxonomy serves as a resource for both human analysts and AI-driven detection systems. To demonstrate its practical application, Chapter 5 illustrates how the taxonomy informs the development of a structured labeling scheme for disinformation datasets. By establishing a rigorous annotation framework, this scheme supports the creation of high-quality training data for AI-based detection models, ultimately contributing to the advancement of more reliable and transparent disinformation detection systems.

The challenge of disinformation detection requires both accuracy and transparency in AI-driven systems. **Part III** *(RQ2)* examines how XAI can enhance the interpretability of these detection mechanisms. Chapter 6 provides a literature review on XAI applications across various domains, analyzing different explainability techniques and their potential adaptation for disinformation detection. Building on these insights, Chapter 7 presents a qualitative user study that explores individual preferences for XAI features and assesses their impact on user trust and comprehension. These findings guide the iterative development of refined prototypes, which are then tested in Chapter 8 through a large-scale online study. This final empirical investigation evaluates the effects of different levels of explainability on user perception, providing critical insights into the balance between transparency and effectiveness in AI-powered disinformation detection.

Addressing the complexities of information disorder requires a holistic approach that considers both technological advancements and human factors. **Part IV** (*RQ3*) broadens the scope by examining the skills and conditions necessary for effectively countering disinformation. Chapter 9 presents a requirement analysis based on qualitative interviews with experts and practitioners, identifying key technical and operational prerequisites for the development and deployment of a forensic tool for deepfake detection. This analysis not only outlines technological specifications but also explores contextual factors that influence real-world implementation. Shifting the focus to individual competencies, Chapter 10 examines the skills required to critically engage with potentially misleading information, emphasizing the importance of data literacy and critical media literacy. This chapter introduces an integrated literacy model that combines these dimensions, providing a framework to guide future educational initiatives and interventions to strengthen resilience against disinformation.

Finally, **Part V** synthesizes the findings from the preceding chapters, offering a comprehensive discussion of their scholarly and practical implications in Chapter 11. By situating the results within a broader academic and applied context, this section highlights their significance while acknowledging the inherent limitations of the research. A critical reflection on these constraints delineates the boundaries of the study and identifies directions for future research. As this dissertation concludes, it also lays the groundwork for further inquiry within the field of IS and beyond. Future research may expand on these findings by advancing AI-based detection models and explainability techniques within IS and computer science, examining user interaction and trust in algorithmic systems through the lenses of human-computer interaction and behavioral sciences, investigating platform dynamics and content governance in media and communication studies, and developing interventions for digital and critical media literacy in educational research. By bridging these domains, subsequent studies can contribute to a more integrated and actionable understanding of digital disinformation.

This monographic dissertation comprises both published and unpublished materials. To maintain transparency and academic rigor, published papers are explicitly labeled as such.

2 Theoretical Background

The rise of digital platforms, algorithmically curated content, and ubiquitous social media has fundamentally transformed how information is produced, disseminated, and consumed. While these developments have enhanced connectivity and broadened access to information, they have also facilitated the rapid spread of disinformation. This phenomenon is shaped by the complex interplay of technological infrastructures, psychological predispositions, and sociopolitical incentives (Hameleers, 2023). As digital disinformation undermines public trust, distorts democratic discourse, and destabilizes institutions, it presents a pressing challenge for researchers and practitioners alike. A nuanced theoretical framework that accounts for the actors, mechanisms, and consequences of disinformation is critical for understanding this threat and developing effective countermeasures (Berger et al., 2024).

Within the IS discipline and beyond, early perspectives on digital technologies were often marked by techno-optimism, emphasizing the potential of information and communication technology (ICT) to strengthen democratic participation (Hacker & van Dijk, 2000; Päivärinta & Sæbø, 2006; Phang & Kankanhalli, 2008). Hacker and van Dijk (2000) articulated a vision in which computer-mediated communication (CMC) could enhance information access, reduce participation barriers, and foster decentralized political communities. ICT was seen as a means to bypass traditional gatekeepers, promote inclusive agenda-setting, and enable more responsive, horizontally structured political systems. Building on this vision, scholars such as Phang and Kankanhalli (2008) developed IS-centric frameworks for e-participation, in which governments strategically used ICT to disseminate policy information, solicit citizen input, and structure public deliberation. These frameworks highlighted the role of ICT in enhancing transparency, civic dialogue, and democratic legitimacy (Stieglitz & Dang-Xuan, 2013).

However, as digital infrastructures evolved – especially with the rise of OSN, algorithmic content delivery, and platform economies – critical reassessments emerged within the IS field (Lindner & Aichholzer, 2020). Scholars began to challenge earlier techno-determinist assumptions, arguing that democratic outcomes are not embedded in technology itself but are contingent on broader sociotechnical configurations, governance models, and platform design (Avgerou, 2010; Sarker et al., 2013). Rather than merely facilitating participation, digital platforms have also enabled the manipulation of information flows, the amplification of falsehoods, and the erosion of institutional credibility – core features of modern disinformation campaigns. More recent IS research has shifted its focus from

digital inclusion to digital manipulation, emphasizing how the affordances of digital infrastructures can be exploited to subvert democratic processes (Gkeredakis et al., 2021; Majchrzak & Markus, 2012). This evolving discourse underscores a growing recognition that information systems are not neutral tools but are deeply embedded in sociopolitical dynamics that shape participation, power, and the construction of truth (Pfeiffer et al., 2024; Weinhardt et al., 2024).

This chapter aims to provide a structured foundation for the analysis of digital disinformation and its mitigation, particularly in relation to emerging AI-driven interventions. The first section explores the conceptual and structural dimensions of disinformation in digital environments. It defines disinformation, distinguishes it from related phenomena such as misinformation and malinformation, and analyzes the roles and strategies of state and non-state actors engaged in information manipulation. It further examines the design of digital platforms and the cognitive mechanisms that render individuals susceptible to false narratives, concluding with an assessment of disinformation's societal and institutional impact.

The second section focuses on combating disinformation in the age of AI. It surveys current intervention strategies, including regulatory frameworks and media literacy initiatives, and explores the growing use of AI tools for disinformation detection and mitigation. Special attention is given to the design of explainable AI systems that prioritize transparency, accountability, and interpretability.

By addressing both the theoretical underpinnings of disinformation and the technological responses to it, this chapter situates disinformation as a critical problem space for IS research. It emphasizes the need for interdisciplinary, sociotechnically aware approaches to understanding and mitigating the far-reaching consequences of digital disinformation.

2.1 Disinformation in the Digital Sphere

In the context of the digital age, the widespread dissemination of disinformation represents a significant and complex challenge with profound implications for societal functioning and individual well-being. A comprehensive understanding of the mechanisms underlying the creation, propagation, and consumption of disinformation is essential for developing effective responses to this phenomenon. This section presents a theoretical framework for investigating key aspects of disinformation in the digital sphere. It begins by defining disinformation and examining the role of malicious actors who intentionally propagate falsehoods. The analysis then turns to the influence of digital platform design on the fragmentation of public discourse, a key factor in enabling the proliferation of

misleading information. Central to this discussion is an exploration of individual susceptibility to disinformation, focusing on the cognitive biases, deficiencies in critical thinking, and gaps in media literacy that render individuals more vulnerable to manipulation. Finally, the section evaluates the broader societal consequences of disinformation, including its effects on democratic processes, social trust, and public opinion. Through this analysis, the section seeks to provide a thorough theoretical understanding of disinformation and its pervasive impact in the digital era.

2.1.1 The Concept of Disinformation

2.1.1.1 Disinformation's Definitional Landscape

A precise conceptualization of disinformation is crucial for both its identification and mitigation. Without a clear definition, efforts to detect and counter disinformation risk being imprecise or ineffective, limiting our ability to address its broader consequences for society and democratic processes. By establishing clear indicators, researchers can develop more effective methodologies for combating its spread (Fallis, 2015). Wardle and Derakhshan (2017) offer a foundational typology that differentiates disinformation from related forms of misleading content, placing them in the so-called *information disorder* (Figure 2). Their framework categorizes misleading information into three distinct types: disinformation, misinformation, and malinformation. While disinformation and misinformation are frequently conflated, the crucial difference lies in intent (Colomina et al., 2021). *Disinformation* consists of verifiable false information that is deliberately fabricated to cause harm to individuals, social groups, organizations, or nations (European Commission, 2018). In contrast, *misinformation* refers to false or misleading information disseminated without the intent to deceive or inflict harm; it often arises from errors, misinterpretations, or unverified claims (Fetzer, 2004; Wu et al., 2019). *Malinformation*, on the other hand, involves factual information that is weaponized to cause harm (Wardle & Derakhshan, 2017).

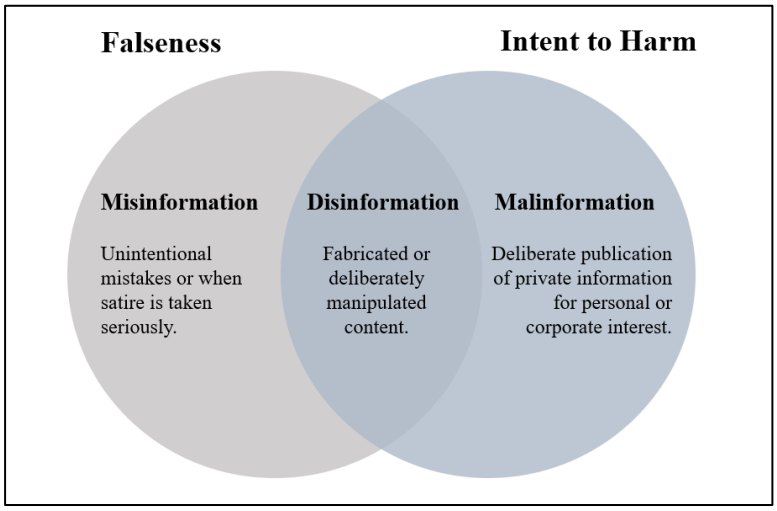

Figure 2. Types of information disorder (Wardle & Derakhshan, 2017).

Within this broader discussion of deceptive communication, *Information Manipulation Theory* (IMT), developed by Steven McCornack, offers a framework for understanding how deception extends beyond overt falsehoods (McCornack, 1992). IMT posits that deception often operates through the subtle manipulation of four dimensions of information: quantity (selective omission of details), quality (provision of false or misleading information), relevance (inclusion of extraneous or misleadingly framed information), and manner (deliberate ambiguity or obfuscation). A central insight of IMT is that even entirely truthful statements can be deceptive when strategically structured to mislead an audience, rendering deception detection particularly challenging (McCornack, 1992). Despite its influence, IMT has been critiqued for its limited empirical testability (Jacobs et al., 1996). In response, McCornack introduced IMT2 in 2014, incorporating insights from cognitive neuroscience and artificial intelligence to refine its explanatory capacity. This revised model remains relevant in analyzing contemporary forms of strategic deception, particularly in the context of digital communication and the algorithmic amplification of misleading content (McCornack, 2015).

A commonly referenced yet imprecise term related to disinformation is *fake news*. Defined by Allcott and Gentzkow (2017) as "news articles that are intentionally and verifiably false and could mislead readers," fake news represents a subset of disinformation. However, the term has been widely criticized for failing to capture the complexity of the phenomenon (Wardle & Derakhshan, 2017). Furthermore, political actors have appropriated the term as a rhetorical strategy to delegitimize media coverage that contradicts their interests. This strategic misuse of *fake news* has enabled the erosion of public trust in journalism and the media more broadly (Marwick & Lewis, 2017). Additionally, unreli-

able sources and politically motivated figures have weaponized the term to dismiss legitimate reporting and undermine fact-based narratives (Haigh et al., 2017). Therefore, the term will not be used in this work.

Disinformation must also be distinguished from *rumors* and *conspiracy theories*, as these forms of communication do not necessarily depend on the veracity of their claims. Rumors derive their influence from social transmission rather than their factual basis (Berinsky, 2017), while conspiracy theories rest on the belief that a hidden and powerful group secretly manipulates societal events and politics (Sunstein & Vermeule, 2009). This distinction is significant because detecting disinformation requires focusing on the intent to mislead rather than solely identifying factual inaccuracies. Indicators of accidental misinformation differ markedly from those of deliberately misleading content, necessitating distinct analytical approaches (Fallis, 2015).

Another critical aspect of disinformation is the transparency of its source. Some disinformation campaigns are overtly acknowledged by their creators (overt disinformation), whereas others obscure their origins to enhance their credibility and impact (covert disinformation) (Fetzer, 2004). Generally, the ambiguous provenance of much disinformation allows it to circulate through traditional media outlets, often gaining legitimacy through repetition – a process described as the *amplifier effect* (Bennet & Livingston, 2018). Furthermore, disinformation frequently exploits and reinforces existing ideological biases. The repetition of dominant narratives and stereotypes, which are often referred to as *deep stories* or *deep frames*, perpetuates, among others, racist, misogynistic, xenophobic, and queerphobic discourses (Phillips & Milner, 2021).

A broader epistemological perspective positions disinformation within the politics of knowledge production. Historically, disinformation functioned as a tool for legitimizing racial hierarchies and maintaining structural inequalities. In the digital era, social media platforms facilitate the resurgence of racial, antisemitic, and colonial tropes, frequently disseminated through memes, hashtags, and algorithmic amplification (Flores-Yeffal et al., 2019; Tuters & Hagen, 2020). The real-world consequences of disinformation are substantial. Events such as racially motivated mass shootings exemplify the dangers posed by strategically crafted falsehoods (Fausset & Bogel-Burroughs, 2021; Kreiss, 2021). Similarly, in India, anti-Muslim disinformation spread via Facebook contributed to violence against Rohingya communities, mirroring the patterns of disinformation that preceded the Rohingya genocide in Myanmar (Equality Labs, 2019). While misinformation can be found across the political spectrum, research indicates that disinformation

is particularly prevalent within authoritarian right-wing movements, where liberal democratic principles are perceived as threats to nationalist and traditionalist ideologies (Bennet & Livingston, 2018).

2.1.1.2 Historical Continuities and Transformations in Disinformation

The phenomenon of disinformation has been central to human communication since antiquity. A prominent example is found in the Roman rivalry between Antony and Octavian (Posetti & Matthews, 2018). Octavian launched a targeted propaganda campaign, using concise slogans on coins to tarnish Antony's image as a debauched womanizer under Cleopatra's influence. This campaign contributed to Octavian's rise as Augustus, demonstrating how the manipulation of public perception through disinformation could destabilize political systems and facilitate the consolidation of power. The invention of the Gutenberg printing press in 1493 significantly accelerated the spread of disinformation, culminating in the first large-scale news hoax – the *Great Moon Hoax* of 1835 (Thornton, 2000). The *New York Sun* published a series of six articles claiming the discovery of life on the moon, complete with illustrations of humanoid bat creatures and bearded blue unicorns (see Figure 3). Throughout history, conflicts, regime changes, and catastrophes have often served as key moments for the spread of disinformation.

Figure 3. An illustration published in the New York Sun in 1835 (Wills, 2017).

While the practice of using falsified or manipulated information to shape public opinion and political alignment shows to have a long history (Shu et al., 2020a), the term disinformation itself is a relatively recent development, gaining widespread recognition only

in the 1950s (Manning & Romerstein, 2004). Whereas access to information has drastically expanded in the digital era, the internet also allows individuals without professional media or journalistic expertise to distribute content, thereby democratizing information flow. However, this increased accessibility presents both opportunities and risks, as the quality and accuracy of information can decline, particularly regarding truthfulness and authenticity. Concerns about the lack of veracity in news have been voiced at least since the 19th century (Appel, 2020).

Historical instances of disinformation from the 20th and 21st centuries include deceptive advertising, government propaganda, doctored photographs, and fake documents. One of the most prototypical examples is Operation Bodyguard, a World War II campaign designed to mislead the Germans about the location of the D-Day invasion. The Allies used various tactics, including fake radio transmissions and fraudulent military reports, to convince the Germans that the invasion would occur in Calais instead of Normandy (Fallis, 2015). Another notable example is Operation INFEKTION, an influential disinformation campaign in the 1980s, which falsely claimed that the United States developed HIV/AIDS as a biological weapon (Boghardt, 2009). Similarly, during the Second Iraq War, the Bush Administration propagated false claims regarding weapons of mass destruction, fabricated heroic stories about U.S. soldiers, and suppressed critical media coverage of war casualties and anti-war movements (Kumar, 2006; Snow & Taylor, 2006). In recent years, however, disinformation seems to have become significantly more prevalent (Fallis, 2015).

Public spheres in many countries are becoming increasingly fragmented and disrupted as key democratic principles – such as trust in authoritative information from social and political institutions – are challenged (Bennet & Livingston, 2018). At the heart of this issue is the erosion of public trust in democratic institutions, particularly in the press and political systems, which has been compounded by the hollowing of political parties and diminished electoral representation. In contrast to the mid-20th century, when there was greater institutional trust and public authorities had more control over information, the current era is marked by a decline in trust and a diversification of media sources (Zimmermann & Kohring, 2018). Figure 4 illustrates the long-term decline in public trust in the U.S. federal government since the 1960s, with fluctuations corresponding to major events, economic conditions, and shifts in party control of the White House. In recent years, trust levels have remained persistently low (Bell, 2024).

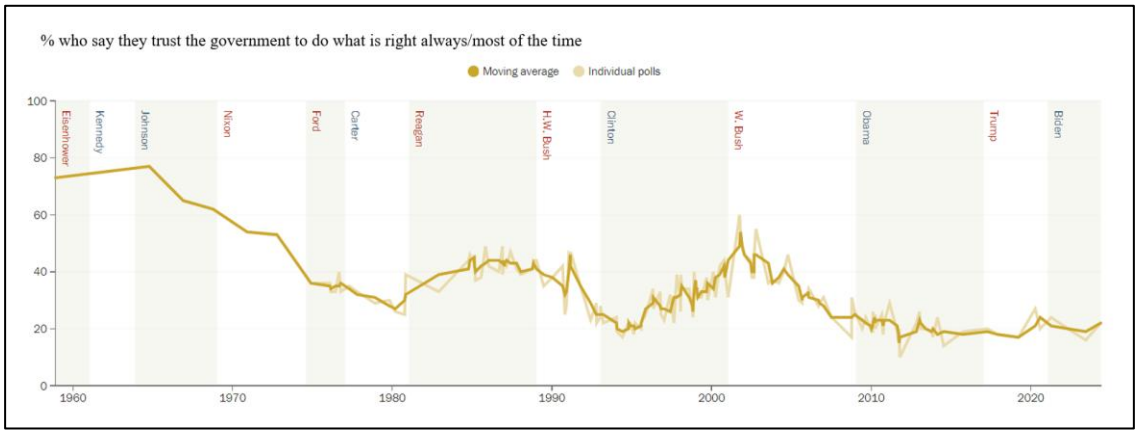

Figure 4. Decline of trust in the U.S. government since 1958 (Bell, 2024).

Technological and economic shifts have played a central role in this transformation, with the decline of local news outlets being a key development. The rise of online news consumption has contributed to the weakening of traditional business models, leaving many digital publishers reliant on advertising revenue rather than subscriptions (Marwick & Lewis, 2017). As social media platforms increasingly surpass traditional television as the primary news source for younger audiences (Aïmeur et al., 2023), the media landscape has become more fragmented. Unlike the era dominated by mass media, today's media environment is characterized by a kaleidoscopic mix of television networks, online newspapers, social media like X (formerly Twitter) and Facebook, as well as alternative sites such as WikiLeaks, radical websites, and disinformation campaigns employing journalistic formats. This proliferation of sources – including bots, troll factories, and anonymous discussion threads on platforms like 4chan – has created a new set of challenges for information integrity and trust (Bennet & Livingston, 2018).

As the diversification of media sources accelerates and traditional gatekeepers lose influence, disinformation benefits from a media environment that prioritizes engagement over accuracy. With a significant share of public discourse now occurring online, the power to shape narratives increasingly rests in the hands of private actors who own and regulate social media platforms (Berger et al., 2024). At the core of this transformation is the attention economy, in which content is primarily valued based on its capacity to generate clicks, shares, and engagement rather than its credibility (Zhang et al., 2021). The vast and unfiltered flow of information online has turned attention into a scarce and highly sought-after resource, incentivizing the proliferation of sensationalist and emotionally charged content, often at the expense of factual accuracy (Marwick & Lewis, 2017).

In recent years, disinformation has increasingly been disseminated through fabricated news outlets that imitate credible and established media sources. This trend contributes to a growing sense of uncertainty regarding the reliability of information, leading many

individuals to disengage from news consumption altogether (Berger et al., 2024). As trust in traditional media declines, online platforms have assumed a central role in shaping public discourse, often filling the informational void left by legacy media institutions (Kuo & Marwick, 2021). In other words, digital platforms have become indispensable to global communication and democratic engagement (Berger et al., 2024).

In addition, the professionalization and commercialization of disinformation highlight its strategic role as a tool for political and ideological manipulation. The increasing prevalence of disinformation as a paid service underscores its integration into the broader media and communications landscape (Rodríguez-Fernández, 2019). Elections, in particular, serve as flashpoints for disinformation campaigns, often prompting the implementation of countermeasures designed to safeguard democratic processes and institutions (Hoxtell, 2023). The disinformation industry has evolved into a highly organized sector, with specialized agencies and consultants designing and executing large-scale campaigns. Many of these agencies rely on economically vulnerable workers, a factor that exacerbates recruitment efforts during periods of economic downturn (Berger et al., 2024).

In the past years, AI has emerged as a transformative force in the production and dissemination of disinformation, significantly intensifying both the scale and sophistication of such campaigns. By enabling the automated generation of text, imagery, and video that closely mimics authentic human communication, AI reduces the barriers to producing persuasive and deceptive content (Zhao et al., 2025). This technological capacity facilitates the widespread diffusion of false information, particularly across social media platforms, as seen during the COVID-19 pandemic, where AI-driven content contributed to the erosion of public trust in health communication (Germani et al., 2024; Menz et al., 2024). A particularly salient manifestation of this threat is the advent of deepfakes – synthetic media created using AI techniques that produce highly realistic yet fabricated audio-visual material (Masood et al., 2023). These tools not only distort public perception but also complicate efforts to verify information, posing significant challenges to democratic institutions and media integrity (Godulla et al., 2021; Vaccari & Chadwick, 2020). As AI technologies become increasingly advanced and accessible, they facilitate the rapid and large-scale dissemination of disinformation, heightening already existing concerns regarding political manipulation, societal polarization, and the destabilization of public discourse (Spitale et al., 2023).

Building on these concerns, the widespread use of AI-generated content – particularly deepfakes – further undermines trust in traditional media sources by accelerating the flow of disinformation and complicating the task of verifying authenticity (Keller et al., 2024). Therefore, addressing the threat that disinformation poses to information integrity requires the development of robust identification techniques and policies aimed at mitigating its spread. However, such measures can only be effective if they are informed by a

deeper understanding of the nature and scope of disinformation itself (Fallis, 2015). Moreover, while technological advancements have reshaped the ways in which disinformation spreads, they also offer promising solutions for mitigating its impact (Schreiber et al., 2021). While no single approach can fully address the problem, technical interventions – such as algorithmic detection, content authentication, and fact-checking automation – play a crucial role in reducing the reach and influence of false information. To maximize their effectiveness, these solutions should be integrated into broader interdisciplinary strategies that account for social, political, and economic dimensions (Washington & Kuo, 2020).

2.1.2 Malign Actors in the Disinformation Ecosystem

Disinformation campaigns involve a diverse array of actors, each contributing to the production, dissemination, and amplification of false narratives. These actors can be categorized into distinct typologies based on their motivations, strategies, and affiliations (Zhang et al., 2021). Due to the deceptive nature of disinformation, direct inquiry into its producers is inherently challenging, as those engaged in such activities often seek to conceal their identities (Guess & Lyons, 2020). This opacity complicates efforts to trace the origins and objectives of disinformation campaigns, highlighting the need for a structured understanding of the various actors involved. To address this analytical need, this subchapter is visually anchored by an original network graph (Figure 5), which conceptualizes the actor landscape of disinformation as a web of interconnected and, in many cases, covertly coordinated entities. This integrated visualization serves two key purposes: First, it surfaces the often-hidden architecture of disinformation networks; second, it provides a heuristic framework through which the typologies discussed in the remainder of this section can be understood relationally rather than in isolation. As such, the figure is intended to make explicit the multi-nodal and dynamic nature of contemporary disinformation, helping readers grasp not only *who* the actors are, but also *how* they may be connected in diffuse, indirect, or hybridized ways.

State-sponsored and State-affiliated Actors. State-affiliated actors are central to modern disinformation ecosystems, often pursuing national security, political, or economic objectives (Mirza et al., 2023). These actors leverage state resources to shape public discourse, influence foreign electorates, and destabilize adversaries. A key subset of these actors includes ‘cyber troops’ (Bradshaw & Howard, 2017; Woolley & Howard, 2016). Prominent cases include Russia’s Internet Research Agency (IRA) and China’s state-controlled disinformation efforts, both of which have sought to disrupt democratic processes, foster division, and promote authoritarian narratives (Colomina et al., 2021; Pamment, 2020).

Political Actors. Political disinformation is frequently deployed by actors with direct or indirect affiliations to domestic political entities, seeking electoral advantages or broader ideological dominance (Mirza et al., 2023). Such campaigns exploit sociopolitical cleavages, deepen polarization, and erode democratic trust (Hameleers, 2023). During election cycles, disinformation serves to discredit opponents, manipulate voter perceptions, and mobilize partisan support. These tactics are utilized across the political spectrum, with both right- and (radical) left-wing movements weaponizing disinformation for strategic gains (Egelhofer & Lecheler, 2019; Nikolov et al., 2021). Examples include manipulated media content, deepfakes, and coordinated inauthentic behavior aimed at swaying public sentiment.

Corporate and Commercial Actors. Disinformation campaigns are not solely politically motivated; corporate actors also engage in deceptive practices to protect brand reputations or enhance financial interests (Mirza et al., 2023). Historical examples include fossil fuel companies funding disinformation to downplay climate change risks (Mulvey et al., 2015). Similarly, hyper-partisan and alternative media outlets frequently amplify misleading content to maximize user engagement and advertising revenue (Hameleers, 2023). The profit-driven model of digital platforms further incentivizes the spread of sensationalist and misleading narratives (Munn, 2020).

Ideological and Activist Actors. Individuals and groups motivated by ideological, religious, or normative beliefs significantly contribute to disinformation ecosystems. These actors deploy emotionally charged and misleading narratives to advance their agendas (Hamm, 2020; Mirza et al., 2023). Conspiracy movements such as QAnon and anti-vaccination networks exemplify how disinformation can be systematically produced and propagated for ideological influence (Mirza et al., 2023). The COVID-19 pandemic illustrated how crises can be exploited to spread falsehoods, such as disinformation surrounding 5G technology or vaccine safety (Hamm, 2020). These actors often operate within echo chambers that reinforce their beliefs and increase their influence over susceptible audiences.

Media and Social Media Actors. Social media platforms and their algorithms play a pivotal role in the amplification of disinformation (Catering, 2018; Colomina et al., 2021; Lukito, 2020). Algorithmic ranking systems prioritize engagement-driven content, often at the expense of accuracy, thereby facilitating the viral spread of misleading narratives (Hameleers, 2023). Individual influencers, fringe networks, and opinion leaders further contribute to this phenomenon, either for financial gain or ideological purposes (Hamm, 2020). The participatory nature of digital media fosters an environment where both intentional and unintentional disinformation thrives (Guess & Lyons, 2020). This dynamic is

further reinforced by the profit-oriented business models of digital platforms, where emotionally charged and sensationalist content generates higher user engagement, increased advertising reach, and ultimately greater financial returns (Munn, 2020).

Non-state Networks and Individual Actors. Decentralized networks and independent actors frequently play a role in disinformation dissemination. Troll farms and bot networks amplify false narratives at scale, while individual actors monetize sensational content through ad revenue or social media virality (Colomina et al., 2021; Mirza et al., 2023). Some individuals engage in disinformation campaigns for personal amusement, disregarding societal consequences (Marwick & Lewis, 2017). The early stages of the COVID-19 pandemic highlighted how opportunistic actors exploited information gaps to spread harmful narratives, often unchecked by platform moderation policies (Mirza et al., 2023).

Hybrid and Networked Actors. Disinformation campaigns frequently exhibit a hybridized nature, involving coordination among multiple actors operating across platforms (Bontcheva & Posetti, 2020). Coordinated inauthentic behavior – such as state-sponsored campaigns revealed by Facebook – demonstrates how governments, political entities, and private consultancy firms collaborate to achieve shared objectives (Colomina et al., 2021). These campaigns integrate diverse actors, including bots, influencers, and grassroots participants, complicating mitigation efforts and obscuring the origins and intentions of disinformation operations (Hameleers, 2023). The increasingly sophisticated networked nature of these activities underscores the necessity of interdisciplinary approaches to detection and counteraction.

Figure 5 illustrates the complex web of influence and possible collaboration among the various actors involved in disinformation campaigns. The arrows depict the directional flow of these interactions, revealing how different entities shape and amplify misleading narratives. For instance, state actors may exert control over media actors by managing state-affiliated outlets and disseminating propaganda, while political actors may strategically engage corporate actors, such as PR firms, to craft and spread persuasive disinformation. Media actors further amplify content from non-state actors, often propelled by algorithmic promotion and virality, increasing its reach and impact. Meanwhile, hybrid and coordinated actors bridge multiple categories, demonstrating how disinformation efforts frequently involve a mix of governmental, political, commercial, and ideological players. These interconnections are reinforced by attention-based digital environments, where engagement itself functions as a valuable resource. By mapping these relationships, the figure highlights the deeply networked nature of modern disinformation campaigns, where various stakeholders contribute to producing and disseminating false or misleading information.

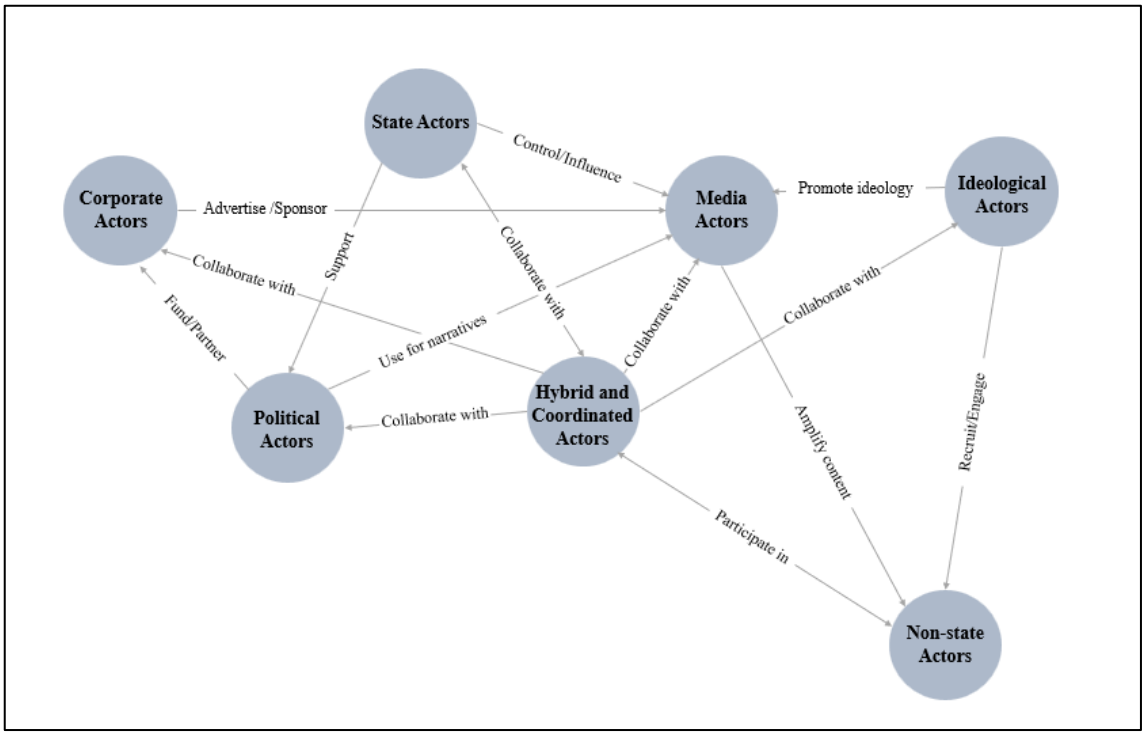

Figure 5. Network graph of disinformation actor relationships.

A comprehensive understanding of the actors involved in disinformation campaigns is essential for developing effective countermeasures. While state-sponsored actors remain prominent, non-state entities, commercial interests, and decentralized networks also contribute to the complexity of the disinformation landscape. Addressing this challenge requires a multi-stakeholder approach that integrates technological, regulatory, and educational strategies to mitigate the harmful effects of disinformation on democratic societies.

2.1.3 Platform Design and the Fragmentation of Public Discourse

In the digital age, the role of online platforms in shaping public discourse has become increasingly complex. While social media was initially heralded as a democratizing force, enabling broad participation in political and social debates, it has also become a conduit for the rapid spread of disinformation. The rise of algorithmic content curation, audience fragmentation, and monetizing strategies has created an environment where misleading information can thrive and propagate with unprecedented speed. Understanding how digital platforms contribute to the amplification and resilience of disinformation is crucial to addressing the broader challenges contemporary information ecosystems pose.

Bennett and Pfetsch (2018) characterize the current era of political communication as one marked by disrupted public spheres, where social media plays a central role in fracturing public debate. A recent report on the democratic governance of digital platforms identifies

key issues, including the large-scale amplification of disinformation and the emergence of echo chambers (Diaz Ruiz, 2023). One mechanism that facilitates these phenomena is decontextualization, which occurs when a message is reproduced in a different conversational context without its original framing. This process is exacerbated by the decentralized nature of online communication infrastructures, where networks can fork conversations, creating fragmented publics (Krafft & Donovan, 2020). While decentralization was initially championed as a foundational value of web architecture, it has increasingly been exploited by disinformation agents who use it to filter dissent and deploy computational propaganda across platforms. This ability to evade scrutiny by shifting conversations into new, isolated contexts makes disinformation particularly resilient (Krafft & Donovan, 2020).

Both state and non-state actors leverage online platforms to manufacture consensus, manipulate public opinion, and suppress ideological opposition (Mirza et al., 2023). The platform filtering effect enables disinformation agents to exploit fragmented conversational contexts, allowing disinformation to persist unchallenged. Even when rational discourse exposes falsehoods in one setting, those debunked narratives can be repackaged and redistributed into new, less critical spaces (Krafft & Donovan, 2020). Empirical evidence further supports the significance of cross-platform content circulation: approximately one-third of tweets in Starbird and Wilson's (2020) dataset contained links to external domains, often leading to other social media platforms, thereby reinforcing the interconnected nature of disinformation dissemination.

The emergence of disinformation has prompted debates on social platforms' accountability, leading to governmental demands for algorithmic interventions to detect and marginalize manipulative content (De Blasio & Selva, 2021). This shift in responsibility from journalistic sources to digital platforms underscores the structural vulnerabilities of the attention economy. Benkler et al. (2018) attribute the success of disinformation campaigns, in part, to social media's economic model, which prioritizes engagement-driven virality. This marks a stark departure from earlier narratives that celebrated the democratizing potential of digital media (Howard & Hussain, 2013).

Although social media companies are not necessarily originators of disinformation, they act as gatekeepers and amplifiers, influencing its reach and impact (Kim et al., 2018). Audience fragmentation, monetization incentives, and data extraction intensify the conditions for disinformation proliferation, as digital platforms profit from highly engaging content (Diaz Ruiz, 2023). The commercial logic of these platforms differs fundamentally from journalistic gatekeeping in quality press, as financial incentives often reward sensationalism over accuracy (Hameleers, 2023). Content creators, influencers, and digital marketers adapt to these dynamics, employing attention-hacking techniques to maximize engagement, which in turn fuels the virality of clickbait and polarizing content (Tellis et

al., 2019). The financial motivation behind disinformation is well-documented: media studies have consistently found that the spread of falsehoods is financially lucrative (Braun & Eklund, 2019). Facebook's own internal reports revealed that its advertising algorithms were leveraged to segment users into ideological micro-communities for targeted political messaging – a strategy infamously exploited by Cambridge Analytica (Walker et al., 2019).

Disinformation campaigns often originate in fringe online spaces such as 4chan and 8chan, where anonymous users develop and amplify politically motivated conspiracy theories (Guess & Lyons, 2020). A key tactic in disinformation dissemination is trading up the chain, whereby narratives emerge in obscure forums before being deliberately escalated to more mainstream platforms and media outlets (Krafft & Donovan, 2020). This process illustrates that disinformation is not merely a byproduct of identity affirmation but rather an intentional strategy to manipulate the broader media ecosystem. Unique features of platforms like 5chan, where threads are ephemeral and constantly regenerated, facilitate narrative reframing, allowing disinformation actors to reshape discourse dynamically (Krafft & Donovan, 2020). These sites operate as interconnected networks, reinforcing and amplifying each other's content. Ultimately, mainstream social media platforms (X, Facebook, YouTube) serve as conduits for disinformation's expansion into broader public discourse (Guess & Lyons, 2020).

The design of social media platforms is a crucial factor in amplifying disinformation. Strategic design choices, such as implementing tracking mechanisms to trace content migration across platforms, could mitigate some of the filtering effects that facilitate the spread of disinformation (Krafft & Donovan, 2020). Starbird and Wilson (2020) emphasize that researchers must adopt cross-platform approaches to fully understand disinformation campaigns, given their transmedia nature. While web decentralization allows for distributed discourse, social media corporations centralize power by determining which content is amplified or suppressed. This dual dynamic means that disinformation can be both decentralized in its production and centralized in its reach. The case of 4chan illustrates how an authorless piece of content can gain authority by leveraging platform design to filter out dissent while strategically moving up the chain to gain legitimacy (Krafft & Donovan, 2020).

Recent developments in social media platform governance reveal a concerning pattern where platform owners are increasingly wielding their considerable power to reshape information flows under the banner of "free speech." This trend is exemplified by Elon Musk's transformation of Twitter into X (Center for Countering Digital Hate, 2024), Mark Zuckerberg's elimination of fact-checkers at Meta (McMahon et al., 2025), and Donald Trump's creation of Truth Social following his de-platforming on Twitter (Zhang

et al., 2024). These shifts represent a significant departure from previous content moderation approaches and raise profound questions about platform accountability, information integrity, and democratic discourse in the digital public sphere. As long as the structure of the web and social media platforms remains unchanged (or even changes for the worse), disinformation campaigns will continue to scale. Not only have adversarial groups learned to align within specific web communities, but they have also developed strategies to exploit online communication infrastructures for audience expansion (Colomina et al., 2021). Platforms, by design, provide fertile ground for the spread of falsehoods, maximizing both reach and profitability (Hameleers, 2023). Given this landscape, it is imperative that stakeholders – including technologists, designers, regulators, researchers, and web users – push for reforms that integrate accountability, transparency, justice, and co-design into platform governance (Frey et al., 2019).

A multi-stakeholder model of co-regulation has been increasingly proposed, wherein platform operators collaborate with non-governmental organizations to monitor and remove harmful content while establishing standardized codes of practice (De Blasio & Selva, 2021). However, current platform interventions remain siloed, lacking the coordinated efforts necessary to effectively counteract disinformation networks (Starbird & Wilson, 2020). Addressing this issue requires platforms to work together in identifying and mitigating disinformation campaigns across the entire digital ecosystem. Only through collective action can we begin to dismantle the infrastructure that enables and sustains the weaponization of digital media for disinformation purposes.

2.1.4 Individual Susceptibility to Disinformation

In the contemporary digital landscape, the rapid expansion of disinformation presents a significant challenge to the ways in which individuals process and engage with information. As the tools for producing and spreading information have become more sophisticated, the ability to critically assess the veracity of content has grown more complex (Appel & Doser, 2020). This changing environment has raised fundamental questions about how information is consumed, interpreted, and trusted, making it essential to understand the factors that shape individuals' susceptibility to disinformation.

2.1.4.1 Cognitive Heuristics in the Processing of Information

Humans, in their day-to-day cognitive processing, frequently rely on heuristics – mental shortcuts that simplify decision-making processes (Metzger & Flanagin, 2013). While these heuristics provide efficiency and reduce cognitive load, they often introduce systematic biases that can compromise the accuracy of judgments (Weber & Knorr, 2020). Such cognitive distortions go back to the approach of motivated cognition, also known as

motivated reasoning, in the scientific literature (Druckman & McGrath, 2019; Epley & Gilovich, 2016; Kahan, 2015). It describes the linking and mutual influence of motivation and cognition. When people prefer a certain result, their thought process is steered unnoticed in the desired direction by systematic errors when retrieving, constructing, or evaluating information. Through these heuristics, motivation (the preferred outcome) therefore influences people's cognitions (the thought process) (Kunda, 1990).

One such heuristic is the *availability bias*, which leads individuals to overestimate the likelihood of an event based on how easily instances of that event come to mind (Tversky & Kahneman, 1974). This bias is particularly relevant in the context of disinformation, where the repetitive exposure to misleading information on social media platforms makes false narratives seem more plausible due to their heightened availability in one's cognitive environment (Thaler & Sunstein, 2008). The frequent repetition of disinformation can give the impression of factual accuracy, even in the absence of objective corroboration.

Another pervasive cognitive bias is *confirmation bias*, prompting individuals to selectively attend to information that aligns with their pre-existing beliefs while disregarding evidence that challenges these beliefs (Kunda, 1990; Wason, 1960). In the context of disinformation, confirmation bias plays a pivotal role in shaping individuals' acceptance of misleading content, particularly within politically polarized environments. Empirical studies have demonstrated that individuals are more inclined to believe and share information that conforms to their ideological orientations, regardless of the veracity of the information (Kahan, 2017; Taber & Lodge, 2006). This bias not only facilitates the acceptance of disinformation but also fuels its propagation, as individuals are less likely to engage critically with content that supports their worldview (Bronstein et al., 2018).

Additionally, the *representative heuristic* contributes to individuals' susceptibility to disinformation by fostering judgments based on perceived similarities between new information and existing stereotypes or prototypes (Akert et al., 2008). In the case of disinformation, individuals may evaluate the plausibility of a narrative based on its emotional appeal or how it aligns with their preconceptions, rather than engaging in a rigorous evaluation of its factual accuracy. This heuristic can result in individuals attributing greater credibility to sensationalist or emotionally charged content, regardless of its truthfulness (Weber & Knorr, 2020).

The *hindsight bias*, wherein individuals perceive outcomes as being more predictable after they have occurred (Christensen-Szalanski & Willham, 1991), further exacerbates the challenge of combating disinformation. Following the debunking of false content, individuals may retrospectively assert that they had always known the information to be false, reinforcing their confidence in their ability to accurately assess future claims. This bias

may undermine the learning process, as individuals fail to critically reflect on the mechanisms that contributed to their initial acceptance.

Taken together, these heuristics illustrate how cognitive shortcuts, while useful for efficient decision-making, simultaneously increase individuals' vulnerability to disinformation by making misleading content appear more plausible, familiar, or aligned with existing beliefs.

2.1.4.2 On the Need for Critical Thinking and Media Literacy

A crucial factor contributing to falling victim to information manipulation is the lack of critical thinking skills, which are essential for navigating the complexities of the digital media environment. As the volume of information individuals encounter has exponentially increased in the digital age, the need for robust critical thinking has become paramount (Metzger & Flanagin, 2013). However, many individuals are ill-equipped to critically evaluate the vast array of content they encounter daily. Studies have demonstrated that a lack of media literacy, defined as the ability to access, analyze, evaluate, and create media in various forms (Aufderheide, 2018), is closely linked to an increased susceptibility to disinformation (Faragó et al., 2023; Sirlin et al., 2021). Inadequate media literacy hampers individuals' ability to discern the credibility of sources and evaluate the reliability of information, thereby facilitating the spread of disinformation (Kellner & Share, 2007).

Moreover, critical thinking, which involves reflective and reasoned analysis of information, is integral to the development of media literacy. Research has shown that individuals who engage in more analytical thinking are less likely to fall prey to disinformation (Bronstein et al., 2018; Pennycook & Rand, 2018a). Consistent with the assumptions underlying *classical reasoning* theory (Kohlberg, 1994), these findings underscore the importance of fostering critical thinking skills, as individuals who actively engage in thoughtful analysis are better equipped to evaluate the trustworthiness of information and resist the persuasive influence of misleading content. Critical thinking enables individuals to identify the biases and heuristics that shape their interpretation of information, thus mitigating their susceptibility to manipulation (Guess et al., 2020; Pereira & Moura, 2019). Unfortunately, many educational systems have yet to integrate media literacy and critical thinking into their curricula comprehensively (Mcdougall, 2019; Reboot Foundation, 2022). Without such training, subjects are left vulnerable to information manipulation and ill-prepared to engage as informed, reflective citizens in a media-saturated world.

Finally, the concept of *third-person perception*, wherein individuals believe that others are more susceptible to media influence than themselves, further compounds the challenge of disinformation (Jang & Kim, 2018). This perceptual discrepancy often leads to

an overconfidence in one's ability to resist misleading content, which can, in turn, diminish the motivation to critically evaluate media. Research has revealed that people tend to overestimate the impact of disinformation on others while underestimating its potential effect on their own judgments (Lazer et al., 2018). This false sense of immunity results in a failure to recognize one's own biases and susceptibility to disinformation, perpetuating the cycle of deceptive information.

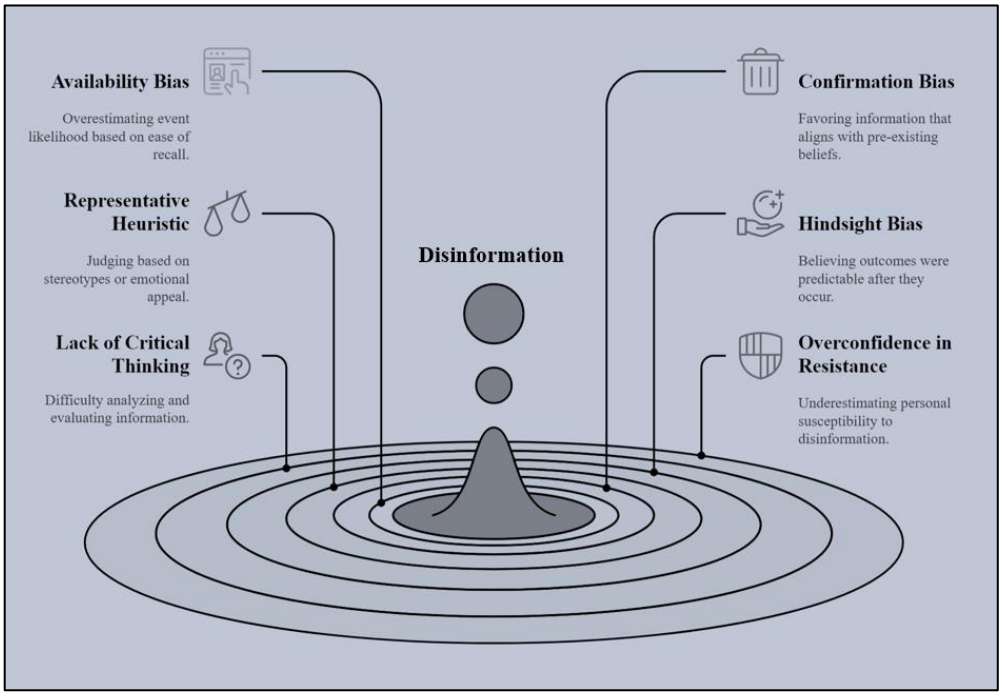

Figure 6. Factors of individual susceptibility to disinformation.

Addressing the challenge of manipulated information in this digital age necessitates a multilayered response. Concluding this section, Figure 6 summarizes the different factors that contribute to individual susceptibility to and, eventually, the successful spread of disinformation, including cognitive biases, overconfidence, lack of critical thinking, and insufficient media literacy. The figure illustrates this process through the metaphor of a drop falling into water, where each ripple represents a factor that may amplify and extend the reach of disinformation. Just as the concentric circles spread outward from a single point of impact, these vulnerabilities interact in ways that allow deceptive content to radiate further into the information environment. This visualization underscores that the influence of disinformation is not static but expands dynamically, with each layer of susceptibility adding momentum to its diffusion. While cognitive heuristics and their resulting biases are deeply rooted in human cognition and thus relatively resistant to change, it is the inflated confidence in one's own ability to detect falsehoods and the underdevelopment of critical thinking skills that offer more accessible levers for intervention. Notably, these latter factors are not only more malleable but also more amenable to influence from

external sources, such as educational programs, institutional policies, and platform-level interventions. In this regard, media literacy education is pivotal in equipping individuals with the critical skills necessary to assess the credibility and reliability of information encountered online. Educational programs must focus not only on how to access and create media but also on how to critically analyze the veracity of information, fostering a more discerning public (Hobbs, 2017; IFLA, 2017). Critical thinking, as an integral component of media literacy (Potter, 2013), helps individuals recognize the biases and heuristics influencing their judgments, thereby enhancing their ability to discern fact from fiction (Pereira & Moura, 2019). By incorporating media literacy and computational thinking in and outside of school, we may better prepare individuals to navigate the complexities of the digital media environment (Soßdorf et al., 2024; Valtonen et al., 2019) and participate in today's society (Marten, 2010), as will be discussed in Chapter 10.

2.1.5 The Impact and Consequences of Disinformation

Disinformation has emerged as a pervasive threat to democratic stability, human rights, and social cohesion (Berger et al., 2024). The rapid expansion of digital media and social networking platforms has facilitated the widespread dissemination of misleading and manipulative content, amplifying its impact on political, economic, and societal structures (Colomina et al., 2021; Khaled, 2022). By exploiting existing societal divisions, disinformation deepens political polarization, erodes trust in public institutions, and compromises electoral integrity. Beyond its political implications, disinformation also poses risks to public health, threatens economic stability, facilitates cybercrime, and exacerbates social unrest. This section examines the multifaceted consequences of disinformation, highlighting its role in shaping public perceptions, influencing decision-making, and challenging the foundations of democratic governance.

Political Polarization. Disinformation campaigns significantly influence information consumers, fostering polarization and thereby escalating societal tensions and instability. Polarization has exacerbated discord over critical global issues, including social justice, immigration, COVID-19 vaccines, Brexit, climate change, and Russia's invasion of Ukraine (French et al., 2024). Disinformation has infiltrated many, if not all, of the contentious topics that drive societal divides, often fueling hostility and mistrust. The political and social ramifications of disinformation-induced polarization can be severe, with far-reaching consequences for governance, democratic stability, and public trust (Qureshi et al., 2021). One of the most profound effects of disinformation is its ability to reinforce ideological biases and create insular echo chambers that restrict exposure to diverse perspectives (French et al., 2024). Disinformation campaigns strategically target individuals based on their pre-existing beliefs, deepening societal fractures and reducing the potential for constructive dialogue. Empirical evidence indicates that over 40% of individuals

worldwide express concern over disinformation's role in amplifying polarization and enabling foreign interference in domestic political affairs (Colomina et al., 2021). Beyond societal dissension, the consequences of polarization extend to organizations, where reputational damage and perceived declines in institutional integrity have been reported as direct outcomes of disinformation campaigns (Mody, 2020). One of the most striking examples of polarization exacerbated by disinformation was the 2016 U.S. presidential election. The combination of an increasingly partisan political climate and the proliferation of misleading content online facilitated the rapid dissemination of false narratives. Reports indicate that in the five months preceding the election, approximately 25% of shared political news on Twitter (now X) contained false or highly biased information (Bovet & Makse, 2019). The persistence of disinformation in subsequent years further eroded trust in democratic institutions, culminating in the violent storming of the U.S. Capitol on January 6, 2021, following the 2020 U.S. election (Cellan-Jones, 2021). These events underscore the profound impact of disinformation-driven polarization on democratic societies, necessitating a deeper understanding of its mechanisms and mitigation strategies.

Undermining Democratic Institutions and Electoral Integrity. Disinformation presents a profound threat to the integrity of democratic institutions and electoral processes. By fostering confusion and skepticism about elections, it erodes public confidence in both electoral systems and political institutions. Election interference can be understood as the deployment of illegitimate and coercive tactics designed to manipulate public opinion and voter choices, thereby undermining citizens' capacity to exercise their political rights freely (Colomina et al., 2021). A fundamental aspect of electoral integrity is the ability to vote without undue influence, ensuring that freedoms of thought, opinion, and privacy are upheld and that deceptive information does not distort political discourse. However, numerous governments have engaged in disinformation campaigns that contravene these democratic principles (French et al., 2024). Coordinated disinformation campaigns have been implicated in several democratic elections, including the Brexit referendum in 2016, the French presidential election in 2017, and the Mexican and Italian elections in 2018 (Rodríguez-Fernández, 2019). These instances illustrate how deceptive narratives are strategically employed to influence voter perceptions, sow distrust in political institutions, and question the legitimacy of electoral outcomes. When disinformation successfully manipulates public opinion, it not only threatens electoral integrity but also diminishes overall confidence in democratic governance, leading to reduced political engagement and increased susceptibility to populist rhetoric (Hooghe, 2018). A prominent example of such influence occurred during the 2016 U.S. presidential election, where millions of individuals engaged with disinformation from unreliable sources on social media (Silverman, 2016). Observers have argued that fabricated news stories may have influenced

electoral outcomes and contributed to Donald Trump's victory (Parkinson, 2016). Research suggests that some of this disinformation was intentionally disseminated on social media to shape voter behavior (Allcott & Gentzkow, 2017; Shane, 2017). The spread of false information persisted into the 2020 U.S. election, further exacerbating political polarization. In response, third-party fact-checking organizations and dedicated platforms were established to help citizens discern credible election news from deceptive content (O'Sullivan et al., 2021). Despite these efforts, disinformation continued to deepen partisan divisions, with supporters of Donald Trump and his opponent, Joe Biden, entrenched in opposing narratives. The consequences of disinformation extended beyond the electoral process. As stated in the previous section, the culmination of these divisions became evident in the storming of the U.S. Capitol, intended to disrupt the certification of President-elect Joe Biden's victory. The attack, driven in part by disinformation-fueled narratives about election fraud, resulted in fatalities, injuries, and extensive damage to both public property and public confidence in democratic institutions (Cellan-Jones, 2021; French et al., 2024).

Erosion of Trust in Media and Public Institutions. A further significant consequence of disinformation is the erosion of trust in mainstream media and public institutions. Empirical research indicates that exposure to false or misleading information undermines confidence in key democratic institutions, including governments, parliaments, courts, and the processes that sustain them, while weakening trust in public figures, journalists, and independent media (Berger et al., 2024). Disinformation campaigns frequently exploit this vulnerability by discrediting professional journalism, often accusing it of bias, collusion, or misinformation, thereby reinforcing skepticism toward traditional news sources (Ognyanova et al., 2020). The decline in media trust has facilitated the expansion of alternative news ecosystems, which frequently lack editorial oversight and prioritize sensationalism to maximize audience engagement (Berger et al., 2024; Colomina et al., 2021). Research suggests that while disinformation generally decreases trust in the media, it can paradoxically bolster trust in government institutions when political narratives align with an individual's ideological leanings (Ognyanova et al., 2020). This dynamic underscores the complex and often contradictory ways in which disinformation reshapes public perceptions, ultimately undermining democratic accountability. The content and framing of disinformation play a crucial role in shaping public trust. Sensationalized and scandal-driven narratives, characteristic both of disinformation and certain tabloid-style reporting, have been shown to erode trust in news organizations (Hopmann et al., 2015; Ladd, 2011). Fraudulent information not only directly undermines the credibility of the press by alleging bias and incompetence but also does so indirectly by contradicting widely accepted claims from reputable media sources. Furthermore, the mere presence of disinformation that mimics legitimate journalism contributes to public skepticism about news media as a whole (Ognyanova et al., 2020). The impact of disinformation extends beyond media

trust, affecting confidence in political institutions with profound implications for democratic engagement. Public trust in government shapes civic and electoral behavior, with disillusioned citizens more likely to disengage from politics and public discourse as a reaction to perceived institutional failure (Hooghe, 2018). While dissatisfaction with governance can sometimes drive civic mobilization, prolonged cynicism and institutional mistrust may lead to political disengagement. The extent to which disinformation erodes trust in political institutions is contingent upon several factors, including the ideological orientation of media sources, the predispositions of individuals consuming the content, and the political context in which such narratives circulate (Ognyanova et al., 2020). Additionally, the characteristics of disinformation evolve over time, potentially altering its impact on public trust.

Human Rights Violations. The dissemination of false information has significant implications for human rights, as disinformation can infringe upon fundamental freedoms, including the right to freedom of thought, privacy, and access to accurate information (Colomina et al., 2021). The right to freedom of thought encompasses protection against covert manipulation of beliefs and opinions; however, disinformation campaigns frequently exploit psychological biases to influence public perception without individuals' awareness (French et al., 2024). Furthermore, privacy violations arise when personal data is harvested for microtargeting, enabling the spread of tailored disinformation that undermines individual autonomy and informed decision-making (Colomina et al., 2021). Beyond its impact on individuals, disinformation also threatens social cohesion by fostering division and intolerance. The strategic dissemination of false or distorted information targeting specific social groups reinforces exclusionary narratives, solidifying the perception of an 'out-group' and exacerbating the societal marginalization of certain groups. Research suggests that disinformation can influence public attitudes toward marginalized communities, particularly migrant populations, by shaping perceptions of their legitimacy and social integration (Szakacs & Bognar, 2021). The far-reaching consequences of disinformation on human rights underscore the urgency of addressing its proliferation. By manipulating public discourse, eroding privacy, and fostering social divisions, disinformation not only undermines democratic institutions but also poses direct risks to individual and collective well-being.

Public Health Risks. In highly polarized environments, particularly during periods of insecurity, individuals are more likely to seek out information that aligns with their preexisting beliefs or political ideology (Weismueller et al., 2024). This tendency was evident during the COVID-19 pandemic, as individuals who harbored historical mistrust toward vaccines gravitated toward sources promoting dubious or unverified alternatives (Modgil et al., 2021). The pandemic underscored the critical role of media as a primary source of health-related information. However, the widespread circulation of false or misleading

content, often disguised as legitimate disease prevention and control strategies, contributed to an overload of disinformation. This, in turn, influenced public behavior and health outcomes, leading to increased social unrest, distrust, and even violent incidents, including attacks on healthcare professionals (Moscadelli et al., 2020). Moreover, the COVID-19 pandemic starkly illustrated the potentially fatal consequences of health-related disinformation. In Iran, for instance, false news about alcohol as a supposed cure for COVID-19 led to approximately 800 deaths and the hospitalization of nearly 6000 individuals due to methanol poisoning (Hassanian-Moghaddam et al., 2020).

Beyond its societal consequences, the proliferation of disinformation has profound implications for mental health. Exposure to misleading or alarmist health narratives has been linked to heightened anxiety, depression, and emotional exhaustion (Lin et al., 2020). Additionally, the spread of false information fosters public panic and undermines confidence in scientific institutions, further exacerbating public health crises (Rocha et al., 2021). The psychological effects of disinformation extend beyond general distress, contributing to specific symptoms such as fatigue, anger, and insomnia (Islam et al., 2020; Radwan et al., 2020; Secosan et al., 2020). These developments highlight the broader consequences of disinformation for public health, particularly in crisis situations. The interplay between disinformation, public perception, and institutional trust can shape both individual health behaviors and collective responses to health emergencies. As false information continues to circulate in digital and traditional media, its potential to influence health-related decision-making and exacerbate public anxiety remains a pressing concern in contemporary societies.

Digital Violence and Hate Speech. Disinformation is frequently intertwined with online hate speech and digital violence, amplifying social divisions and aggravating harm against vulnerable communities. The deliberate dissemination of false information targeting minority groups has fueled xenophobia, racism, and discrimination, particularly during periods of crisis (Ognyanova et al., 2020). For example, during the COVID-19 pandemic, Roma communities were unjustly scapegoated for spreading the virus in parts of Europe, resulting in discriminatory policies and heightened social stigmatization (Szakacs & Bognar, 2021). Similarly, coordinated hate speech campaigns have contributed to acts of violence against minorities and human rights defenders, demonstrating the broader societal risks associated with digital disinformation (Colomina et al., 2021). The concept of cyber-violence encompasses a spectrum of coercive and abusive behaviors, including cyberstalking, social media harassment, and the non-consensual dissemination of intimate images. Perpetrators of digital violence include both state and non-state actors, as well as private individuals and organized groups (Colomina et al., 2021). The proliferation of disinformation has intensified these forms of online aggression, with digital tools increas-

ingly used to harass, intimidate, and manipulate individuals (French et al., 2024). Governments, political entities, and other interest groups exploit social media to discredit opponents, circulate defamatory narratives, and incite targeted online harassment. The convergence of disinformation, hate speech, and digital violence emphasizes the complex challenges posed by online manipulation. As digital platforms continue to serve as conduits for harmful content, the interaction between disinformation and online aggression raises pressing concerns about the social and political ramifications of scarcely regulated digital spaces.

Threats to Cybersecurity. As digital platforms become increasingly central to public discourse, disinformation has emerged as a significant cybersecurity threat. Malicious actors, including state-sponsored entities, strategically employ disinformation as a tool of cyber warfare to destabilize governments and manipulate international relations (Petratos, 2021). Cyber-enabled disinformation campaigns have targeted critical infrastructure, financial markets, businesses, and national security institutions, often causing considerable disruption (French et al., 2024). By spreading misleading narratives, adversaries can erode public trust in governmental and economic systems, thereby weakening national stability. Beyond its role in geopolitical conflicts, disinformation is increasingly exploited for cybercriminal activities. Cybercriminals utilize deceptive tactics such as phishing scams, fraudulent advertisements, and fabricated news stories to manipulate individuals and exploit financial systems (Khaled, 2022). These schemes not only facilitate financial fraud but also compromise personal data security, contributing to broader concerns regarding digital safety and sovereignty. The erosion of control over national information infrastructures and the manipulation of digital public spheres by foreign or anonymous actors pose significant challenges to a state's ability to protect its digital territory and maintain informational autonomy (Kachelmann & Reiners, 2023). The intersection of disinformation and cybersecurity highlights the evolving nature of digital threats. As online disinformation tactics become more sophisticated, their implications extend beyond political manipulation to encompass economic vulnerabilities and individual data protection. This underscores the growing need to address disinformation as both an informational and a cybersecurity challenge.

Economic Damage. The relationship between disinformation and corporate communication has not been explored as extensively as its impact on institutional and political discourse (Rodríguez-Fernández, 2019). However, in the digital economy, companies increasingly exploit disinformation to enhance their online presence and gain a competitive edge. In the short term, such strategies are often aimed at increasing social media engagement and improving brand visibility. A common tactic involves manipulating consumer reviews on platforms such as Amazon and TripAdvisor, where fabricated testimonials artificially enhance a company's reputation and influence purchasing decisions

(Rodríguez-Fernández, 2019). The fabrication of online reviews has evolved into a structured industry, with specialized firms offering deceptive promotional services. For instance, in 2013, it was revealed that Samsung had paid Taiwanese bloggers and students to produce misleading content, discrediting its competitor, HTC (Fiorenza et al., 2018). While efforts have been made to develop tools capable of detecting fraudulent content, the prevalence of digital disinformation continues to shape consumer opinions. These deceptive practices reveal the broader economic implications of disinformation. Beyond reputational manipulation, the widespread use of misleading corporate strategies raises ethical concerns and challenges the integrity of digital marketplaces.

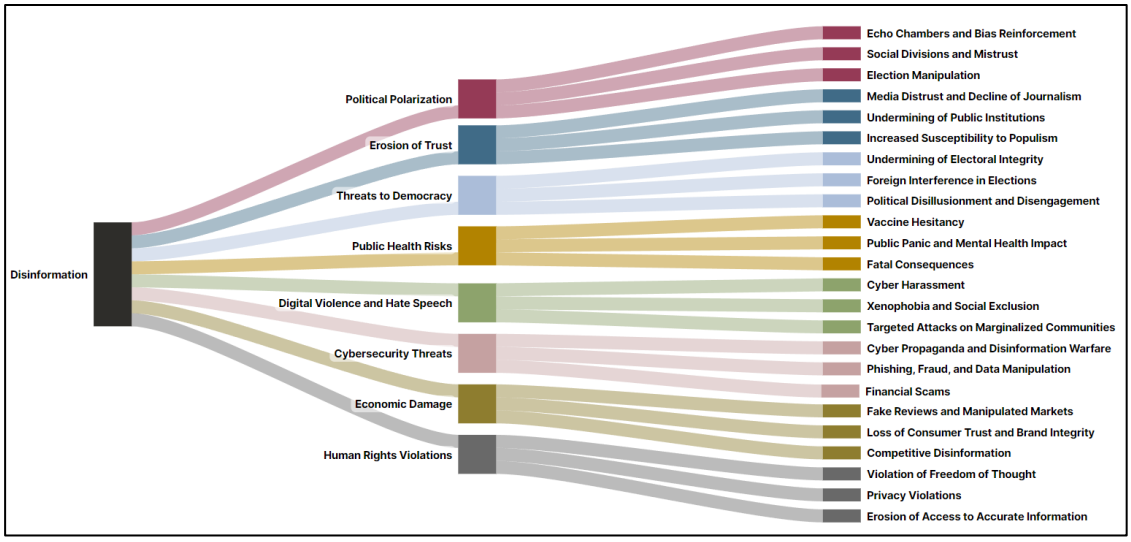

Figure 7. Overview of disinformation's consequences.

The pervasive influence of disinformation underlines its role as a destabilizing force in modern societies (see Figure 7). From deepening political polarization and eroding trust in democratic institutions to facilitating cyber threats and economic deception, disinformation extends beyond the digital sphere to shape real-world outcomes. The entanglement of false information with hate speech, digital violence, and electoral manipulation illustrates the complexity of contemporary information warfare, where disinformation serves as both a tool of influence and a catalyst for societal fragmentation. Moreover, its economic implications – ranging from corporate disinformation to fraudulent market practices – highlight the extent to which deception is embedded within digital economies. As regulatory measures and counter-disinformation strategies continue to evolve, understanding the mechanisms and consequences of disinformation remains crucial for addressing its threats to democracy, social cohesion, and institutional integrity.

2.2 Combating Disinformation in the Age of Artificial Intelligence

Addressing the proliferation of disinformation requires a multifaceted approach that integrates technological innovation, regulatory frameworks, and strategies to foster civic engagement. AI has emerged as both a contributing factor and a potential solution to the spread of disinformation. While AI-driven technologies can be exploited to generate and disseminate misleading information, they also offer robust tools for detecting, mitigating, and preventing the spread of falsehoods. Nonetheless, the application of AI in counteracting disinformation raises profound ethical and practical challenges, particularly in relation to issues of transparency, accountability, and the reliability of automated systems. This chapter investigates key interventions and mitigation strategies in the battle against disinformation, with a specific emphasis on the role of AI. It explores structured frameworks for counter-disinformation efforts, the deployment of AI-based detection technologies, and the significance of explainable AI (XAI) in promoting user trust and ensuring system accountability. Through a detailed analysis of these approaches, this chapter seeks to provide a thorough understanding of how AI can be responsibly employed to combat disinformation, while also addressing challenges associated with its use.

2.2.1 Interventions and Mitigation Strategies

The spread of disinformation presents a complex and evolving challenge that demands a multidisciplinary and systematic response. The DISARM Framework (DISARM, 2023) provides a structured approach to categorizing the diverse strategies used to counter disinformation. Building on this foundation, this section systematically organizes these mechanisms into coherent categories and further enriches them with additional approaches identified in the literature. These strategies span multiple disciplines and objectives, including regulatory interventions, technological solutions, and initiatives aimed at fostering civic resilience and constructive discourse. By systematically classifying these approaches, this section outlines key counter-disinformation efforts and their underlying theoretical foundations. Despite their varied methodologies, these strategies can be organized into seven overarching categories (Figure 8), each reflecting distinct objectives and mechanisms.

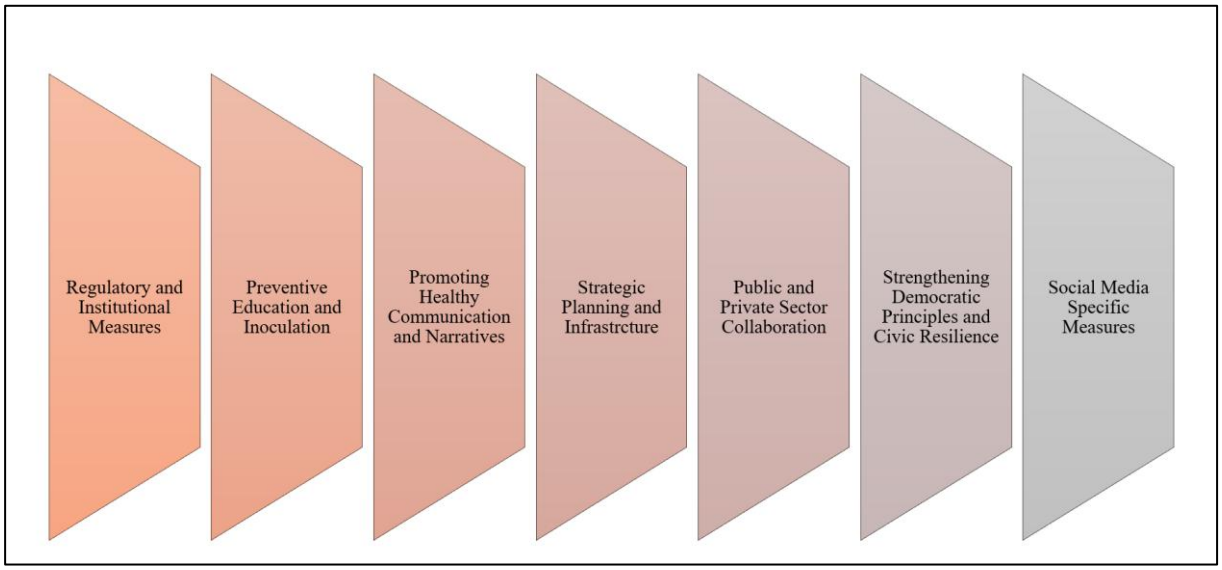

Figure 8. Categories of intervention and mitigation strategies.

Regulatory and Institutional Measures. Formal governance and institutional oversight play a critical role in mitigating disinformation. This category includes regulatory frameworks designed to increase transparency on digital platforms, privacy legislation aimed at curbing manipulative microtargeting, and initiatives to safeguard the independence of free media (De Blasio & Selva, 2021). Strengthening trust in credible institutions is another key component, as it ensures that reliable information sources remain accessible and authoritative. Governments and digital platforms have adopted voluntary commitments, such as the EU Code of Practice against Disinformation, to enhance transparency and cooperation (European Court of Auditors, 2020; Hoxtell, 2023). However, enforcement and monitoring remain significant challenges, often limiting the effectiveness of these measures. Additionally, regulatory interventions risk unintended consequences, such as censorship concerns or the concentration of power in regulatory bodies (European Court of Auditors, 2020). The scalability of institutional measures largely depends on international cooperation and the willingness of digital platforms to comply with evolving governance frameworks (Hoxtell, 2023).

Preventive Education and Inoculation. Preventive strategies focus on equipping individuals with the cognitive tools necessary to resist disinformation. Media literacy programs serve as a foundational approach, enhancing critical thinking skills and empowering individuals to assess the credibility of online content (Lim & Tan, 2020; Schmitt et al., 2020). These initiatives often emphasize source evaluation, fact-checking techniques, and awareness of manipulative tactics. Complementing media literacy, inoculation-based strategies preemptively expose individuals to weakened forms of disinformation, fostering psychological resistance to disinformation tactics (Lewandowsky & van der Linden, 2021). Drawing from inoculation theory, which likens cognitive resistance to the immune

system's response to vaccines, these interventions introduce individuals to misleading arguments in a controlled setting, allowing them to develop counterarguments (Compton, 2013; McGuire & Papageorgis, 1962). Empirical studies indicate that inoculation messages can effectively enhance resilience across various domains, including political and health-related disinformation (Banas & Rains, 2010; van der Linden, 2019).

Promoting Healthy Communication and Narratives. Constructive discourse plays a crucial role in mitigating the impact of disinformation. Strategies in this category focus on fostering inclusive, identity-neutral narratives, promoting in-person engagement to rebuild social trust, and encouraging balanced representations of diverse perspectives. These efforts seek to reduce societal fragmentation and cultivate a public sphere resilient to divisive disinformation campaigns. A key mechanism in this domain is the strategic use of social norms to counter disinformation. Research suggests that social influence can discourage individuals from endorsing or disseminating false information by reinforcing prevailing attitudes within a given community (Kozyreva et al., 2024). This approach distinguishes between *descriptive norms* – which reflect the majority's disapproval of spreading disinformation – and *inductive norms*, which frame such actions as morally unacceptable (Cialdini et al., 1991). By shaping normative beliefs about information-sharing behavior, social norms interventions can contribute to reducing the spread of misleading content (Kozyreva et al., 2024).

Strategic Planning and Infrastructure. Effective counter-disinformation efforts require robust planning and infrastructure to enable rapid and coordinated responses to emerging threats (Colomina et al., 2021). This category includes the development of intelligence and monitoring frameworks, crisis response protocols, and mechanisms for identifying systematic vulnerabilities in the information ecosystem (French et al., 2024). By strengthening institutional preparedness, these measures enhance resilience against both organic and coordinated disinformation campaigns.

Public and Private Sector Collaboration. Given the interdisciplinary nature of disinformation challenges, cross-sector collaboration is essential. Effective countermeasures rely on partnerships between governmental bodies, private entities, and civil society organizations to facilitate detection, reporting, and enforcement mechanisms (Colomina et al., 2021). International coalitions further enhance the scalability of interventions, ensuring a unified response across jurisdictions. In addition to institutional efforts, research indicates that citizens actively participate in identifying and correcting false information online, demonstrating that disinformation is not only a challenge of dissemination but also one of response (Golovchenko et al., 2018). This underscores the importance of collaborative frameworks that integrate both top-down regulatory measures and grassroots corrective actions.

Strengthening Democratic Principles and Civic Resilience. Reinforcing trust in democratic institutions is a critical component of disinformation mitigation. This category encompasses initiatives such as civic education programs, the promotion of pro-democracy narratives, and the strategic use of information as a tool for safeguarding liberal values (French et al., 2024). By enhancing public confidence in democratic processes, these efforts seek to reduce the susceptibility of target audiences to manipulative tactics employed in disinformation campaigns. Empowerment-based approaches further strengthen civic resilience by equipping individuals with the skills and knowledge necessary to evaluate information critically. Research suggests that fostering political literacy and encouraging engagement with credible news sources can mitigate the influence of false narratives while promoting informed decision-making (Colomina et al., 2021).

Social Media-Specific Measures. Social media platforms play a central role in the spread of disinformation (Shu et al., 2020a), making platform-specific interventions a crucial component of counter-disinformation efforts. These measures include increasing transparency in algorithmic decision-making, developing automated detection systems, and establishing shared fact-checking databases. Additionally, privacy-focused initiatives, such as offering paid alternatives to data-driven advertising models, aim to reduce the financial incentives that contribute to disinformation proliferation. One widely debated intervention is *deplatforming* – the removal or restriction of accounts that systematically disseminate disinformation (Kleemann, 2024). While this strategy can effectively limit the reach of disinformation campaigns, it often results in the migration of affected actors to less regulated platforms (Hoxtell, 2023; Kleemann, 2024). Moreover, research suggests that deplatforming can lead to short-term amplification effects, as removed content gains increased visibility due to media attention (Kleemann, 2024). The long-term efficacy of this approach remains subject to ongoing debate, particularly given the high costs of enforcement and the absence of a standardized cross-platform strategy (Hoxtell, 2023).

Beyond specific interventions, it is important to recognize higher-level conceptual approaches that guide the design of counter-disinformation efforts. Among these, prebunking and debunking represent two complementary strategies that respectively aim to prevent and correct exposure to misleading information. *Prebunking*, or attitudinal inoculation, aims to proactively expose individuals to weakened forms of disinformation, equipping them with cognitive defenses before encountering manipulative narratives (Lewandowsky & van der Linden, 2021). This approach has demonstrated efficacy in reducing susceptibility to disinformation by fostering critical awareness (Tay et al., 2022). Conversely, *debunking* involves the correction of false information after it has been disseminated. Research indicates that effective debunking requires more than simple fact-checking; it is most successful when it offers alternative explanations and highlights inconsistencies within disinformation narratives (Lewandowsky & van der Linden, 2021).

Studies suggest that corrections are more persuasive when delivered by trustworthy sources, framed with explicit refutations, and supplemented with explanatory context (Ecker & Antonio, 2021; Kendeou et al., 2019; Swire et al., 2017). Optimized debunking formats that incorporate these principles have been shown to outperform standard fact-checking approaches (Ecker & Antonio, 2021; MacFarlane et al., 2021). Despite advancements in prebunking and debunking methodologies, the *continued influence effect*, where disinformation persists even after correction, remains a significant challenge (Tay et al., 2022). Further research is needed to refine corrective interventions and develop adaptive strategies that address the evolving tactics of disinformation actors (Stray, 2019; Tay et al., 2022).

While these typologies and interventions provide a foundation for combating disinformation, they also raise important challenges. Regulatory and corporate measures risk consolidating power in ways that undermine pluralism and freedom of expression (Colomina et al., 2021). Similarly, interventions targeting algorithmic systems may inadvertently reinforce existing inequalities, as data-driven decision-making disproportionately affects marginalized communities (Mensah, 2023). Although increased transparency can help uncover algorithmic biases, it does not necessarily lead to equitable outcomes, particularly for communities that already face visibility suppression or disproportionate content moderation (Chaka, 2022).

A further challenge lies in assessing the effectiveness of different countermeasures (Dowse & Bachmann, 2022). Despite growing empirical research, comparative evaluations of prebunking, debunking, and regulatory interventions remain limited (Tay et al., 2022). This lack of empirical clarity complicates the development of evidence-based strategies, highlighting the need for ongoing interdisciplinary research. Ultimately, the landscape of disinformation mitigation reflects the complexity of the challenge itself. Effective responses must balance accountability with freedom of expression, systemic reform with individual empowerment, and regulatory oversight with technological adaptability. As disinformation tactics continue to evolve, so too must the strategies designed to counter them.

2.2.2 Artificial Intelligence in Disinformation Mitigation

Artificial Intelligence (AI)-driven tools have become essential in the fight against disinformation, offering scalable solutions to the challenges posed by the rapid spread of false content across social media platforms. These tools, grounded in machine learning (ML) and deep learning (DL) algorithms, are increasingly deployed to detect and mitigate disinformation, offering the potential for both large-scale detection and real-time interven-

tion. This aligns with several categories outlined in the DISARM Framework, demonstrating both their contributions and limitations within existing counter-disinformation strategies. AI-based systems are particularly effective in enhancing *social media-specific measures* by automating the detection and mitigation of false or manipulative content. Given the vast volume of content circulating on digital platforms, manual detection is both laborious and inefficient, making AI-driven automation essential (Abdullah All Tanvir et al., 2019; Pérez-Rosas et al., 2018). AI models, particularly ML and DL algorithms, excel at processing large datasets and identifying patterns much more quickly than human experts (Aïmeur et al., 2023). In addition, AI's capacity to provide warnings and contextual insights positions these tools as key components of *preventive education and inoculation*, potentially advancing media literacy by alerting users to the presence of disinformation (Bezzaoui et al., 2022). However, while AI tools offer significant advantages in terms of scalability and speed, the integration of AI into counter-disinformation efforts raises critical concerns regarding transparency, accountability, and the ethical implications of delegating such tasks to automated systems.

The challenge of disinformation, particularly in the digital era, illustrates the need for advanced technological solutions. While manual detection remains possible, it requires specialized expertise, significant time investment, and human resources. Furthermore, psychological theories suggest that humans are not inherently adept at identifying false information, as disinformation often targets cognitive biases, emotional vulnerabilities, and pre-existing beliefs (Galli et al., 2022). This highlights a crucial limitation: humans may inadvertently fall prey to the very mechanisms that disinformation seeks to exploit. In this context, AI tools may become vital in offering a faster, more systematic approach to combating disinformation.

AI has demonstrated significant efficacy in a myriad of classification tasks, including image recognition, speech processing, and natural language analysis, rendering it a promising candidate for disinformation detection (Granik & Mesyura, 2017). The increasing availability of large-scale datasets, coupled with advancements in computational capabilities, has facilitated the refinement of ML and DL algorithms in distinguishing between authentic and fabricated content. Prominent approaches include classical ML techniques – such as decision trees, random forests, Naïve Bayes, and support vector machines – as well as more sophisticated DL architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Al-Asadi & Tasdemir, 2022). These methodologies enable AI to identify textual and contextual markers indicative of falsehoods with a degree of precision that surpasses traditional detection mechanisms.

Nevertheless, despite these technological strides, AI-based disinformation detection remains fraught with challenges. Chief among these is the paucity of high-quality, repre-

sentative datasets essential for training robust AI models (Nyow & Chua, 2019). The efficacy of AI in identifying disinformation is contingent upon access to extensive and diverse corpora that encapsulate various manifestations of disinformation (Deepak et al., 2021; Hangloo & Arora, 2022; Lange & Lechterman, 2021). However, existing datasets are often skewed towards genuine news, thereby impeding the model's capacity to generalize effectively across different forms of deceptive content (Parthiban & Peter, 2022). Moreover, the rapid evolution of online narratives renders many datasets obsolete, as models trained on past instances of disinformation struggle to adapt to emergent tactics and rhetorical strategies (Hakak et al., 2020). This challenge is further exacerbated by the fact that early-stage news reports frequently lack contextual completeness, complicating efforts to ascertain veracity (Agrawal et al., 2021).

To address the limitations posed by data scarcity and outdated corpora, researchers have proposed several innovative strategies. One such approach involves the development of dynamic knowledge bases that are continuously updated to reflect the most recent news articles, thereby ensuring that AI models remain adaptable to evolving disinformation tactics (Sharma et al., 2019). Additionally, synthetic datasets can serve as valuable supplements to real-world data, mitigating privacy concerns while enhancing model robustness (Shahid et al., 2022). Semi-automated methods for data curation, leveraging trusted sources and verified fact-checking agencies, may further bolster dataset reliability.

In addition to dataset limitations, AI-driven detection is also susceptible to biases introduced during data annotation and model training. The classification of news as 'true' or 'false' is often inherently subjective, particularly when addressing politically sensitive or ideologically contentious topics (Gupta et al., 2022). Biases embedded within training data can thus influence model outputs, potentially reinforcing existing disparities and marginalizing alternative discursive communities (Lange & Lechterman, 2021). Ensuring the fairness and impartiality of AI-based interventions necessitates ongoing scrutiny of both training data and algorithmic decision-making processes.

Furthermore, the multifaceted nature of disinformation complicates AI-driven classification efforts. Deceptive information does not exist as a monolithic construct but rather manifests along a spectrum, encompassing outright fabrications, misleading interpretations, and selectively curated distortions of factual information (Hangloo & Arora, 2022). The delineation between misinformation, disinformation, and malinformation remains inherently fluid, presenting a formidable challenge for AI models predicated on binary classifications. This ambiguity is particularly pronounced in politically charged contexts, where the distinction between opinion, satire, and deliberate deception is often blurred (Choudhary et al., 2021). Thus, refining AI methodologies to incorporate a more nuanced understanding of deceptive content is paramount in enhancing detection accuracy.

The refinement of feature selection and classifier mechanisms remains a critical avenue for improving AI's detection capabilities. Sentiment analysis, for instance, has emerged as a powerful tool in identifying disinformation, as manipulative content often elicits strong emotional responses such as fear, anger, or misplaced trust (Torgheh et al., 2021). By analyzing the emotional and contextual underpinnings of deceptive content, AI models can effectively distinguish between disinformation and genuine news (Farhoudinia et al., 2024). Additionally, hybrid detection systems – combining multiple ML and DL techniques – can enhance model robustness, particularly in cases requiring multimodal analysis of textual, visual, and audio-based content (Shae & Tsai, 2019). The integration of blockchain technology has also been proposed as a means of ensuring the verifiability of news content, requiring peer-to-peer validation before publication (Aïmeur et al., 2023; Shahid et al., 2022).

Moreover, the contemporary disinformation landscape extends beyond text-based content, encompassing increasingly sophisticated multimodal fabrications, including manipulated images, videos, and synthetic media (Swapna & Soniya, 2022). Traditional text-based detection techniques are ill-equipped to contend with these emergent threats, necessitating the development of multimodal AI architectures capable of analyzing both textual and visual elements in tandem. Emergent methodologies include GAN fingerprinting, adversarial AI defenses, and blockchain-based verification to track content authenticity. However, research in this domain remains nascent, with a dearth of comprehensive multimodal datasets posing a significant impediment to progress (Akhtar, 2023). The advent of deepfake technology and AI-generated synthetic media further exacerbates these challenges, as it enables the seamless creation of hyper-realistic yet entirely fictitious content, rendering conventional detection mechanisms increasingly obsolete (Gupta et al., 2022).

The velocity with which disinformation propagates across digital platforms further compounds the complexity of detection. AI systems must operate in real-time to curtail the rapid dissemination of falsehoods before they attain widespread traction (Hangloo & Arora, 2022). However, the computational demands associated with training and deploying AI models at scale often result in latency, diminishing their efficacy in responding to nascent disinformation campaigns (Barrutia-Barreto et al., 2022). Optimizing AI architectures to enhance real-time detection capabilities is, therefore, imperative in mitigating the temporal advantage leveraged by disinformation actors.

In addition to technical limitations, the integration of AI into disinformation detection frameworks raises broader epistemological and ethical concerns. While AI models can ascertain the probability of a given piece of content being false, their decision-making processes frequently lack transparency. This opacity undermines public trust in automated systems, necessitating the adoption of explainable AI (XAI) methodologies that

elucidate the rationale underlying algorithmic determinations (Bailer et al., 2021). By fostering greater interpretability, XAI may enhance user confidence in AI-driven disinformation detection and facilitate more informed engagement with digital content (Schmitt et al., 2024).

Integrating user-based information into AI detection models has demonstrated significant potential in identifying the sources and dissemination patterns of disinformation. Features such as account age, number of posts, follower networks, and social media behavior can provide crucial indicators of disinformation campaigns (Deepak et al., 2021; Mridha et al., 2021; Shahid et al., 2022). However, the use of such data introduces ethical concerns regarding user privacy and data security, necessitating a balance between effective detection and individual rights (Shahid et al., 2022). Furthermore, AI models capable of verifying the credibility of news authors and publishers may enhance trust by offering transparency into content origins (Choudhary et al., 2021; Tanwar & Sharma, 2021).

Another fundamental challenge in AI-driven disinformation detection is ensuring cross-national and cross-cultural consistency. Ensuring that AI models generalize effectively across diverse sociocultural and linguistic contexts remains a persistent challenge, as models trained on specific datasets may exhibit reduced efficacy when applied to novel domains, such as political discourse or public health disinformation (Deepak et al., 2021). Addressing these concerns requires a commitment to ethical AI development, emphasizing inclusivity, transparency, and accountability. Given the diverse sociopolitical landscapes and linguistic intricacies across different regions, AI models must be capable of detecting disinformation in ways that transcend cultural and national boundaries. However, existing detection mechanisms often exhibit biases rooted in the datasets upon which they are trained, which are frequently dominated by content from Western contexts (Gupta et al., 2022). This discrepancy hinders the generalizability of AI models, as the markers of disinformation may vary significantly depending on the sociocultural and political environment in which they emerge (Shu et al., 2020a). Moreover, in authoritarian or politically polarized contexts, AI-based fact-checking tools risk being weaponized to suppress dissenting voices, further complicating their ethical implementation (Colomina et al., 2021).

Beyond linguistic and cultural inconsistencies, regulatory disparities between nations pose additional hurdles to the effectiveness of AI-driven disinformation mitigation. While some governments implement stringent content moderation policies, others adopt more lenient or ambiguous regulatory frameworks, creating a fragmented approach to disinformation governance. The EU Digital Services Act (DSA) mandates transparency and accountability in content moderation, affecting how AI-driven systems identify and mitigate false content. In contrast, U.S. regulations, particularly Section 230 of the Communications Decency Act, continue to shield platforms from liability, raising debates over AI's

role in moderating disinformation. A notable risk is that restrictive interventions may inadvertently push dissatisfied users towards alternative, less regulated platforms where disinformation can propagate with even greater ease (Lange & Lechterman, 2021). Consequently, a globally coordinated effort is necessary to ensure that AI-driven solutions are not only technically robust but also socially and ethically attuned to the nuances of different cultural and regulatory environments.

Besides its technical applications, AI presents significant opportunities for education and research. One promising avenue is the use of gamification techniques to improve public awareness of disinformation tactics, thereby fostering greater digital literacy and critical engagement with online content (Bezzaoui et al., 2022; Sharma et al., 2019). By exposing users to interactive simulations of disinformation campaigns, such interventions may reduce the susceptibility of individuals to manipulative narratives (Washington, 2023).

In the research domain, advancements in methodological approaches have the potential to enhance the reproducibility and reliability of AI-driven disinformation detection. Open-source tools, standardized experimental setups, and publicly accessible datasets can facilitate the development of more rigorous and transparent evaluation frameworks (Agrawal et al., 2021; Akhtar, 2023). Ensuring the reproducibility of results is particularly crucial in this field, as inconsistencies in model performance can undermine the credibility of AI-based counter-disinformation initiatives.

Despite the promise of AI in combating disinformation, it is evident that the field remains at an early stage, with significant challenges still outweighing the available solutions. A dominant issue is the overemphasis on technical feasibility, often at the expense of addressing the broader social, political, and ethical dimensions of disinformation. As technological innovations frequently outpace regulatory frameworks and societal adaptation, a cautious approach is necessary to ensure that AI-driven interventions do not inadvertently exacerbate existing inequalities or contribute to the suppression of free expression.Moreover, the increasing sophistication of multimedia disinformation necessitates the development of multimodal AI models capable of simultaneously analyzing text, images, and videos. As deepfake technology becomes more pervasive, traditional text-based detection mechanisms will become increasingly inadequate (Swapna & Soniya, 2022). Therefore, the evolution of AI-driven solutions must prioritize adaptability and real-time responsiveness to effectively counteract emerging threats.

Social and ethical concerns, while often sidelined in technical discussions, must also be central to future research endeavors. The potential of AI models to be weaponized for surveillance or ideological gatekeeping underscores the necessity of transparency and accountability in algorithmic decision-making. Efforts to enhance XAI methodologies may help mitigate concerns regarding algorithmic opacity while fostering greater public trust

in automated content moderation systems (Bailer et al., 2021). As research in this domain advances, interdisciplinary collaboration between technologists, policymakers, and social scientists will be vital in navigating the multifaceted landscape of AI-driven disinformation detection.

2.2.3 Enhancing Disinformation Detection with Explainable Artificial Intelligence

As AI becomes increasingly integrated into various domains, understanding how users interpret algorithmic features and comprehend algorithm-based systems is crucial (Shin et al., 2020). Whenever individuals encounter algorithmic decision-making, they must determine whether, how, and to what extent they trust AI-based services (Wölker & Powell, 2021). However, as AI systems grow more complex, they often function as 'black boxes' (Figure 9), making their decision-making processes opaque to users (Castelvecchi, 2016). This opacity presents challenges, particularly for non-expert users who lack the technical knowledge required to interpret AI-generated outcomes (Shin, 2021). The increasing complexity of AI models results in diminished transparency, which can negatively impact user trust and confidence in algorithmic decisions (Weitz et al., 2019).

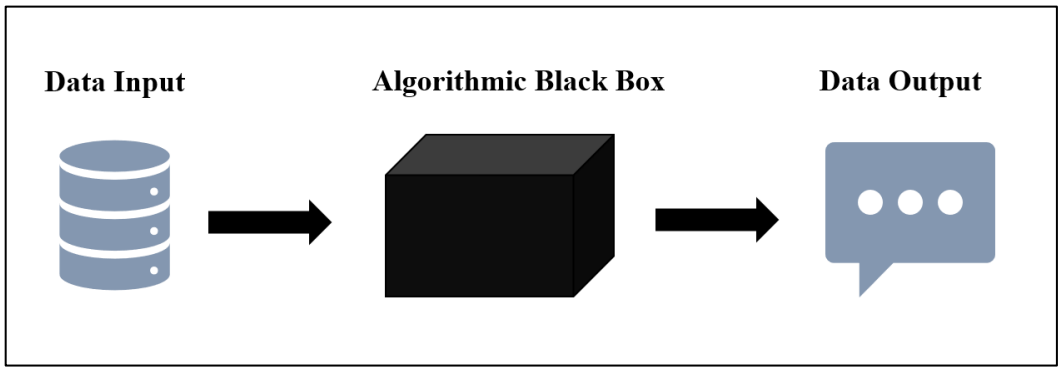

Figure 9. The black box problem.

To address these concerns, explainable AI (XAI) has emerged as a crucial area of research. XAI refers to machine learning and AI technologies that provide human-understandable justifications for their outputs or processes (Gunning et al., 2019; Meske & Bunde, 2020). While there is no universally accepted definition of explainability in AI, it is generally conceptualized as the ability to articulate how an algorithm operates and why it produces specific results (Arrieta et al., 2020; Weitz et al., 2019). Research indicates that AI systems providing explanations enhance user confidence and foster trust in algorithmic outcomes (Lipton, 2018; Rosenfeld & Richardson, 2019). Trust serves as a bonding mechanism between humans and AI, playing a pivotal role in the development of human-centered AI systems (Shin et al., 2020). Furthermore, the presence of explanations

ensures AI accountability by making decisions more transparent, verifiable, and rule-compliant (Shin, 2021).

Growing concerns regarding AI opacity have led to increasing regulatory pressure to ensure transparency in AI decision-making. The European Union (EU) has taken a significant step in this direction through its AI Act, which outlines harmonized rules for AI. Article 13, titled 'Transparency and Provision of Information to Users', mandates sufficient transparency to enable providers and users to understand AI systems' functions and recommendations (Schmitt et al., 2024). Additionally, the 'Right to Explanation' established under the General Data Protection Regulation (GDPR) has fueled efforts to develop explainable and transparent AI models, emphasizing fairness, trust, and comprehensibility (Gongane et al., 2024). Despite notable advancements, concerns remain that explanation methodologies primarily cater to AI experts while neglecting the needs of end-users (Weitz et al., 2019). Thus, further research is necessary to enhance explainability methods that are accessible and useful to non-expert users. In the context of combating disinformation, XAI plays a crucial role in fostering trust and reliability in AI-driven detection systems (Schmitt et al., 2024). Given the EU AI Act's mandate for incorporating meaningful explanations into AI systems, ensuring transparency in disinformation detection is imperative. However, defining what constitutes a 'meaningful explanation' remains challenging, as its scope and applicability vary across domains and tasks. XAI has demonstrated significant potential in addressing disinformation by demystifying AI-based classification processes and enhancing public trust in automated detection systems (Longo et al., 2024; Speith & Langer, 2023). A key aspect of XAI in this domain is its ability to promote digital literacy and media accountability. By providing clear and comprehensible explanations regarding why certain content is flagged as disinformation, XAI empowers users to critically assess the information they consume, fostering a more informed and discerning audience (Ngueajio et al., 2025). This is particularly relevant as AI-powered fact-checking tools increasingly influence the way information is disseminated and verified in digital spaces.

Various techniques have been explored to enhance explainability in disinformation detection. These include visualization-based explanations (Yang et al., 2019) and interactive interfaces that allow users to interrogate AI decision-making processes (Chien et al., 2022). Additionally, XAI-driven tools can highlight key textual components contributing to AI predictions, thereby aiding users in assessing content credibility and increasing their confidence in AI-based detection models (Rosso et al., 2024). Empirical research has demonstrated that user performance in evaluating claims improves when exposed to accurate AI explanations, which, in turn, strengthens their trust in AI-assisted fact-checking (Mohseni et al., 2021). Moreover, explanations that provide an appropriate level of detail

enhance the utility of AI systems in assessing news veracity, though they require additional time and cognitive effort from users (Linder et al., 2021).

The effectiveness of disinformation warnings is also influenced by the presence of explanations. Epstein et al. (2022) found that explanations enhance the effectiveness of disinformation warnings, although they do not necessarily increase self-reported trust in the warning labels. More broadly, AI assistance improves lay users' performance in content verification tasks, and when provided with free-text explanations, non-experts can achieve accuracy levels comparable to those of experts (Schmitt et al., 2024). Despite these advantages, the success of XAI in disinformation detection remains contingent upon the quality and diversity of datasets used in training and implementation (Ngueajio et al., 2025).

XAI is fundamental in addressing the challenges posed by the opacity of AI systems, particularly in the domain of disinformation detection. As AI-driven detection tools become more prevalent, ensuring transparency and explainability is essential to building user trust, fostering media literacy, and promoting ethical AI deployment. Regulatory frameworks such as the AI Act and GDPR have accelerated the push toward human-centric and transparent AI models (Pfeiffer et al., 2024), underscoring the importance of providing meaningful explanations to end-users. Despite significant progress in XAI research, challenges persist in designing explanations that are comprehensible and actionable for non-expert users.

In sum, this chapter has laid a comprehensive theoretical foundation for understanding digital disinformation by integrating insights from IS research, cognitive psychology, and sociotechnical studies. It has highlighted how technological infrastructures, platform design, and human cognitive mechanisms interact to facilitate the spread and impact of false information. At the same time, it has shown that countering disinformation requires not only technological solutions but also educational, regulatory, and societal interventions that acknowledge these interdependencies. By systematically addressing the theoretical foundations of this dissertation, this chapter lays the groundwork for subsequent empirical and design-oriented research, highlighting the value of interdisciplinary perspectives that account for the complex interplay of human, technological, and sociopolitical factors in digital disinformation.

Part II

Conceptualizing Online Disinformation

3 Navigating Democracy's Challenges: A Review of Research Projects on False Information and Hate Speech[2]

3.1 Introduction

In recent years, society has experienced what can be considered a poly-crisis (Henig & Knight, 2023) – while the climate crisis leads to natural hazards, there are multiple global wars, such as the Russian invasion of Ukraine in 2022 and the Israel-Hamas war since 2023. Meanwhile, social media platforms are utilized to spread disinformation and hate (Shahi et al., 2024). This might cause harm to society, e.g., due to false health advice, such as in the COVID-19 pandemic (Naeem et al., 2021), placing our democracies under significant strain. Further, hate speech poses risks for individuals psychologically (Bilewicz & Soral, 2020). Both issues relate to polarizing societies (Vasist et al., 2024) and, therefore, constitute a threat to trust in society (Weinhardt et al., 2024).

To address the challenges facing the public sphere in the digital age, it is essential for researchers to critically engage with the design, governance, and regulation of digital platforms. This includes analyzing algorithmic biases, handling information manipulation, fostering trust in digital artifacts, and proposing design principles that align with democratic values. Today, however, large platform providers such as X (formerly Twitter) increasingly restrict the possibility of researching platform mechanisms and collecting data, thus making it more difficult for researchers to access the domain (Ledford, 2023). Suggesting the establishment of six research areas for Digital Democracy research, Weinhardt et al. (2024) call for Information Systems (IS) researchers to engage in research exploring how platforms influence human behavior and social cohesion in order to recognize their broader impact beyond business models and interfaces. As networks originally meant to inform and connect individuals are now increasingly being used to

[2] This chapter comprises an article that was published by Isabel Bezzaoui, Kai Schewina and Georg Voronin in the following outlet with the following title: Navigating Democracy's Challenges: A Review of Research Projects on False Information and Hate Speech. In Wirtschaftsinformatik 2024 Proceedings. 122, 2024. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated.

spread hate and disinformation (Aïmeur et al., 2023), interdisciplinary research across Information Systems, Computer Science, Political Science, Sociology, Communication Science, Law, and others has become crucial (Sample et al., 2020). Information Systems researchers are called upon to prioritize transparency, inclusion, and literacy, focusing on innovative ways to preserve and promote democracy (Weinhardt et al., 2024). To build resilient democracies, research is essential in the areas of disinformation and hate speech to identify mechanisms and evaluate countermeasures (Bennet & Livingston, 2018). One research area introduced by Weinhardt et al. (2024) focuses on the foundation of democratic engagement: trust. It examines how various forms of misinformation, disinformation, malinformation, and hate speech influence the political landscape and trust. Therefore, it is critical to assess and map out the current efforts within the discipline of Information Systems research regarding the impact of these phenomena on democracy. For this reason, we formulate the following research question:

> *RQ: How do current publicly funded research projects in Germany and the EU address the impact of false information and hate speech on (digital) democracies, and what gaps exist that information systems researchers can fill to enhance the resilience of democratic societies in the digital age?*

By understanding what research is currently being undertaken, we can identify gaps and areas that require further exploration. This evaluation can help create future projects, ensuring they address the most pressing issues and contribute effectively to preserving and promoting democratic values in the digital age. Thus, we aim to provide an overview of the current state of publicly funded research on these topics. To do so, we consider all projects that are currently funded by the German Federal Ministry of Education and Research (BMBF), the German Research Foundation (DFG), and projects sponsored by the European Union (EU). These three are among the most important sources of third-party funding in Germany (Hornbostel, 2001). We identify several gaps in current research that need to be addressed by federal and international organizations to ensure the resilience of our democratic society.

The remainder of the chapter is structured as follows: Section 2 provides the theoretical background, exploring the relevance of false information and hate speech in the context of digital democracy. Section 3 details the methodology for systematically reviewing ongoing research projects. Section 4 presents the results, starting with a descriptive analysis followed by a qualitative content analysis to synthesize the key findings. Section 5 discusses the role of IS research in addressing these issues, highlighting the interdisciplinary potential of IS to contribute to the understanding and mitigation of false information and hate speech. Finally, Section 6 concludes the chapter with a summary of the findings and suggestions for future research directions.

3.2 Theoretical Foundation

In today's digital age, the rapid proliferation of information has transformed the way individuals communicate and access news. This chapter delves into the critical theoretical notions necessary to understand the phenomena of false information and hate speech, two pervasive issues that significantly impact societal discourse and public opinion.

3.2.1 False Information

The contemporary capability for virtually anyone to publish and share content online not only enhances opportunities for social participation but also generates new avenues for the dissemination of false information (Appel, 2020; Shu et al., 2017). Presently, research on detecting manipulated information is a rapidly expanding domain that spans multiple disciplines, including Computer Science, Information Systems, Media Studies, and Social Science (Kapantai et al., 2021; Mahyoob et al., 2020; Verma et al., 2021; Yu & Lo, 2020). It is critical to distinguish between the terms false information, misinformation, disinformation, and malinformation. False information pertains to "verifiably false information", with disinformation and misinformation being subcategories dependent on the intent. While misinformation refers to "false information that is shared without the intention to mislead or cause harm", disinformation is defined as "false information that is shared to intentionally mislead" (Aïmeur et al., 2023). Further, malinformation is defined as "genuine information that is shared with an intent to cause harm" (Aïmeur et al., 2023), therefore differentiating itself from the other terms by the genuine property of its authenticity. These concepts are crucial as they relate to the potential erosion of trust in society (Weinhardt et al., 2024), which can be severely undermined by negative experiences, such as deception through disinformation (Schwerter & Zimmermann, 2020).

The use of technology may support the spread of misleading or deceptive information. Social bots offer the opportunity to spread news at high frequency. However, it is often humans who voluntarily spread false information, especially via social media such as X (formerly known as Twitter) or Facebook (Wardle & Derakhshan, 2017). In this context, the question also arises as to who is particularly vulnerable to deceptive information. Some studies suggest that, rather than partisan bias, too little analytical thinking is a significant risk factor. The higher the ability to think critically, the less individuals appear to believe in false news (Bronstein et al., 2018; Faragó et al., 2023; Pennycook & Rand, 2018b). Therefore, it is essential to develop a comprehensive understanding of the phenomena related to false information while simultaneously devising systematic methods to counteract them (Bezzaoui et al., 2022a).

3.2.2 Hate Speech

Kansok-Dusche et al. (2023) define hate speech as derogatory expressions based on assigned group characteristics, intended to harm, and capable of causing harm on multiple levels (individual, communal, societal). This includes negative stereotyping, dehumanization, and expressions of violence (Paasch-Colberg et al., 2021). Bäumler et al. (2024) add that, unlike cyberbullying, hate speech can be subtle or humorous, targeting individuals and social groups vicariously. Online hate speech significantly impacts democracy by polarizing society and undermining democratic discourse (Weinhardt et al., 2024).

The public sphere, as described by Habermas (1962), is a space for rational discourse and public opinion formation. Social media platforms have the potential to be such spheres. However, hate speech on platforms often excludes marginalized groups from the dominant public sphere, leading them to form counter-publics – alternative spaces for expressing experiences and advocating for change (Fraser, 1990). While online hate speech normalizes discriminatory behavior and increases societal polarization (Cialdini et al., 1990; Soral et al., 2020), counter-publics provide platforms for marginalized groups to organize, support each other, and engage in activism, fostering collective resilience and challenging discriminatory norms (Eckert et al., 2021). A democratic discourse that includes marginalized individuals is crucial, as the discourse in the public sphere underpins common social values of coexistence and democratic legal norms. Excluding social groups means that these values and norms may no longer be supported by all parts of society, potentially leading to discrimination against minorities. Addressing online hate speech and including minorities from counter-publics is essential for maintaining democratic discourse and societal cohesion. Research on the mechanisms of hate speech dissemination and the effectiveness of counter-narratives is thus vital to ensure the resilience of democratic societies.

3.3 Methodology

Although there is ample methodological guidance for conducting structured literature reviews, limited instruction is available on how to review practical artifacts such as research projects. For this reason, we make use of Gnewuch and Mädche's (2022) approach to reviewing software artifacts and adapt their seven-step method to our context of a structured project review. We adapt their seven steps as follows:

1. *Problem Formulation*. The review's main objectives are determined, focusing on the project's characteristics, properties, or features central to the review. Additionally, it is crucial to establish the scope of the review. The scope is defined by the inclusion of three project sponsors and a focus on currently ongoing projects. This study focuses on research projects in the EU and Germany as an example of investigating

research projects on a federal level. The EU is one of the largest political and economic entities globally, comprising 27 member states with a combined population of over 440 million people. Its policies and regulations often set standards that influence global norms, particularly in digital governance, data protection, and media regulation. Germany is not only the largest economy in the EU but also a key player in shaping EU policies. Its actions and approaches often have a significant impact on the direction and effectiveness of EU-wide initiatives (European Union, 2024).

2. *Software Artifact Search*. Relevant projects are searched for via the internet and relevant databases, and decisions are made about their suitability for the review. The pre-defined keywords for projects on false information were "disinformation", "Desinformation", "fake news", "Falschinformation", "false information", and "misinformation". For projects regarding hate speech we searched for "hate speech", and "Hassrede", respectively. We extracted data from the BMBF, DFG, and EU websites. For DFG, we conducted a search in the database GEPRIS for the pre-defined search terms and filtered for ongoing projects. In the second step, the details of the consortium and further information on the identified projects were conducted through an additional web search. For the EU, we searched the database of the Community Research and Development Information Service (CORDIS) for the defined search string and filtered for ongoing projects. Subsequently, the project consortium and individual members were identified in order to further categorize the projects based on their relation to the field of Information Systems. For BMBF, as there is no central database that lists and categorizes projects, we use a search engine as well as the website search functionality to identify disinformation and hate speech-related projects. Further, once identified, we consider the respective line of funding.
3. *Screening for Inclusion*. Projects are screened based on predetermined criteria to determine their relevance, resulting in a list of 79 eligible projects. All projects were screened in terms of the project title, project focus, project description, involved countries, sponsors, consortium, duration, involved disciplines, and target groups we only included projects that are currently running and whose main object of research is either false information or hate speech.
4. *Quality Assessment*. The quality of the selected projects may be assessed based on practical relevance or target group feedback. As this step explicitly does not include the scientific quality (Gnewuch & Maedche, 2022) and the analysis' scope is of an empirical rather than normative nature, we exclude this step from our review.
5. *Data Extraction*. Applicable information is extracted from each project by examining the information provided by the relevant databases and search results based on our predetermined criteria.

6. *Documenting and Archiving*. The project information and any related material used as an additional source of information in the review are documented, stored, and archived in an Excel sheet.
7. *Data Analysis and Synthesis.* The evidence extracted from the included projects is collated, summarized, aggregated, organized, and compared, with the findings presented in a consequential manner. We aggregated related target groups to higher orders of abstraction (e.g., "scientists" and "researchers", or "users", "citizens" and "general society" to "users"), as well as for disciplines ("natural language processing", "computer and information science", and "computational linguistics" to "computer science and adjacent"). Further, we classify the non-research consortial partners according to NGOs and other non-profit organizations, for-profit organizations, and public bodies, drawing from the classification by the EU CORDIS database. Through an additional qualitative content analysis after Mayring (2015), the projects' main focal points, as addressed in their descriptions, are analyzed and compared. Proceeding inductively during the empirical analysis, relevant categories are derived directly from the project descriptions. This approach follows a conventional content analysis in which codes are defined during data analysis. The main focus lies on a synthetic creation of categories displaying complex content-related evidence instead of only functioning as markers for certain passages. By going through the material, former categories are either subsumed or a new category is formulated. After working through 50 percent of the data, all categories are revised and eventually reduced to main categories. Following Mayring's method of summary content analysis, the original material is summarized. The aim is to demarcate text elements without distorting the textual core of the data. Through this kind of reduction, more transparency shall be created that still corresponds to the material's basic form (Mayring, 2015). Table 1 displays the final category system applied for qualitative data analysis with distinct definitions of each code and respective anchor examples. The data for our analysis is available via OSF.

No	Category	Definition	Anchor Example
1	Tool Development	The project involves the development of a practical tool.	"The aim of the joint project [...] is to develop a software-based analysis tool that helps experts to better assess disinformation campaigns."
2	Digital Platforms	The project focuses on digital platforms and the online realm as a subject of research.	"Second, the project will address criminal liability in digital networks, not only for users but also for hosts and service providers."
3	Machine Learning	The project focuses on employing methods of machine learning.	"[The project] will integrate the structured knowledge provided by social and human sciences into natural language processing tools and deep learning algorithms, so as to develop new hybrid intelligence systems."
4	Policy Advice and Frameworks	The project focuses on formulating policy advice or theoretical (legal) frameworks for implementation.	"Finally, the project results will be made available in concrete recommendations for action for citizens, the media and politicians."
5	Social Media	The project involves research on social media or uses social media data.	"This project deals with punishable behaviors in social media, with a focus on expression offenses."
6	Qualitative Research and Mixed Methods	The project conducts qualitative research or employs mixed methods approaches.	"It will adopt a mixed-methods approach using archival and social media analysis, interviews, cross-sectional and longitudinal surveys, and experiments [...]."
7	Open Access	The project actively makes (part of) its results accessible to the general public.	"The dataset, the model, the training workflow, and the software for operating the service will be made openly available whenever possible and thus can be utilized for other subject areas as well."
8	Fact-Checking	The project focuses on applying or developing methods of fact-checking.	"For this purpose, they are trained in scientific and ethical working methods for well-founded fact-checking."
9	Science and Healthcare	The project focuses on topics such as health, medicine, and science.	"Thus, the project combines methods of transfer learning, information extraction, and fact-checking for the biomedical domain."

Table 1. Category system for content analysis following Mayring (2015). Categories are sorted by frequency.

3.4 Results

This chapter provides an examination of the primary findings from our study, focusing on the analysis of 79 identified projects that address false information and hate speech. The investigation is divided into two sections, Descriptive Analysis and Qualitative Content Analysis, each utilizing a different analytical approach to uncover key insights.

3.4.1 Descriptive Analysis

The following section presents a summary of the primary findings derived from a descriptive analysis of the characteristics of the 79 projects addressing false information and hate speech.

False Information

Overall, we identified 60 ongoing research projects regarding disinformation and related constructs. Of those, eight projects involve Information Systems researchers (i.e., professors or doctoral employees with a degree or PhD in Information Systems and/or work at an Information Systems institute), and further 23 projects involving researchers from adjacent disciplines such as Computer Science, Data Science, Information Science, or Computational Linguistics. Of the eight projects involving Information Systems, six are funded by the BMBF, one by the EU, and one by the DFG. Correspondingly, most of the institutions involved stem from Germany, and the EU project covers 15 countries. The projects run for three (BMBF, DFG) to five years (EU). The target groups of the involved projects are diverse, including authorities and organizations with security tasks (3), healthcare workers and the healthcare system (2), users (4), researchers and innovators (1), and platforms (1). Involved disciplines include Information Systems (8), Computer Science and adjacent (4), Communication Science (2), Information Science (1), Sociology (1), Economics (1), Law (1), and Ethics (1). Overall, the projects involve six non-profit organizations and eight for-profit organizations, most of which are software development or consulting companies, about half of which are part of one EU project, and the remaining from different BMBF projects. Of those eight projects involving Information Systems researchers, seven (87.5%) are interdisciplinary projects involving multiple of the disciplines outlined above.

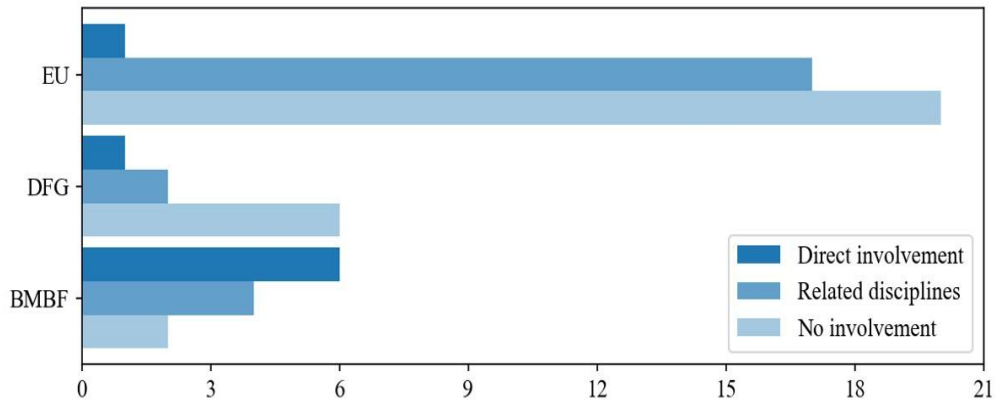

Figure 10. Sponsors of projects in the false information dataset by involvement of Information Systems.

Figure 10 shows the distribution of false information projects across the EU, the DFG, and the BMBF according to the involvement of disciplines related to Information Systems (Computer Science, Data Science, Information Science, or Computational Linguistics). We found 23 ongoing research projects from related fields. Of those, four are funded by the BMBF, 17 by the EU, and two by the DFG. The European projects cover more than 30 countries. The projects span two to five years and target researchers and innovators (10), citizens and the general public (9), human resources (2), health care workers (1), data analysts (1), journalists (1), news institutions (1), and authorities and organizations with security tasks (1). The projects involve 12 NGOs and 16 public, non-research organizations, many of which are public news institutions, organizations, or public bodies, such as ministries of interior or police, and NGOs for gender and sexual diversity organizations. Further, 54 for-profit organizations are involved, many of which are private news institutions. Most non-research partners are involved in European projects. Of those 23 projects, 17 (73.9%) are interdisciplinary.

Hate Speech

Through our analysis, we determined 19 ongoing projects connected to hate speech. Of those, only one includes Information Systems researchers and seven adjacent disciplines. The IS-related project is funded by the BMBF, takes action for three years until July 2026, and specifically targets investigative and law enforcement authorities. They interdisciplinarily combine Information Systems with Computer Science researchers and involve one for-profit organization for software development. A further seven projects include researchers from adjacent disciplines. One is funded by the DFG, and seven by the EU. They span from 1.5 (EU) to 5 years (EU) and involve researchers from 14 European countries. They target users (3), authorities (2), research (2), and community managers (1). Researchers stem from a variety of disciplines, such as Computer Science and similar fields (7), Communication Science (1), Political Science (2), Linguistics (2), Human-

Computer Interaction (1), and Humanities (1). Overall, there are two NGOs, nine for-profit organizations, and five public bodies involved. Out of those seven projects, five (71.4%) are interdisciplinary projects involving the disciplines listed above.

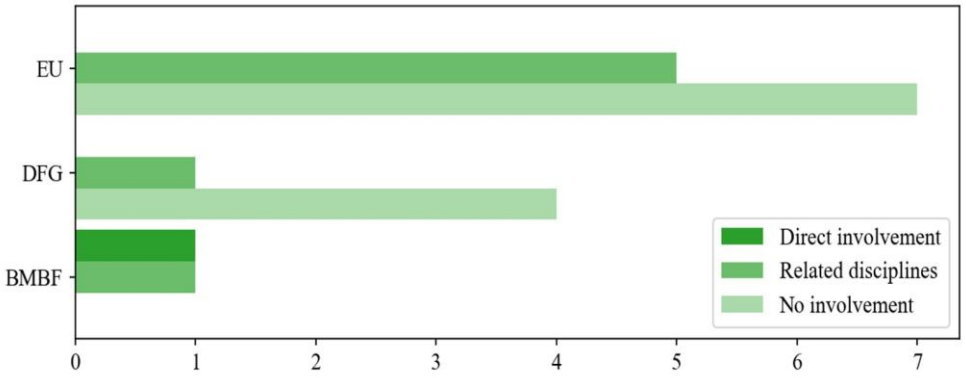

Figure 11. Sponsors of projects in the hate speech dataset by involvement of Information Systems.

Figure 11 depicts the distribution of hate speech projects among the EU, the DFG, and the BMBF according to the involvement of disciplines related to Information Systems. Of the eight ongoing projects from the field of Information Systems and adjacent fields, five are funded by the EU, two by the BMBF, and one by the DFG. Two of the EU-funded projects are registered only in Germany, one only in Italy, and the other two span 12 other European countries, targeting scientists (3), investigative and law enforcement authorities (3), social media users (2), online community and comment section managers (1), the general public (1), police authorities (1), and minority language users (1).

3.4.2 Qualitative Content Analysis

The following section presents a summary of the primary findings derived from a content analysis (Mayring, 2015) of the descriptions of the 79 identified projects addressing false information and hate speech.

False Information

Out of the 60 identified research projects, 21 projects focus on formulating policy advice and/or theoretical (legal) frameworks for implementation. Specifically, nine projects develop policy recommendations for national and international legislators and create new legal frameworks. The other 12 projects propose theoretical models or frameworks addressing notions of disinformation, related phenomena, and educational concepts. Additionally, 21 projects concentrate on developing practical tools. These include mobile applications for detecting manipulated content, analysis tools for experts, dashboards for discourse tracking, and collaborative platforms. Digital platforms are a common focus,

with 19 projects targeting them and specifically investigating social media (14). Here, the primary aim is to analyze the spread of disinformation, moderate digital networks, and detect manipulative content on online platforms and messenger services. Machine learning methods are employed in 18 projects to develop tools or analyze data, frequently using natural language processing for text categorization and information extraction systems. These approaches often include solutions for human-machine interaction. Public accessibility is a key concern for eight projects, which make their tools available via APIs and consider users with diverse backgrounds. Fact-checking is a focus for seven projects, combining automated and human-based methods. Another seven projects specifically target disinformation in science and healthcare, particularly concerning COVID-19, vaccinations, and pseudoscientific conspiracy theories. Lastly, six projects utilize qualitative methods or mixed-methods approaches, predominantly through expert interviews as well as content and discourse analyses. These qualitative methods are often combined with quantitative, computational approaches for comprehensive insights. Figure 12 depicts the frequency of codes applied in the dataset of projects on false information, offering a glimpse into the most prominent focal points within this area of research.

Figure 12. Distribution of codes in the false information dataset.

In examining the role of the IS discipline within this research area, we observed that out of nine projects on false information, the majority focus on developing tools (8) and applying machine learning methods (8), rather than creating theoretical frameworks or policy advice (1). These projects often investigate digital realms (5) and social media (4), with some effort to make results open access (4). Fact-checking methods (0) and qualitative or mixed-methods approaches (1) are rarely included. While two projects focus on science and health, most (7) adopt a holistic, domain-independent perspective on false information.

Hate Speech

Among the 19 research endeavors focusing on the topic of hate speech, eight projects employ machine learning methods, primarily using natural language processing and deep learning for detecting hate speech and analyzing digital hate. Seven projects focus on digital platforms, with three of them specifically targeting social media. These studies primarily analyze the occurrence and spread of digital hate and political hostility, as well as their implications for criminal liability, frequently mentioning Facebook, Telegram, and X (formerly Twitter). Six projects involve developing tools such as AI-based tools for managing online communities, and dashboards as well as browser extensions for analyzing cyber abuse content. Five projects apply qualitative or mixed-methods approaches, using interviews and discourse analysis, often combined with computational analysis. Two projects aim to make their results accessible to the general public, offering them free of charge and focusing on "low-resource" countries. Finally, one project focuses on creating policy advice, proposing a model of accountability mechanisms guided by a civic code of conduct. Figure 13 displays the frequency of codes applied in the dataset of projects on hate speech, providing insights into the most prevalent focal points within this area of research.

Figure 13. Distribution of codes in the hate speech dataset.

Among hate speech research projects, the only one involving IS researchers focuses on digital platforms and social media, developing a tool for detecting and addressing cyberbullying and hate speech. Unlike other projects that use machine learning and qualitative or mixed methods, this project lacks specific methodological details, though it mentions a participatory development process.

3.5 Discussion

Comparing IS projects to the broader landscape of initiatives addressing false information and hate speech in our dataset reveals distinct trends and gaps within the discipline. IS research prominently addresses these issues by developing digital tools and focusing on digital environments. This technological focus has led to the creation of various digital artifacts, such as applications and dashboards, designed to detect and mitigate the spread of false information and hate speech. However, this emphasis on practical, digital solutions has the potential to overshadow the development of theoretical outcomes, such as policy advice or educational frameworks, which are crucial for a holistic approach to these problems. Moreover, the methodological approaches within the IS discipline show a clear preference for quantitative, macro-level studies, frequently employing analysis of big data. This preference results in a limited adoption of qualitative methods, which are essential for understanding the nuanced, human aspects of how false information and hate speech propagate and affect individuals and communities. Our examination of ongoing projects in Germany and the European Union highlights that while there are numerous initiatives addressing false information and hate speech, the involvement of IS research remains relatively limited. Instead, many of these projects are driven by the field of Computer Science, with a strong emphasis on algorithm development. This indicates a significant opportunity for IS researchers to contribute more robustly to the current discourse and efforts against false information and hate speech. The interdisciplinary nature of IS, which inherently blends technological and social perspectives, positions it uniquely to address these complex issues. This is underlined by our identified IS projects being more frequently interdisciplinary projects than those involving related disciplines, although the sample size is small. By incorporating socio-technical perspectives, IS research can bridge the gap between purely technical solutions and the broader societal implications. This involves integrating insights from ethics, law, and other relevant fields to effectively evaluate and implement mechanisms and countermeasures in real-world applications, particularly within governmental and regulatory authorities.

Despite the current limitations, the projects addressing false information and hate speech cover a wide variety of target groups and countries, underlining the global importance of these issues. This diversity in focus underscores the need for comprehensive solutions that are adaptable to different cultural and social contexts. The IS discipline's strong focus on technological solutions provides valuable tools for combating false information and hate speech. However, to enhance the impact of this research, there is a critical need to integrate theoretical frameworks, policy advice, and qualitative methods. By embracing a more balanced and interdisciplinary approach, IS researchers can make significant contributions to building resilient democracies. These democracies would be better informed, more inclusive, and more capable of countering the challenges posed by false information

and hate speech in the digital age. Eventually, the IS discipline should feel encouraged to heed the call for action, particularly in the area of hate speech, where its contributions have been sparse. By leveraging its interdisciplinary strengths and adopting a socio-technical perspective, IS research can not only advance the understanding of false information and hate speech but also develop more effective strategies to combat these issues, ultimately fostering a more informed and cohesive society.

3.6 Conclusion

To build and preserve resilient democracies, it is essential to evaluate the current state of publicly funded research on false information and hate speech. By mapping out existing efforts, we can identify gaps and areas requiring further exploration. This evaluation may guide future projects, ensuring they address the most pressing issues and contribute effectively to preserving and promoting democratic values. Our project review presented in this paper reveals that the IS discipline's current research landscape on false information and hate speech, while interdisciplinary, is heavily oriented toward technological solutions, with an emphasis on digital tools and machine learning. While this reflects the discipline's strengths, there is a notable gap in theoretical, policy-oriented, and qualitative research. Addressing these gaps could lead to more comprehensive strategies for combating false information and hate speech, ultimately fostering a more informed and sage digital democracy. Additionally, Information Systems as a discipline is underrepresented in projects funded by the DFG and the EU, implying there are still opportunities for IS to be involved in other types of projects. Finally, hate speech is rarely researched in projects by Information Systems researchers, although as a discipline, we might be able to provide valuable insights for theory and practice.

The insights provided by this research have some minor limitations. For practical reasons, only publicly funded projects listed in the BMBF, EU, and DFG databases could be taken into consideration. Still, these funding sources cover the most important organizations (Hornbostel, 2001). Additionally, this research adopted a particular emphasis on Germany and the EU. Expanding the geographic focus, especially towards the global south, would be beneficial in capturing a more diverse range of projects and insights.

Reflecting on the call by Weinhardt et al. (2024) to establish novel areas for Digital Democracy research, there is a clear need for IS researchers to broaden their focus beyond technological solutions to include the exploration of how digital platforms influence human behavior and social cohesion. Interdisciplinary research across Information Systems, Computer Science, Political Science, Sociology, Communication Science, and Law is crucial to understanding and mitigating the broader negative impacts of platforms in our

democracies. IS researchers are encouraged to prioritize transparency, inclusion, and literacy, developing innovative ways to preserve and promote democratic values. By focusing on trust, the foundation of democratic engagement, researchers can examine how misinformation, disinformation, malinformation, and hate speech influence the political landscape and public trust.

4 Decoding Deception: A Taxonomy of Online Disinformation in Data Classification[3]

4.1 Introduction

As today's primary news sources, social media and news platforms suffer from inaccurate reporting and the distribution of unfounded opinions (Shu et al., 2017). Especially in times of crises, the viral spread of disinformation poses a central threat to political processes and social cohesion, as the United Nations recently addressed in their disinformation report (United Nations, 2022). Disinformation is defined as false information and, unlike misinformation or malinformation (Wardle, 2019), is spread with the intention to deceive (Shu et al., 2020a). Therefore, automated systems detecting disinformation on digital platforms are indispensable tools in the ongoing effort to maintain the integrity of information, protect democratic processes, and foster a more informed and cohesive society. Research on disinformation detection using machine learning (ML) and natural language processing (NLP) is a rapidly expanding field that spans various disciplines, including computer science, social science, psychology, and information systems (Azevedo et al., 2021; Mahyoob et al., 2020; Yu & Lo, 2020). Most techniques focus on extracting multiple features, incorporating them into classification models, and then choosing the best classifier based on performance (Alsaidi & Etaiwi, 2022; Bozarth & Budak, 2020). Data suggests that disinformation content is difficult to identify (Kapantai et al., 2021) due to a variety of stylistic devices used in disinformation, creating a barrier for purely quantitative approaches to the problem (Rosińska, 2021). The deceptive nature of disinformation, where the aim is to make the information appear to be authentic, may help to

[3] This chapter comprises an article that was published by Isabel Bezzaoui, Jonas Fegert and Christof Weinhardt in the following outlet with the following title: Truth or Fake? Developing a Taxonomical Framework for the Textual Detection of Online Disinformation. In International Journal on Advances in Internet Technology, 15 (3/4), 53-63, 2022, and an article currently under revision by Isabel Bezzaoui, Pavlos Fafalios, Jonas Fegert, Achim Rettinger and Konstantin Todorov in the following outlet with the following title: Decoding Deception with TAXODIS – A Taxonomy of Disinformation Cues for Fine-Grained Text Labeling. In Semantic Web Journal, 2025. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated.

explain this difficulty (Abonizio et al., 2020). Nevertheless, empirical evidence on the structure of disinformation demonstrates that legitimate and deceptive content differ significantly in their substance and sentiment (Hamed et al., 2023; Horne & Adali, 2017). Thus, recognizing the need for a comprehensive understanding, this research delves into the clustering of linguistic features, creating a robust foundation for the empirical training of detection models. Accordingly, we are guided by the following research question:

How can a taxonomy of online disinformation characteristics be designed to facilitate text classification in automated disinformation detection?

We further specify three related sub-questions (*RQs*):

- *RQ1: What linguistic cues of (online) disinformation are reported in the empirical literature?*
- *RQ2: How can these cues be clustered and structured into a taxonomy?*
- *RQ3: How can this taxonomy be made available to facilitate automated detection of disinformation?*

In doing so, we aim to contribute to a shared understanding of disinformation at a linguistic level, providing a nuanced perspective that goes beyond conventional binary detection methodologies. The focal point of this paper is the development, demonstration, and evaluation of the Taxonomy of Online Disinformation (TAXODIS). Proposing a structured taxonomy as a tool for automated detection systems offers scientific guidelines for a more fine-grained annotation of disinformation datasets for training classifiers. We ground the construction of this taxonomy in the principles and technology of the semantic web, offering means to publish and maintain shared and actionable resources of knowledge.

The paper is structured as follows. In section 2, we review related work before giving an overview of TAXODIS in section 3. The methodology of building the taxonomy is given in section 4, while examples of using and linking the resources to existing knowledge graphs are provided in section 5. Several use case scenarios are presented in section 6, before we conclude in section 7.

4.2 Related Work

Recent research addresses both the benefits and drawbacks of different detection methods, as well as their underlying theories (Ansar & Goswami, 2021; Rohera et al., 2022; Zhou et al., 2019). Nevertheless, many disinformation classifiers presented in empirical papers lack explanations on how they were trained or how the datasets used for training were labeled (Akinyemi et al., 2020; Fayaz et al., 2022; Lasotte et al., 2022). Although these explanations are crucial to the transparency and traceability of the research process,

only little research has accounted for this issue (Meel & Vishwakarma, 2020; Molina et al., 2021). Creating a succinct taxonomy that covers the wide-ranging attributes of disinformation regardless of the specific event while also being detailed enough to precisely categorize deceptive content may enhance the transparency of the manual classification process of disinformation datasets.

In the past years, there have been various endeavors to capture the phenomenon of disinformation with taxonomical frameworks. Alexander and Smith (2010) base their approach to taxonomy development on a communication model to illustrate how disinformation is spread to deceive its audiences. While they discuss illustrative examples of different strategies for modifying or distorting messages to subvert their initial meaning, the authors do not suggest a concise taxonomy providing a structured overview of indicators that help identify disinformation in social media. Tambini (2017), on the other hand, provides generic categories that lead to overlapping definitions. The proposed categories encompass a wide range of sociopolitical phenomena such as "falsehood to affect election results" and "news that challenges orthodox authority". These aspects primarily serve a descriptive rather than explanatory purpose, implying a need for more precision in classification. Parikh and Atrey (2018) delineate disinformation features by relying on technical attributes or the structural format of news items. These categories encompass visual elements such as photoshopped images, user-based components involving fake-accounts, and style-based aspects, among others. Their technical approach primarily introduces types of data in news, disinformation detection methods, and common disinformation datasets. While this approach proves valuable for developing automated detection tools, its technical orientation poses challenges when attempting to integrate it with broader frameworks equally focused on non-technical aspects of disinformation. In adopting a detection-oriented approach to the issue, Kumar and Shah (2018) present four broad categories: opinion-based, fact-based, misinformation, and disinformation, without delving into the finer nuances of the domain, such as clickbait, propaganda, and trolling. Their focus is limited to specific domains and they position the terms *disinformation* and *misinformation* at a more granular level, in contrast to the common practice of treating them as overarching umbrella terms. In their taxonomy, Lemieux and Smith (2018) categorize disinformation and misinformation alongside more specific phenomena like hoaxes and rumors, placing them at a similar hierarchical level. Furthermore, they introduce the term "mal-information" as an overarching category, on par with disinformation and misinformation. This approach makes it difficult to assign sub-phenomena, such as conspiracy theories, to overarching phenomena, such as disinformation. Molina et al. (2021) differentiate various types of disinformation by employing four operational indicators: message, source, structure, and network. This approach extends beyond content-based methods and conventional definitions, instead centering on the dissemination of online information and offering insights into potential detection solutions. Their study

provides an extensive overview of the characteristics of fabricated news. However, the proposed taxonomy lacks concision, resulting in nine extensive tables that are neither precise nor concise enough for handling large amounts of data (Nickerson et al., 2013). Kapantai et al. (2021) have designed a succinct taxonomy framework characterized by three fundamental dimensions: motive, facticity, and verifiability. These dimensions and their associated metrics prove crucial in the categorization of disinformation that has been previously identified as such, enabling differentiation between specific manifestations such as clickbait, trolling, and fake reviews. It is essential to note, however, that this taxonomy does not furnish discernible indicators intended to facilitate the proactive identification of disinformation content by human users. Finally, the DISARM framework provides an overview of several sub-frameworks for practitioners to describe and understand different parts of disinformation, including its actors, tactics, and countermeasures. While the framework is intended to help track and counter misinformation (DISARM, 2023), it does not provide a hands-on and scientifically grounded scheme that can be applied to the recognition of disinformation via granular features and characteristics referring to language and content.

None of the mentioned efforts above propose a shared semantic model that would help lead toward a uniform and common understanding of the various categories of features. In that respect, several structured datasets with schemas have been proposed to deal with the specific task of fact-checking or disinformation detection. The MultiFC (Augenstein et al., 2019) and the ClaimsKG (Gangopadhyay et al., 2023, 2024; Tchechmedjiev et al., 2019) datasets both provide structured data of and about claims coming from established fact-checking portals, where claims are stored together with contextual metadata (such as authors, sources, claim reviews and other contextual information, including veracity labels). The two datasets are complementary in some respects. MultiFC focuses on evidence-based fact-checking in terms of downstream tasks, where via the Google Search API the ten most highly ranked search results per claim are retrieved and stored. ClaimsKG, on the other hand, provides a rich data model (an RDFS ontology) to represent check-worthy or fact-checked claims and related metadata, which is an important effort towards standardization and enables federated access to distributed data, where a specific search engine is provided[4] in addition to a public Sparql endpoint (Gasquet et al., 2019). MultiFC contains data in English, while ClaimsKG is multilingual, harvesting data from fact-checking portals in about 10 languages. These datasets can be used to provide a pool of verified claims with additional metadata for fact-checking applications and to extract links to claims that are mentioned in fact-checking articles. However, they do not delve

[4] https://data.gesis.org/claimskg-explorer/home

into the problem and nature of the linguistic and textual features that define disinformation.

In these terms, an important effort for annotating text with general linguistic features is the Linguistic Inquiry and Word Count tool (LIWC). LIWC is a gold standard for word-level text analysis, which has been used in large amounts of scientific publications[5]. It has also proven to be well-suited for web claim-related tasks (e.g., Martinez-Rico et al. (2022) ranked second at the CheckThat! 2022 Fake News Detection Challenge and used LIWC in their pipeline). LIWC extracts features by using over 100 built-in dictionaries that encompass social and psychological states, emotional tones, linguistic properties, cognition processes, analytic speech patterns, punctuation marks, and several word-count-related features. Each dictionary can contain a list of words, a list of word stems, emoticons and other specific word constructions. The LIWC features can be divided into seven distinct categories: syntactic, analytic, sentiment, social, perceptual, informal language, and topic. However, although useful in claim-related analyses for disinformation detection, LIWC has a more general focus. A specific subset of its features can be used to annotate disinformation-related data, but this selection has to be made manually, where this is additionally hindered by the fact that the vocabulary is not formally structured and queryable. In addition, access to LIWC is granted upon request, making it less easy to apply, as it is not openly available. In contrast, the proposed taxonomy in this paper is tailored to disinformation in particular, contains more specific and fine-grained categories and types of features for related downstream tasks, in addition to it being fully open and structured following the semantic web principles.

The current state of the art shows that what is missing so far is a fundamental but concise empirical overview of linguistic detection cues supporting the creation of labels for transparently annotating datasets on a granular level. By implementing a taxonomy encompassing such an overview, a classifier not only produces an output providing indications of content veracity but also furnishes more comprehensive information about prevalent characteristics in disinformation. The novel taxonomy is shaped and made openly available as a (SKOS-based) RDFS resource, which enhances re-usability, interoperability and fairness in general, with advantages such as easy access and federated queries over the vocabulary and the annotated datasets. Finally, this approach aims to enhance digital literacy among both annotators and end-users of the developed classifier.

[5] See https://www.liwc.app

4.3 Taxonomy Overview and Open Availability

Figure 14 depicts the TAXODIS taxonomy. The taxonomy contains (currently) 66 concepts, of which 48 are *leaf* concepts (concepts with no narrower terms), organized in a hierarchical (tree-like) structure of maximum depth four. Its top concept is *disinformation characteristic*, which describes characteristics that are indicative of disinformation in a piece of content. This top term has three narrower terms: i) *detection feature*, which classifies the piece of content based on linguistic or stylistic features that are indicative of the detection of disinformation (e.g., length of the headline, lexical and contentual poorness, level of semantic incoherence, lack of new information, level of topicality, etc.), ii) *categorization*, which classifies the piece of content based on its theme or content type (e.g., social (theme), conspiracy theory (content type)), and iii) *veracity*, which classifies the piece of content based on its veracity (e.g., mostly false, mixture, etc.). A detailed explanation of the narrower terms of these three broad terms is provided in the next section.

We implemented the TAXODIS taxonomy as a SKOS vocabulary/thesaurus. SKOS[6] is a data model designed for the representation of thesauri, classification schemes, taxonomies, and other types of controlled vocabularies. It is a W3C recommendation built upon RDF and RDFS, and its main objective is to enable easy publication and use of controlled vocabularies across the web. The SKOS representation of TAXODIS provides for each term/concept: i) its preferred label in English (using the property *skos:prefLabel*), ii) its definition in English (using the property *skos:definition*), iii) its broader terms, if any (using the property *skos:broader*), iv) its narrower terms, if any (using the property *skos:narrower*), v) its notation, used to uniquely identify the term within the scope of a given concept scheme (using the property *skos:notation*), vi) the scheme (vocabulary/thesaurus) in which the term belongs to (using the property *skos:inScheme*).

[6] https://www.w3.org/2004/02/skos/

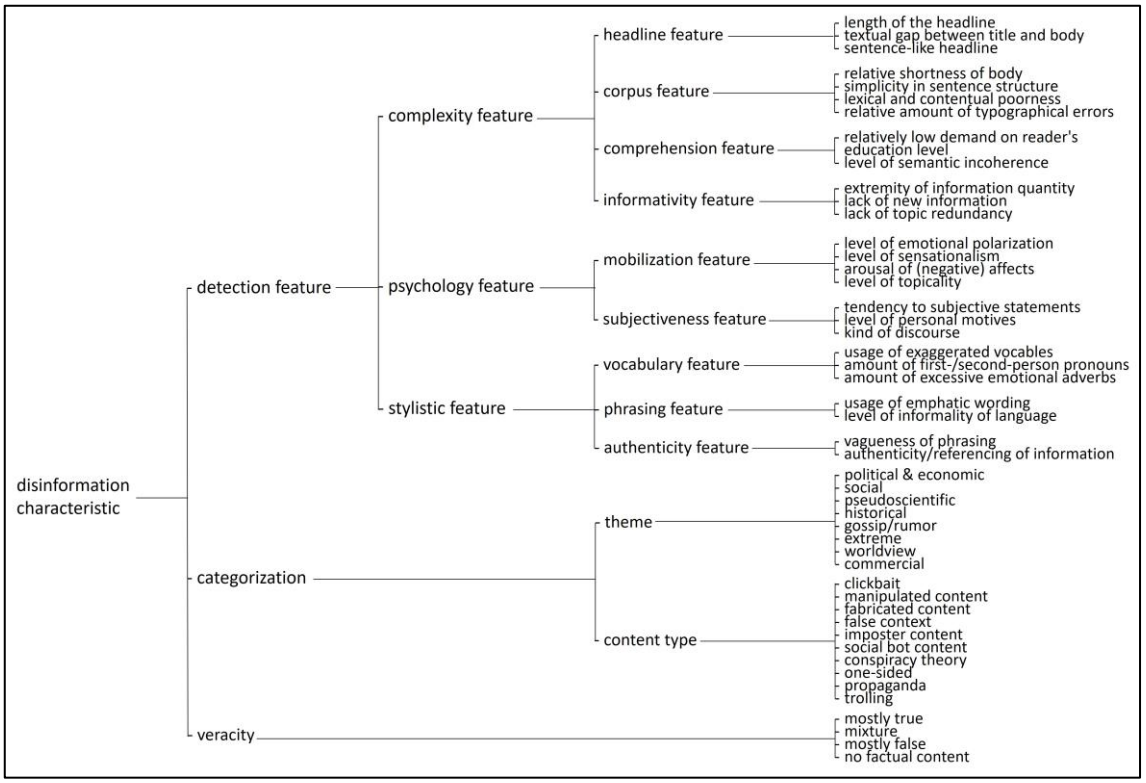

Figure 14. The TAXODIS taxonomy.

We also provide the following metadata using properties of RDFS, DCMT (Dublin Core Metadata Terms) and other widely-used vocabularies: i) the title of the taxonomy (using the properties *rdfs:label* and *dct:title*), ii) the description of the taxonomy (using the properties *rdfs:comment* and *dct:description*), iii) the taxonomy's usage license (using the properties *dct:license* and *cc:license*), iv) the taxonomy's creation date (using the property *dct:issued*), v) the taxonomy's last modification date (using the property *dct:modified*), vi) the taxonomy's version (using the properties *owl:versionInfo* and *owl:versionIRI*), vii) the creators of the taxonomy (using the property *dct:creator*), viii) the taxonomy's namespace URI (using the property *Vann:preferredNamespaceUri*), and ix) the taxonomy's namespace prefix (using the property *vann:preferredNamespacePrefix*). The RDFS file (in Turtle format) of the SKOS implementation of TAXODIS is publicly available under a creative commons license at: https://zenodo.org/records/14264593 (DOI: https://doi.org/10.5281/zenodo.14264593). The (resolvable) namespace of the taxonomy is https://hop.fzi.de/taxodis/.

4.4 Building TAXODIS, the Taxonomy of Online Disinformation

4.4.1 Methodology

Our iterative approach consists of two major parts, integrating insights from multiple disciplines to construct a robust taxonomy. Initially, by conducting a systematic literature review (Webster & Watson, 2002), we gather a comprehensive range of linguistic features of online disinformation from various fields of study. This allows us to capture diverse perspectives on how disinformation manifests across different contexts. Subsequently, we cluster the empirical results in groups, supporting a linguistic-based disinformation detection approach. Categorizing objects aids in understanding and analyzing complex environments, making the creation of taxonomies essential for research and development (Nickerson et al., 2013). Nickerson et al. (2013) provided the first and well-conceived taxonomy-building methodology. Their approach has served as a blueprint for numerous taxonomy projects across various domains (Kundisch et al., 2022). Building on these interdisciplinary foundations, we propose a novel six-dimensional taxonomy based on the categorization criteria identified from the existing empirical literature.

4.4.1.1 Systematic Literature Review

To comprehensively address our first research question, we conducted a systematic literature review following Webster and Watson's (2002) methodological guidelines. A thorough review encompasses pertinent literature on the subject and is not confined to a particular research approach, set of journals, or geographical area (Webster & Watson, 2002). Hence, we utilized large interdisciplinary databases to access all relevant research fields for our project. Upon careful examination of the literature concerning linguistic features and disinformation detection characteristics, we synthesized an overview of frequently used descriptions referring to various types and characteristics of disinformation content. However, the ad hoc definitions introduced by each study may give rise to conflicts or overlaps. Accordingly, the overarching objective of our literature review is to consolidate the existing knowledge on categorizing disinformation and to discern patterns and key concepts within the literature. Our aim is to advance prior research by synthesizing this knowledge into a cohesive taxonomy. To achieve this goal, we followed a structured procedure for our review: Initially, we identified our sources from digital libraries and defined our search terms, which were subsequently applied to the selected sources. Afterward, we refined our selection of primary studies by employing inclusion and exclusion criteria on the search results. To further enhance the comprehensiveness of our review, we conducted both backward and forward searches based on the selected primary studies. An automated search was executed across five prominent scientific databases to

identify relevant publications: IEEE Xplore Digital Library, Scopus, ACM Digital Library, Web of Science, and Springer Link. Initially, we conducted several pilot searches on our research topics to compile a preliminary list of papers. Based on these searches, we defined search terms that aligned with our research objectives. The selected search phrases, limited to abstract and title, were as follows: linguistic 'disinformation' OR 'fake news' AND 'classification' OR 'detection'. For the next phase of our research, the following three inclusion and exclusion criteria were formulated: We excluded sources that solely address the issue of disinformation from a computational perspective, advocating technical solutions reliant on machine learning and statistical models to automatically categorize news articles into predefined categories, such as fake or real. Additionally, we omitted sources that primarily conducted performance evaluations of such models. Publications that mention specific categories or characteristics of false information without attempting systematic classification or providing explanations for the proposed categories were excluded. This criterion was applied to sources where the disinformation phenomenon is not a central concept, such as papers that incidentally use terms like 'fake news', or those that discuss specific types of false information without integrating them into a comprehensive framework, rendering them non-exhaustive or merely indicative. In the interest of promoting common scientific understanding, only papers written in English were included in our review. Our search yielded 29 primary studies across six different disciplines (e.g., computer science, linguistics, psychology, and media studies) introducing linguistic frameworks for disinformation detection. The selection process encompassed records obtained through database searching as well as those identified through additional backward and forward searches based on the initial records.

Figure 15 provides a detailed overview of the selection process, encompassing records obtained through database searching as well as those identified through additional backward and forward searches based on the initial records.

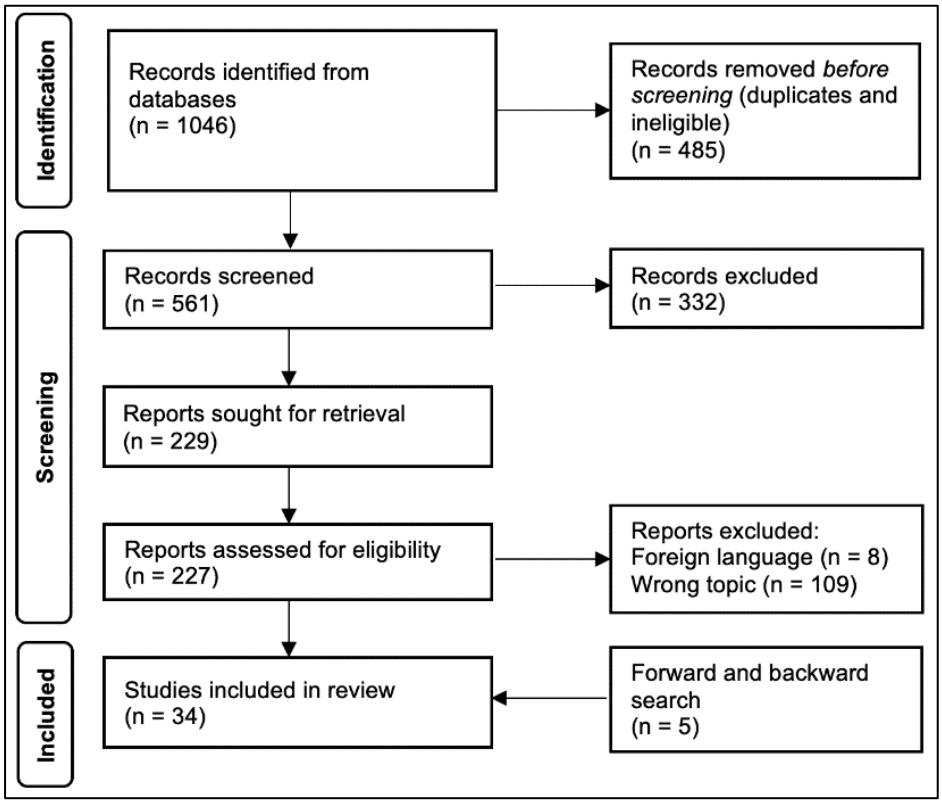

Figure 15. PRISMA flow diagram.

In total, 34 papers were included in our review. Our initial objective was to identify linguistic cues of online disinformation in the empirical literature (RQ1). Subsequently, we extracted the identified features of disinformation and organized them into clusters based on similarity to prepare our findings for addressing RQ2.

4.4.1.2 TAXODIS' Features

Our overall goal is to create a taxonomy of online disinformation that helps create a common understanding of what constitutes disinformation from a linguistic viewpoint, provides a list of categories and detection characteristics and can be used to develop labels that can be applied to diverse datasets (RQ3). After examining the findings from RQ1, we clustered them along their similarities into a schema (RQ2), considering a more granular level of the proposed features from the literature. We observed many commonalities but also differences at both the category and dimension levels. In order to make sense of the patterns and contradictions, we applied several general rules during the processing of the data. First, we removed types and definitions that are either too generic (e.g., yellow press) or too technical (e.g., deep fakes). Second, we removed duplicates and synonyms to avoid repetitions and overlaps. Lastly, any types and definitions that were incorrectly

categorized as disinformation (e.g., misinformation) were removed. After our fifth iteration, we did not identify any new characteristics and dimensions from the reviewed studies.

Our final framework (Table 2) consists of six dimensions, since, in our case, all ending conditions (Nickerson et al., 2013) were satisfied. The first dimension covers complexity features (1) that help to evaluate the complexity and readability of the text, splitting into headline, corpus, comprehension, and informativity. It allows TAXODIS users to evaluate the informational content and textual structure of the content under consideration. Our second dimension contains psychology features (2) that describe attitudes, behaviors, and emotions. This dimension, which splits into mobilization and subjectiveness, aids in illuminating and quantifying the cognitive process and individual concerns that underlie the writings. We added a third dimension, stylistic features (3), to reflect the writer's style and the syntax of the text, such as the number of verbs and nouns used, as well as the use of specific terminologies. This dimension splits into vocabulary, phrasing, and authenticity. The fourth and fifth dimensions help to categorize disinformation content, as themes (4) contain categories such as *pseudoscientific* or *historical*, and content type (5) allows differentiating between different types of content. Moreover, disinformation content can differ strongly in its deceitfulness. For this reason, our last dimension accommodates grades of veracity (6) to facilitate the evaluation of different kinds of disinformation corresponding with our fifth dimension. Below, we provide details for the individual features.

meta-characteristic	dimension	subdimension	feature	code	characteristic value	
					0	**1**
detection	complexity features	headline	length of the headline	headlength	low	high
			textual gap between title and body	headgap	no	yes
			sentence-like headline	headsent	no	yes
		corpus	relative shortness of body	corpshort	no	yes
			simplicity in sentence structure	corpsimpl	no	yes
			lexical and contentual poorness	corplex	no	yes
			relative amount of typographical errors	corperror	low	high
		comprehension	relatively low demand on reader's education level	compeduc	no	yes
			level of semantic incoherence	compincoh	low	high
		informativity	extremity of information quantity	infoextrem	low	high
			lack of new information	infonewinfo	no	yes
			lack of topical redundancy	infotopredun	no	yes
	psychology features	mobilization	level of emotional polarization	mobpolar	low	high
			level of sensationalism	mobsensat	low	high
			arousal of (negative) affects	mobaffect	low	high
			level of topicality	mobtopical	low	high
		subjectiveness	tendency to subjective statements	subjtenden	low	high
			level of personal motives	subjmotiv	low	high
			kind of discourse	subjdiscours	knowledge-based	opinion-based
	stylistic features	vocabulary	usage of exaggerated vocables	vocexagg	no	yes
			amount of first-/second-person pronouns	vocpronoun	low	high
			amount of excessive emotional adverbs	vocadverb	low	high
		phrasing	usage of emphatic wording	phrasemph	no	yes
			level of informality of language	prhasinformal	low	high
		authenticity	vagueness of phrasing	authvague	low	high
			authenticity/referencing of information	authrefer	frequently referenced	poorly referenced
categorization	themes	political & economic		thempoleco	no	yes
		social		themsoc	no	yes
		pseudoscientific		themscience	no	yes
		historical		themhisto	no	yes
		gossip/rumor		themgoss	no	yes
		extreme		themextrem	no	yes
		worldview		themworld	no	yes
		commercial		themcommer	no	yes
	content type	clickbait		typclick	no	yes
		manipulated content		typmanipul	no	yes
		fabricated content		typfabric	no	yes
		false context		typfalse	no	yes
		imposter content		typimpost	no	yes
		social bot content		typbot	no	yes
		conspiracy theory		typconspir	no	yes
		one-sided		typonesid	no	yes
		propaganda		typpropa	no	yes
		trolling		typtroll	no	yes
veracity grade	mostly true			vtrue	no	yes
	mixture of true and false			vtruefalse	no	yes
	mostly false			vfalse	no	yes
	no factual content			vnofact	no	yes

Table 2. The TAXODIS Taxonomy of Disinformation.

Complexity Features

Headline. Unreliable sources try to convey as much information as possible in the title to draw the reader's attention. Thus, they use a higher amount of plain text or words in the headline (Gruppi et al., 2018) and often display a lower textual similarity between the body of the article (Biyani et al., 2016). Titles of fake content often present sentence-like claims about people and entities associating them with actions (Fernandez, 2019; Horne & Adali, 2017).

Corpus. Unreliable sources tend to have a lower level of plain text or number of words in relation to real articles (Kumar & Shah, 2018), and their sentences exhibit a lower complexity in structure and a relatively low amount of words (Gruppi et al., 2018; Horne & Adali, 2017). Fake articles tend to have less diversity at the lexical and content level (Azevedo et al., 2021) and empirically exhibit a higher amount of typographical errors (Zhou et al., 2004).

Informativity. Fake articles often correspond with either a considerably low amount of information or a remarkable overload of information (Zhou et al., 2019). The body of fake articles adds relatively little new information but serves to repeat and enhance the claims made in the title (Azevedo et al., 2021; Horne & Adali, 2017). Valid articles about a particular topic contain several direct or indirect references to this subject. One can interpret those as a kind of contextual redundancy which fake sources are usually missing (Badaskar et al., 2008).

Psychology Features

Mobilization. Unreliable sources tend to use more emotionally persuasive language in general, leading to high levels of emotional polarization (Ribeiro Bezerra, 2021; Wang et al., 2019). Providing sensationalist content, fake articles tend to be written in a hyperbolic way to attract the reader's attention, i.e., with high usage of all-caps words or exclamation marks (Gruppi et al., 2018; Jeronimo et al., 2019). To cause an arousal of (negative) affects, fake content uses a higher degree of words related to emotional actions, states, and processes (Azevedo et al., 2021; Markowitz & Hancock, 2014). Legitimate sources tend to report on past events, whereas fake articles often focus on highly recent topics (Fernandez, 2019).

Subjectiveness. Exhibiting a tendency to subjective statements, fake articles are often written from a more personal view (Jeronimo et al., 2019). Creators of fake content are frequently driven by personal motives like raising profit, promoting ideology, and psychological aims (Kapantai et al., 2021). Words and expressions of fake articles relate to a more argumentative discourse aiming to convince the reader of a specific point of view (Azevedo et al., 2021).

Stylistic Features

Vocabulary. Unreliable sources more often use hyperbolic words such as superlatives and subjectives (Fernandez, 2019; Mahyoob et al., 2020) and display more first-person and second-person pronouns than legitimate articles (Fernandez, 2019; Rashkin et al., 2017). To lure readers to the content, disinformation displays a higher amount of excessive emotional adverbs (Biyani et al., 2016; Mahyoob et al., 2020).

Phrasing. Unreliable sources use a high level of exclamation marks, swear words, and visual references, and are slightly more prone to emotional tones and higher polarity (Azevedo et al., 2021; Ribeiro Bezerra, 2021). The language of fake content tends to be less formal than reliable articles (Horne & Adali, 2017).

Authenticity. Fake articles use a higher amount of vague phrasing or hedging words to achieve a more indirect form of expression (Mahyoob et al., 2020), while legitimate sources are considerably better referenced than unreliable articles (Kumar et al., 2016).

Themes

The category *political and economic* refers to content about specific politicians, or legal, political or economic actions. Content about social events, activists, public benefit, and minority organizations, as well as dangers or threats to human and animal health, is incorporated in the *social* category. *Pseudoscientific* content calls on supposedly scientific research or reputable institutions without identifying concrete sources or by manipulating them to create a false theory. Content about historical events or the distant past of public figures is subsumed under the theme *historical*. In addition to that, gossip or rumors may be spread about public figures without a political or activist profile. *Extreme* themes cover drastic, catastrophic or brutal events. The feature *worldview* is applied to content about religion, faith, and spiritual figures as well as various non-religious ideologies, views, and beliefs. Themes can also be *commercial*, such as false product reviews, advertising campaigns, or commercial clickbait aimed at accumulating views, likes, and comments (Rosińska, 2021).

Content Type.

Clickbait refers to sources that intentionally use exaggerated, misleading, or unverified headlines or thumbnails to attract readers to open the webpage (Kapantai et al., 2021). *Manipulated content* involves altering information or an image to deceive the recipient, who receives it without being aware of its misuse (Wardle & Derakhshan, 2017). *Fabricated content* encompasses entirely false stories lacking a factual basis, with the intent to deceive and cause harm. Particularly severe forms of fabrication mimic the style of legitimate news articles to mislead recipients (Kapantai et al., 2021). Real information may

be presented in a false context, where the recipient acknowledges its truth but remains unaware that the context has been altered (Wardle & Derakhshan, 2017). *Imposter content* involves genuine sources being impersonated by false, made-up sources to support a false narrative. This can include abusing a journalist's name, a logo, or a website (Kapantai et al., 2021). A *social bot* is a computer algorithm that automatically produces and posts content, interacting with legitimate users and other bots to emulate and possibly alter their behavior (Ferreira et al., 2022; Wang et al., 2019). *Conspiracy theory* applies to stories without a factual basis that usually explain important events as secret plots by governments, powerful groups, or individuals (Kapantai et al., 2021). *One-sided content* is heavily biased, promoting division and polarization. It features imbalance, inflammatory, and emotionally charged information, often containing a mix of true and false or mostly false details (Kapantai et al., 2021). *Propaganda* is information created by a political entity to influence public opinion and gain support for a public figure, organization, or government (Tandoc et al., 2018). *Trolling* is the intentional posting of offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation (Kapantai et al., 2021).

Grade of Veracity

Following Potthast et al. (2017), *mostly true* indicates that a piece of content is based on factual information and accurately depicts it. This rating excludes unsupported speculation or claims. *Mixture of true and false* describes content with some accurate and some inaccurate elements. It applies when speculation or unfounded claims are combined with real events, numbers, or quotes. *Mostly false* is used when the majority or all of the information in a content piece is inaccurate. This rating also applies when the central claim is false. *No factual content* is for posts expressing pure opinion, comics, satire, or anything without a factual claim. This adopted gradation follows a similar approach to knowledge graph 'ClaimsKG' (Tchechmedjiev et al., 2019), where the different veracity labels are mapped to four basic categories (i.e., true claims, false claims, mixture claims, other claims).

4.4.2 An Example

Consider the article published on Before It's News entitled "RFK Jr: Fauci Must Be Prosecuted for 330K Murders, As Mass Graves Found Outside NYC (Video)"[7]. This article has the following values on the TAXODIS detection features (manually annotated):

[7] https://beforeitsnews.com/alternative/2024/09/rfk-jr-fauci-must-be-prosecuted-for-330k-murders-as-mass-graves-found-outside-nyc-video-3821353. html (accessed on October 30, 2024)

length of the headline = high, textual gap between title and body = no, sentencelike headline = yes, relative shortness of body = yes, simplicity in sentence structure = no, lexical and contentual poorness = yes, relative amount of typographical errors = yes, relatively low demand on reader's education level = yes, level of semantic incoherence = high, extremity of information quantity = high, lack of new information = yes, lack of topical redundancy = yes, level of emotional polarization = high, level of sensationalism = high, arousal of (negative) affects = high, level of topicality = high, tendency to subjective statements = low, level of personal motives = high, kind of discourse = opinion-based, usage of exaggerated vocables = yes, amount of first-/secondperson pronouns = high, amount of excessive emotional adverbs = high, usage of emphatic wording = yes, level of informality of language = high, vagueness of phrasing = high, authenticity/referencing of information = poorly referenced. As regards the categorization features, the article falls under the themes *political & economic* and *extreme*, and the content type *fabricated content*, while its veracity grade is *mostly false*.

4.5 Taxonomy Usage and Linking to Related Vocabularies

The taxonomy can be used together with existing, established vocabularies for the annotation of (disinformation) resources. We suggest the exploitation of the Web Annotation Data Model[8], which is a W3C recommendation for the structured representation of annotations that can be shared and reused across different platforms. In this model, an annotation (instance of class oa:Annotation) is considered to be a set of connected resources, typically including a body (instance of class oa:Body) and a target (instance of class oa:Target), and conveys that the body is related to the target. The exact nature of this relationship changes according to the intention of the annotation, but the body is most frequently somehow "about" the target (the intention of the annotation can be represented using the class oa:Motivation). In our case, the body of the annotation is a taxonomy term, accompanied by a value (level or degree) for the terms that are under *detection feature*, and the target is a disinformation piece of content or resource. Figure 16 shows an example in which an article (instance of class oa:Target) is linked to two annotations: one which categorizes the article as of social theme (taxodis:themsoc) and one which categorizes the article as having *high* topicality level (taxodis:mobtopical). The intention (motivation) of both annotations is classification (oa:classifying). Notice that the first annotation is directly linked to the taxonomy term taxodis:themsoc through multiple instantiation (the term is an instance of both oa:Body and skos: Concept). This annotation

[8] https://www.w3.org/TR/annotation-model/

method can be applied for all taxonomy terms that are under *categorization* and *veracity*, since these terms do not accept a degree value or level like the terms that are under *detection feature*.

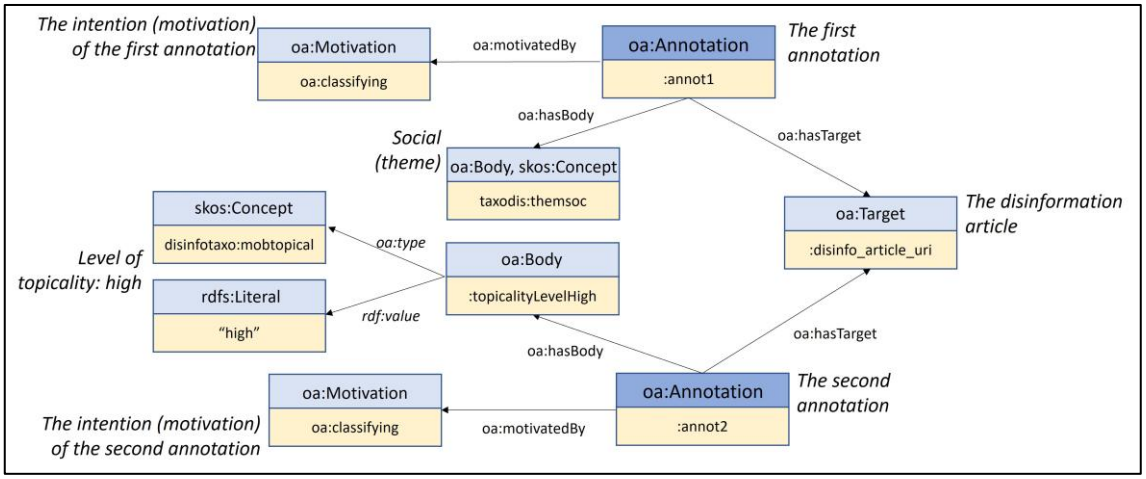

Figure 16. An annotation example using TAXODIS together with the Open Annotation Data Model in which an article is categorized as of social theme and as having a high topicality level.

Figure 17 shows how we can link the annotated resource with rich (meta)data using another established vocabulary, namely schema.org. Schema.org[9] is a collaborative, community activity with a mission to create, maintain, and promote schemas/vocabularies for structured data on the web. It provides classes and properties for embedding structured data to web resources. In the example of Figure 17, the annotated article is both an instance of oa:Target and an instance of schema:CreativeWork. This allows using properties of schema.org for providing more information about the article, such as its URL (instance of schema:URL), its publication date (instance of schema:DateTime), its headline (instance of schema:Text), its author (instance of schema:Person), and its content (instance of schema:Text). We can also link the article with entities of different types mentioned in it, such as persons, places, etc., using the property schema:mentions. In addition, we can link claims (instances of schema:Claim) to the articles using the property schema:appearance. A claim can then be linked to its text, video/audio (if any), and author (using the properties schema:text, schema:video/schema:audio, and schema:author, respectively), as well as with claim reviews (instances of schema:ClaimReview). In a similar way, a claim review can be linked with related data such as its author, URL, publication date, headline, review body, etc.

[9] https://schema.org/

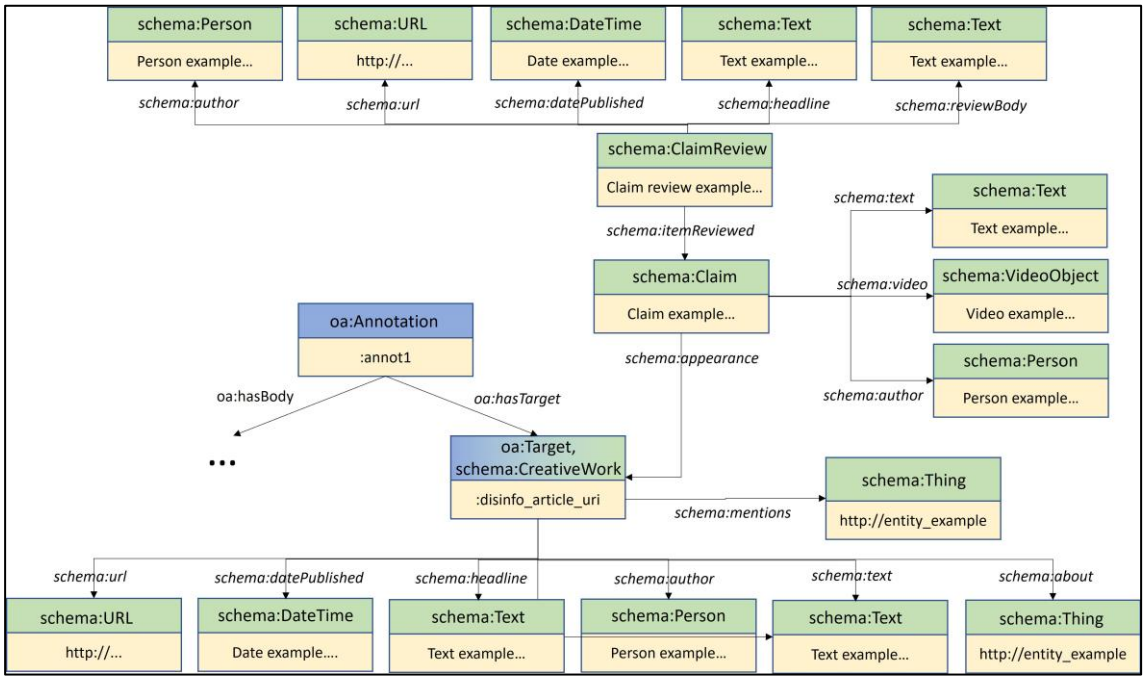

Figure 17. Enriching the annotated resource with rich information using schema.org.

Another well-known vocabulary that can be used together with the taxonomy is the SIOC Core Ontology[10], a data model that provides the main concepts and properties required to describe information from social media and online communities. Linking to such established vocabularies supports the integration of annotation data with existing knowledge bases that make use of the same data models, such as ClaimsKG (Tchechmedjiev et al., 2019) and TweetsKB (Fafalios et al., 2018).

Queries that can be answered using TAXODIS annotations include:

- Retrieve all resources classified as of social theme and which have a high level of emotional polarization
- Retrieve all resources with imposter content together with the values of all features that are under *psychology feature*
- Retrieve the number of resources per content type having high usage of emphatic wording
- Retrieve all resources published on a specific time period containing claims that have been reviewed and have received the veracity score *mostly false*

[10] https://www.w3.org/submissions/sioc-spec/

- Retrieve all resources mentioning a specific person that are mostly false, together with the values of all features that are under *detection feature*

The first query of the above list is translated to SPARQL as follows:

```
PREFIX taxodis: <https://hop.fzi.de/taxodis/>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX schema: <http://schema.org/>
SELECT ?resourceUri ?resourceHeadline ?resourceAuthor WHERE {
?annot1 oa:hasTarget ?resourceUri ; oa:hasBody taxodis:themsoc .
?annot2 oa:hasTarget ?resourceUri ; oa:hasBody ?annot2Body .
?annot2Body oa:type taxodis:mobpolar ; rdf:value "high" .
OPTIONAL { ?resourceUri schema:headline ?resourceHeadline }
OPTIONAL { ?resourceUri schema:author ?resourceAuthor } }
```

Obtaining the Feature Values

For a given piece of content, we can estimate the value of each feature either manually or using dedicated software. Each approach has its pros and cons. The manual approach provides annotations of very high quality. However, it is very laborious, time-consuming, and not scalable (the annotation time is proportional to the number of texts/documents we want to annotate and the number of considered features). On the contrary, using a software system, we can obtain annotations for large corpora with no human effort. However, the accuracy of the annotations is questionable and depends on several factors, such as the overall quality and performance of the software system, the availability of training data, the language used in the input texts, etc. Furthermore, there might be a monetary cost for using the system.

Existing software systems that can be used to obtain values for one or more of the TAXODIS features include: i) linguistic and word usage analysis tools (such as LIWC (Boyd et al., 2022) for the *detection* features, ii) topic and theme extraction tools (Dhar et al., 2021) for the *categorization* features, and iii) fact-checking, disinformation detection, and claim linking tools (such as ClaimLinker (Maliaroudakis et al., 2021)) for the *veracity* features. Moreover, if enough training (annotation) data is available, dedicated classifiers per feature can be built and used for larger text corpora. Surveying such software systems and evaluating their performance is out of the scope of this paper.

4.6 TAXODIS Evaluation and Use Cases

4.6.1 Taxonomy Use and Evaluation

The taxonomy was initially introduced through an internal workshop in 2023, during which a group of interested researchers (from sociology, computer science, and political science) and practitioners (from NGOs and industry) utilized it to create labels for identifying different types of disinformation for the research project DeFaktS. During the workshop, the participants applied TAXODIS to scrutinize real-world data, i.e., numerous social media posts derived from various platforms (e.g., Telegram and Twitter/X) containing disinformation. These labels were then used in annotating a comprehensive dataset for training a classifier to detect deceptive messages (Ashraf et al., 2024). The workshop, focusing on textual detection of disinformation, involved 15 researchers and practitioners from relevant fields. They used TAXODIS to assess whether a given content was disinformation or not and utilized it as a baseline to create suitable labels for data annotation. Two groups of workshop participants approached the task in different ways, testing the taxonomy's usefulness during group work. Jointly, they generated a list of 15 polar labels, 13 of which were selected either directly from the taxonomy or created with its assistance, such as by merging two features into one label for the annotation process. To enhance the robustness and reliability of our annotations conducted through the annotation platform Doccano (Nakayama et al., 2018), we implemented a cross-annotation process. Specifically, a subset of 767 data samples underwent independent annotation by two teams, each consisting of two annotators. This approach ensured a comprehensive evaluation of the labeling process facilitated by the taxonomy. Subsequently, we computed the inter-annotator agreement (McHugh, 2012) to assess the level of concordance between the annotators. To quantify this agreement, we utilized Cohen's Kappa metric, revealing a substantial score of 0.72. This result confirms the strength and dependability of the annotations throughout the dataset, establishing a robust foundation for training a model based on TAXODIS.

4.6.2 Use-Case Scenarios

4.6.2.1 Computer Science and AI

In the field of computer science, and in particular AI and supervised learning, the resource can be of use to build and/or fine-tune language models to perform various downstream tasks related to disinformation detection and analysis. The taxonomy enables fine-grained annotation of text with relevant linguistic features, while the use of standards and semantic web technology allows to query and access specific sub-sets of annotated data in a centralized manner, even if they come from different sources. In that way, this technology

provides the possibility of extracting data for precisely training or fine-tuning of machine learning models that correspond to specific criteria, according to a specific selection of TAXODIS labels. The automatic extraction of the taxonomy features from text via dedicated tools could facilitate annotation. Certain features from the taxonomy can be linked to some of the features from the LIWC vocabulary (discussed above), for which LIWC provides tools for their automatic extraction. However, since the majority of the vocabulary terms are specific to the disinformation context, dedicated tools for their extraction need to be created. Taking it a step further, the taxonomy can facilitate the annotation of new text with reduced reliance on human labor by incorporating examples into prompts for generative AI systems.

The features can contribute to contextualizing the outcomes and predictions in tasks, such as disinformation detection. Indeed, the resource can be useful in enhancing the explainability of language models, such as BERT. A language model fine-tuned on corpora annotated by TAXODIS can be applied to perform various downstream tasks, such as classifying texts as disinformation or not. However, the model as such will struggle to provide an interpretation of its prediction, where understanding why a specific piece of information is classified as flawed or not is crucial for journalists or social scientists (cf. below), as well as ordinary users. A major challenge in AI research is indeed the interpretation of the features used by language models, e.g., by extracting the most predictive tokens (Malkiel et al., 2022; Szczepański et al., 2021), or by understanding the implicit semantics carried by the embedding layers (Chersoni et al., 2021). In our case, if the corpora that are used to train/fine-tune the model are annotated by the high-level linguistic features coming from TAXODIS, one could conduct an explicability analysis by identifying the taxonomy features that contribute most to a specific class prediction. In addition, the vocabulary can help to match the low-level BERT (or BERT-like model) features to high-level, meaningful, and human-curated linguistic features, hence contributing largely to the explainability challenge of language models. A potential way of conducting that analysis is performing independent classification by using a language model with automatically embedded features and then by using a simple binary classifier (like a decision tree) by using the TAXODIS features only and then applying an explainability system, such as SHAP (Lundberg & Lee, 2017) on both in order to identify groups of features on both sides that contribute most to the specific classification outcome.

4.6.2.2 Social Science

In the field of social sciences, understanding and analyzing online disinformation is crucial for examining its impact on public opinion, behavior, and societal dynamics (Allcott & Gentzkow, 2017; Freelon & Wells, 2020). Researchers studying the effects of disinformation on social behavior can utilize the taxonomy to systematically categorize and analyze linguistic features within disinformation content. This structured approach allows

for more precise measurement and comparison of how different types of disinformation affect various demographic groups and societal segments. Computational social scientists often rely on annotated datasets to train models and conduct analyses (Alassad et al., 2021; Rauh & Schwalbach, 2020). The taxonomy's comprehensive framework aids in the consistent labeling of disinformation instances, ensuring that datasets are uniformly annotated. This uniformity may enhance the reliability of statistical analyses and the generalizability and long-term validity of research findings.

Furthermore, providing a standardized taxonomy may facilitate collaboration between social scientists and computational experts. Researchers can leverage the resource to align their qualitative insights with quantitative analyses, fostering interdisciplinary studies that combine linguistic features with social theories. Finally, social scientists can use insights derived from the taxonomy to inform policy recommendations. Understanding the specific linguistic markers of disinformation enables the development of targeted interventions and strategies for mitigating the adverse effects of disinformation on public discourse and democratic processes (Lutz et al., 2024; Munn, 2020).

4.6.2.3 Journalism

In journalism, the taxonomy may serve as a practical tool for improving the accuracy and effectiveness of disinformation detection and fact-checking. Journalists and fact-checkers can use the taxonomy to streamline their verification processes. By referring to the taxonomy's linguistic features, they can more effectively identify and analyze disinformation in news content, ensuring that false claims are quickly and accurately addressed. Additionally, the taxonomy may support journalists in analyzing patterns of disinformation across different media sources. By categorizing linguistic features, journalists can detect recurring themes and tactics used by disinformation campaigns, leading to more informed reporting and deeper investigative insights (Kebede et al., 2022). In education, the taxonomy may provide a valuable resource for training journalists and media professionals. Offering a clear, empirically grounded guide to recognizing disinformation, the resource may equip journalists with a tool needed to navigate complex information environments and maintain high standards of journalistic integrity. Finally, journalists may use the taxonomy to create educational content that raises public awareness about disinformation. By demonstrating how specific linguistic features indicate false or misleading information, they can help readers become more discerning consumers of news and reduce the spread of disinformation.

4.7 Conclusion

The widespread phenomenon of disinformation, understood as deliberately deceptive or false information, presents important risks to political stability and social cohesion, particularly during times of crisis. Automated disinformation detection systems, leveraging machine learning and natural language processing, are essential in the fight against disinformation as tools assisting journalists and social scientists in their efforts. Given the complex and nuanced nature of disinformation, this study contributes a structured taxonomy, named TAXODIS, to aid automated systems in annotating corpora and recognizing linguistic markers of disinformation with high precision. TAXODIS is presented as a SKOS vocabulary, leveraging the semantic web technology and principles. It is, hence, the first resource of its kind that is openly available, reusable, and interoperable, aiming to play the role of a standard, useful for annotation and classification tasks, fostering both scholarly and practical advancements in automated disinformation detection in fields such as computer science, journalism, and social sciences.

5 A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing[11]

5.1 Introduction

In the contemporary information era, the rapid proliferation of online platforms has reshaped communication paradigms. Social platforms have democratized information dissemination, ensuring real-time data sharing. This accessibility, however, is a double-edged sword. On one hand, it promotes knowledge sharing; on the other, it has become a conduit for the spread of disinformation (Shu et al., 2017). The implications of unchecked disinformation are severe. Beyond the obvious erosion of public trust in media and institutions, disinformation can sway public opinion, influence election outcomes, and even catalyze real-world harm (Groshek & Koc-Michalska, 2017; Strömbäck, 2005). In the face of these challenges, ensuring the veracity of digital content has become imperative. Empirical findings underscore the intricate complexity of disinformation, which, with its deceptive nature, strives to cloak itself as legitimate information, making its detection notably elusive (Shu et al., 2020b). While studies emphasize that authentic and deceptive news articles demonstrate substantial disparities in their substantive content (Abonizio et al., 2020; Horne & Adali, 2017), the nuanced and multifaceted characteristics of disinformation amplify the challenge (Rosińska, 2021). Moreover, the lexical and structural features of disinformation often tend to be event-specific, suggesting that classifiers trained on one type of event or topic may underperform when faced with deceptive content derived from a different context (Shu et al., 2017). This multi-dimensional complexity and subtlety of disinformation necessitate innovative approaches that can navigate through its nuanced landscapes, offering a more holistic understanding and detection mechanism.

In the realm of disinformation research, while English has been the primary focus, other significant languages like German have not received equivalent attention. This oversight

[11] This chapter comprises an article that was published by Shaina Ashraf, Isabel Bezzaoui, Ionut Andone, Alexander Markowetz, Jonas Fegert and Lucie Flek in the following outlet with the following title: DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing. In The 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation, 2024. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated. Details of the author's individual contributions to this publication are provided in the appendix.

is particularly evident in the scarcity of comprehensive annotated datasets dedicated to the German language, especially in the domain of disinformation analysis (Schreiber et al., 2021). Furthermore, Germany itself faces a pronounced challenge with disinformation, as indicated by its high number of QAnon members, ranking second globally outside of English-speaking countries (Amadeu Antonio Foundation, 2020). The unique linguistic characteristics and cultural contexts of German differentiate from English, and the limited availability of annotated datasets for German compounds the complexities of disinformation detection in this language. This study navigates through these challenges by presenting a comprehensive approach to understanding and mitigating disinformation, especially within the German linguistic context, through three pivotal contributions:

1. Introducing a richly curated and annotated dataset that encompasses a diverse array of topics and keywords from the German media, thoroughly annotated with binary and fine-grained labels to serve as a foundational resource for developing and evaluating disinformation detection algorithms.
2. Recognizing the complex nature of disinformation, we propose a comprehensive and fine-grained taxonomy-based annotation scheme encompassing linguistic, semantic, psychological, and authenticity features formulated to facilitate a detailed and structured approach to analyzing and labeling tweets.
3. The study further presents experiments employing both classical machine learning models and transformer-based models, providing initial insights into the dataset's utility and serving as a starting point for subsequent research endeavors to develop and refine disinformation detection models in the German language.

5.2 Related Work

Recent efforts in combating disinformation have largely centered around leveraging advanced machine learning techniques and developing datasets to facilitate the training and evaluation of models designed to discern the veracity of information disseminated online. Ali et al. (2022) focused on Arabic disinformation detection related to COVID-19 on Twitter (now X) and Facebook. The authors introduced a new Arabic COVID-19 dataset and applied two pre-trained classification models, AraBERT and BERT base Arabic. Abd Rahim and Basri (2022) introduced MalCov, a dataset containing false and valid news articles related to COVID-19 in the Malay language. The dataset, which comprises articles from social media platforms and has been validated by local authorities, was utilized to build classifiers using machine learning models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression. Suryavardan et al. (2023) introduced Factify 2, a multimodal fact-checking dataset that enhances its predecessor, Factify 1, by incorporating new data sources and adding satirical articles. Factify 2 categorizes data into

three broad categories (support, no evidence, and refute) and further subcategories based on the entailment of visual and textual data, providing a rich resource for developing and evaluating multimodal disinformation detection models. Ciora and Cioca (2022) developed RoCo-Fake, a Romanian COVID-19 disinformation dataset, aggregating various online resources like tweets, news titles, and fact-checking news sites. RoCo-Fake addresses the scarcity of resources for disinformation detection in the Romanian language, providing a valuable resource for the medical domain. Carrella et al. (2023) emphasized the importance of developing language-specific datasets and models to address the challenge of disinformation in Italian. Plepi et al. (2022) conducted an in-depth analysis of users' time-evolving semantic similarities and social interactions, revealing that these patterns can be indicative of disinformation spread. Building on these findings, they proposed a dynamic graph-based framework that capitalizes on the fluidity of user networks to isolate disinformation spreaders. Fatima et al. (2023) introduced YouFake, a multimodal dataset that includes both images and texts collected from popular YouTube channels, providing a comprehensive platform for developing and evaluating models that can handle multi-modal data (text, image, and video) for disinformation classification.

These studies underscore the global and multilingual nature of the disinformation challenge, highlighting the importance of developing datasets and models that cater to various linguistic and cultural contexts. While these datasets provide valuable insights and resources for disinformation classification (Sakketou et al., 2022), it is evident that there is a gap in the availability of German-specific datasets for disinformation detection, highlighting a potential area for contribution and development in the field. Moreover, the available datasets often exhibit a lack of diversity in topics and news categories, frequently concentrating on specific themes or health crises like the COVID-19 pandemic (Mattern et al., 2021). This limitation potentially restricts the generalizability and applicability of models trained on such datasets to a broader spectrum of topics and contexts. Furthermore, there is a noticeable scarcity of datasets that provide transparent and comprehensive annotation schemes for labeling disinformation (Murayama et al., 2022). The meticulousness and granularity in labeling are pivotal for developing models that can discern and understand the nuanced and multifaceted nature of disinformation. Many existing datasets (Ahuja & Kumar, 2023; Vogel & Jiang, 2019) do not offer fine-grained labels or employ polar labeling schemes that enable annotators to adeptly identify and categorize various dimensions and spectrums of disinformation.

In response to these gaps, we introduce *DeFaktS*[12], a dataset uniquely designed for German media. Our dataset not only offers a comprehensive understanding of disinformation within this specific linguistic context but also brings forth a novel approach in its annotation and structure. *DeFaktS* is meticulously curated, emphasizing granularity in labels and ensuring that various dimensions of disinformation are adeptly captured. The annotation scheme and, correspondingly, the labels utilized are designed based on the *Taxonomy of Online Disinformation* developed by Bezzaoui et al. (2022b). Combining empirical findings from various fields such as computer science, linguistics, psychology, and media studies, the taxonomy gathers the many underlying linguistic features of disinformation into a schematic framework. Our annotation framework's strategy revolves around addressing three key research endeavors: First, the identification of specific linguistic cues that signify online disinformation, as highlighted in the empirical literature (Abonizio et al., 2020; Horne & Adali, 2017; Molina et al., 2021). Second, the organization of these linguistic features into a coherent and comprehensive schema. Third, the integration of these dimensions and categories into a clearly defined, structured taxonomy. This positions *DeFaktS* not just as another dataset but as an advanced contribution to the ongoing global effort to curb the influence of disinformation.

5.3 Dataset

Twitter (now X) is a primary hub for real-time news dissemination. Its influence, coupled with the potential for spreading deceptive content that can mold public opinions, underscores its significance (Li & Su, 2020; Zhou et al., 2021). Therefore, we chose it as our primary data source.

5.3.1 Data Collection

Our *DeFaktS* dataset is carefully crafted, focusing on the German media domain, ensuring a robust and comprehensive collection suitable for in-depth analysis of various news topics. Initially, we compiled a list of 129 pertinent and diverse news topics, which were predominantly trending at the time of data collection. This included a range of controversial and high-impact topics such as elections, the energy crisis, lockdown measures, the war in Ukraine, the gender pay gap, immigration, climate, and inflation, among others. A word cloud depicting the prominence of these topics within our dataset can be seen in Figure 22. In order to establish the topics, we started with a set of related keywords. We

[12] https://github.com/caisa-lab/DeFaktS-Dataset-Disinformation-Detection

then collected German-language tweets that contained these keywords and added the first 2000 tweets that fit our criteria to our database. Given Twitter's (now X) dynamic nature and the prevalence of retweets, we removed duplicate entries to avoid any potential biases in our subsequent analyses.

5.3.2 Data Annotation

5.3.2.1 Fine-Grained Labels Annotation Scheme

The primary objective of the data annotation was to scrutinize the tweets, identifying and highlighting instances indicative of disinformation. In pursuit of this, a detailed annotation framework was designed, which has general category labels and more nuanced polar labels, each dissecting distinct facets of the tweets and pinpointing specific features potentially signaling disinformation. To ensure that current empirical knowledge is taken into account, the annotation framework is based on the *Taxonomy of Online Disinformation* (Bezzaoui et al., 2022b). The taxonomy synthesizes scientific evidence from various disciplines into a concise overview covering dimensions ranging from more granular characteristics, such as semantic aspects (Cardoso, 2021) of disinformation, to broader aspects for categorization, such as various content types.

The *DeFaktS* annotation scheme was specifically developed to dissect and identify framing techniques utilized in the dissemination of disinformation through German social media. Our comprehensive labeling approach is geared towards detecting nuanced ways in which information is framed, which can influence perceptions and propagate disinformation. Our annotation process is rooted in four principal dimensions: content type, authenticity, semantic, and psychological features, each chosen for its empirical association with disinformation. Semantic features help to analyze the content for meaning and consistency, as disinformation is often riddled with contradictions or repeated content lacking new insights (Azevedo et al., 2021; Horne & Adali, 2017). Psychological features encompass tactics like polarization, emotionalization, and sensationalism. These features construct narrative frames that manipulate emotional biases to enhance engagement and dissemination (Gruppi et al., 2018; Jeronimo et al., 2019; Ribeiro Bezerra, 2021; Vicario et al., 2019; Wang et al., 2019). Authenticity features assess the authenticity of references and the clarity of phrasing, helping to determine whether the information is framed within a reliable context or crafted to mislead by obfuscating facts (Fernandez, 2019; S. Kumar et al., 2016; Mahyoob et al., 2020). Content type features address the thematic framing of content, including pseudo-scientific claims, forged content, and propaganda. Such framing shapes audience perception and is an integral part of disinformation strategies (Bąkowicz, 2019; Kapantai et al., 2021; Rashkin et al., 2017; Rosińska, 2021; Tandoc et al., 2018).

Category	Dimension	Feature	Code	Description
Polar Labels	Semantic Features	level of semantic inconsistency	infoincon	Disinformation exhibits a higher degree of contentual inconsistencies like semantic contradictions or logic errors throughout the text.
		lack of (new) information	infonewinfo	The body of unreliable articles adds relatively little new information, but serves to repeat and enhance the claims made at the beginning.
	Psychology Features	level of polarization	psychpolar	Unreliable articles frequently narrate in terms of a clear friend-foe-distinction with regard to specific national, ethical, or religious groups or elites as foes or perpetrators. The opposing group (often framed in a common "we", "ourselves", "the government") takes the part of the victim who needs to be protected.
		level of emotionalization	psychemo	Unreliable sources incline to use a more emotionally persuasive language and touch more often sensible subjects (like children, death and burial).
		level of sensationalism	psychsensa	Fake articles tend to be written in a hyperbolic way to attract the reader's attention, i.e. with a high usage of all-caps-words, exclamation marks or a general sentiment wording.
		level of abasement	psychabas	Disinformation frequently entails stereotype narratives and ressentiments to denigrate targeted groups.
	Authenticity Features	level of topicality	authtopic	Legitimate sources tend to report about past events whereas fake articles focus on highly recent topics.
		vagueness of phrasing	authvague	Fake articles use a higher amount of hedging words (like 'possibly', 'usually', 'tend to be') to achieve a more indirect form of expression. Also they evoke a feeling of uncertainty by addressing the vagueness of information directly.
		authenticity/referencing of information	authrefer	Legitimate sources are considerably better referenced than unreliable articles. Unreliable sources tend to use none, false or wrong contextualized references.
	Content Type Features	pseudoscientific	typpseudo	Content that calls on supposedly scientific research or reputable institutions without identifying concrete sources or by manipulating them to create a false theory.
		forged content	typforged	Stories that lack any factual ground or manipulated information or image. The intention is to deceive and cause harm. Could be text or visual media.
		false context	typfalcontex	Real information is being presented in a false context. The recipient is aware that the information is true, but he does not realize that the context has been changed.
		conspiracy theory	typconspir	Stories without factual basis which usually explain important events as secret plots by government or powerful individuals. By definition their truthfulness is difficult to verify. Evidence refuting the conspiracy is regarded as further proof of the conspiracy.
		propaganda	typpropa	Information that is created by a political entity to influence public opinion and gain support for a public figure, organization or government.
		no factual content	typopinion	This rating is used for posts that are pure opinion, comics, satire, or any other posts that do not make a factual claim. This is also the category to use for posts that are of the "Like this if you think..." variety.
General Labels			corpkeyword	Keywords used to search tweets. This label does not indicate polarity but marks the span of text containing the search keyword.
			catposfake	Category indicating possible disinformation. This label is attributed to posts that receive one of the polar labels.
			catneutral	Indicates neutral posts where there is no indication of disinformation. Such posts should never have any other polar labels.

Figure 18. Fine-grained annotation framework

To ensure fidelity and uniformity in our annotations, domain experts from the *Center for Monitoring, Analysis, and Strategy* (CeMAS) conducted a rigorous training workshop. Here, annotators were equipped with guidelines and engaged in activities using sample data, which honed their ability to recognize text passages containing deceptive indicators aligned with our polar labels. Figure 18 and Figure 19 illustrate the framework and provide examples of tweets annotated with these labels, demonstrating the application of our method and underscoring the role of each feature in pinpointing disinformation.

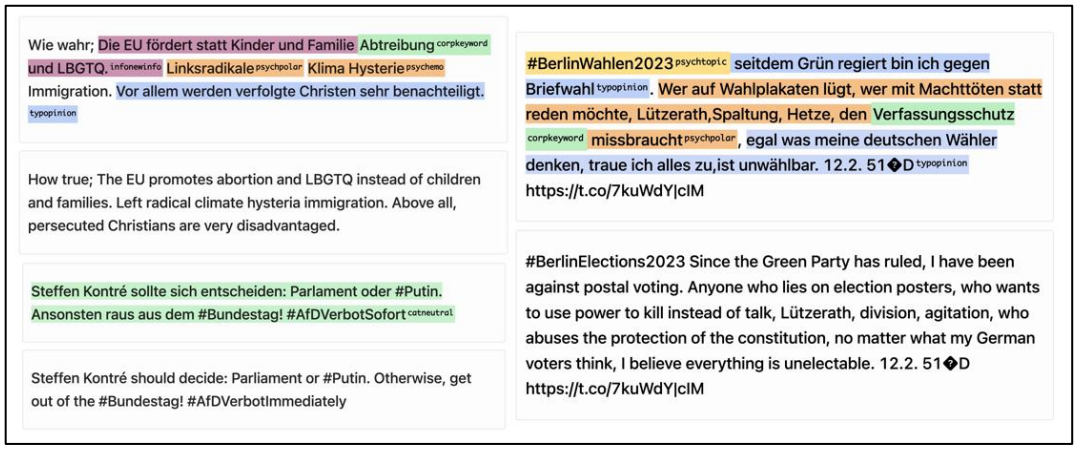

Figure 19. Annotated samples: original German and translated English text for three tweets.

5.3.2.2 Binary Labels

In addition to the multi-label annotation scheme that categorizes posts into an array of polar and general labels, a binary classification scheme is also employed to demarcate between two primary categories:

- *Real News* is dedicated to posts that are regarded as neutral in nature. Specifically, posts under this umbrella contain exclusively the label "catneutral".
- *False News* represents posts that exhibit traits indicative of potential disinformation or bias. Posts allocated to this category contain at least one of the polar labels but are devoid of the label "catneutral".

5.3.2.3 Annotation Platform

In this study, we utilized *Doccano* (Nakayama et al., 2018), an open-source annotation tool, to facilitate our annotation process, primarily owing to its user-friendly interface and capability to streamline collaborative efforts. *Doccano* is well-equipped with features tailored to our task requirements, thereby making it an apt choice for managing our annotation activities. The project was configured as a sequence labeling task, enabling the annotators to select specific text spans and assign labels to them, supporting multiple labeling. Furthermore, annotators had the flexibility to select the entirety of the text to assign general category labels. Prior to uploading the data to *Doccano*, default labels with the code "corpkeyword" were assigned to highlight keywords within the text, which were initially used for filtering tweets during the data collection process (as also mentioned in the annotation scheme). Additionally, comprehensive annotation guidelines were uploaded to the platform, serving as a readily available reference for annotators during the text annotation process, thereby ensuring consistency and adherence to the specified labeling criteria.

5.3.2.4 Cross Annotation

To fortify the robustness and dependability of our annotations, we undertook a process of cross-annotation. A subset of 767 samples was independently annotated by two annotators, ensuring a thorough examination of both our fine-grained and binary labels. Consequently, inter-annotator agreement (IAA) (McHugh, 2012) was computed for both labeling methods to gauge the level of concordance between the annotators. In the cross-annotation subset, we observed disagreements across the labels: 53 for binary labels and 95 for fine-grained labels. Given that the fine-grained labels span 17 categories, higher contradictions were seen compared to binary labels. To quantify the IAA, we employed Cohen's Kappa metric, unveiling a substantial agreement with a score of 0.72 for binary labels. For fine-grained labels, which naturally present a more complex annotation scenario, the average score across multiple labels was 0.56, indicating a moderate level of agreement. In an additional layer of evaluation, and to assess the similarity in the sets of fine-grained labels assigned to the annotators for each instance, we calculated the Jaccard Similarity Score, achieving a noteworthy score of 0.88. This score, paired with Cohen's Kappa metric, affirms the robustness and reliability of the annotations across our dataset, ensuring a solid foundation for the subsequent experiments and analyses.

5.3.2.5 Dataset Statistics

The dataset comprises a total of 105,855 posts, where 20,008 tweets are labeled with the class distribution of 11,776:8,232 of *Real News* and *False News,* respectively. The dataset encapsulates a variety of attributes for each tweet, enabling analyses related to temporal patterns, identifying topics, trends, and user engagements. A general overview of the dataset's statistical characteristics is shown in Table 3.

Data	**Statistics**
Unique Users	44,486
Average Tweet Length (characters)	187
Average Tweet Length (words)	24
Average Likes	22
Average Retweets	4
Average Replies	3
Average Quotes	0.4
Average Tweets/User	3
Number of Tweets with URLs	65,889

Table 3. Basic data statistics

Upon curating the *DeFaktS* dataset, a thorough exploratory data analysis was conducted to comprehend the underlying patterns and characteristics inherent to the collected attributes. All the polar labels have varying counts associated with them, the most frequently

associated polar label is *typopinion* with 5,354 occurrences, followed by *psychsensa* with 2,056 occurrences. There are no specific polar labels associated with *Real News* in the dataset. This means that the dataset's *Real News* entries do not have any of the polar labels from the annotation guidelines, which aligns with the notion that these polar labels are indicators of fake or unreliable information. The label *typopinion* has the highest occurrence, suggesting that many disinformation tweets in the dataset are opinion-based without factual content. Labels like *psychsensa* (indication of sensationalism) and *psychemo* (Indication of emotionalization) also have significant occurrences, indicating common features of sensationalism and emotional language in disinformation. Given this analysis, we can infer that disinformation in the dataset frequently exhibits features such as sensationalism, emotionalization, lack of proper referencing, and more. To better understand this, we can visualize a bar graph of the polar labels distribution for the tweets (Figure 21) as well as the distribution across binary labels (Figure 20).

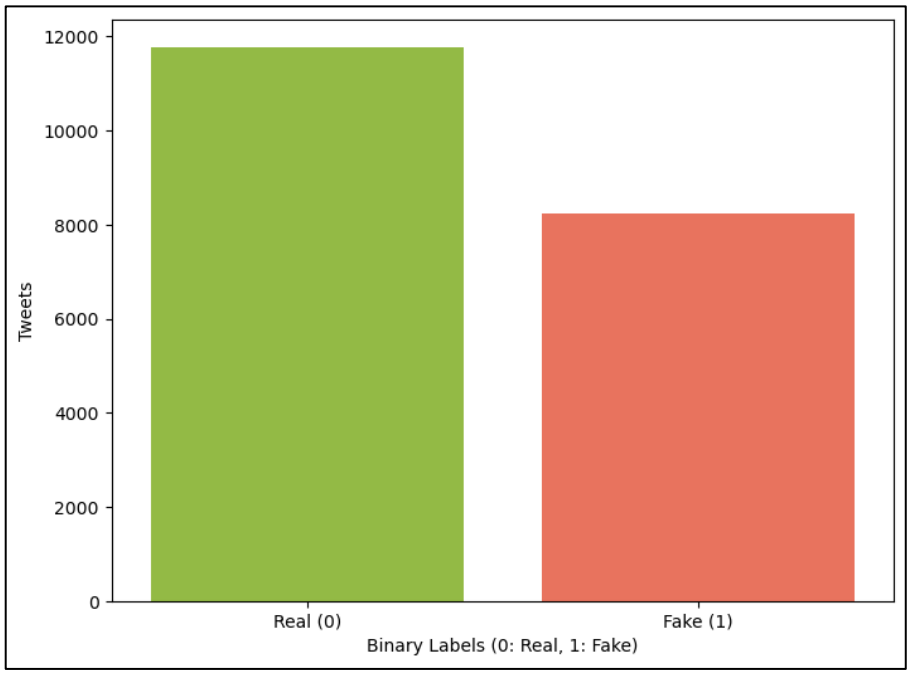

Figure 20. Distribution of binary labels.

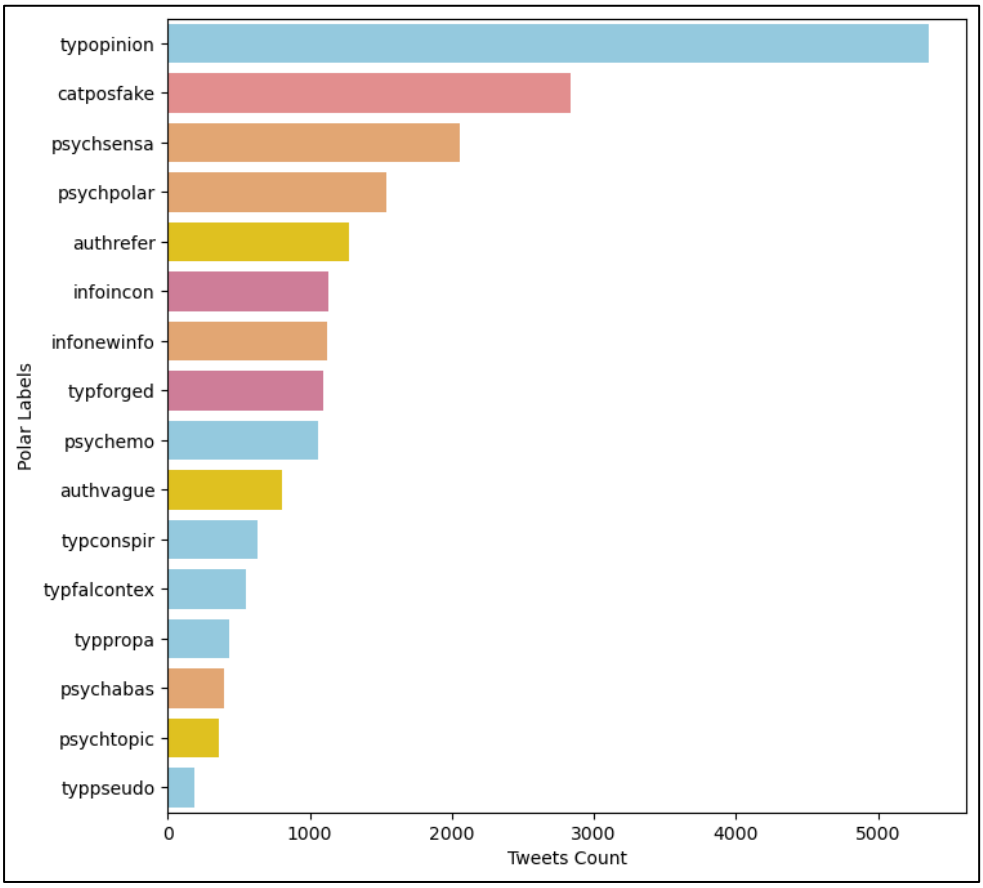

Figure 21. Distribution of polar labels in "False News".

5.4 Methodology

5.4.1 Preprocessing

In our research, preprocessing was crucial to mitigate noise and ensure data quality. We executed several steps, including stop word removal, lower-case conversions, tokenization, and lemmatization. Additionally, we stripped URLs to eliminate potential source link biases, ensuring a cleaner dataset for feature extraction and model training.

5.4.2 Features and Text Encoding

To represent our text data, the following features and embeddings were utilized for model training:

- **Bag of Words (BOW)**: A vector representation counting word occurrences, ignoring grammar and word order (Qader et al., 2019).

- **Term Frequency-Inverse Document Frequency (TF-IDF)**: Highlights word frequency in a document relative to its frequency across all documents, offering a measure of its importance (Havrlant & Kreinovich, 2017).
- **Word2Vec**: Embeddings that capture semantic meanings of words, using pre-trained models on German Wikipedia with 100-dimensional representations (Yamada et al., 2020).
- **GerVADER Sentiment (GVSent)**: Sentiment-based features derived using GerVader (Tymann et al., 2019) to determine word polarity, providing overall sentiment scores of tweets.

5.4.3 Traditional ML Classifiers

We utilized the following classical machine learning models in our baseline experiments:

- **Support Vector Machines (SVM)**: A supervised algorithm recognized for its effectiveness in text classification by finding the optimal hyperplane for data separation (L. Wang, 2005).
- **Random Forest (RndFor)**: Constructs multiple decision trees for high accuracy and can handle large datasets as well as missing values.
- **Logistic Regression (LogReg)**: Commonly used for binary classification, but adaptable for multilabel tasks using methods like the one-vs-rest (OvR) approach.

5.4.4 Deep Learning Models

Acknowledging the prowess of language models in diverse Natural Language Processing (NLP) tasks and their ability to grasp the contextual relationships between words, we fine-tuned several state-of-the-art pre-trained language models using our dataset.

- **BERT-Base**: Pretrained on English data, it is recognized for capturing deep contextual word relationships (Devlin et al., 2019).
- **BERT-Multilingual:** Trained on 104 languages, this variant of BERT is adept at handling linguistic diversity, making it suitable for diverse languages, including German (Pires et al., 2019).
- **BERT-German**: Tailored for German, it captures linguistic nuances specific to the language while also understanding cross-lingual patterns.
- **Xlm-RoBERTa**: An advanced BERT variant trained on a vast corpus known for its high performance in various NLP tasks (Conneau et al., 2020).

5.5 Experimental Setup

In our experiments, we evaluated the models across two distinct classification paradigms: binary (distinguishing between *False* and *Real News*) and fine-grained (categorizing across 17 labels). Confronted with a pronounced class imbalance in our dataset between *Real* and *False News* instances, we resorted to downsampling the *Real News* category. This strategy was instrumental in ensuring parity in representation between *Real* and *False News* categories, a balance we maintained for both classification tasks. However, when transitioning to the fine-grained classification, we refrained from further downsampling. Given the varied distribution across the 17 labels, additional downsampling could risk discarding valuable data, particularly for polar labels with limited samples. As the next step in our process, we employed a consistent preprocessing pipeline across all models. We established a 5-fold cross-validation for our classical ML models to assess their performance and ensure robustness in our analysis. For features like BOW and TF-IDF, the vectorizer was restricted to a maximum of 5000 features, considering both unigrams and bigrams. For our transformer-based models, we partitioned the dataset into training (80%), validation (10%), and test (10%) sets. The training set was utilized to fine-tune the pre-trained models, the validation set to tune hyperparameters and prevent overfitting, and the test set to evaluate the model performance. The models were trained using a batch size of 32 across 10 epochs. We employed early stopping, monitoring the validation loss. Training would halt if no loss improvement was observed over 3 consecutive epochs. The AdamW optimizer was utilized, configured with a learning rate of $2e-5$.

5.6 Results

To evaluate the effectiveness of both our classical and transformer-based models, we computed several metrics, including accuracy, precision, recall, and F1-score. The F1-scores for our experiments are presented in Table 4 and Table 5.

Baselines			
Binary Class			
Features	**SVM**	**RndFor**	**LogReg**
TF-IDF	0.76	0.74	0.81
BOW	0.78	0.72	0.80
GVSent	0.47	0.52	0.46
Word2 Vec	0.64	0.44	0.58
Fine-grained Class			
Features	**SVM**	**RndFor**	**LogReg**
TF-IDF	0.40	0.48	0.54
BOW	0.50	0.48	0.54
GVSent	0.23	0.27	0.29
Word2 Vec	0.27	0.28	0.29

Table 4. F1-scores for experiments with feature-based models.

	Binary	**Fine-grained**
BERT-Simple	0.78	0.49
BERT-Multi	0.80	0.61
BERT-German	0.86	0.65
Roberta	0.82	0.58

Table 5. F1-scores for experiments with deep-learning models.

5.6.1 Binary Classification

Using feature-based models, the best performance for the binary classification task was achieved with TF-IDF representations, closely followed by BOW. This indicates that count-based representations effectively capture distinguishing features between Real and Fake categories. Transformer-based models, particularly BERT-German, outperformed feature-based models, highlighting their robust ability to discern *Real* from *False News* in German content. The detailed classification report reveals that the model is adept at identifying disinformation instances (evident from a high recall) but occasionally misclassifies other content as disinformation.

5.6.2 Fine-Grained Classification

Feature-based models like TF-IDF and BOW exhibited satisfactory performance in the fine-grained classification task, albeit lower than their binary classification counterparts. This drop in performance is anticipated due to the intricate nature of distinguishing among numerous categories. A closer examination of the detailed classification report reveals that labels like *catneutral* and *typopinion* are predicted with higher precision and recall, suggesting these categories possess distinct features easily identifiable by the model. However, classes such as *psychsensa, psychpolar,* and *authrefer*, despite having ample

instances, did not fare as well. This might hint at these classes sharing overlapping features with others or being inherently more challenging to classify. Sparse classes, like *typconspir* and *psychabas*, predictably struggled, emphasizing the challenges of classifying underrepresented categories. Transformer-based models, especially BERT-German, continued to outpace feature-based models in the fine-grained classification task. However, a detailed label-wise analysis uncovers significant performance variance across labels. For instance, while labels like *infonewinfo* and *typfalcontex* were accurately predicted, others such as *typpseudo* and *psychemo* encountered difficulties. This discrepancy might arise from dominant overshadowing subtler ones in multi-label contexts.

5.6.3 Analysis and Discussion

The empirical results underscore the unparalleled advantages provided by language-specific models, such as BERT-German. Their adeptness at understanding linguistic intricacies, grammar, and vocabulary specific to the German language is pivotal. The timeless efficacy of TF-IDF and BOW representations was evident even when combined with classical models. However, the sentiment scores from German Vader (GerVader) underperformed compared to other features. The brevity of tweets, often filled with slang and abbreviations, can impede accurate sentiment analysis. Tools like Vader provide generalized sentiment features, which may be inadequate for intricate tasks like disinformation detection. Exploring sentiment computation using advanced language models might offer more nuanced insights.

It is evident from our results that binary classification, while challenging, is simpler than fine-grained classification. All models, both feature-based and deep learning, exhibited superior performance in binary classification. This observation is in line with expectations, as discerning between two broad categories (*False* vs. *Real*) is intuitively simpler than distinguishing among 17 nuanced categories. The model has found challenges in categorizing them, as some classes might have overlapping features with other classes, making it hard for the model to distinguish between them. For example, *psychpolar* and *psychsensa* both deal with emotional or sensational content in the text. The potential overlap in their features might be causing misclassifications. Some labels might differ in very nuanced ways which are hard to capture with the given features. For instance, *authrefer* and *authvague* both deal with the authenticity of the content, but one might be about poor referencing while the other is about vague claims. Capturing such subtle differences is challenging.

Incorporating external knowledge from knowledge graphs, ontologies, or trusted news databases is essential for validating claims and providing the necessary context, especially for aspects concerning authenticity and references. While models such as BERT-German

have shown effectiveness, the integration of advanced Large Language Models (LLMs) can take this a step further. LLMs, renowned for their excellence in context learning and prompting-based techniques, can tap into their extensive linguistic capabilities and world knowledge to cross-reference and validate claims against established facts. By fine-tuning these models or employing precise prompts that reflect the context and intent of the content, LLMs become powerful tools for uncovering subtle disinformation cues that may bypass more traditional detection methods.

5.7 Linguistic Analysis

The word cloud representation in Figure 22 depicts the frequency of news topics within the tweets from our dataset, offering a glimpse into the most prominent themes and discussions within the German media. The size of each word indicates its frequency in the tweets, with larger words appearing more frequently.

Figure 22. Distribution of topics in the dataset.

Upon analyzing the textual content of the tweets, we notice that tweets classified as *Real News* tend to be slightly more extensive, both in terms of character length and word count, compared to disinformation, as depicted in Figure 23.

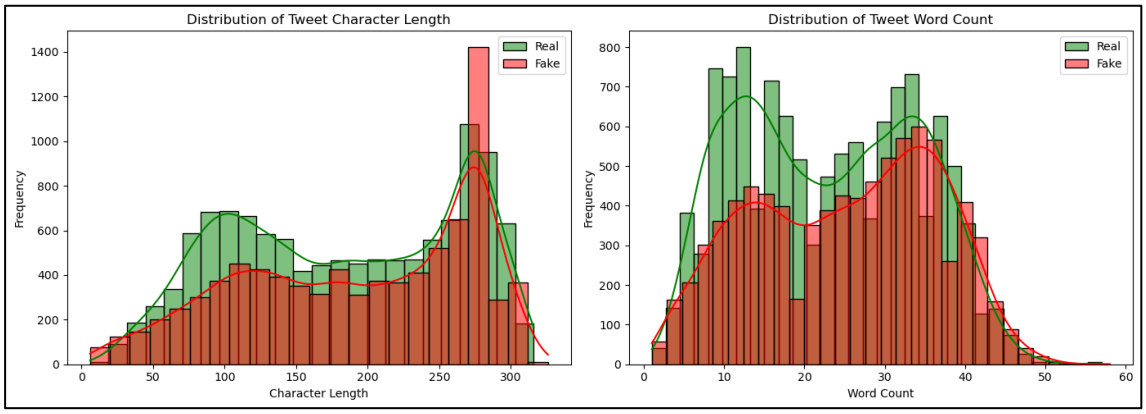

Figure 23. Textual distribution in "Real" vs. "False News".

This might suggest that *Real News* endeavors to provide more detailed and thorough information, possibly requiring additional words or incorporating URLs to convey accurate information. Conversely, a peak in character usage in disinformation indicates that such posts might occasionally employ a more verbose narrative compared to *Real News*, potentially crafting a compelling, albeit deceptive, storyline.

5.8 Conclusion

In this research, we presented *DeFaktS*, a unique dataset tailored for disinformation analysis within the context of German political discussions on Twitter (now X). Through a comprehensive annotation scheme, our dataset facilitates the precise identification and labeling of deceptive content. Beyond binary labels of *Real* and *False News*, *DeFaktS* incorporates fine-grained labels that signify polarized information in textual spans. Our experimental benchmarks, established using both traditional ML classifiers and state-of-the-art deep learning methods, highlight the efficacy of transformer-based models, especially the BERT-German variant, in discerning disinformation patterns. The insights derived from our study pave the way for further nuanced analysis and the development of more robust detection methodologies in the domain of disinformation. Overall, *DeFaktS* serves as a resource for the German media research community, promoting further exploration into refined analysis and detection techniques against disinformation.

5.9 Ethical Considerations and Limitations

Our research heavily relies on tweets, a publicly accessible form of data. While this data is public, ensuring the anonymity of the individuals and preventing potential misuse is paramount. All user data is kept separately on protected servers, linked to the raw text

and network data solely through anonymous IDs. This precaution ensures that any personal information, such as user handles or profile details, is isolated from the research data, thereby respecting user privacy and safeguarding against potential breaches. It is important to note that conducting further analyses on Twitter (now X) data for future research endeavors is not limited to the greatly restricted access for researchers to data generated and distributed by the platform. Additionally, engaging human annotators for the labeling of data containing mentally and emotionally harmful content displays a challenge that researchers should handle responsibly. In the context of this project, to safeguard the annotators' well-being, different safety measures, such as group meetings and mood polls, were applied. While our research aims to detect and combat disinformation, there is potential for misuse. The tools and methods could be appropriated to suppress genuine information or target certain narratives. We emphasize that the primary goal is to detect disinformation and not to suppress freedom of expression.

Part III

Detecting Disinformation through Explainable Artificial Intelligence[13]

[13] This part comprises an article that was published by Isabel Bezzaoui, Carolin Stein, Christof Weinhardt and Jonas Fegert in the following outlet with the following title: Explainable AI for Online Disinformation Detection: Insights from a Design Science Research Project. In Electronic Markets 35, 66, 2025. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated.

6 Opening the Black Box: How Explainable AI Enhances Trust in Disinformation Detection Systems

6.1 Introduction

The manipulation of information through online disinformation represents a profound threat to the integrity of the digital public sphere and the functioning of liberal democracies (Del Vicario et al., 2016). This challenge has been increasingly acknowledged in Information Systems (IS) research (Weinhardt et al., 2024), especially as the rapid proliferation of manipulated content – exacerbated by the capabilities of generative artificial intelligence (AI) (Hanley & Durumeric, 2023) – has escalated beyond electoral contexts, becoming a pervasive societal issue (Truong et al., 2024; Williams et al., 2024). With digital platforms now central to public discourse, ensuring the accuracy and trustworthiness of information is more critical than ever. In response to this threat, advancements in AI offer promising approaches for moderating disinformation (Ansar & Goswami, 2021; Shu et al., 2020b; Wei et al., 2019). However, deploying AI in such a sensitive domain presents new challenges, particularly regarding the transparency, reliability, and user acceptance of algorithmic decisions. In 2018, the European Commission enacted the General Data Protection Regulation (GDPR), which mandates a right for explanations to end-users directly impacted by an algorithmic decision (Voigt & Von dem Bussche, 2017). This legal framework highlights the importance of designing AI systems that can provide clear and understandable reasoning for their decisions, particularly in contexts where these systems operate autonomously (Mohseni et al., 2019).

Explainable AI (XAI), while not universally defined (Thiebes et al., 2021), encompasses diverse efforts to enhance the transparency and trustworthiness of AI by making its decision-making processes more understandable to users (Adadi & Berrada, 2018). The XAI research domain is expansive and interdisciplinary (Brasse et al., 2023), encompassing the fields of IS, human-computer interaction (HCI), and social sciences, involving collaboration among researchers and practitioners across diverse disciplines (Miller, 2019). The application of XAI holds particular relevance in high-stakes situations or use cases where a model output directly impacts human decision-making (Blackman & Ammanath, 2022; Confalonieri et al., 2021).

Disinformation – i.e., the intentional dissemination of false or misleading information to deceive the public (European Commission, 2018) – can greatly impact individuals and society. It has become a means of hybrid warfare attacking liberal societies from within (Shu et al., 2017) and was, therefore, rated as the most severe threat anticipated over the next two years (World Economic Forum, 2024). These dynamics can have significant political repercussions, influencing elections and spreading disinformation during crises such as the COVID-19 pandemic and conflicts in regions like the Levant (Bessi & Ferrara, 2016; Murphy, 2023; Pennycook et

al., 2020). The intentional nature of disinformation requires detection systems that go beyond technical accuracy. Effective detection tools must not only identify harmful content but also provide interpretable, evidence-based explanations for their decisions to establish trust and credibility (Stitini et al., 2022). This need is particularly critical for disinformation because its contentious nature often provokes skepticism regarding interventions, raising concerns about political bias, censorship, and fairness. Unlike misinformation, where user misunderstandings can often be remedied with corrections (Vraga & Bode, 2020), disinformation interventions must address deliberate attempts to manipulate or polarize, heightening the demand for XAI to justify the system's outputs. Therefore, XAI represents a strategic tool not only for enhancing algorithmic transparency but also for safeguarding platform governance and business sustainability in an increasingly complex information environment (Lehrer et al., 2018; Maedche et al., 2019).

The dissemination of disinformation through Online Social Networks (OSN) underscores the urgent need for automated detection systems that respond swiftly and effectively. However, in online discussions, interventions such as moderation are often perceived as controversial, raising concerns about transparency and potential censorship (Mathew et al., 2020). Introducing AI-based moderation software for disinformation detection could exacerbate these concerns, as algorithms are frequently viewed as unreliable and opaque (Gorwa et al., 2020; Suzor et al., 2019). Integrating XAI-based models could help break the black box effect by providing necessary context, allowing end-users to evaluate the veracity of news content independently and reliably. Despite growing interest in XAI, the intersection of explainability and disinformation remains underexplored (Guo et al., 2022; Rjoob et al., 2021). Current research primarily focuses on technical accuracy and detection efficacy, with limited attention to the user-centric design principles necessary for building transparency in AI-based disinformation detection systems (Wells & Bednarz, 2021). By focusing on disinformation rather than misinformation, this study emphasizes the heightened technical and social complexities of disinformation detection, where transparency, user trust, and contextual explanations are paramount. Specifically, the objective is to create a user-centric foundation for developing an XAI model applicable to digital platforms and social media channels. Guided by the principles of Design Science Research (DSR) (Hevner et al., 2004; Thuan et al., 2019), the study is driven by the following research question (RQ):

> *RQ: How should an (X)AI-based tool for detecting online disinformation be designed to foster user trust, comprehension, and usability by leveraging explainability and transparency?*

This research advances theoretical understanding by integrating user-centric principles into designing XAI systems for disinformation detection, focusing on how user feedback and contextual explanations can enhance trust, comprehensibility, and usability. Specifically, we extend prior work in IS and HCI by identifying design principles that balance transparency with user perception, challenging the assumption that greater transparency always improves user experience (Gunning & Aha, 2019; Haque et al., 2023). Using a DSR approach, Chapters 6 to 8 detail

two iterative design cycles aimed at developing an XAI-based disinformation detection tool. These cycles synthesize insights from a structured literature review (Chapter 6), empirical user feedback (Chapters 7 and 8), and theoretical perspectives on responsible AI design (Chapter 8). The key contribution of this study lies in its development of actionable guidelines for creating XAI systems that are not only technically robust but also aligned with user expectations in sensitive and high-stakes domains. Our findings underscore the importance of integrating user feedback early in the design process and highlight the nuanced trade-offs between transparency and user experience in XAI design. This study offers a foundation for future studies seeking to advance the theoretical and practical understanding of XAI application in the disinformation domain.

6.2 Research Background

In recent years, the rapid advancement and integration of AI into critical applications have raised significant concerns regarding transparency, trust, and usability. XAI has emerged as a promising response, aiming to make AI systems more understandable to human users by providing insights into their decision-making process. At its core, XAI seeks to open the "black box" of AI models, offering meaningful, interpretable, and actionable explanations for various stakeholders (Angelov et al., 2021). However, despite its potential, much remains to be explored in effectively operationalizing XAI features and addressing the challenges of balancing transparency with user-centric design (Adadi & Berrada, 2018; Minh et al., 2022). These challenges are particularly salient in digital platform contexts, where AI-powered decision-making intersects with economic, regulatory, and ethical considerations (Alt, 2021; Herm et al., 2022). In such environments, trust-building is not only a technical concern but also a business imperative.

Explanations delivered via XAI systems are operationalized through explainability features, which supply reasoning for a model's decisions. These features can be classified based on their method of generation and their scope of explanation. A key distinction is made between model-agnostic and model-specific approaches. Model-agnostic methods, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations), are versatile tools capable of explaining the behavior of any black-box model by emphasizing feature importance in classifications and predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). In contrast, model-specific methods, like Grad-CAM (Selvaraju et al., 2017), tailor explanations to the unique characteristics of particular algorithms. Moreover, explainability features differ in scope: local explanations focus on individual outputs, while global explanations elucidate the model's overall behavior (Confalonieri et al., 2021; Linardatos et al., 2020). Regarding transparency, both types of features play complementary roles, with local explanations often addressing immediate user concerns and global explanations enhancing broader trust and understanding. Recent work has proposed frameworks that combine technical explanation methods with business model implications, identifying XAI archetypes applicable to online platforms (Gerlach et al., 2022).

The increasing prevalence of disinformation has underscored the need for transparent AI systems, particularly in the context of detection and intervention. Research has shown that tailored explanations can significantly enhance trust and perceived reliability (Schmitt et al., 2024). However, challenges persist, as overly detailed explanations can lead to cognitive overload (Linder et al., 2021) and overreliance on incorrect system outputs (Gorwa et al., 2020; Mohseni, Yang, et al., 2021). Furthermore, the effectiveness of XAI in improving users' mental models and decision-making has yet to be fully explored. Studies like those of Nguyen et al. (2018) and Mohseni et al. (2021) demonstrate that XAI can enhance users' ability to assess AI predictions. However, the practical implications for real-world systems remain unclear.

A central challenge in designing XAI lies in balancing transparency with usability. Transparency – revealing a system's inner workings – is a prerequisite for understandability but does not guarantee user comprehension (Haque et al., 2023). Effective explanations must account for the target audience's cognitive abilities, expertise, and expectations (Adadi & Berrada, 2018; Gilpin et al., 2018). Research suggests that a user-centric approach, emphasizing interpretability over mere transparency, is particularly critical for non-expert users (Cirqueira et al., 2020). In disinformation detection, this challenge is amplified by the inherent complexity of the task and the ethical considerations surrounding content moderation. Researchers have proposed various explanation modalities, such as attention-based visualizations and natural language explanations, to address concerns about fairness and censorship (Guo et al., 2022). These concerns are especially relevant for digital platforms that rely on algorithmic content curation, where platform legitimacy and business model sustainability depend heavily on users' trust in moderation systems (Wanner et al., 2022). While these efforts align with regulatory frameworks like the European Union's AI Act, the real-world impact on user understanding and trust has yet to be comprehensively evaluated. Moreover, most existing studies on XAI focus on technical metrics such as fidelity, feature importance accuracy, or computational efficiency (Wells & Bednarz, 2021). These metrics, however, do not adequately address how users perceive explanations in real-world contexts. There is a clear gap in the literature regarding comprehensive evaluation frameworks incorporating user-centered metrics such as comprehensibility, trust, and usability. Additionally, few studies consider the influence of demographic or social background on how explanations are understood and trusted. As highlighted by Binder et al. (2022), integrating linguistic rules or domain-specific context can enhance explainability in real-world systems like online review platforms, offering a parallel to disinformation detection tools.

To address these research gaps, this study designs and evaluates an XAI artifact tailored to disinformation detection, guided by theoretically grounded design principles and rigorous user feedback. By combining qualitative and quantitative evaluations, including a large-scale online study, we aim to contribute new insights into how explainability features can be more effectively communicated and evaluated from a user-centered perspective. Our work builds on existing XAI frameworks but emphasizes the importance of integrating user feedback into the design and evaluation process to ensure that AI systems are transparent but also comprehensible and trustworthy.

6.3 Research Approach

As a problem-solving paradigm, DSR focuses on the creation of artifacts to provide both descriptive and prescriptive knowledge and innovative solutions (March & Smith, 1995; Vom Brocke et al., 2020). In the HCI community, DSR is an established method to support the iterative development of technical artifacts focusing on effective human use (Adam et al., 2021; Herm et al., 2022). With their six-step research procedure, Peffers et al. (2007) introduce a structured approach to problem-centered DSR projects. To thoroughly answer our research question, we conduct two DSR cycles following their established procedure of problem identification, definition of objectives, design and development, demonstration and evaluation, and communication (Peffers et al., 2007). While our first DSR cycle focuses on the artifact's relevance (Hevner, 2007), rigorously evaluating the problem space by conducting a structured literature review and an in-depth qualitative analysis of user feedback on initial design guidelines (Gurzick & Lutters, 2009), the second cycle strengthens the evaluative rigor (Hevner, 2007) by quantitatively evaluating refined design guidelines and associated hypotheses in an online experiment (Peffers et al., 2012) with fully-functioning XAI prototypes (see Figure 24).

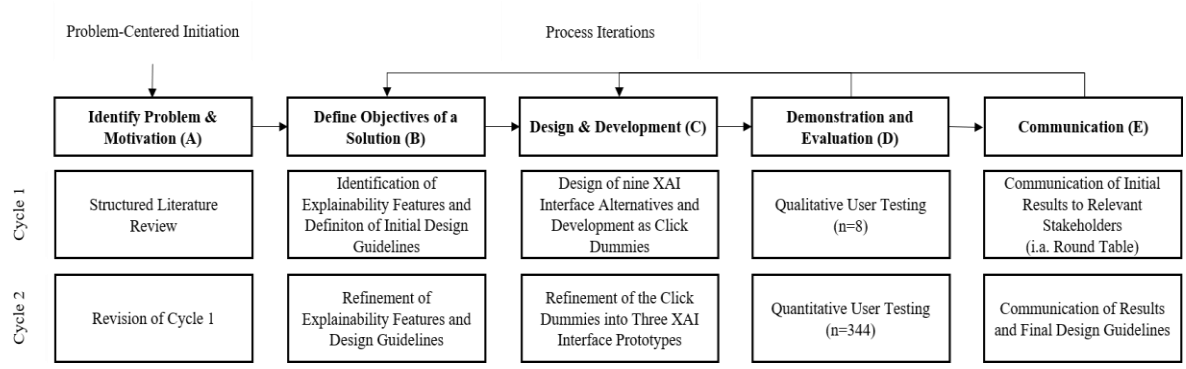

Figure 24. Overview of the DSR approach.

6.3.1 Conduction of the First DSR Cycle

In this chapter, to evaluate the problem space thoroughly and motivate potential solutions (Peffers et al., 2007), we conduct a structured literature review following Webster and Watson (2002) (A). Implementing the PRISMA workflow (Page et al., 2021), we structurally identified and screened literature dealing with applying XAI in front-end design, resulting in the analysis of 57 literature endeavors. The literature review's results inform the second and third research activities of our first cycle: To define preliminary objectives for a solution (Peffers et al., 2007), we derive initial design guidelines for developing a disinformation detection tool on digital discussion platforms (Gurzick & Lutters, 2009), emphasizing the critical role of end-user perspectives in the successful design and adoption of such systems (B). The design and development of DSR artifacts comprise the derivation of functionality and architecture based on solution objectives and the artifact's creation (Peffers et al., 2007). Thus, we implement the guidelines (Lukyanenko et al., 2017) in nine mockups for an XAI disinformation detection tool (C). In Chapter Seven, to demonstrate the artifact's usability and evaluate the extent to which

the solution objectives are met (Peffers et al., 2007), we cover the fourth and fifth DSR activities simultaneously in the conduction of an on-site qualitative user study in the form of a focus group (Tremblay et al., 2010) with n=8 users (D). We conclude the first DSR cycle by communicating the initial findings to practicing professionals (Peffers et al., 2007), among other things, through a practitioners' round table (E).

6.3.2 Conduction of the Second DSR Cycle

Following the iterative nature of DSR research (Hevner, 2007; Peffers et al., 2007), we revise our initial DSR cycle and its insights from the qualitative study and the practitioners' feedback (A) to refine our solution objectives (B). Building on our revised design guidelines (Prat et al., 2015), we further develop the XAI interface click-dummies into three fully functioning XAI prototypes (C). In Chapter Eight, we then set out to quantitatively demonstrate and evaluate our solution artifact (Peffers et al., 2012; Venable et al., 2016) by designing and conducting an online experiment with n=344 participants (D). Using a between-subject experimental design (Sonnenberg & vom Brocke, 2012), the online study compares the artifacts' suitability to improve comprehensibility, usability, and trust compared to a baseline AI system with no explanations. Finally, the study's findings inform the development of integrated design guidelines for XAI-based systems in disinformation detection (E).

6.4 Problem Awareness (A)

In this work, we set out to design an XAI-based system to foster user trust, comprehension, and usability in online disinformation detection. Research has shown that XAI offers promising opportunities to provide interpretable insights into AI decision-making processes. However, evaluations predominantly emphasize technical metrics, such as fidelity and computational efficiency, while overlooking how human users perceive and use explanations in the frontend. This gap is especially pressing in disinformation detection, where explanations must balance transparency with usability while navigating ethical concerns like bias and fairness. Moreover, current evaluation frameworks inadequately address how frontend designs influence user understanding, trust, and satisfaction. To address these critical gaps, there is a need to systematically investigate how frontend designs of explainability features can be optimized to support responsible and user-centric AI systems. This study responds to this need by focusing on designing and evaluating explainability interfaces tailored to disinformation detection.

To gain a structured overview of the current state of frontend design in XAI research and its application for disinformation detection, we conducted a structured literature review based on Webster and Watson (2002). Figure 25 represents the workflow implemented in this paper, resulting in 57 papers included in the final review. An overview of the results will be given below.

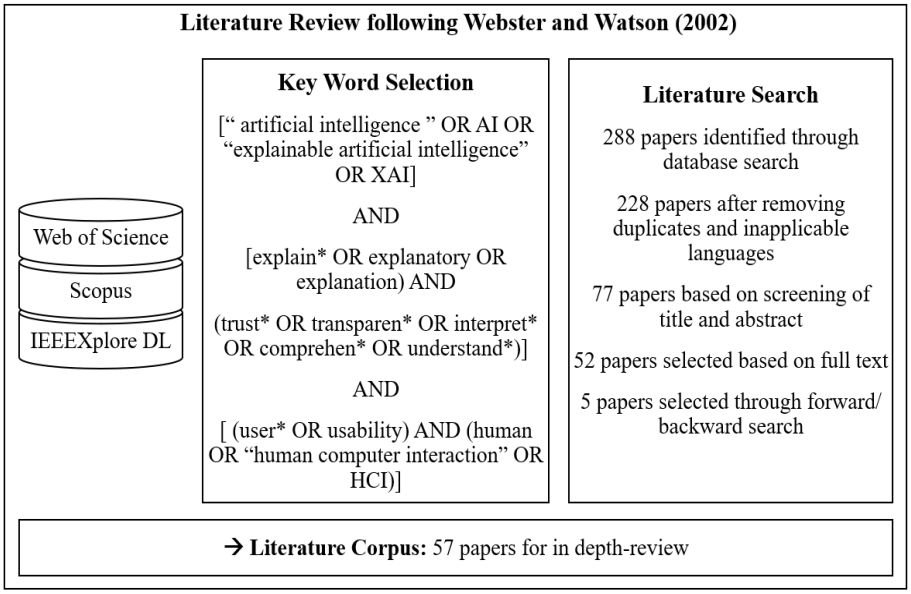

Figure 25. Workflow guiding through the review process.

The scientific domain of XAI, particularly the front end, is highly recent, with over 70% of relevant articles published between January 2021 and August 2023. Healthcare (15.8%) and deception detection (14.0%) are the most prominent domains. However, only three studies in the latter domain focus on the detection of online disinformation. Image classification tasks receive special emphasis, while textual data classification is sparse. Visual explainability features are the most common (50.9%), followed by multimodal (29.8%) and textual (15.8%) features. Local explanations (52.6%) are more prevalent than global explanations (7.0%), with 40.4% of sources combining both. In line with this paper's focus, 80.7% of the literature targets inexperienced end-users.

Subsequently, the literature corpus was analyzed and organized into systematic clusters based on the sources’ main foci. Table 6 summarizes the investigated literature and highlights key findings. 14 explainability features are presented as representatives of their variations and individual modifications, along with a brief description.

Explainability Feature	Description	Cue Type	Literature
Heatmap (saliency)	Regions of images, functions, or text are highlighted graded after their importance	Visual	Selvaraju et al. (2017); Hudon et al. (2021); Rieger and Hansen (2020); Kim et al. (2020); Lewis et al. (2021); Kumar et al. (2021)
Display and exploration of similar instances	Depending on the type of data, similar instances offer additional insights into the feature space	Any	Rjoob et al. (2021); Hwang et al. (2022)
Superpixels (saliency)	Meaningful segments of images or functions are emphasized	Visual	Guillemé et al. (2019); Trinh et al. (2019); Heimerl et al. (2020); Baur et al. (2020); Apicella et al. (2021)
Simple plots (feature importance, partial dependence, other)	Visualizing relationships between input and classification	Visual	Banerjee et al. (2023); Ekanayake et al. (2023); Nguyen et al. (2018); Dey et al. (2021)
Counterfactuals	Visualizing the extent of alteration required in input features to change the model's output	Visual	Confalonieri et al. (2021), Le et al. (2023); Cheng et al. (2020); Vermeire et al. (2022); Singla et al. (2023)
Confidence score	Certainty of a specific classification	Textual or visual	Le et al. (2023)
Concepts (normative, comparative, other)	Cluster of pixels that conveys an idea	Visual	Cai et al. (2019); Huang et al. (2022)
Generative representations (various)	Individually visualized relationships or impact of input features	Visual	Kumar and Sharma (2021); Alves et al. (2020); Kubat and Kubat (2017); Linse et al. (2022); Schreiber and Bock (2019)
Natural language explanations (template, predefined, collaborative)	Human-like textual or auditory explanation of the model's workings	Textual	Das et al. (2023); Mencar and Alonso (2019); Wang et al. (2019); Dong et al. (2021); Zhang et al. (2022); Zhang et al. (2023)
Conversational agent	Multi-way interactive natural language explanations	Textual or auditory	Malandri et al. (2023); Hepenstal et al. (2021); Khurana et al. (2021)
Auditory	Spoken natural language explanations	Auditory	Schuller et al. (2021)
Haptic (feature importance, rules)	Physicalization of other explainability features	Haptic	Colley et al. (2022)
Multimodal (combination through an interface)	Individual combination of multiple explainability features	Any combination	Chromik et al. (2021); Schultze et al. (2023); Park et al. (2022); Finzel et al. (2021); Weitz et al. (2021); Zytek et al. (2021); Kadir et al. (2023); Cirqueira et al. (2020); Hoque and Mueller (2021); Tamagnini et al. (2017); Salako et al. (2021); Zhu et al. (2022); Kerzel et al. (2022); Mohseni et al. (2021)
Keyword contribution	Degree of contribution of a single keyword on the classification	Textual or visual	Mohseni et al. (2019); Linder et al. (2021)

Table 6. Summary of the literature review's key findings.

Among other things, the reviewed literature examines various XAI models designed for detecting forms of deception. A notable commonality among these approaches is the absence of visual data as input features, except for one approach tailored explicitly to identifying deepfake videos (Trinh et al., 2021). The emphasis on the textual dimensions is apparent, leading to the recommendation to prioritize this input type when developing an XAI approach for disinformation detection. Consequently, visual data is not considered for heatmap overlays, which are exclusively applied to highlight the contribution of keywords in the textual input. The systematic, iterative software development approach, as proposed by Basil and Turner (1975), advocates beginning with a relatively simple application and gradually introducing new features and enhancements iteratively. This iterative process ensures the delivery of high-quality solutions. The initial focus is on the textual dimension, with the potential implementation of extensions or enhancements in subsequent iterative cycles. Moreover, Mohseni et al. (2021) highlight the importance of carefully balancing explanations in terms of simplicity and information content. Overly dense explanations may lead to rejection by end-users, potentially harming a trustworthy human-machine relationship. This observation further supports the advocated systematic software development process.

The representation of confidence in a prediction is deemed simple and valuable for building trust between humans and AI (Le et al., 2023). However, the relevance of the confidence score may be significant only when it surpasses a specific threshold, especially for inexperienced end-users. Therefore, low scores indicating low confidence in a classification may be streamlined. Shu et al. (Shu et al., 2017) propose incorporating diverse metadata input features into a disinformation classification model to improve performance. While the expected benefit of input metadata on performance is acknowledged, it remains uncertain whether end-users perceive an explainability feature relying on metadata as helpful and contributive. Thus, in line with Basil and Turner (1975), it is suggested that metadata explainability features be excluded in the initial approach. Natural language explanations fully expand only on demand and summarize the most influential features in a classification that aligns with the criteria outlined by Mohseni et al. (2021) for simple yet effective explanations. While often expected to emulate human behavior, conversational agents may face challenges when primarily dedicated to specific applications due to their limited functionalities and knowledge (Brendel et al., 2020; Hepenstal et al., 2021). Consequently, the potential for user frustration arises, which may be detrimental to trust in human-machine interaction. In the context of disinformation, Mohseni et al. (2019) distinguish two kinds of interpretability: algorithmic interpretability and human interpretability. Algorithmic interpretability assists machine learning experts in visualizing model parameters, inspecting behavior, and improving performance. Human interpretability aims to provide transparency for inexperienced end-users by offering comprehensible explainability features to elucidate how a model works and how decisions are made. This form of interpretability is crucial for fostering trust in the human-machine relationship, aligning with the objectives of this work.

In summary, the literature underscores the importance of simplicity and clarity in explainability features to build trust among inexperienced end-users. However, this focus reveals a gap in user-centered research on how these explanations are best delivered and experienced on the front end of XAI applications. Addressing this gap is crucial for developing XAI systems that are not only transparent but also user-friendly across diverse application domains, including disinformation detection.

6.5 Solution Objectives (B)

Building upon the literature review, the findings can be distilled into solution objectives in the form of design guidelines (Gurzick & Lutters, 2009), generalized for constructing an XAI model to detect disinformation on digital platforms:

1. *Preserve the original GUI*. Maintain the existing platform's GUI to ensure a seamless transition for users and uphold their established interaction habits. This helps avoid disruption and maintains usability and comfort (Garaialde et al., 2020).
2. *Balance simplicity and clarity*. Strive to balance simplicity with an effective explanation of the model's decisions. Use iterative evaluations to refine explanations, ensuring they are clear and comprehensible without becoming overly complex (Mohseni et al., 2021).
3. *Empower inexperienced users*. Design features to be accessible to inexperienced users, ensuring they retain decision-making authority and can effectively navigate and understand content. This supports user empowerment and fosters trust (Mohseni et al., 2019).
4. *Supplement confidence scores*. Use confidence scores as a supplementary feature to indicate prediction certainty. Simplify the presentation to avoid overwhelming users while providing essential information (Le et al., 2023).
5. *Implement colored saliency for critical insights*. Highlight significant keywords in the text. Use clear color schemes and balance complexity to maintain clarity (Chromik, 2021; Selvaraju et al., 2017).
6. *Provide expendable natural language explanations*. Design natural language explanations to be concise and initially hidden, expanding upon user interactions. This approach keeps the interface clean while allowing users to access detailed information as needed (Das et al., 2023).
7. *Exclude conversational agents*. Avoid integrating conversational agents in the initial model to prevent potential user frustration. Focus on delivering clear and direct explanations through other features (Brendel et al., 2020; Hepenstal et al., 2021).
8. *Evaluate explainability features.* Conduct practical evaluations of explainability features to assess their effectiveness and impact in real-world scenarios. This evaluation is essential for understanding how well the features meet user needs and improve the overall user experience (Mohseni et al., 2021).

9. *Develop and improve iteratively.* Follow an iterative development approach to enhance the application continuously. Incorporate user feedback and adapt to evolving needs and technological advancements to ensure ongoing improvement and high quality (Basil & Turner, 1975).

6.6 Click-Dummies of an XAI Interface (C)

In the subsequent phase of our design process, we systematically implemented the solution objectives outlined for developing our XAI-based disinformation detection tool. This process began with preserving the platform's original GUI to ensure a smooth integration of new features (Guideline 1). For embedding the system's initial warning, we designed three alternatives (Figure 26): An overlay hiding the classified post, a banner above the post, and a banner below the post – all expandable upon desire (Das et al., 2023).

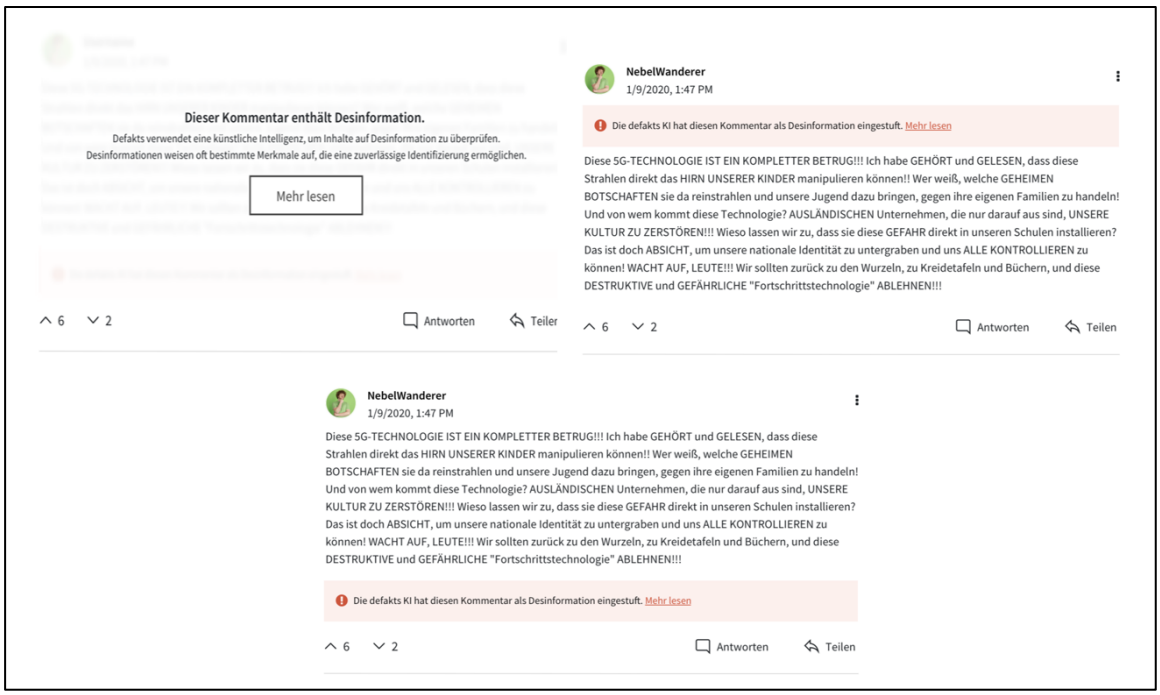

Figure 26. Design alternatives for different flaggings of classified posts.

We balanced simplicity with clarity, ensuring that each explainability feature was effective and easy to understand for users with varying levels of expertise (Guidelines 2 & 3). The development focused on integrating confidence scores and text highlighting to enhance the transparency of the model's predictions (Guidelines 4 & 5). Our designs for the display of confidence scores (Figure 27) either showed a display in percentages (Schmidt et al., 2020) or, to provide an even more simplified concept that may cater to especially inexperienced users, a gradation of "low", "medium", and "high" (Mohseni et al., 2019; Mohseni et al., 2021).

Figure 27. Design alternatives for confidence score displays.

In order to emphasize text parts that were relevant for the system's prediction (see Figure 28), we prepared a design displaying highlighted parts directly in the classified post (Chromik, 2021; Selvaraju et al., 2017). As an alternative, another design suggests citations of relevant passages in the explanatory text to keep the initial post clean and simple.

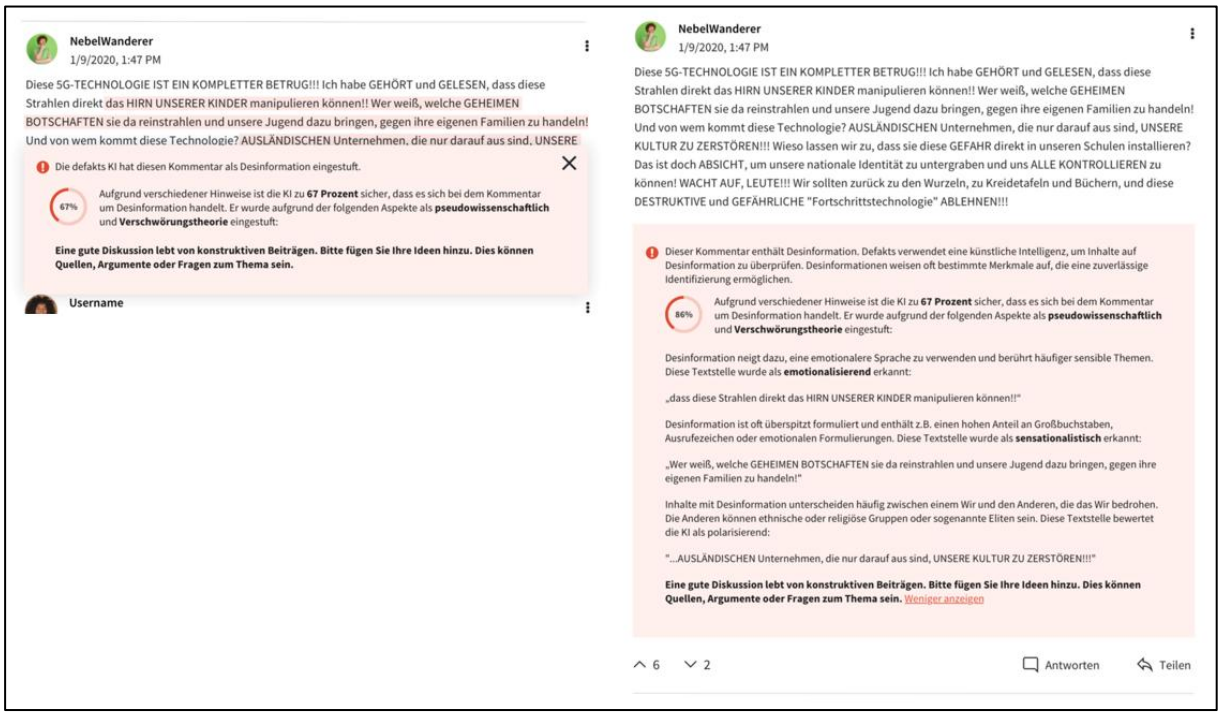

Figure 28. Design alternatives for highlighting parts relevant for the system's classification.

To address user needs for understandable explanations, we designed expandable natural language explanations (Guideline 6). To ensure the provision of critical information while striving to avoid information overload, a longer, more detailed explanation and a shorter explanation were developed (see Figure 29).

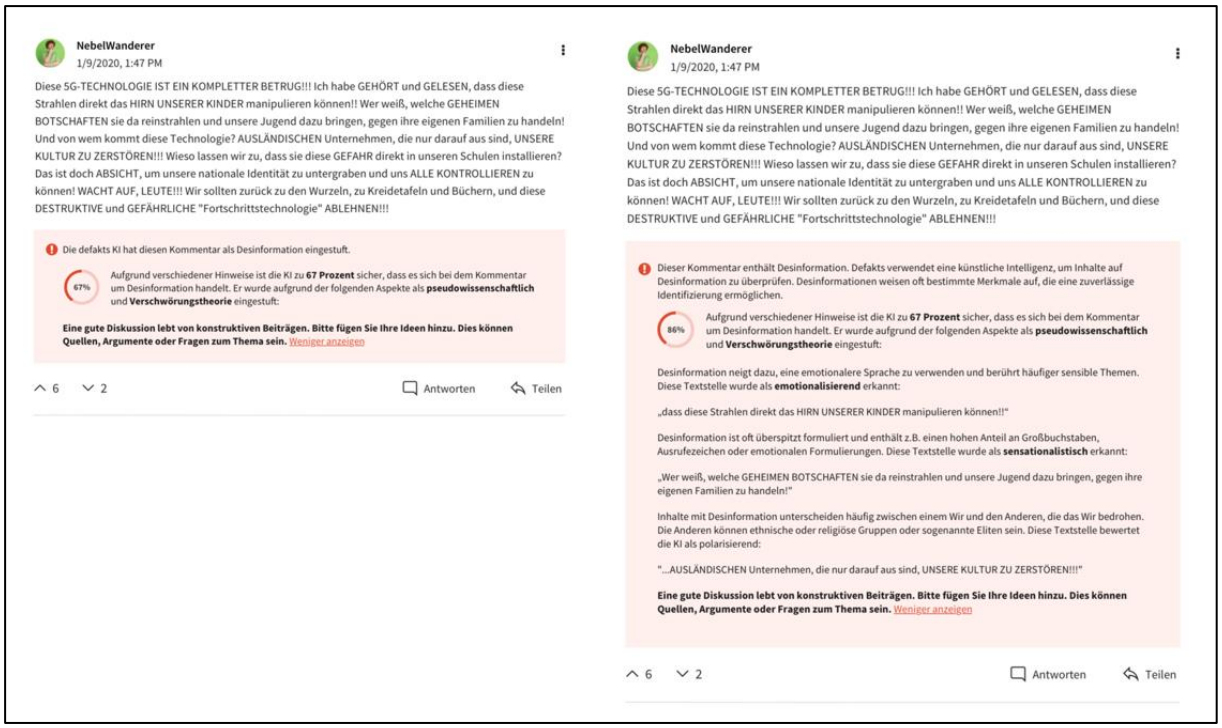

Figure 29. Design alternatives for different explanation lengths.

In alignment with the literature review's findings, conversational agents were excluded from the initial design to prevent potential frustration (Guideline 7). These considerations culminated in nine distinct design suggestions, which were visualized in mockups to illustrate the proposed features and their integration into the XAI systems. The mockups serve as a foundation for further refinement and practical evaluation in qualitative user testing (Guideline 8), guiding the ongoing development of a robust and user-centric disinformation detection tool (Guideline 9).

7 Preliminary Insights into User Preferences for Disinformation Detection Systems: A Qualitative Approach

7.1 Qualitative User Testing (D)

To demonstrate and evaluate the effectiveness of an XAI tool detecting online disinformation, it is essential to understand the perspectives of end-users, which are crucial for the successful application of such tools. By focusing on the target group's perspectives in a qualitative focus group (Tremblay et al., 2010), we seek to ensure that the design of these systems aligns with their preferences and enhances their trust and understanding. Such alignment is pivotal for the responsible development and effective integration of AI-based disinformation detection tools. In this chapter, we will first elaborate on the design and conduct of the study before presenting its results in detail.

7.2 Procedure

We conducted qualitative user testing to evaluate design preferences for our developed XAI mockups. The goal was to gain an in-depth understanding of how diverse users perceive and interact with the system's output. The study involved eight participants, equally divided by gender and aged 24 to 64, recruited via the recruiting platform TestingTime to ensure diversity in demographics and professional backgrounds. Two on-site sessions were held in February 2024, each lasting two hours with four participants. Led by two researchers and two practitioners, these sessions assessed responses to our nine different design options for the AI system's output display. The sessions followed a structured format:

1. *Introduction and Briefing:* Participants were briefed on the study's purpose and the confidentiality of their participation.
2. *Design Presentation:* Nine designs were sequentially presented, with explanations of each format's rationale.
3. *Individual Questionnaire:* Participants completed a questionnaire capturing their initial reactions and preferences, with the freedom to review the designs as needed.

4. *Joint Discussion:* A moderated group discussion explored participants' thoughts, aiming to uncover deeper insights into usability and preferences.

Data collection included questionnaires, observational notes, and discussion transcripts, which were analyzed using evaluative qualitative content analysis (Kuckartz, 2012). This method involved reviewing and summarizing the data through inductive category formation (Mayring, 2015), focusing on identifying key themes, user preferences, and potential concerns to inform the tool's further development.

7.3 Results

These findings provide initial insights into participants' encounters with disinformation, their familiarity with AI technologies, and preferences regarding the presentation of warnings in relation to posts. The following section delves into further details derived from these responses.

In response to the question "Have you already encountered disinformation? If so, where?" five out of the eight participants confirmed that they have encountered instances of disinformation. The platforms most frequently cited for encountering disinformation include social media platforms such as Facebook, X (formerly Twitter), YouTube, and Instagram. Participants noted that these encounters primarily revolved around political discourse, occurring in both public forums and private discussions. When asked about their experience with AI-based systems, specifically where they have consciously gained experience, six out of the eight participants indicated that they have used AI-based systems before. Common experiences cited include interacting with generative AI (ChatGPT) and other chatbots.

Before receiving an explanation of the system's classification, users were provided with a brief warning indicating that a post was labeled as potential disinformation. Regarding the placement of the warning messages in relation to classified posts, participants were asked, "Where should the warning be placed (before the post, after the post, or post hidden)?". Six out of the eight participants expressed a preference for having the warning displayed above the post. When asked to choose between brief and detailed explanations for AI classifications, all eight participants preferred the longer version of the text. Common explanations for this strong preference were the increased trust and understandability provided by more comprehensive explanations. Participants claimed that it "should be possible to find out why the AI classified a post in this way" [TN2] and that "it makes the reference more credible, and this strengthens trust in AI" [TN6]. However, it was also posited that detailed explanatory texts could potentially induce fatigue over extended periods:

> "What may also be annoying for some – not for me – is the length of the text. If it feels like it pops up with every post, you definitely lose interest at some point. But on the other hand, I wouldn't really know how to minimize that." [TN4]

Furthermore, participants exhibited diverse perspectives regarding the display of confidence scores in the context of disinformation detection. Several participants expressed a consistent preference for the utility of confidence scores, with four individuals finding them consistently helpful. Conversely, three participants indicated that they never found confidence scores helpful, while four others believed they were only beneficial if they exceeded a specific threshold. The threshold for what constitutes a helpful confidence score varied considerably among participants, ranging from as low as 20% to as high as 80%. Although the concept of confidence scores was explained to all participants during the briefing and in the questionnaire, it became evident that the comprehension of confidence scores poses challenges for laypersons, rendering them prone to misinterpretation. Consequently, this factor impacts the perceived utility of displaying such scores and the perceived usefulness of the provided information. One participant raised concerns about the clarity of low percentage scores without concrete examples [TN2], while another participant made the following statement:

> "I don't think measuring in percentage is a suitable unit of measurement for comments in a forum. In reality, every post on the forum will not be 100% compliant, and it becomes visually annoying that an AI is checking people" [TN7]

Here, it becomes obvious that confidence scores can be easily misunderstood as to what they actually refer to. If users assume, for example, that such a score evaluates the credibility of a person instead of the system's own confidence in its prediction, one can expect a corresponding rejection of its display. Additionally, participants were asked which variant of confidence score display (as a percentage or gradation in low, medium, and high) they preferred if such a score were to be shown. Here, a clear preference became visible: Seven out of eight individuals favored a percentage display. One person stated that they would find a display in percentages "clear and comprehensible – 'medium' is kind of vague so I would rather interpret it, hm, that's a bit unclear now. Whereas with '67%' I would have the feeling that I have clear information. Seems precise, convincing as if the AI knows what it's doing." [TN2]. Other participants expressed similar sentiments, claiming that they "can visualize the probability better with percentages" [TN3] and that a display of gradation does "not provide me personally with a basis on which I want to rely" [TN1]. However, one participant offered a contrasting viewpoint, preferring simpler classifications:

> "It's a simpler classification with three levels. At up to 100% everyone assesses the situation for themselves. Some find 60% completely reliable and some only

> from 90% for example. The percentage variant offers too much scope for interpretation and making decisions based on gut feeling." [TN4]

These varied responses illustrate a general preference for percentage-based confidence scores, although some participants see value in more straightforward classification. The divided opinions on the usefulness of a confidence score stand in stark contrast to the consensus regarding the importance of highlighting text passages relevant for classification: All eight participants found it helpful to display the text passages that the AI considers indicative of disinformation. In this context, individuals indicated that the highlighting serves multiple functions for them, going beyond the direct interaction with the system:

> "This also makes the AI's advice reliable and ensures that it is given more credence. At the same time, it sensitizes the reader to recognize disinformation more easily in the future." [TN6]

Other participants added that the highlighting helps to "understand and comprehend things better" [TN8], making it "transparent how the AI has assessed what has been classified as 'red' and I can check for myself what I think of it." [TN7]. These unanimous responses highlight the importance of transparency in AI assessments, as displaying specific text passages helps users understand and trust the system's conclusions. Finally, participants were asked how they preferred the AI to display text passages that indicated disinformation five out of eight individuals favored color highlights in the original posts instead of citing relevant text passages in the explanatory text. One attendee explained their preference for colored highlights in the original post as follows:

> "Striking colors are an eye-catcher. It also reminds me a bit of my school days: important information was marked with a highlighter, here too. So why make it complicated and quote the article again in a large block instead of making the info text short and concise and simply using and including the existing post?" [TN4]

Another participant supported this preference, claiming that "readers are shown even more clearly which passages and statements are involved" [TN5] and "text passages can be found much more quickly" [TN5]. In contrast, participants who preferred citations of the passages in the explanatory text argued that this variant is better structured. One person stated that they find colorful highlights "too confusing, as you have to constantly open pop-up windows for an explanation." [TN6]. Consequently, they claimed this "would discourage me from reading the explanations and thus deprive me of the opportunity to gradually recognize disinformation myself" [TN6]. These statements highlight the participants' general preference for color highlights in the original posts, as this method is seen

as more intuitive and clear. However, a significant minority preferred citations in the explanatory text for better structure and ease of understanding.

7.4 Stakeholder Communication (E)

Hevner et al. (2004) stress the importance of effective communication of DSR research results "both to a technical audience (researchers who will extend them and practitioners who will implement them) and to a managerial audience (researchers who will study them in context and practitioners who will decide if they should be implemented within their organizations)" (Hevner et al., 2004, p. 82). We followed this approach and presented our artifact through various presentations at practitioners' conferences and through media outlets to provide critical insights into user interactions with XAI-based systems for disinformation detection. Furthermore, we communicated and discussed results focusing on an expert audience, conducting a round table format together with 16 researchers from various disciplines and practitioners from domains including politics, citizen participation, communication science, machine learning, and fact-checking in March 2024. Through these discussions, we identified key areas that require attention in the development of XAI tools, specifically emphasizing the importance of comprehensibility, user-friendliness, and trustworthiness. The expert feedback underscored the necessity for designing systems that are not only effective in detecting disinformation but also comprehensible and reliable from a user perspective.

8 Validating User Preferences for Disinformation Detection Systems: A Quantitative Study

8.1 Revision of the First Cycle and Objective Refinement (A, B)

The qualitative user testing's findings and the round table discussion underscore the importance of designing AI-based disinformation detection systems that are transparent, user-friendly, and trustworthy. By addressing user preferences and concerns early on, developers can create more effective tools that not only detect disinformation but also educate and empower users to navigate digital spaces more critically. This approach is essential for fostering a more informed and resilient digital public. Accordingly, alongside our formulated guidelines, the results discussed at the round table inform the further development of our prototype by deciding which design choices can be implemented directly (indicated by the participants' consensus) and which design choices may need further testing in the future (indicated by the participants' disagreement or varying preferences). Therefore, the initial warning appears above the post. Users can view the explanation by clicking "Read more" (Guidelines 1 & 2). The system provides a detailed explanation: The note explains the characteristics on which the classification is based and which text passages the AI is referring to (Guidelines 3 & 6). Although there is a slight observable preference for displaying a confidence score, it may only be displayed above a certain value. Accordingly, several prototype variants are designed to address these diverse needs. One variant will offer explanations without a confidence score, while the other will include it (given in percentage) (Guideline 4). Furthermore, the specific text passages of a post that are relevant to the AI's classification shall be displayed. Participants favored both the option of color highlighting in the original post and the citation of the text passages in the explanatory text of the classification. As there was a slight tendency towards color highlighting in the original post, this tendency will be reflected in the design of the prototypes for the quantitative study (Guideline 5).

Our next step is to test these prototypes through an online study to evaluate the effectiveness of these design choices based on user interactions (Guidelines 8 & 9). To frame our design process within a broader context, we draw on empirical literature examining the impact of XAI on user perceptions. XAI has emerged as a pivotal approach to bridge the gap between complex AI models and user comprehension. Understanding AI systems'

working principles is crucial for users to make informed decisions in various contexts (Haque et al., 2023). In particular, XAI plays a vital role in enhancing its understanding. Understandability specifies whether the features and attributes of a model are easily recognizable by users without knowing its inner composition. XAI ensures that AI systems are not just accurate but also interpretable and transparent, making their operations more comprehensible to users (Arrieta et al., 2020). When explanations are presented appropriately, user understandability significantly increases (Bussone et al., 2015; Cai et al., 2019; Eiband et al., 2019; Hudon et al., 2021). Experimental research shows that a user's knowledge about the system's interactions results in better understandability of the system (Bove et al., 2021; Branley-Bell et al., 2020; Cheng et al., 2020). Users who can grasp how an AI system functions are more likely to find it user-friendly and reliable (Górski & Ramakrishna, 2021). For non-technical stakeholders, clear, concise, and comprehensive information is essential to avoid cognitive overload (Hudon et al., 2021). Properly labeled and explained attributes, along with well-reasoned decisions, are crucial for increasing user understandability (Li et al., 2021). Accordingly, we hypothesize the following:

H_1: XAI leads to a higher degree of perceived understandability compared to AI without an XAI component.

Trust in AI systems can be bolstered by providing contextual information and transparent decision-making processes (Bove et al., 2021; Cirqueira et al., 2020; L. Wang et al., 2019). Moreover, a high confidence level for predictions helps users build trust in the system (Bussone et al., 2015; Ehsan et al., 2021). The explanation should contain enough details regarding the prediction and decision-making procedure so that users can feel confident and trust the system. Too much information could create cognitive overload and decrease users' understanding and trust (Cramer et al., 2008; Hudon et al., 2021; Schmidt et al., 2020). To promote trust in the system, it is recommended to reduce the knowledge gap between the user and the system by collaborating with users during the XAI development lifecycle (Chromik, 2021; Hong et al., 2020; Park et al., 2021). Therefore, we formulate the following hypothesis:

H_2: XAI leads to a higher degree of trust in the system compared to AI without an XAI component.

XAI systems are also shown to positively impact usability (Oh et al., 2018), potentially leading to higher technology acceptance (Davis & Grani, 1989; Venkatesh & Davis, 2000). Furthermore, to increase usability, accessible and interactive interfaces should be designed and developed for non-technical stakeholders (Andres et al., 2020; Brennen, 2020). Involving the stakeholders in the development lifecycle may also increase a system's usability (Chromik, 2021). Therefore, we formulate the following hypothesis:

H_3: XAI leads to a higher degree of perceived usability compared to AI without an XAI component.

These findings and the iterative development process help refine our prototype and establish a framework for the subsequent demonstration and evaluation phase. As we proceed with an online study to assess the impact of these design choices, this approach is situated within the broader context of XAI's influence on user perceptions. Through systematic testing and iteration, the aim is to critically assess the final system's effect on the transparency metrics of understandability, trust, and usability (Haque et al., 2023).

8.2 Prototypes of an XAI Interface (C)

Building on the qualitative study's user feedback, we refined our interface prototypes to enhance user experience and functionality. This iterative development process has led to the creation of three interactive design prototypes for a discussion platform. Each prototype integrates an AI-based system that monitors contributions to a digital discussion and flags suspicious content as potential disinformation. The prototype (Figure 30), used for the first treatment, serves as our baseline system. It shows the system's binary classification of suspicious content without providing further explanations on the system's reasoning behind its prediction.

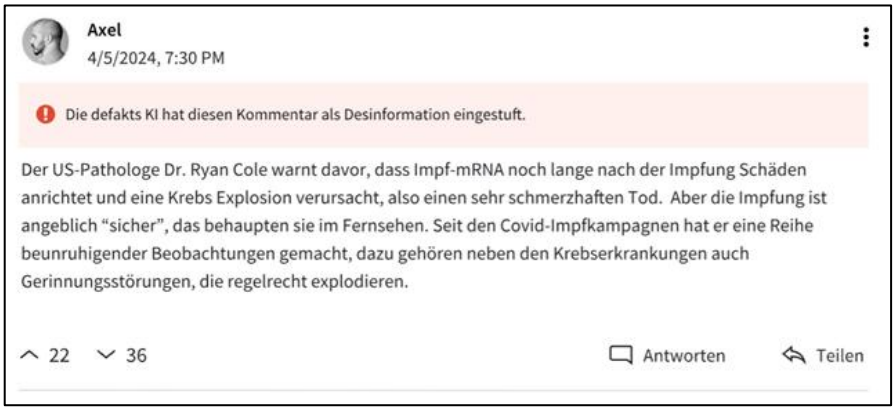

Figure 30. First design prototype without explanations.

The second prototype (Figure 31), used for the second treatment, shows users the system's classification and provides them with additional explanations in an expandable window. Similar to our baseline prototype, a banner appears above each classified post. Upon clicking on "Read more", users are provided with an explanation of why the system recognized the post as disinformation, as well as how the recognized characteristics are indicative of disinformation.

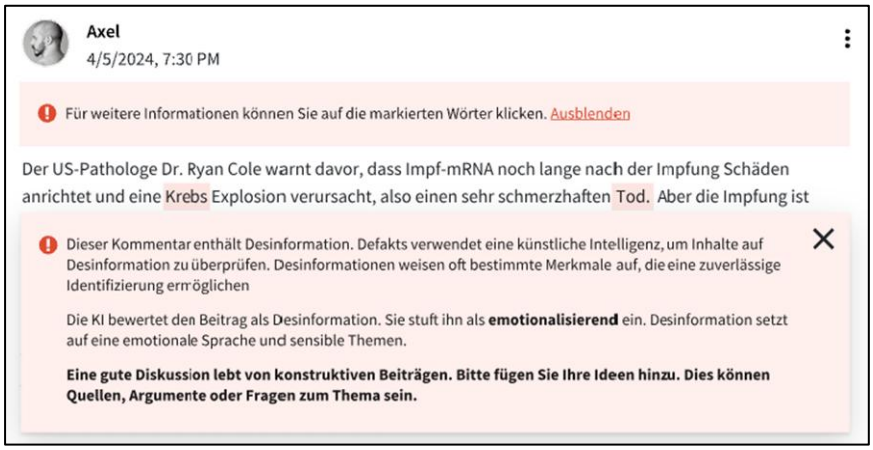

Figure 31. Second design prototype with explanations.

Our third prototype (Figure 32), later used for the third treatment, provides explanations identical to those of our second prototype but supplements them with a confidence score. In this part of the explanation, the system communicates its confidence in its own prediction.

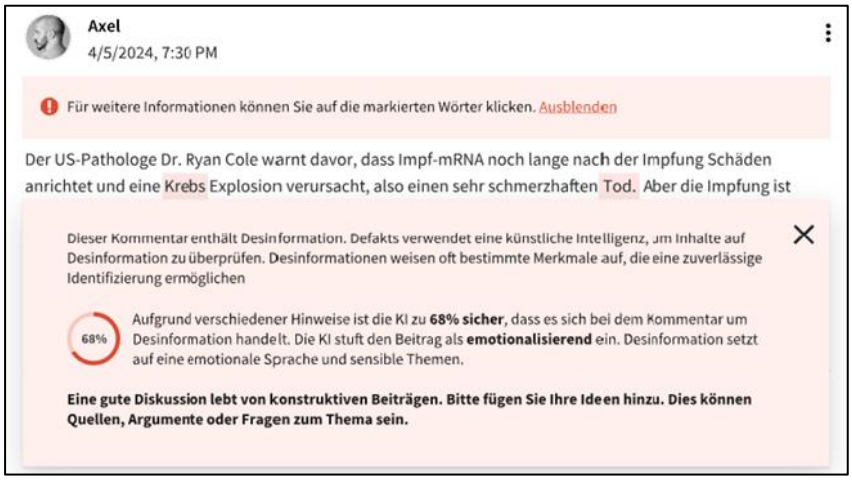

Figure 32. Third design prototype with explanations and confidence score.

8.3 Quantitative Online Study (D)

Strengthening the evaluative rigor of the first DSR cycle, we demonstrate and evaluate our solution artifact in a quantitative experimental approach. In the following, we will first present our approach to designing and conducting the online study before delving into its results.

8.3.1 Procedure

The study, conducted in July 2024, involved the recruitment of 400 participants through the online panel provider Prolific. To ensure data quality, a pre-test was conducted and two attention-check (AC) questions in the form of instructional manipulation checks (IMCs) were included in the questionnaire. These questions were designed to identify inattentive respondents. The first AC question was positioned in the middle of the questionnaire, while the second was placed toward the end. Participants who failed one or both of these questions were excluded from further analysis, resulting in the removal of 56 respondents. After this exclusion process, a total of 344 participants remained in the dataset for analysis. Participants were selected based on their demographic diversity with considerations for age (mean = 32.02, SD = 10.38) and sex (171 male and 173 female). Each participant was presented with one clickable prototype. Participants were informed that the study aimed to investigate user perceptions of an AI-supported tool for detecting disinformation on digital platforms, such as discussion forums. They were provided with a brief overview of the study, including the expected duration of approximately 30 minutes, and were encouraged to respond to the questionnaire honestly and carefully via the initial instructions in the questionnaire. In this study, the prototypes featured an online discussion on the topic of new vaccination technologies in the context of the COVID-19 pandemic (Figure 33). The comment section displayed six comments, two of which were classified as disinformation. The study followed a between-subjects design (Charness et al., 2012) and included three experimental treatments, each aligning with one of our three prototypes.

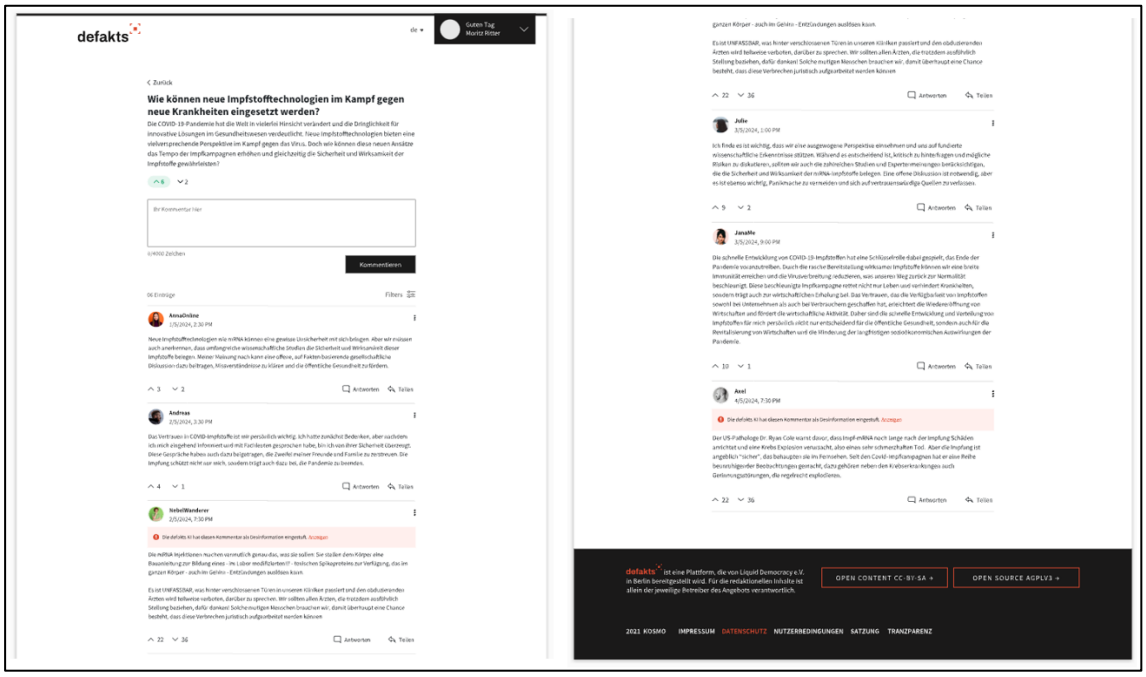

Figure 33. Clickable user interface of the discussion with two classified posts.

Participants were randomly assigned to one of the three groups to ensure that any observed differences in outcomes could be attributed to the experimental manipulation rather than pre-existing differences among participants. The key concepts under investigation included participants' trust in the presented AI-based system, perceived understandability of the provided information, and their overall usability experience. On the basis of established theoretical constructs, these concepts were measured on a 1-7 Likert scale (fully disagree to fully agree). Additional measures were taken to assess participants' demographic characteristics, their propensity to trust, and their prior experience with AI. In order to explore the effects of the experimental conditions on participants' perceptions and behaviors while also accounting for demographic variables, we conducted Kruskal-Wallis tests complemented by Dunn-Bonferroni post-hoc tests and linear regression analyses.

Reliability analyses were conducted for each of the constructs used in the study (see Table 7). The *understandability* construct, consisting of five items (Madsen & Gregor, 2000), showed very good reliability. Further, the *trust* construct, consisting of six items (Merritt, 2011), indicated excellent internal consistency among the items. The *usability* construct, measured by five items (Benbasat & Wang, 2005) demonstrated good reliability. These findings indicate that the items within each scale are sufficiently consistent to be considered reliable measures of their respective constructs.

Construct	Number of Items	Cronbach's Alpha	Average Inter-Item Correlation	Guttman's Lambda 6	Standard Error
Understandability	5	0.88	0.60	0.87	0.010
Trust	6	0.91	0.62	0.90	0.008
Usability	5	0.83	0.50	0.80	0.015

Table 7. Summary of reliability analyses for the measured constructs.

Before conducting the primary analyses, the normality of the data was tested using the Shapiro-Wilk test. The results ($p < 0.001$) indicated a significant deviation from normality, violating one of the key assumptions required for parametric tests such as ANOVA. Given this violation, non-parametric tests were used for the main analyses. To compare the effects of the three treatment conditions, the Kruskal-Wallis test was employed as a non-parametric alternative to the one-way ANOVA. This test was used to assess whether there were statistically significant differences in the dependent variables (e.g., trust, understandability, usability) across the three experimental groups. In addition, multiple linear regression analyses were conducted to explore the impact of demographic and personal factors (e.g., age, educational background, and previous AI experience) on the dependent variables. These analyses allowed for examining how these characteristics might influence participants' responses independent of the treatment effects. Informed consent was obtained from all participants before their involvement in the study, and they were assured of the confidentiality and anonymity of their responses.

8.3.2 Results

The three explainability levels' effect on participants' perceptions of the system's trustworthiness, perceived usability, and understandability, and, as an additional insight, their overall agreement with the displayed classifications were analyzed (see Table 8 and Figure 34 to Figure 37) and are presented in the following.

Statistic	Understandability	Trust	Usability	Classification Agreement
Median scores	Treatment 1: 5.00	Treatment 1: 5.17	Treatment 1: 5.40	Treatment 1: 6.00
	Treatment 2: 5.00	Treatment 2: 4.58	Treatment 2: 5.60	Treatment 2: 5.00
	Treatment 3: 5.00	Treatment 3: 4.67	Treatment 3: 5.40	Treatment 3: 5.00
χ^2(df)	$\chi^2(2) = 1.03$	$\chi^2(2) = 7.91$	$\chi^2(2) = 0.76$	$\chi^2(2) = 15.64$
p-value	p = 0.60	p = 0.02	p = 0.68	p < 0.001
$e^2_{ordinal}$ (95 CI)	$e^2_{ordinal}$ = 0.003 (95% CI [0.000446, 1.00])	$e^2_{ordinal}$ = 0.02 (95% CI [0.0474, 1.00])	$e^2_{ordinal}$ = 0.002, (95% CI [0.000441, 1.00])	$e^2_{ordinal}$ = 0.05 (95% CI [0.02, 1.00])
Post-hoc test Treatment 1 vs 2	Z = -0.09, p.adj = 1.00, d = -0.12	Z = 2.71, p.adj = 0.14, d = 0.26	Z = -0.82, p.adj = 1.00, d = -0.09	Z = 3.15, p.adj = 0.0048, d = 0.43
Post-hoc test Treatment 1 vs 3	Z = -0.92, p.adj = 1.00, d = -0.12	Z = 2.00, p.adj = 0.14, d = 0.26	Z = -0.67, p.adj = 1.00, d = -0.07	Z = 3.65, p.adj = 0.0008, d = 0.43
Post-hoc test Treatment 2 vs 3	Z = -0.83, p.adj = 1.00, d = -0.09	Z = -0.72, p.adj = 1.00, d = -0.06	Z = -0.15, p.adj = 1.00, d = 0.01	Z = 0.48, p.adj. = 1.00, d = 0.12

Table 8. Summary statistics of Kruskal-Wallis test and post-hoc analyses (Dunn-Bonferroni test, Cohen's d).

Understandability. A Kruskal-Wallis test indicated no significant difference in understandability among the three treatment groups with a negligible effect size (Figure 34). Median scores were identical at 5.00 across all treatments. These findings suggest that the inclusion of XAI components does not significantly enhance understandability compared to a basic AI system. Given the consistent median scores and small effect sizes, we cannot confirm hypothesis H_1, which proposed that XAI components would improve understandability.

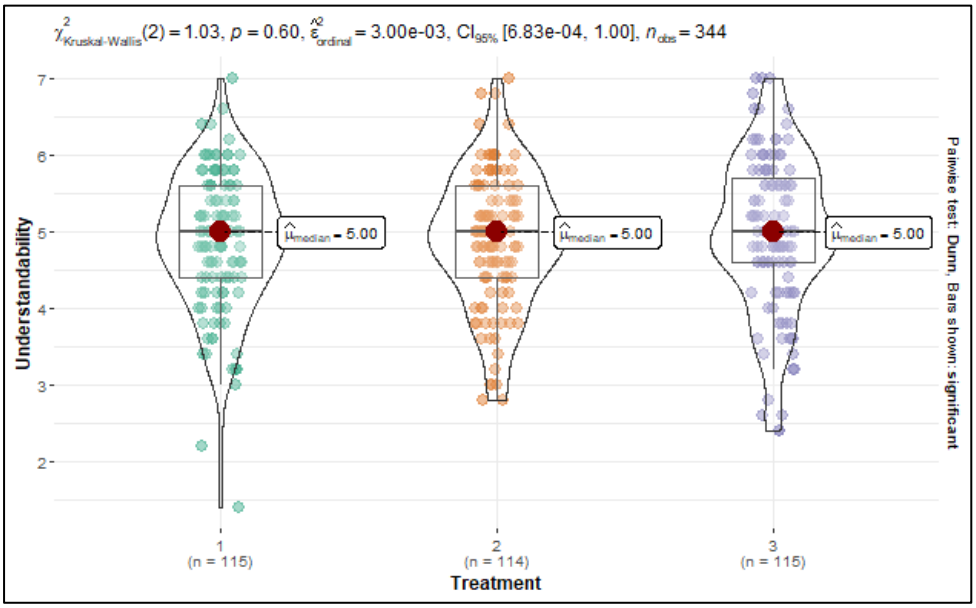

Figure 34. Kruskal-Wallis test of perceived understandability.

Trust. The Kruskal-Wallis test revealed a significant difference in trust scores among the three treatment groups with a small effect size (Figure 35). Median trust scores were 5.17 for treatment one, 4.58 for the second treatment, and 4.67 for treatment three. Dunn-Bonferroni post-hoc tests showed a significant difference between treatment one and treatment two with a small effect size. However, the difference between treatment one and treatment three was not significant. Furthermore, no significant difference was found between the second and third treatment with a negligible effect size. These results indicate that while there is a small but significant difference in trust between the control group and the group with explanations, the presence of XAI components does not lead to higher trust overall. Consequently, we cannot confirm H_2.

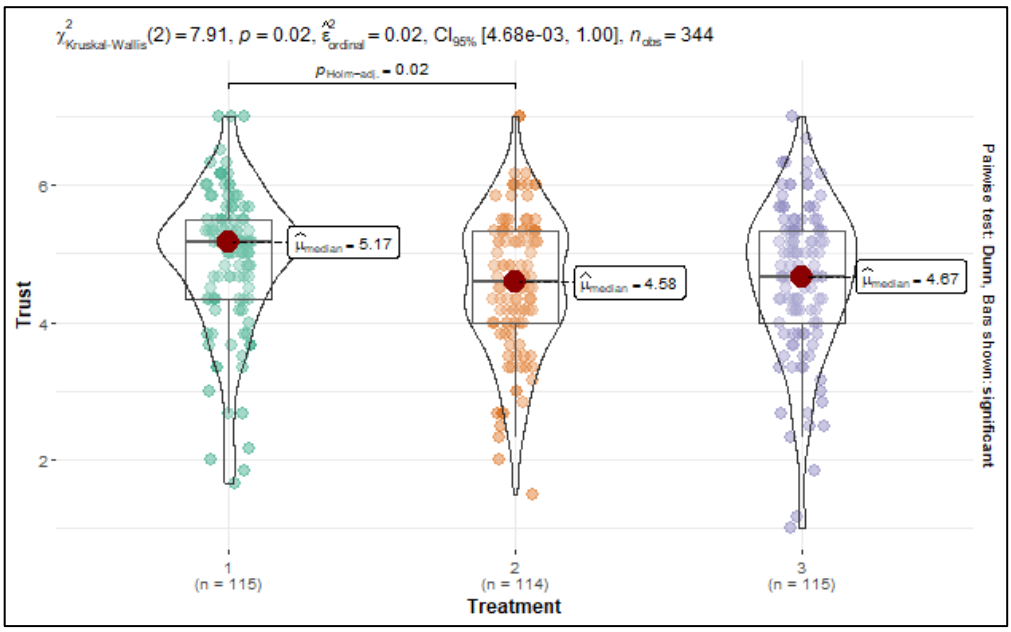

Figure 35. Kruskal-Wallis test of trust in the system.

Usability. The results of the Kruskal-Wallis test show no significant differences in usability scores across the treatment groups with a negligible effect size (Figure 36). Median usability scores were 5.40 for treatment one, 5.60 for treatment two, and 5.40 for treatment three. Confirming the initial observation, Dunn-Bonferroni post-hoc tests also found no significant pairwise differences between treatment one and two, treatment one and three, and treatment two and three. These observations suggest no significant differences in perceived usability among the treatment groups, and the presence of XAI components does not enhance usability over a basic AI system. As such, we cannot confirm H_3.

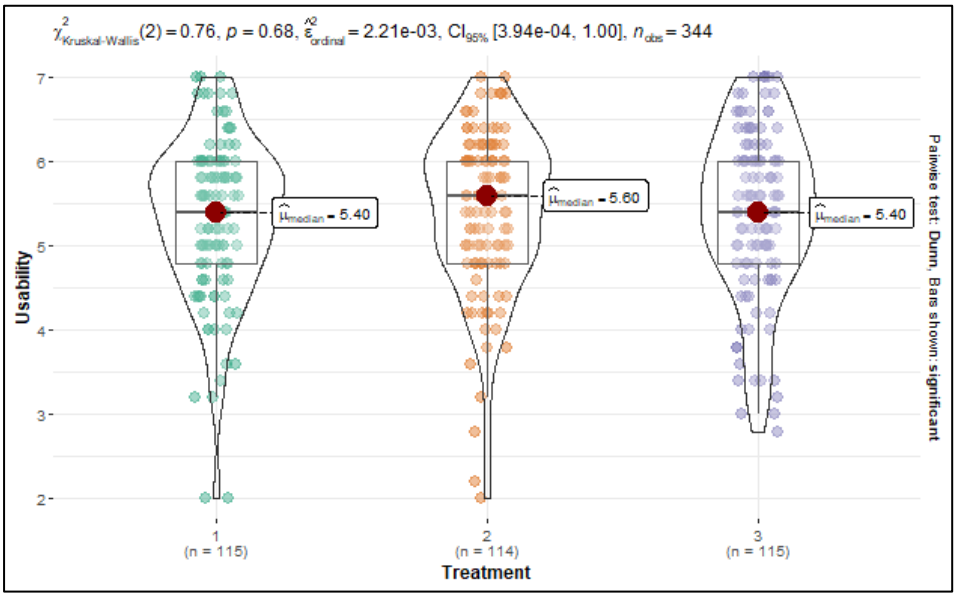

Figure 36. Kruskal-Wallis test of perceived usability.

Classification agreement. For additional insights, we investigated potential differences between the treatments regarding the participants' overall agreement with the displayed classifications (Figure 37). The Kruskal-Wallis test revealed a significant effect with a moderate effect size. Median classification agreement was highest in the control group (6.00), while both treatment groups scored lower (5.00). Dunn-Bonferroni post hoc tests showed that the control group had significantly higher agreement scores compared to both treatment two and treatment three, suggesting small to moderate differences. No significant difference was found between the two XAI treatment groups, with a minor effect size. These findings indicate that the introduction of explanations, with or without confidence scores, actually reduced participants' agreement with the system's classifications.

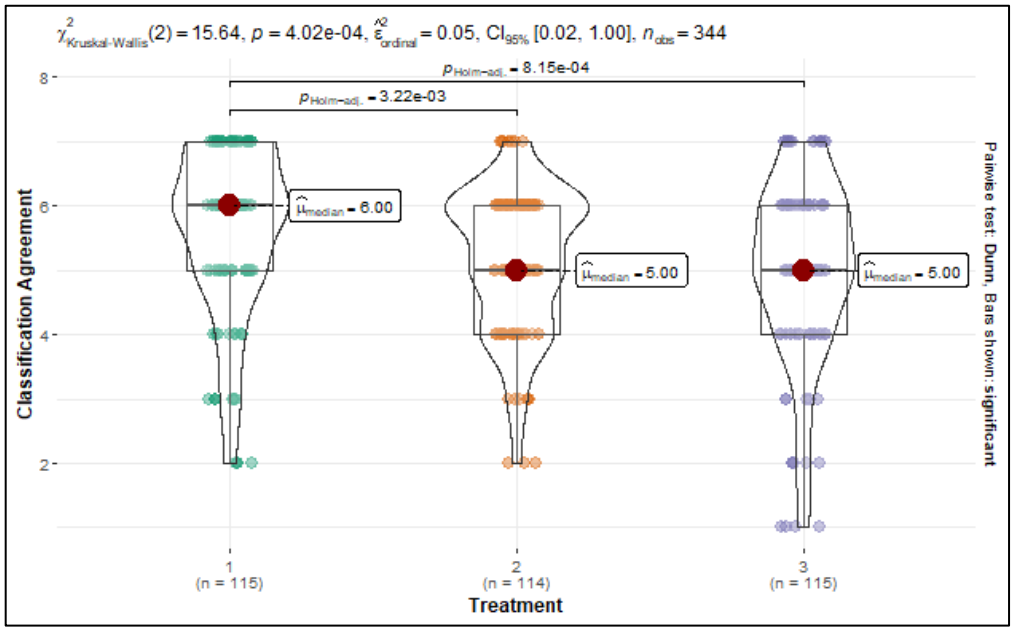

Figure 37. Kruskal-Wallis test of classification agreement.

Impact of demographic and personal characteristics. In previous analyses, we examined the impact of the different treatments, varying in their degrees of explainability, on the measured constructs. This analysis extends our understanding by employing linear regression to explore additional factors associated with these constructs. Table 9 presents the results of four separate linear regressions, each assessing the relationships between various predictors and our four distinct dependent variables: understandability, trust, usability, and classification agreement. The models include the predictors age, gender, academic background, prior experience with AI, individual propensity to trust, and treatment membership. The model for understandability (1) suggests that older individuals tend to perceive understandability as lower (β = -0.014, $p < 0.01$). Similarly, individuals with higher levels of general trust (propensity to trust) are more likely to perceive greater understandability ($\beta = 0.101$, $p < 0.01$). Other variables, including gender, academic background, and AI experience, are not significantly associated with understandability in this model. For trust (2), older individuals report lower levels of trust in the system (β = -0.016, $p < 0.01$), while individuals with a higher propensity to trust exhibit higher levels of reported trust ($\beta = 0.178$, $p < 0.001$). Gender, academic background, and prior AI experience are not significantly associated with trust in this model.

	Dependent Variable:			
	Understandability (1)	**Trust (2)**	**Usability (3)**	**Classification Agreement (4)**
Age	-0.014**	-0.016**	-0.015**	-0.018**
	(0.005)	(0.006)	(0.005)	(0.007)
Female	-0.022	-0.014	-0.339***	-0.303*
	(0.098)	(0.111)	(0.101)	(0.134)
Academic Status	-0.192	-0.022	-0.103	-0.231
	(0.103)	(0.117)	(0.106)	(0.142)
AI Experience	0.306	-0.088	0.151	-0.070
	(0.208)	(0.237)	(0.216)	(0.287)
Trust Propensity	0.101**	0.178***	0.084*	0.200***
	(0.036)	(0.041)	(0.038)	(0.050)
Treatment Two	0.089	-0.263	0.112	-0.345*
	(0.120)	(0.136)	(0.124)	(0.165)
Treatment Three	0.155	-0.214	0.087	-0.546***
	(0.119)	(0.136)	(0.123)	(0.164)
Constant	4.712***	4.707***	5.509***	5.604***
	(0.314)	(0.357)	(0.324)	(0.432)
Observations	341	341	341	341
R2	0.069	0.086	0.077	0.118
Adjusted R2	0.050	0.067	0.058	0.101
				Note: *p<0.05; **p<0.01; ***p<0.001

Table 9. Results of our linear regression.

For usability (3), age is negatively associated with perceived usability ($\beta = -0.015$, $p < 0.01$), suggesting that older individuals report a lower perception of usability than younger participants. Notably, gender also shows a significant association, with female participants reporting lower levels of usability compared to male participants ($\beta = 0.339$, $p < 0.001$). In contrast, an individual's propensity to trust has a positive association with usability ($\beta = 0.084$, $p < 0.05$), while academic background and prior experience with AI do not exhibit significant associations. Regarding the overall agreement with the displayed classifications (4), age is negatively associated with agreement with the system's classifications ($\beta = -0.018$, $p < 0.01$). Individuals with a higher propensity to trust exhibit higher agreement levels ($\beta = 0.200$, $p < 0.001$), while female participants report lower levels of agreement compared to male participants ($\beta = -0.303$, $p < 0.05$). Neither academic background nor prior experience with AI is significantly associated with classification agreement. Interestingly, the treatment conditions do not exhibit significant associations with perceived understandability, trust, or usability. However, both treatment two ($\beta = -0.345$, $p < 0.05$) and treatment three ($\beta = -0.546$, $p < 0.001$) are significantly associated with lower classification agreement compared to the control group. Participants

provided with explanations from the system exhibited lower agreement rates with its predictions, particularly in treatment three, where the explanatory text included a confidence score.

8.3.3 Discussion

Our study explored how different degrees of explainability, including explanations with or without confidence scores, impact user perceptions across multiple constructs, such as understandability, trust, usability, and classification agreement. The findings reveal that these treatments had a minimal effect on participants' evaluations, suggesting that additional underlying factors, such as demographic and individual characteristics, play a more significant role in shaping user experiences and perceptions (Schemmer, 2022).

The analysis indicated that the presence of explanations, whether with or without a confidence score, did not significantly affect understandability among the different treatment groups. Consequently, the results do not support the notion that XAI components improve understandability compared to a basic AI system without such components. The observed lack of improvement in understandability may be attributed to cognitive overload and issues with the relevance of the explanations provided (Liu et al., 2021; Sanneman & Shah, 2022; Tsai et al., 2021). Specifically, the data suggests that as age increases, perceived understandability tends to decrease, possibly because older participants may find complex or technical explanations more challenging. In contrast, higher levels of general trust are positively associated with greater perceived understandability, indicating that those who are more trustful are likely to find the system's explanations clearer. Additionally, factors such as gender, academic background, and AI experience did not show significant effects on understandability, suggesting that the effectiveness of explanations may be more closely related to cognitive factors and trust rather than demographic or experience-based differences. These findings reinforce the need for platform operators to carefully tailor explanation formats to user profiles to maintain accessibility and perceived value across diverse user segments (Gregor & Hevner, 2013; Rai, 2020).

Moreover, the analysis reveals that the presence of explanations without confidence scores was associated with a lower level of trust compared to the control group. Further, adding confidence scores to the explanations did not significantly enhance trust compared to the control group, indicating that confidence scores alone may not effectively enhance trust unless combined with other supportive elements (Hamm et al., 2023; Schmidt et al., 2020). This suggests that participants might have perceived the explanations as less straightforward or more confusing than simply receiving no explanations at all (Papenmeier et al., 2019; Poursabzi-Sangdeh et al., 2021). The presence of explanations,

whether with or without a confidence score, did not significantly affect perceived usability among the treatment groups. The negligible effect sizes and similar median usability scores across groups suggest that the treatments had no meaningful impact on participants' perception of usability.

Consequently, the results do not support the notion that XAI components improve usability compared to a basic AI system without such components. Despite being informed by qualitative user testing, the lack of significant impact on perceived usability from explanations, whether with or without a confidence score, may be attributed to several factors. First, the explanations, even when designed based on user feedback, may not have effectively addressed all aspects of usability or aligned with users' specific interaction needs, e.g., when explanations seem unintuitive (Mohseni et al., 2021; Schmidt et al., 2020). Second, it is possible that these explanations might not have sufficiently altered users' overall experience or efficiency with the system (Schemmer, 2022; Wanner et al., 2022). These results emphasize the broader challenge in integrating AI-driven features within platform interfaces without disrupting core user flows, a critical concern in the business design of digital platforms (Lyytinen et al., 2021). Currently, the existing XAI literature lacks a comprehensive set of methodologies and metrics for effectively assessing the quality of explanations (Sanneman & Shah, 2022).

The analysis of participants' classification agreement suggests that the presence of explanations was associated with lower classification agreement compared to the control group. This finding underscores the complex interplay between explainability and user agreement. The significant differences between treatment one and both treatments two and three indicate that the explanations provided in these treatments may have introduced additional uncertainty or complexity (Sanneman & Shah, 2022). Specifically, the inclusion of confidence scores in treatment three and the detailed textual explanations in both treatments may have made the system's decision-making process more transparent but also more challenging to interpret, particularly for users without prior familiarity with AI-based systems. One plausible explanation for this decrease in agreement is that overly detailed or technical explanations might have prompted users to scrutinize the system's classifications more critically, leading to increased doubt or skepticism (Ferguson et al., 2022). While this can be seen as a positive outcome in contexts where critical engagement with AI decisions is desirable, it may not align with the goal of fostering trust and usability in disinformation detection tools. Furthermore, explanations that incorporate probabilistic or confidence information can introduce cognitive overload for users who may lack the expertise to interpret such data effectively, exacerbating uncertainty. This observation aligns with prior research suggesting that user trust and agreement can be undermined when explanations are perceived as too complex or ambiguous (Miller, 2019). In

platform settings, this could translate into reduced conversion, churn, or lack of confidence in AI-generated outputs, particularly in high-stakes domains like e-commerce or content moderation (Benbya et al., 2020; Rai, 2020).

Our linear regression analysis elucidates several significant determinants impacting the constructs of understandability, trust, usability, and classification agreement, independent of treatment variations. The results consistently demonstrate that age has a negative relationship with each of the dependent variables, indicating that older individuals generally displayed higher aversion when interacting with our system. This trend may be attributed to age-related cognitive and perceptual changes, which could affect how older individuals process and evaluate information (Salthouse, 1992, 1994; Zahodne et al., 2011). Older adults might experience greater difficulty in understanding new concepts, trusting new technical systems, or experiencing high usability due to accumulated experience or changes in cognitive functions (Miller & Bell, 2012; Peters et al., 2008; Salthouse et al., 1999).

Conversely, an individual's propensity to trust exerts a positive influence across all constructs, underscoring the role of individual trustfulness not only in enhancing trust in the system but also in perceived understandability, usability, and agreement with the classifications provided by the (X)AI. This finding highlights the importance of inherent trust levels in shaping perceptions. People who naturally exhibit higher trust are likely to approach information and systems with a more positive outlook, which could enhance their overall experience and evaluation (Fan et al., 2020).

Notably, gender differences are evident in our findings as being female is associated with a lower perceived usability and lower classification agreement. This indicates that, with regard to some elements, female participants potentially perceive the system less favorably compared to their male counterparts. The observed discrepancy may stem from varying expectations, experiences, or societal factors that affect how different genders interact with and evaluate systems (Reeder et al., 2023). Further research is needed to explore the underlying causes of these gender-related differences, including potential biases in system design or differences in interaction styles. In contrast, whether someone has an academic degree or prior experience interacting with AI has no significant influence on any of the constructs. This may imply that educational background and prior exposure to AI-based systems are less influential in shaping user experience than other individual characteristics, such as age and trust propensity. From a platform design perspective, these insights suggest that adaptive personalization, based on traits like age and trust propensity, may help mitigate usability barriers and enhance engagement across heterogeneous user bases (Berente et al., 2021; Lyytinen et al., 2021). In summary, this analysis extends our understanding of how individual differences shape user perceptions, highlighting the significance of age and trust propensity while indicating the need for further exploration

into gender-related differences. This broader perspective complements our findings related to treatment variations, offering a more comprehensive view of the factors influencing user evaluations while hinting at the need for the design of adaptive systems (Kabudi et al., 2021).

8.4 Integrated Design Guidelines (E)

Finalizing our second DSR cycle by building on our previous design guidelines, we further refine and expand our approach to developing responsible XAI systems for disinformation detection. Considering our empirical findings, the integrated guidelines (Gurzick & Lutters, 2009) highlight user needs and the importance of maintaining a balance between simplicity, clarity, and adaptation while also addressing demographic and individual differences:

1. *Integrate explanations seamlessly into the user experience.* Ensure that explanations are integrated in a way that enhances, rather than disrupts, the overall usability of the system. Since the addition of confidence scores did not significantly improve trust or usability, focus on how explanations are presented and ensure they contribute positively to the user experience without causing confusion.
2. *Simplify explanations to avoid cognitive overload.* Ensure that explanations provided by the XAI system are clear and not overly complex. Given that explanations did not significantly impact understandability, it is crucial to avoid introducing unnecessary complexity. Tailor explanations to be straightforward and relevant to the user's current context to prevent cognitive overload.
3. *Prioritize trustworthiness in design to build credibility for inexperienced users.* Even though confidence scores alone did not significantly enhance trust, ensure that explanations are part of a broader strategy to build system credibility. Develop supportive elements that reinforce trust and reliability, ensuring users perceive the system as trustworthy and effective in detecting disinformation.
4. *Make explanations optional by offering customizable explanation features.* In line with the principle of user empowerment, explanations should be an optional feature, allowing users to access additional details only when needed. This approach respects the user's autonomy and avoids unnecessary complexity in the overall user experience.
5. *Consider user trust and cognitive factors.* Recognize that inherent trustfulness and cognitive factors may significantly influence how users perceive explanations. Account for cognitive differences, such as those related to age, by simplifying explanations for older users who may struggle with more technical content.

6. *Address demographic and individual differences through adaptability in design.* Design explanations adaptable to different user profiles, acknowledging that factors such as age and trust propensity affect user perceptions, and be mindful of potential biases in system design as well as differences in how various demographic groups interact with the system. Consider conducting targeted user research to tailor explanations effectively.
7. *Refine and test explanation mechanisms continuously.* Continuously refine explanation mechanisms based on user feedback and iterative testing. The findings suggest that explanations alone might not improve usability or classification agreement. Regularly test and adjust explanations to better align with user needs and enhance the system's effectiveness.

By adhering to these guidelines, responsible XAI systems for disinformation detection may be developed to better meet user needs, enhance usability, and improve overall effectiveness in combating false information on digital platforms.

8.5 Conclusion

8.5.1 Summary

This study addressed the research question of how a responsible XAI-based system for detecting online disinformation should be designed to foster user trust, understandability, and comprehension. By leveraging a Design Science Research (DSR) approach (Peffers et al., 2007), we developed and evaluated explainability features tailored to the high-stakes, sensitive domain of disinformation detection. Through a comprehensive literature review, iterative design cycles, and empirical user testing, we provide both practical design guidelines and important theoretical insights into the limitations and potential of explainable AI (XAI) in real-world applications.

From a theoretical standpoint, this study contributes to an underexplored intersection between XAI and disinformation detection by shifting the focus from purely technical accuracy toward user-centric design principles (Rjoob et al., 2021; Wells & Bednarz, 2021). While transparency is widely recognized as a cornerstone of XAI (Haque et al., 2023), our findings challenge the assumption that greater transparency inherently leads to improved user trust, comprehension, or usability (Schmidt et al., 2020). Contrary to common expectations, the inclusion of XAI components did not significantly enhance participants' understanding or trust in the system, and in some cases even introduced confusion or reduced agreement with system outputs. These results emphasize the importance of de-

signing explanations that are not only technically accurate but also cognitively appropriate for the target user group. By demonstrating that explanations can inadvertently increase cognitive load, our study refines existing cognitive load theory and highlights the contextual and individual variability in how users perceive and benefit from XAI. We show that user demographics—particularly age—and individual characteristics like trust propensity significantly influence the effectiveness of explainability features. Older users, for example, reported lower levels of trust, usability, and understanding, suggesting a need for adaptive XAI systems that account for users' cognitive and experiential diversity. Additionally, our application of the DSR methodology underscores the value of integrating theoretical and empirical insights into the iterative development of XAI systems. This study contributes to IS and HCI literature by offering a framework for embedding user feedback early and systematically in the design process, revealing the nuanced trade-offs between transparency, usability, and user trust. Practically, our findings translate into actionable design guidelines for developing responsible, user-aware XAI systems in the disinformation space. These include simplifying explanations to minimize cognitive overload, tailoring them to users' demographic and cognitive profiles, and offering explanations as optional features to preserve user autonomy. Furthermore, we advocate for combining XAI with other trust-enhancing mechanisms, such as user feedback loops, to foster engagement and reliability.

In conclusion, this research advances both theoretical understanding and practical implementation of explainable AI by uncovering the complex interplay between user characteristics, contextual factors, and design choices in disinformation detection systems. While explainability does not universally improve user perceptions, our contributions provide a foundation for future studies to build more adaptive, context-sensitive, and trustworthy XAI systems, crucial for navigating the evolving challenges of disinformation and responsible AI governance in the digital age.

8.5.2 Limitations

While this study offers valuable insights into the responsible design of XAI systems for disinformation detection, some limitations must be acknowledged to fully contextualize the findings and guide future research. The structured literature review, though comprehensive, is inherently limited by the selection criteria and databases. The focus on specific keywords or publication types may have excluded relevant studies that could provide additional insights or counterpoints. The qualitative user testing's sample allowed for an in-depth exploration of participants' experiences and perspectives; nevertheless, it may not fully represent the diversity of views within the population. We therefore conducted a quantitative study to test the results with a broader range of backgrounds and present more

generalizable results. The online study's design was cross-sectional, capturing user perceptions at a single point in time. Longitudinal studies would be beneficial to assess the long-term impact of explainability features on user perceptions. The study observed a reduction in agreement with the system's classifications when explanations were provided. Investigating the content and format of the explanations could reveal whether they contribute to misunderstandings or if alternative presentation methods might improve agreement.

Furthermore, focusing on the design of the explanations, rather than also considering their content and providing a broader array of examples with varying textual features, may not fully capture the range of disinformation features users might encounter. Future studies could expand on this by offering participants more diverse examples, which could help identify how different types of explanations interact with varying content and how they affect user perceptions. Finally, our study focuses on the perception of explainability features. Other aspects of algorithmic transparency (such as model accuracy) are also crucial for how users perceive the system and should be considered in future research to develop a more comprehensive approach to responsible AI design. By acknowledging these limitations, future research can deepen our understanding of how to effectively design and implement XAI systems for disinformation detection and other high-stakes applications. Such research can ultimately support platform providers of OSNs in responsibly adopting and integrating AI-based systems for disinformation detection, fostering a more trustworthy and accountable digital platform ecosystem.

8.5.3 Future Work

The ethical deployment of AI in cyberspace governance, especially for disinformation detection, requires a thorough examination to safeguard transparency and fairness on digital platforms. Future research may explore several avenues to build on our findings. First, further studies may investigate a broader range of explanation types and their interactions with various user demographics to identify which formats are most effective in different contexts. Second, longitudinal studies could provide insights into how users' perceptions of AI systems develop over time and whether continuous exposure to explanations affects their experience. Third, investigating the integration of explanations with other trust-enhancing features, such as transparency mechanisms and user feedback systems, could offer a holistic approach to improving user interactions with AI in the combat of online disinformation. In conclusion, while explainability is a critical component of responsible AI, its effectiveness in promoting usability, user trust, and comprehension requires careful consideration and tailored implementation. Our study underscores the importance of a nuanced approach to integrating explainability features and highlights the need for ongoing research to refine these mechanisms and better align them with user needs. Building

on our findings, future work can contribute to the development of more effective and trustworthy AI-based systems for disinformation detection and beyond.

Part IV

Tackling Information Manipulation: Skills and Requirements

9 Designing Deepfake Detection Systems: Practitioner Requirements Across Sectors[14]

9.1 Introduction

In recent years, deepfakes — synthetic media generated through artificial intelligence (AI) (Masood et al., 2023) — have received considerable attention across public discourse, academia, and policy arenas, as they exemplify a transformative shift in the creation and perception of digital content (Almars, 2021; Fabuyi et al., 2024; Wang et al., 2020). Their potential to disrupt information ecosystems, fuel disinformation, and erode institutional trust has sparked widespread concern (Fernández Gambín et al., 2024). Yet, while the narrative surrounding deepfakes has been shaped by strong assumptions about their societal threat (Abdullah et al., 2024; Albahar & Almalki, 2019; Westerlund, 2019), we still know surprisingly little about how professionals who encounter these phenomena in practice, such as journalists, security agencies, non-governmental organizations (NGOs), and industry actors, actually assess their relevance and impact in their organizational contexts (Durães et al., 2023; Godulla et al., 2021). Despite advances in the technical sophistication of deepfake detection techniques, much of the existing research remains technology-centric and thereby prioritizes algorithmic performance over contextual relevance and user-centered design. Moreover, few studies propose concrete tools that integrate multiple methodological approaches in ways that align with real-world professional workflows (Ben Aissa et al., 2024; Sharma et al., 2024). However, detection techniques can only fulfill their potential if they are designed in alignment with the expectations, work practices, and trust conditions of those who are meant to use them (Schlichtkrull et al., 2023; Warren et al., 2025). From a human-AI interaction perspective, this raises critical questions: How do practitioners currently view the impact of deepfakes in their respective fields? What criteria must detection systems fulfill to gain trust, be clearly understood, and support effective decision-making? This study takes an Action

[14] This chapter comprises a paper conditionally accepted at ICIS 2025 by Isabel Bezzaoui, Louis Jarvers, Jonas Fegert and Christof Weinhardt with the following title: Designing Deepfake Detection Systems: Practitioner Requirements Across Sectors, 2025. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated

Design Research (ADR) perspective to investigate how domain experts assess the practical implications of deepfakes and articulate requirements and design principles for supportive detection tools. By conducting structured expert interviews across diverse sectors, we derive design-relevant insights and translate them into a requirement analysis that informs the development of multimodal detection systems. Our goal is to contribute to the design knowledge for artifacts that can meaningfully support practitioners in evaluating the authenticity of digital content. We address the following research questions:

RQ1: How do practitioners perceive the current relevance of deepfakes, and what role do they see for automated detection systems in their organizational contexts?

RQ2: How should multimodal deepfake detection tools be designed to meet practitioners' needs and support trust in digital content?

Our study reveals sector-specific variations in how practitioners assess the relevance of deepfakes. For instance, law enforcement experts emphasize the growing significance of deepfakes for jurisdictional authority, while representatives from the financial sector acknowledge the increasing awareness but note the limited direct impact so far. Furthermore, practitioners across all sectors express skepticism toward detection systems that function as "black boxes", providing binary results without offering transparency into how those conclusions are reached. By synthesizing the interview data and analyzing the requirements of practitioners in diverse contexts, we offer a cross-sectoral perspective that bridges technical capabilities with user-centric design considerations. This study contributes to the human-AI interaction and algorithmic experience literature by moving beyond abstract threat narratives toward a grounded understanding of trust, organizational expectations, and socio-technical design in the context of AI-assisted deepfake detection. We contribute to IS research in three ways: we provide empirical insights into how practitioners across sectors perceive and respond to deepfakes (1); we identify design-relevant requirements and design principles that support the development of transparent, usable, and context-sensitive detection tools (2); and we extend the application of ADR to the domain of AI-generated media, showing how practice-informed insights can guide the design of socio-technical systems in emerging problem spaces (3).

The remainder of this paper is structured as follows: Section 2 reviews prior work on deepfake detection and human-AI interaction. Section 3 outlines our methodological approach. Section 4 presents our empirical findings. Section 5 discusses implications for system design, and Section 6 concludes with contributions and directions for future research.

9.2 Research Background

9.2.1 Deepfakes: Definitions and Societal Relevance

Deepfakes are defined as synthetically manipulated media generated through AI, employing methods like generative adversarial networks (GANs) to create hyper-realistic audio and video that misrepresents real individuals and events (Odeh, 2024; Vaccari & Chadwick, 2020). Their rise has garnered significant attention across public discourse and academia, primarily due to their potential to disrupt information ecosystems, fuel disinformation, and erode institutional trust (Noreen et al., 2022). Reports have highlighted that deepfakes can propagate mistrust among consumers, as they often blur the lines between authentic and manipulated content, leading to a pervasive skepticism toward digital media (Twomey et al., 2023). As news media becomes increasingly rich with deceptive content, the potential for deepfakes to undermine journalists' credibility is particularly alarming, calling our dependence on visual media as an indicator of authenticity into question (Doss et al., 2023; Sandoval et al., 2024). Despite growing concerns, there remains a gap in understanding how various stakeholders, such as journalists and investigators, evaluate the relevance and impact of deepfakes in their professional contexts (Qureshi & Khan, 2024). Existing literature emphasizes the need for a more nuanced approach that considers not only the technical capabilities of detection tools but also their alignment with user needs and expectations within specific organizational environments (Qureshi & Khan, 2024; Vaccari & Chadwick, 2020). Thus, exploring the perceptions of professionals encountering deepfakes in practice is crucial for a holistic understanding of this phenomenon, enabling more profound insights into the sociocultural implications they may carry.

9.2.2 Deepfake Detection Methods: Trends and Multimodal Approaches

The rapid advancement in deepfake technologies has prompted concurrent developments in detection methodologies, which predominantly utilize machine-learning techniques (Kaur et al., 2024). Initial detection efforts have often relied on traditional approaches focusing on a single modality, such as independently analyzing video or audio inputs (Heidari et al., 2024; Rowan & Pears, 2022; Zhang et al., 2021). However, a growing consensus within research advocates for multimodal detection strategies, integrating various data types (audio, video, image, and text) to enhance robustness against sophisticated deepfake manipulations (Cai et al., 2023; Chen & Tan, 2021; Park et al., 2024; Rana et al., 2022). Recent advancements in deepfake detection emphasize the effectiveness of spatiotemporal models, such as Convolutional Neural Networks (CNNs) and long short-term memory (LSTM) networks, which utilize both spatial and temporal cues from video

data (Almars, 2021; Shelke et al., 2023; Vaishnavi et al., 2023). These multimodal architectures demonstrate higher accuracy compared to unimodal approaches, as they can better capture the complex inconsistencies within deepfake content (Rowan & Pears, 2022; Vaishnavi et al., 2023). Furthermore, developments in proactive defense mechanisms aimed at disrupting the generation of deepfakes are also being explored as complementary measures to support detection technologies (Juefei-Xu et al., 2021; Park et al., 2024). As detection methodologies evolve, there is an increasing recognition of the necessity to embed user perspectives and requirements into the development of these systems. Effective interaction between humans and detection tools hinges upon transparent functionalities that resonate with users' contextual needs (Alanazi & Asif, 2024; Chen & Tan, 2021; Groh et al., 2022; Lyu, 2024).

9.2.3 Professional Practice and the Design of Detection Tools

Kumar et al. (2024) focused on integrating human judgment with computational techniques to detect deepfake images, using an intelligence augmentation approach that considers the beliefs and intentions of the observer. They identified exogenous cues that may help humans detect deepfakes and proposed a foundation for combining human and computational methods in future direction efforts. Akinyemi et al. (2024) explored the influence of AI-generated content labels on users' perceptions and sharing behavior. Their experimental study assessed whether disclosure labels could reduce the spread of deepfakes by altering users' inherent trust in the content. This highlights the potential of labeling as an intervention in combating disinformation. While these studies contribute valuable insights, many focus on technical detection methods. Vasist and Krishnan (2022) observed that much of the literature centers on computational challenges in deepfake detection, overlooking the social, ethical, and psychological implications. Kaur et al. (2024) identified key challenges in detection, but their focus on the technical requirements of building detection models leaves out the practical needs of end-users. Similarly, Trinh et al. (2021) developed an interpretable framework to improve the trustworthiness of deepfake detection, yet their focus remains on theoretical aspects rather than practical, real-world applications. Moreover, research on professionals' needs highlights that while experts from journalism recognize the threat of deepfakes, the actual impact on their practices remains limited. Sohrawadi et al. (2020) explored the requirements of journalists and found that effective detection tools should be user-friendly, integrate seamlessly into editorial workflows, and provide clear, actionable explanations. They identified a gap between existing detection technologies and the needs of media professionals for real-time, intuitive verification tools. Similarly, Weikmann and Lecheler (2024) examined the

implications of deepfakes for fact-checkers, showing that while deepfakes are acknowledged as a potential threat, their direct impact on journalistic efforts has been limited thus far.

Despite these contributions, a key gap persists in understanding how professionals across different sectors assess the practical relevance of deepfakes and what requirements they have for detection tools (Abbas & Taeihagh, 2024). Most existing studies focus on technical solutions without addressing the contextual factors that influence the adoption and effectiveness of these tools (Fernández Gambín et al., 2024). Our study aims to fill this gap by exploring the perceptions and expectations of professionals from diverse domains, offering insights for developing user-centric, context-sensitive detection systems that align with real-world needs.

9.3 Methodology

To derive actionable design knowledge for the development of deepfake detection tools, this study adopts a qualitative, explorative research design situated within the Action Design Research (ADR) paradigm. ADR emphasizes a design research approach that both supports the development of innovative IT artifacts within real organizational contexts and enables learning from the intervention to address practical, real-world problems (Mullarkey & Hevner, 2019; Sein et al., 2011). In line with this orientation, our study seeks to understand how domain professionals perceive the relevance of deepfakes and what they expect from detection systems designed to support decision-making within their organizational contexts.

The motivation for this approach is grounded in the need to better understand sociotechnical dynamics that shape the use, interpretation, and institutional integration of AI-based media verification tools. As emphasized in qualitative IS research, one of the key benefits of such an approach is its ability to capture the cultural and social context in which decisions are made (Benbasat et al., 1987). In domains where stakeholders interact with or are affected by deepfakes, such as journalism, civil society, or national security, these contextual factors play a significant role in shaping decision-making (Myers, 2019). Given the complexity and contextual embeddedness of deepfake use and detection, qualitative interviews offer a robust method for uncovering domain-specific requirements, underlying assumptions, and unarticulated needs (Niebert & Gropengiesser, 2013). Prior design research case studies (Vom Brocke et al., 2020) have underscored the relevance of qualitative interviews for eliciting design-relevant insights, particularly when examining emergent phenomena or system requirements. To guide the structure and analysis of our interviews, we draw on Kaiser's (2014) methodological framework for expert interviews in political science, which is well-suited for exploring institutional perspectives,

power dynamics, and practical constraints — all of which are critical for designing human-centered AI systems in high-stakes domains. To investigate these dynamics, we employed semi-structured expert interviews, which allow for deep insights into personal experiences, contextual constraints, and tacit knowledge that may otherwise remain inaccessible (Kaiser, 2014; Myers, 2019). The interview guideline provided a consistent framework of core questions while allowing the interviewer to explore relevant topics in more depth as they arose, ensuring both comparability and richness of data.

9.3.1 Data Collection

The interviews were conducted with 15 practitioners from journalism, civil society organizations, security services, and industry. These individuals were selected based on a prior stakeholder mapping process and their professional exposure to disinformation and media verification. Practitioners were chosen through purposive sampling to ensure representation across key domains while minimizing functional overlap and ensuring comprehensive coverage of deepfake detection use cases across professional contexts. Table 10 outlines the roles and domains of the practitioners interviewed for this study, categorizing their prepositions according to specific sectors such as media, security, and civil society. Potential interviewees were contacted via email and received an overview of the research project and its goals in advance, ensuring transparency and that the interviewee felt confident enough to answer the interview questions. All 15 interviewees were informed that they would be kept updated about future development steps and expressed willingness to be contacted again for follow-up activities such as iterative requirements elicitation or validation through survey or additional interviews. The interviews were conducted in German between February and April 2025 via the GDPR-compliant DFNconf video conferencing tool. Before each interview, participants gave their informed consent (Payne & Payne, 2004). At the outset of the interview, the researchers introduced themselves and the project to clarify the scope and purpose of the conversation. Interviews concluded with an open-ended opportunity for additional comments and follow-up questions. Each session lasted between 27 and 50 minutes, with an average duration of 39 minutes. All interviews were audio-recorded, then transcribed verbatim to enable systematic analysis. To ensure consistency and comprehensibility, the interview guideline was pre-tested and revised for clarity and coherence. The interviews followed a flexible protocol with open-ended questions centered around four key themes: perceptions of deepfake relevance, experiences with media manipulation, evaluation criteria for detection tools, and practical constraints within stakeholders' domains.

Domain	Role	ID
Administration	Press and communication officer	P01
	Crisis communication expert	P05
Media and Journalism	Fact-checking team lead	P02
	Investigative reporter	P09
	Fact-checking analyst	P12
Industry	AI and security consultant	P03
	Finance risk analyst	P04
	Innovation strategist	P08
Security and Defense	Digital forensics specialist	P06
	Cyber operations officer	P07
	Investigator	P10
	Intelligence analyst	P11
Civil Society and NGOs	Media literacy trainer	P13
	Senior policy expert	P14
	Editorial director	P15

Table 10. Roles and domains of interviewees.

9.3.2 Data Analysis

The transcribed data were analyzed using MAXQDA, a state-of-the-art tool in qualitative research (Kuckartz & Rädiker, 2019), following a multi-stage thematic content analysis that combined deductive and inductive coding strategies (Kaiser, 2014; Mayring, 2015). A deductive coding frame was initially constructed based on the research questions and interview themes. This was then expanded through inductive coding, allowing new patterns and concerns to emerge directly from the data. All three researchers collaboratively reviewed and iteratively refined the resulting codebook to ensure conceptual clarity and alignment with the research goals. Codes were grouped into subdimensions relevant to the analysis, including: perceived threats and opportunities, criteria for tool trustworthiness, organizational integration, and contextual constraints. This framework enabled a nuanced interpretation of cross-cutting expectations and domain-specific variations in how deepfake detection tools are understood and assessed. Data saturation was considered achieved when no new inductive codes or themes emerged from additional interviews, ensuring a comprehensive representation of participant perspectives.

9.3.3 Deriving Design Knowledge

Following data analysis, we conducted a requirement analysis by systematically translating coded segments into design-relevant requirements. This process was informed by the principles of design research, which emphasize iterative reflection and abstraction from empirical data (Vom Brocke et al., 2020). Identified requirements were then synthesized and clustered into broader categories based on thematic overlap and system design relevance. To derive design principles (DPs) from the collected data (Schacht et al., 2015), we followed a layered abstraction logic commonly employed in design research: first, we identified recurring meta-requirements (MRs) across stakeholder groups, which reflect generalized user needs grounded in context (Walls et al., 1992). These MRs were then synthesized into overarching design principles that provide prescriptive guidance for developing human-centered deepfake detection tools. Each DP is thus empirically grounded in stakeholder expectations while abstract enough to inform design choices across systems and domains.

9.4 Results

This section presents the findings of a qualitative content analysis conducted to explore the perspectives of practitioners regarding the perceived relevance of deepfakes and their automated detection, as well as the design and functional expectations of deepfake detection tools. Based on 15 semi-structured interviews with professionals across domains such as journalism, law enforcement, public administration, civil society, and cybersecurity, the analysis engages with the methodological framework of Mayring (2015) to extract and interpret central themes. In what follows, our two research questions will be answered consecutively.

9.4.1 Deepfake Relevance and the Case of Automated Detection

9.4.1.1 Increasing Relevance and Technological Advancement

The interview findings demonstrate that practitioners universally perceive deepfakes as a rapidly evolving technological challenge. A technology executive from an innovation lab described the remarkable advancement trajectory: “super dynamic technology… within the last 24 months from a comic-like image that was immediately recognizable as fake, to today’s synthetic media content… where you really have to look closely to determine if it’s real or not.” (P08). A security expert from a federal agency articulated a critical inflection point: “We’re currently at our borderline, I believe. I’m convinced that anyone interested can still recognize deepfakes today.” (P05). However, this same expert projected a narrow timeframe before human detection capabilities would be overwhelmed:

"But these technical inadequacies that help us recognize deepfakes today with human experience – these inadequacies will soon no longer exist." A police official offered a similar assessment, noting that while current relevance in policing remains limited, "I estimate that it will continue to grow. Looking at the AI trend, where it's developing, suddenly there are thousands of different providers making AI voice changers, thousands of people making videos" (P06). Still, not all practitioners perceived deepfakes as an immediate or dominant threat. A specialist from a press agency noted: "The proportion of deepfakes in disinformation is very low. Until now… That doesn't mean it's not there. There are certainly some, but there are many more cheap fakes" (P02). This perspective highlights that while deepfakes are technically advanced, simpler forms of media manipulation, such as image modification or text-based disinformation, currently remain more prevalent.

While the sentiment toward deepfakes was overwhelmingly negative, frequently framed as a threat to trust, authenticity, and verification, one interviewee also pointed to the technology's constructive potential: "But it is also a technology that can be used for good", they noted. "Think of movie productions. You could also use deepfakes in the news with presenters if someone is absent." (P15). This outlier perspective illustrates that, although rare among respondents, there is an awareness of possible beneficial applications, particularly in controlled or creative environments.

Rather than viewing deepfakes as merely manipulated videos, interviewees consistently conceptualized the threat as spanning multiple content types. A technology executive explicitly framed this perspective: "We see deepfakes as a multimodal use case. Text is simple, but video, audio, images, all of these uses can involve deepfakes, either completely synthetic media generation or augmentation and modification of existing data." (P08). An editor from a major media outlet highlighted verification challenges: "How can we even verify if this audio recording is a real audio recording or is it AI-generated?" (P09). This multimodal concern was echoed by a fact-checker who noted that "AI-generated content plays a quite significant role for us, because an additional layer has been added to our verification work." (P12).

9.4.1.2 Sector-Specific Relevance Assessments

The interview data reveal variation in how practitioners from different sectors assess deepfake relevance to their specific operational contexts. For example, a law enforcement expert emphasized the growing importance of deepfakes in the context of jurisdictional authority and geolocation tasks: "Very high and already very high and increasingly important" (P11). Conversely, a representative from a financial institution acknowledged the growing awareness but noted the limited direct impact thus far: "We have not been affected so far. It was dormant for a while. But now there are the first cases in the Asian

region where deepfakes are being used in connection with CEO fraud." (P04). The military and defense sector also recognized the potential risks posed by deepfakes, with one expert noting that they are "known to everyone dealing with this subject area," including the recognition that deepfakes can be deployed "against allies or own forces" for "destabilization of democracy" (P07).

On the other hand, a government communications professional observed that deepfakes remain more of a novelty at present: "the deepfakes we deal with are mostly funny things," suggesting a spectrum of applications from benign to malicious (P01). The issue of deepfakes eroding trust in information was also highlighted by experts in media literacy and policy. A media literacy expert described deepfakes as "democracy-eroding," specifically when they shape political opinion formation: "To what extent does it go in a direction that politicizes or emotionalizes me so strongly that I then form a false or manipulated political opinion, which I actually wouldn't get if I were to form an opinion based on true information." (P13). A senior policy expert in platform regulation and online hate pointed out that deepfakes "could undermine trust in credible information when people no longer know what's real and what's not." (P14). This concern was particularly acute regarding younger individuals, representing "another factor that significantly complicates this political opinion formation, especially among adolescents" (P13). Nevertheless, a fact-checker noted public anxiety about deepfakes: "We are quite in touch with the audience, and we notice a huge uncertainty regarding this topic. Because especially this 'I can no longer trust my own eyes' really concerns people" (P12).

9.4.1.3 Necessity for Automated Detection Systems

Across sectors, practitioners underscored the declining viability of human-only detection methods. A security specialist provided a clear timeline, stating, "We're talking about a year or so. Then the technology will be so sophisticated that it will simply become more difficult or no longer possible to recognize deepfakes simply through human experience. And that's when technology comes into play." (P05). This view was further reinforced by a policy expert, who acknowledged that "the creation of deepfakes has now advanced very far. And the tips that were provided earlier might not necessarily be sufficient anymore to be able to manually recognize whether it's a deepfake or not" (P14).

The perceived need for automated detection is particularly pronounced due to the time-sensitive nature of the damage caused by deepfakes. The innovation lab executive emphasized this, noting that: "audio messages can now be falsified so easily, high quality, and for bad actors, the advantage is that they scatter something into the public and until it's identified as not real, the damage has actually already happened" (P08), highlighting detection speed as crucial. Practitioners framed detection not merely as a technological

solution but as an organizational capability (P08), describing a "horizon scanning" approach to anticipate future developments and prepare appropriate responses before direct impacts manifest (P04). A corporate security consultant observed that systems for deepfake detection have relevance in private sector companies recruiting employees, citing a case where "someone applied and conducted an application and employee interview, and the person who appeared wasn't actually the person in the background" (P03).

Figure 38 provides a conceptual overview of the key themes identified in the practitioner interviews. It illustrates how technological advancements, sector-specific use cases, and rising public concern converge to intensify the perceived urgency of the deepfake threat. These drivers, such as the declining effectiveness of human detection and the multimodal nature of synthetic content, underscore the need for automated detection systems. Notably, the figure also emphasizes that the ultimate response extends beyond technical solutions, highlighting the significance of organizational readiness, anticipatory monitoring, and public communication strategies.

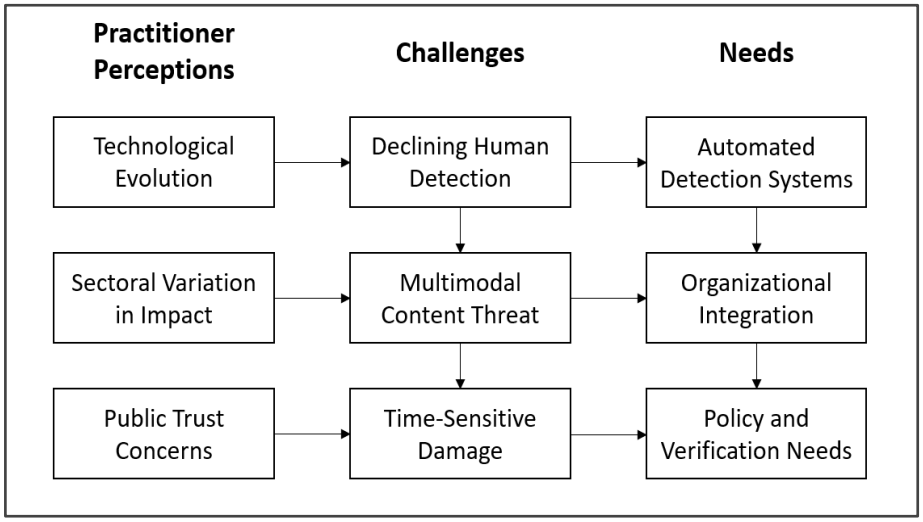

Figure 38. Practitioner perceptions of deepfake relevance and detection needs.

So far, these discussions reveal that the relevance of deepfake detection extends beyond traditional security contexts, influencing areas such as recruitment, media verification, and public trust. As practitioners point out, the effectiveness of detection systems in these diverse areas will play a pivotal role in managing the broader societal and organizational impacts of deepfakes. In light of these insights, it is clear that the challenge of deepfakes is widely recognized across sectors, with practitioners acknowledging their increasing relevance despite current variations in their impact. While sophisticated deepfakes remain relatively rare, many foresee a critical inflection point when human detection will no longer be sufficient. The growing necessity for automated detection systems stems from

a combination of declining human capabilities, the urgency of responding to time-sensitive damage, and the need for integrated organizational responses. These findings suggest that, although deepfakes have not yet reached the prevalence some predicted, there is a broad consensus on the need for more robust, integrated detection systems to address both technological advancements and organizational needs.

9.4.2 Practitioner Requirements for Deepfake Detection Tools

9.4.2.1 Usability and Accessibility: Designing for Everyday Expertise

Across all interviews, there was a striking convergence on the expectation that deepfake detection tools must be easy to use, fast, and low-threshold. The demand for a high degree of usability was articulated most emphatically by actors from journalism, civil society, and public education — groups who routinely engage with manipulated media but often lack access to technical expertise. However, all participants expressed clear preferences for “intuitive tools” that “don’t require prior training,” as an interviewee from a cybersecurity company put it (P03). The ideal tool was described as "as straightforward as possible" and “operable in everyday routines without much explanation” (P03). Several respondents stressed that detection tools must not appear overly technical or abstract, as this would deter use by less technologically literate practitioners. A law enforcement expert remarked: “If I don’t understand what the tool is doing, I won’t trust it. And I won’t use it.” (P11). Here, usability and trust appear closely linked: a low-threshold interface is not merely a matter of convenience, but a precondition for credibility. In addition, practitioners emphasized that tools should not require mandatory registration or third-party resources but rather operate on local hardware. A fact-checking expert noted: “I don’t want to have to register anywhere or upload sensitive material to some unknown server. I just want to check quickly whether something is suspicious.” (P02). This reflects a strong sensitivity towards data protection, anonymity, and fast integration into existing work routines. The findings suggest a central design imperative: deepfake detection tools must be tailored to non-specialists, with minimalistic and context-sensitive interfaces, allowing for quick decisions under pressure. This category can be summarized as a clear call for accessible, low-complexity interaction design.

9.4.2.2 Transparency and Explainability: Legibility Over Black Box Certainty

A second core theme concerned the transparency and explainability of the detection process. Practitioners from all sectors expressed skepticism toward tools that function as “black boxes,” delivering binary results (e.g., real/fake) without offering insight into the reasons behind them. This was seen as problematic not only for individual trust but for institutional accountability. A crisis communication expert from a federal agency emphasized: “If I’m supposed to present this in court, I need to be able to explain where the

result came from. Otherwise, it's useless." (P05). Even outside judicial contexts, respondents insisted on some form of human-understandable feedback. An innovation lab expert from industry noted: "If the tool just says 'fake' — that doesn't help me. I want to know what signals it found: was it the voice? The blinking? Something in the metadata?" (P08). There was a clear preference for tools that not only provide a probability score but also contextual explanations — visual overlays, annotated features, or textual descriptions of why a piece of content might be suspect. Several interviewees also called for "confidence levels" or leveled indicators rather than absolute values. This reflects a sophisticated understanding among practitioners that detection technologies are probabilistic and context-dependent. A media analyst in fact-checking warned against binary interpretations: "It must be clear that this is not a 100% judgment. The user should understand that it's a likelihood, not a final decision." (P12). These statements underscore the importance of transparency not just as an ethical principle, but as a functional requirement for professional use. Tools must help users understand, evaluate, and reflect on the outputs rather than simply act upon them. Interpretation and contextualization of the results remain a crucial feature throughout the entire analysis process.

9.4.2.3 Multimodality and Contextual Analysis: Complexity of Input, Coherence of Output

A particularly notable insight from the analysis was that practitioners already expect detection tools to incorporate multiple modalities. Participants repeatedly emphasized that today's deepfakes often span visual, auditory, and textual domains — and thus require detection mechanisms that do the same. As a military cybersecurity consultant stated: "The good fakes are always multimodal. The ones that get shared a lot — they combine voice, video, and subtitles. If you only analyze the video, you miss the bigger picture." (P07). This observation was echoed across contexts. Civil society actors mentioned memes with fabricated subtitles; journalists referenced TikTok videos with manipulated voiceovers; law enforcement pointed to forensic cases where metadata, timestamps, and inconsistencies in speech were critical. Thus, the capacity to synthesize and compare across modalities was seen not as a luxury but as a baseline requirement. Moreover, participants expressed the need for contextual analysis that goes beyond technical features. A federal law enforcement specialist put it: "Sometimes the content doesn't look fake at all — but something about the context is off. Like, the person says something they'd never say. The tool should help me notice that." (P10). This expectation introduces a new layer of complexity: practitioners are not only looking for pixel-level or signal-based anomalies but are also attentive to semantic inconsistencies and behavioral implausibilities. Taken together, the findings indicate that users want multimodal, context-aware tools capable of evaluating the alignment between different information layers. Such tools

would not merely process inputs in isolation, but reflect the complexity of how media is produced, consumed, and trusted in real-world environments.

9.4.2.4 Integration and Operational Fit: Embedding Tools in Institutional Routines

Another recurring theme across sectors was the need for deepfake detection tools to integrate into existing workflows and infrastructures. Standalone web platforms were seen as inadequate for operational use, particularly among law enforcement, policy, and government interviewees. A senior policy expert in platform regulation and online hate noted: "We can't just use any random website. The tool has to be usable in a secure internal environment, ideally with no internet access and under clear legal conditions." (P14). Similar sentiments were echoed by public-sector media analysts and administrative units. The legal and data protection frameworks in which these actors operate impose strict requirements on software usage, especially when it involves the upload or processing of media data. For these users, features like local deployment, audit logs, and compliance with GDPR or internal IT standards were not negotiable. Journalists and civil society actors, while less constrained legally, also emphasized the importance of workflow compatibility. A fact-checking specialist from a press agency shared his vision: "You have your own editorial CMS or your system, all the images go into it, then you have the tool that automatically evaluates them directly." (P02). These accounts reflect a shared desire to reduce friction and avoid media disruptions. Deepfake detection, in the eyes of many, is not a specialized task but an increasingly common step in everyday media work. Consequently, detection functionality must be integrated into broader ecosystems — editorial, forensic, educational — in ways that align with sector-specific logics and limitations.

9.4.2.5 Legal and Ethical Boundaries: Compliance, Consent, and Caution

Practitioners, particularly those in public administration and law enforcement, were acutely aware of the legal constraints surrounding AI tools. Chief among these were concerns around GDPR compliance, especially regarding data retention, user tracking, and the interpretability of automated decisions. An interviewee working in a law enforcement unit made this point explicitly, requiring: "that it [the tool] is GDPR-compliant, [...], that the data is stored securely, that it is only stored by the authority, or that it is guaranteed in writing that everything is legally compliant. That's always important, if I'm going to put internal authority images somewhere, whether it's just for analysis or recognition, everything has to be properly secured." (P06). This reflects a broader theme of legal operability, where technical features must be subordinated to regulatory constraints. Practitioners also voiced concern about the impending AI Act and its implications for explainability, risk classification, and institutional liability. A related set of concerns focused on the potential misuse of detection tools themselves. Several participants raised the risk that governments or private actors might exploit such tools to suppress dissenting content or

conduct surveillance under the guise of authenticity verification. An administrative security specialist warned: “And if you look at what's going on geopolitically at the moment, this idea is perhaps not so far-fetched. Especially if you look at who has the ability to create such tools, at least financially and technologically, these are the same people who are currently very interested in shifting the boundary between truth and falsity, at least in the USA.” (P08). This dual concern — compliance on one side, ethical restraint on the other — reveals the normative terrain in which practitioners operate. Tools are not evaluated purely for their functionality, but also for their alignment with broader principles of democratic accountability, individual rights, and institutional transparency. These are not peripheral concerns; they are embedded in the interpretive schemata through which practitioners make sense of new technologies.

9.4.2.6 Governance and Trust: The Political Economy of Detection Tools

The final major theme that emerged from the analysis was that of governance, specifically, who develops and maintains the detection tool. Trust was not automatically extended to technology providers, particularly large private firms. Instead, participants consistently expressed a preference for tools developed through public or hybrid models, with transparent oversight. A government communications professional emphasized: “Ideally, it should be open source — so we can check what it’s doing and who’s behind it.” (P01). Others pointed to universities or trusted public research institutes as potential developers. A few supported public-private partnerships, provided that core functionalities remained auditable and accessible. This expectation speaks not only to governance in the narrow sense but to the political economy of technological trust. Practitioners considered a tool’s credibility to hinge on its provenance and institutional alignment, with functional features assessed through ethical and political lenses. The implications of this theme extend beyond procurement. They suggest that transparency, explainability, and usability cannot be treated as purely technical features. They are also functions of who builds the tool, under what conditions, and for whose benefit. Table 11 provides a summary of the key concerns expressed by each interviewee.

ID	Usability & Accessibility	Transparency & Explainability	Multimodality	Integration & Workflow	Legal & Compliance	Trust & Governance
P01	✓	✓	✓	✓	✓	✓
P02	✓	✓	✓	✓		✓
P03	✓	✓		✓	✓	✓
P04			✓	✓		✓
P05		✓	✓		✓	
P06		✓	✓	✓	✓	✓
P07	✓		✓		✓	✓
P08	✓	✓	✓	✓		✓
P09		✓	✓			✓
P10		✓	✓	✓		✓
P11	✓	✓	✓	✓		✓
P12	✓	✓	✓	✓		
P13	✓	✓				✓
P14	✓	✓		✓	✓	✓
P15	✓	✓	✓	✓	✓	✓

Table 11. Summary of individual interviewee concerns (✓) by key themes.

9.4.2.7 Synthesizing Functional and Normative Expectations

The interviews analyzed in this study do not simply list features; they articulate a set of deeply interwoven expectations, grounded in practical experience and normative reflection. The practitioners' perspectives reveal that effective deepfake detection is not a matter of technical accuracy alone, but of social embedment, epistemic transparency, legal conformity, and moral trust. What emerges is not a fixed checklist but a field of tensions: between simplicity and complexity, between automation and human judgment, between privacy and accountability. In Mayring's terms, the derived categories reflect both the manifest content of practitioner discourse and the latent structures of professional ideology. Multimodal deepfake detection tools, if they are to be embraced by their intended users, must navigate this terrain with both technical precision and social intelligence. Based on the qualitative findings, we derived a set of ten meta-requirements (MRs) (Walls et al., 1992) that reflect the expectations, constraints, and practical needs articulated by stakeholders across domains such as journalism, public education, law enforcement, and

civil society. These MRs were then synthesized into five overarching design principles (DPs) (Schacht et al., 2015), which guide the development of multimodal deepfake detection tools in a way that is both technically robust and socio-organizationally appropriate. Each DP addresses a cluster of related MRs, ensuring that the principles are grounded in empirical user needs while remaining generalizable for future design contexts. Figure 39 summarizes this mapping, illustrating how the DPs systematically respond to practitioner expectations.

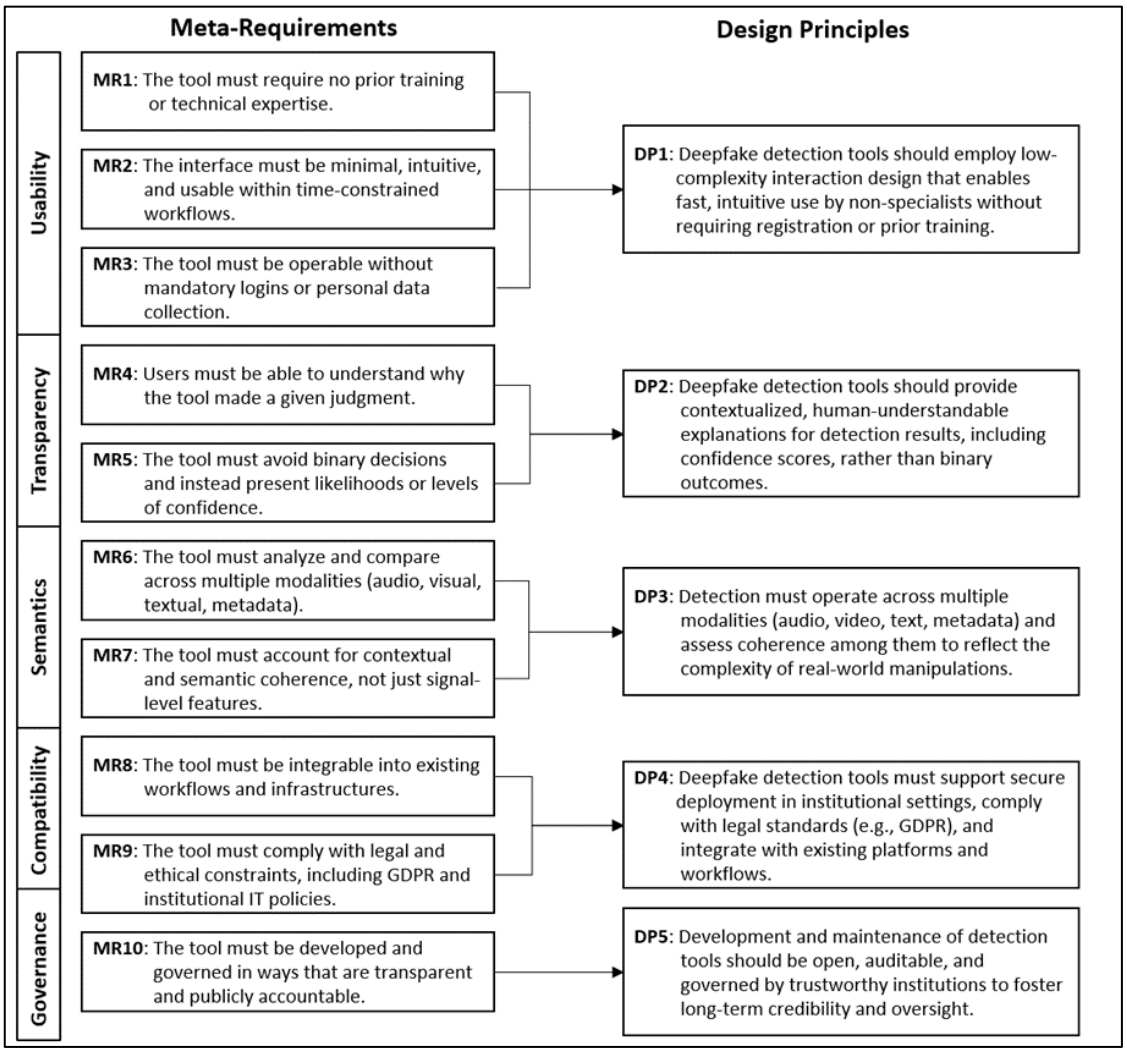

Figure 39. Meta-requirements (MR) and design principles (DP) for deepfake detection tools.

Figure 40 provides an overview of the analytical process of applying codes to specific text segments of the interview transcripts, leading to the formation of specific meta-requirements and design principles.

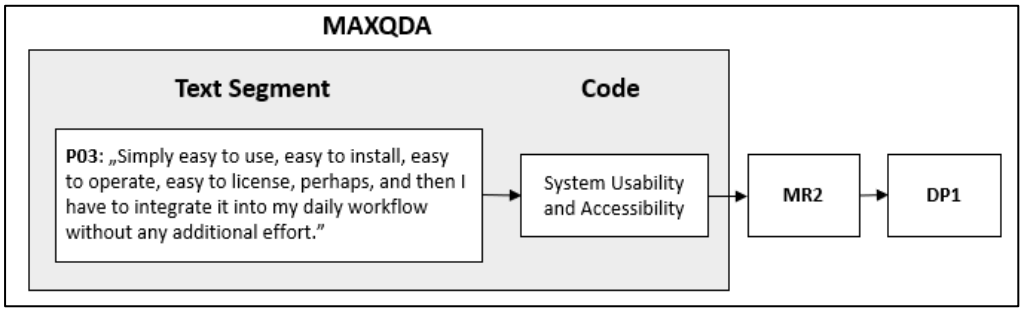

Figure 40. Summary of the analytical workflow.

9.5 Discussion

The growing concern around deepfakes in professional domains is not rooted in their current prevalence but in the anticipation of their future disruptive potential. Interviewees consistently emphasized a sense of urgency: while deepfakes are not yet widespread in their respective sectors, many foresee a tipping point. This forward-looking concern reflects a broader shift in risk perception – from reacting to immediate threats toward preemptively designing systems for emerging ones. Such anticipatory governance aligns with the literature on proactive cybersecurity (Bada et al., 2019) and speculative design in information systems (Auger, 2014), which emphasize that tools must be developed in advance of crises, not in their wake. This urgency coexists with a second finding: a strong consensus among practitioners that human judgment alone is increasingly insufficient to detect synthetic media. As generative AI tools outpace lay perceptual abilities, the need for automated detection has become not just apparent but inevitable. This shift toward automation echoes developments in other epistemic infrastructures (Frauenberger, 2019; Larkin, 2013; Star & Ruhleder, 1996), where human discretion gives way to computational assessments. In the context of media authentication, this raises critical questions about how trust is constructed and maintained. If detection tools become an established part of the media's trust infrastructure, their design must account for technical precision and epistemic legitimacy.

Interviewees also emphasized that deepfakes are inherently multimodal phenomena, combining manipulated video, synthetic audio, and even falsified text. This complexity undermines the utility of single-modality detection systems and directly supports MR6, which requires tools to "analyze and compare across multiple modalities." The need for multimodal detection aligns with research on integrated data streams (Mehta et al., 2018) and cross-modal analysis architectures (Baltrušaitis et al., 2018), suggesting that future detection tools must be capable of analyzing diverse content types simultaneously. Additionally, practitioners highlighted the importance of semantic coherence (MR7), noting

that technical features alone are insufficient for detection. These requirements are synthesized in DP3, which stipulates that "detection must operate across multiple modalities and assess coherence among them to reflect the complexity of real-world manipulations."

Across interviews, the demand for transparency and explainability in detection systems emerged as a critical theme. Practitioners explicitly rejected "black box" tools, particularly in sectors such as law, journalism, and public administration, where credibility is paramount. As articulated in MR4, "users must be able to understand why the tool made a given judgment," and in MR5, tools should "avoid binary decisions and instead present likelihoods or levels of confidence." This resonates with growing concerns in IS literature about algorithmic opacity (Burrell, 2016) and the need for explainable AI (XAI) (Doshi-Velez & Kim, 2017). These requirements directly inform DP2, which calls for "contextualized, human-understandable explanations for detection results, including confidence scores, rather than binary outcomes."

Moreover, practitioners emphasized accessibility and usability as non-negotiable requirements. As captured in MR1, tools must "require no prior training or technical expertise," with MR2 specifying "minimal, intuitive" interfaces suitable for "time-constrained workflows." The importance of avoiding mandatory logins or personal data collection (MR3) was also highlighted, particularly by practitioners concerned with privacy and rapid deployment. These requirements coalesce in DP1, which prescribes "low-complexity interaction design that enables fast, intuitive use by non-specialists without requiring registration or prior training." This design principle reflects not merely convenience but recognizes that detection tools must be democratically accessible to be effective in countering the spread of deepfakes.

Practitioners also noted that detection systems must fit into existing workflows and organizational structures, as captured in MR8 and MR9. Legal experts pointed to evidentiary standards and GDPR compliance, administrative practitioners noted internal CMS constraints, and media professionals cited editorial processes. These comments point to the importance of situated use – a concept well-established in socio-technical systems literature (Orlikowski, 2000; Suchman, 1987). DP4 addresses these concerns by requiring tools to "support secure deployment in institutional settings, comply with legal standards, and integrate with existing platforms and workflows." This tension between generalizability and contextual sensitivity suggests the need for modular architectures that allow customization without sacrificing analytic rigor.

The governance and political economy of detection tools also surfaced as a key concern, articulated in MR10 regarding transparent and publicly accountable governance models. Interviewees expressed skepticism toward proprietary solutions developed by large tech firms, citing fears about ethics, sovereignty, and opacity. Many advocated for open-

source or hybrid governance models, where tools are publicly accountable and community-vetted. This aligns with emerging discourses on digital sovereignty (Pohle & Thiel, 2020) and public-interest technology (Schank & McGuinness, 2021). DP5 addresses this by specifying that "development and maintenance of detection tools should be open, auditable, and governed by trustworthy institutions to foster long-term credibility and oversight." The normative implication is clear: if detection tools are to support democratic processes and public trust, their ownership and governance must reflect those values.

Finally, the findings point to a crucial role for the IS research community. The development of deepfake detection tools is not merely a technical challenge – it is an institutional, ethical, and political one. The IS field, with its longstanding engagement in socio-technical system design (Hevner et al., 2004), is uniquely positioned to shape this emerging infrastructure. This requires interdisciplinary collaboration between computer scientists, organizational scholars, IS researchers, legal experts, ethicists, and affected users (Weinhardt et al., 2024). It also requires a shift in orientation, from descriptive studies of existing systems to normative engagement with what these systems ought to be. In sum, this study illustrates that designing deepfake detection tools is not just about building better algorithms. It is about rethinking how truth is infrastructurally supported in digital societies. The interviews underscore that effective systems must be anticipatory, explainable, multimodal, workflow-sensitive, ethically grounded, and publicly governed. These are not just technical requirements; they are democratic imperatives.

9.6 Conclusion

9.6.1 Summary

The aim of the study was to investigate the views and needs of practitioners from various fields regarding the significance of deepfakes and the features they expect from detection systems, in order to gain insights for the development of user-friendly and trustworthy tools to ensure information integrity. In summary, our study provides empirically grounded insights and design principles that can directly inform the development of deepfake detection tools to make them more user-friendly, transparent, context-sensitive, and, ultimately, trustworthy. This helps to better address the challenges posed by deepfakes in various professional domains and protect the integrity of digital information. Our results reveal that practitioners across sectors view deepfakes as a rapidly advancing threat that will increasingly disrupt information ecosystems and erode trust. While currently less prevalent than simpler “cheap fakes”, deepfakes are growing in sophistication across video, audio, and images. The impact varies by sector, but common concerns include

erosion of public trust and influence in political opinions, especially among younger populations, as well as the potential for criminal abuse (e.g., fraud). There is consensus that human detection alone will soon be insufficient, making automated systems necessary for timely responses to potential harm. For detection tools, practitioners require solutions that are easy to use with minimal technical expertise, while providing transparent explainable results rather than black-box outcomes. These tools must analyze multiple data types with contextual understanding and integrate seamlessly into existing workflows and secure infrastructure. Legal compliance, particularly with GDPR, is essential, as is trustworthy governance with preference for open-source or transparent oversight models rather than proprietary solutions from large tech companies. These requirements reflect both technical needs and socio-organizational considerations for effectively addressing the growing deepfake challenge.

9.6.2 Limitations

As with any qualitative research, certain limitations should be acknowledged to contextualize the findings and inform their interpretation. This study is based on qualitative, semi-structured expert interviews. While this approach provides rich, in-depth insights into practitioners' perspectives, the findings are not directly generalizable to a broader population. The sample size of 15 interviewees is appropriate for qualitative research (Hennink & Kaiser, 2022), yet it does not offer statistical representativeness for all professionals interacting with deepfake technologies. Furthermore, all interviews were conducted in German, which may limit the diversity of perspectives and affect the transferability of the findings to other linguistic and cultural contexts. Although the sample includes experts from various sectors, the specific needs and challenges within these sectors may be more nuanced than the study's overarching analysis captures. Additionally, the study centers on the perceived relevance of deepfakes and the requirements for detection tools. It does not directly assess the actual implementation or effectiveness of current or emerging deepfake detection technologies in practice. Moreover, the interviews took place between February and April 2025. Given the rapid evolution of deepfake technologies and corresponding countermeasures, practitioners' perceptions and requirements may shift in the near future. Finally, while the study derives a set of requirements and design principles for deepfake detection tools, these have not yet been translated into a concrete artifact design nor evaluated in practice. As such, it remains an open question to what extent the proposed tool would effectively address real-world problems within specific organizational contexts. This represents a crucial next step in the iterative design research process and a key opportunity for future research.

9.6.3 Future Work

Building on the insights of this study, several directions for future research emerge. One important step is to broaden the scope of inquiry by including more diverse linguistic, cultural, and geographical contexts. This would allow for a more nuanced understanding of how deepfakes are perceived and managed across different professional environments. Our semi-structured interviews provided valuable initial insights into practitioner perspectives. Building on these, formal and iterative requirements engineering will help refine and validate the requirements to ensure clarity and alignment among stakeholders. Further research could also adopt a more sector-specific focus, exploring in greater depth the particular challenges faced by fields such as journalism, law enforcement, education, or cybersecurity. Complementary quantitative studies could help assess how widespread certain perceptions or practices are and whether they change over time. In this context, longitudinal studies may offer valuable insights into how the perceptions and requirements of practitioners evolve in response to technological developments. As deepfake technologies – and their countermeasures – continue to advance rapidly, regular re-evaluation will be crucial to keep research aligned with real-world needs. Finally, future work could shift from perception-based analysis to empirical evaluation of detection tools in practice. As a next step in design research, studying their implementation, usability, and actual effectiveness across contexts would help bridge the gap between technological development and practical application. Such an approach would move beyond theoretical derivation toward practical validation, shedding light on whether and how the proposed tool can effectively address the challenges practitioners face when dealing with deepfakes. Taken together, these directions highlight the need for an interdisciplinary, adaptive, and ongoing research agenda to keep pace with the evolving deepfake landscape.

10 Literacies Against Disinformation: Examining the Role of Data Literacy and Critical Media Literacy to Counteract Disinformation[15]

10.1 Introduction

The "digital condition" (Stalder, 2018) places contemporary societies and individuals in a tension between the community-creating potential of the Internet and the risks social media poses to democracy. Digital platforms have become a primary forum for promoters of far-right ideologies and disinformation. Nowhere is it easier for them to reach their own followers as well as a broader audience. Their goal is to directly link their racist and anti-democratic messages with current sociopolitical discourses and life worlds (Glaser et al., 2017; Liang & Cross, 2020). And by using the internet, they meet the younger generation where they are. Although portraying their platforms as a kind of youth movement in which patriotically minded people spontaneously meet and exchange ideas, these right-wing ideologues are, in reality, employing a strategic concept for the ideological seizure of power in the social sphere. Right-wing extremists have been using the Internet and especially community organizing platforms for propaganda for some time, often disguised by subcultural elements ranging from music and games to vegan cooking. In this sense, they are active users, interpreters, and influencers who contribute to the digital condition. They capture attention and establish rapport before introducing their extremist ideas. This occurs through ideologically driven texts, links to niche communities, and the promotion of events by radical organizations (Glaser et al., 2017). At the core of right-

[15] This chapter comprises an article that was published by Anna Soßdorf, Carolin Stein, Isabel Bezzaoui and Jonas Fegert in the following outlet with the following title: Literacies Against Fake News: Examining the Role of Data Literacy and Critical Media Literacy to Counteract Disinformation. In MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung, 59, 55-76, 2024. Note: Tables and figures were renamed, reformatted, and newly referenced to fit the structure of the dissertation. Chapter and section numbering and respective cross-references were modified. Formatting and reference style were adapted and references were updated. Details of the author's individual contributions to this publication are provided in the appendix.

wing radicalization, however, is disinformation and propaganda (Lewandowsky & Yesilada, 2021). As these groups use subtle exposure to their ideas through memes and disinformation to shape discussions (Liang & Cross, 2020), individuals need special literacy skills to navigate the digital space and avoid falling victim to strategic disinformation and propaganda.

To address the challenges of dealing with disinformation in social media, this paper aims to show how important competencies could be fostered to counter deceptive information. We examine a distinct set of competencies, beginning with a comprehensive understanding of media competencies (Soßdorf, 2023; Trültzsch-Wijnen, 2020). We focus first on critical media literacy, which enables a critical and reflective approach to structures, processes and content in social media (Allen et al., 2022). Second, we focus on data literacy, which we define as the ability to understand how data and numbers are represented as well as a capacity for data-driven autonomy of action in dealing with disinformation in a competent way.

The article elaborates how these two literacies can be interwoven in a three-step process of awareness, reflection, and empowerment (Schmitt et al., 2018), and how their interrelation can be further developed into a model to create synergies empowering people to stand up against disinformation. In our *Synergistic Literacy Model Against Disinformation*, we argue that individual literacies together contribute to the shaping of a comprehensive empowerment for living in a digitally driven culture by using media responsibly, critically examining media forms, exploring media effects, and finally deconstructing alternative media (Kellner & Share, 2005). In the long run, to combat online disinformation, an examination of the interplay of media and data literacy competencies is crucial for educators, learners, and developers of media tools. We argue in a broader sense that such emerging sets of competencies – if they are encouraged by a digital infrastructure offering learning opportunities – facilitate participation in modern society (Marten, 2010). Ultimately, they may stabilize democracy and thus contribute not only to digital literacy in general but also to civic literacy and participatory citizenship.

10.2 Theoretical Background on the Challenges of the Digital Condition

In the digital condition, the "multiplication of cultural possibilities" (Stalder, 2016, p. 10) becomes permanent and maintains a constant presence in our everyday lives comprised of three central dimensions: referentiality, communality, and algorithmicity. Whereas referentiality encompasses the infrastructure and social action on the Internet in which actors access, refer to, modify, remix, and create new content from existing digital products to

(co-)shape cultural meaning, communality refers to collaboratively created content. Finally, algorithmicity involves the digital landscape in which automated decision-making processes reduce and shape the information overflow. This approach facilitates the extraction of information from the expanding pool of data available to individuals, subsequently serving as a foundation for both individual and collaborative actions. On the one hand, these ideas of a constantly present digital ecosystem offer numerous opportunities for community engagement using digital tools and platforms (Kaplan & Mazurek, 2018). On the other hand, precisely these community-building tools are used by different audiences to spread and amplify populism and disinformation and thereby foster societal polarization (Glaser et al., 2017). This raises two major challenges requiring society to examine different concepts of digital competence.

The first challenge is that several peculiarities of social media, such as its basic modes of representation and interaction, promote certain developments in the course of discussions. Youth-oriented approaches have gained particular momentum through the stylistic tools of the social web. Multimedia forms of presentation, emotionalization, and sarcasm are employed by right-wing extremists, among others, to ensure the rapid dissemination of deceptive content. Hostile attitudes toward marginalized groups are also incited through targeted disinformation that spreads quickly on the Internet. Disinformation is defined as false information, spread with the intention to deceive (European Commission, 2018). Under the guise of serious reporting, right-wing extremists publish reports that are either completely invented or based on news from reliable media outlets but distorted by racist and anti-democratic messaging (Glaser et al., 2017). The origins of these articles are usually difficult or impossible to trace, as the authors rely on inconsistencies being lost in the flood of information and statements not being checked for their truthfulness (Conway et al., 2019). In this context, being able to distinguish facts from disinformation requires a developed and specific set of competencies as well as critical thinking (Bezzaoui et al., 2022a; Chu & Lee, 2014). Guess et al. (2020) demonstrated that improved media literacy can, for example, assist individuals in more precisely assessing the authenticity of online content. The results of their study indicate that the absence of sufficient critical media literacy plays a significant role in individuals' susceptibility to disinformation.

A second challenge is that the digitization of society goes hand in hand with increasing datafication (Schüller et al., 2019). Technological advances enable larger amounts of data to be collected and stored (Clarke, 2016; Twidale et al., 2013) just as new methods of data and information processing and retrieval are emerging (Hambarde & Proenca, 2023). Although these developments allow users to make powerful claims and inferences, they also fuel inequality and exploitation. Data ownership and literacy skills restrict who can use data to their advantage (D'Ignazio, 2017). Increasing efforts to publish data in publicly accessible portals is not sufficient to ensure the usability of the data by the lay citizen

(Simonofski et al., 2022; Twidale et al., 2013). In the absence of the necessary knowledge and skills, the mere publication of decontextualized data can contribute to the propagation of fallacies. Simultaneously, data products increasingly find their way into media, where they are expressed, contextualized, and interpreted by authors (Schüller et al., 2019). As such, critical engagement with articles published in the media frequently depends on the recipient's ability to extract and evaluate underlying data (Debruyne et al., 2021; Schüller et al., 2019), just as searching for, selecting, evaluating, and interpreting essential information becomes more difficult (Mahyoob et al., 2020; Shu et al., 2020b; Verma et al., 2021).

Based on this theoretical background and the challenges presented above, different literacy concepts will be discussed in the following to lay out the argumentation for our new *Synergistic Literacy Model Against Disinformation.*

10.3 Countering Right-Wing Extremist Disinformation Requires Literacies

Media competencies and literacies, both in general terms and in regard to specific competencies, have been a broad field of research in recent decades (Fischer et al., 2020; Kerres, 2020; Livingstone, 2004; Potter, 2010; Reddy et al., 2020; Trültzsch-Wijnen, 2020). Various studies have been conducted to evaluate the importance of media skills and appropriate frameworks, such as the Frankfurt Triangle, the 4Cs, and the Digital Competence Framework for Citizens (DigComp) (Brinda et al., 2020; Carretero et al., 2017; Pfiffner et al., 2021; Rasi et al., 2019). Livingstone et al. (2012) contend that formulating a general definition of media literacy with universal criteria is challenging due to the diverse contexts and target groups involved. Concurrently, Hug (2011) observes an ongoing trend toward the emergence of new literacy concepts with a broad focus. He asserts that these concepts must be precisely defined and critically examined. Nevertheless, the current understanding of media literacy can be summarized as skills in "accessing, analyzing, evaluating, and creating media messages," the application of "creative and playful forms of multimodal media content production," "abilities to reflect on one's communication behavior, to act and participate in society," and finally, the capacity "to promote one's digital well-being" (Rasi et al., 2019, p. 1). In terms of frameworks, Zorn (2011) summarizes the core elements of the various frameworks and models as the development of skills, including the "selection, production, usage, and evaluation of media" (Zorn, 2011, p. 187). In the German-speaking discourse, definitions of media literacy range from the ability to use various media for one's own communication and activity (Baacke, 1999) to the ability to use media in a self-determined, creative, and socially responsible way as well as to move in media contexts (Tulodziecki, 1998). Since these

early definitions, different debates have arisen around the meaning of the term (Aufenanger, 2001; Hugger, 2008; Schorb et al., 2017; Spanhel, 2011; Tulodziecki, 2015).

Around 2010, a broad discussion unfolded around the scope of the term media literacy. This discussion was marked by the ambivalence inherent in the term, often perceived in various approaches as both a "general requirement or significant quality for action in the media field" as well as an "objective in the sense of a desired level of competence" (Tulodziecki, 2011, p. 22). It has also become imperative to refine the conceptualization of the term in the era of digitalization and widespread access to digital media, and thus to transcend the understanding of media literacy that evolved in the analog era (Zorn, 2011). Given that the discourse has revolved primarily around the educational dimensions of media literacy, authors have proposed a distinction between an "administrative-pedagogical perspective," a "pedagogical-practical theoretical perspective", and an "educational-theoretical-reflective perspective" (Jörissen, 2011, p. 228).

One recent promising perspective summarizes various literacies under the two dimensions of media literacy and information literacy in order to capture the current debate and to cluster the different individual skills within a structural concept (Trültzsch-Wijnen, 2020). In this context, Trültzsch-Wijnen describes media literacy as the ability to critically understand and evaluate media content, and information literacy as the technical skills of usability, knowledge about access, and identification of application strategies (Soßdorf, 2023; Trültzsch-Wijnen, 2020). Critical media literacy (CML) goes beyond the notion of classical media literacy, strongly emphasizing critical engagement with power dynamics and ideologies shaping media content and representation in media discourse (Kellner & Share, 2007). Simultaneously, expanding the understanding of information literacy, data literacy (DL) addresses the promotion of skills necessary to navigate an increasingly datafied information environment (Carmi et al., 2020; Schüller et al., 2019). Following this division, the two literacies addressed in this paper, CML and DL, represent these two approaches to media by a) looking at the media contexts and b) referring to skills in the use of information data.

A critical perspective toward the media recognizes that the presentation of information incorporates power imbalances. To foster a critical comprehension of both manipulative communication and the internet as a distribution medium, individuals must have broad knowledge and a deeper understanding of (social) media functionalities (Rieger et al., 2017). Consequently, a thorough investigation of media content must also examine how the media typically influence audiences in interpreting and navigating messages related to factors that favor dominant groups (Higdon, 2020). In view of the current impact of phenomena such as hate speech, filter bubbles, and disinformation and how these affect the functioning of our society, it is crucial to understand CML as a key competency (Peissl

et al., 2018). This competency encourages people to consider why a message was sent and where it came from (Kellner & Share, 2007).

Ganguin and Sander (2015) define CML as the ability to analytically, reflexively, and ethically evaluate and judge media content. Following Kellner and Share (2005), CML entails the development of skills in analyzing media codes and conventions, and the ability to critique stereotypes and ideologies as well as the competence to interpret media texts' multiple meanings. Therefore, CML goes beyond analyzing the content of media and delves into understanding the power dynamics associated with the creation and dissemination of that content. Additionally, it assists individuals in responsibly consuming media, including discerning and assessing media content, critically examining media forms, exploring media effects, and, based on those abilities, deconstructing alternative media. In the context of teaching CML, dealing with disinformation is undoubtedly important (Maloy et al., 2022; Peissl & Sedlaczek, 2022). It is crucial to highlight that media culture may contribute to the promotion of racism, ethnocentrism, and various forms of prejudice. It may also endorse disinformation, problematic ideologies, and questionable values. Thus, advocating for a dialectical approach to the media and questioning ideology, bias, and connotation of content are essential to CML (Kellner & Share, 2005). The notion of ideology critique embedded in CML education can, among other things, equip individuals to quickly recognize right-wing extremist maneuvers such as the spread of disinformation and hostility towards specific social groups in the digital space.

DL is among the newer competencies that developed structurally out of the term information literacy, which was introduced in the context of libraries and the corresponding need to deal with collected information (Carmi et al., 2020; Schüller et al., 2019). The term DL was coined with increased digitalization and datafication to describe competencies necessary to address these developments. Yet, demarcations between multiple literacies, such as information, statistical, and digital literacy, remain blurred (Bhargava et al., 2015; Gould, 2021; Schüller et al., 2019). As such, DL is subject to multiple definitions, ranging from the definition of concrete skill sets as the "ability to read, work with, analyze, and argue with data as part of a larger inquiry process" (D'Ignazio & Bhargava, 2016, p. 84) to a more general empowerment of individuals to navigate and engage with their own data-based environments (Bhargava et al., 2015; Schüller et al., 2019). Importantly, these definitions underscore the multifaceted nature of the term, including a call to action based on acquired literacies (Bhargava et al., 2015; D'Ignazio & Bhargava, 2016). Multiple frameworks have sought to capture the facets of DL and their implications for a data-literate society (Bhargava et al., 2015; Carmi et al., 2020; Schüller et al., 2019). However, societal and technological developments constantly add new aspects to the field. Advances in computational analytics and artificial intelligence create new opportunities and challenges to data value creation and lead to the emergence of terms such

as data science literacy and big data literacy (D'Ignazio & Bhargava, 2016; Sander, 2020; Schüller et al., 2019). Likewise, the increase in online dis-, mis-, and malinformation requires a revision of the DL concept (Carmi et al., 2020; Koltay, 2022). As such, discussions on critical DL concepts emphasize the ability to critically evaluate data and datafied environments in terms of their backgrounds, intentions, and modes of operation (Koltay, 2022; Sander, 2020).

10.4 Synergetic Linkage of Critical Media Literacy and Data Literacy

In the academic discourse, multiple efforts have been undertaken to link or distinguish different literacy fields. Kellner and Share (2005) use the term "multiple literacies" to refer to the many different competencies needed in today's society to access the social public sphere and to be able to interpret, criticize, and participate. Koltay (2022) argues that the ongoing technological conversion of media, information, and communication systems encourages the combination of different sets of literacies, hypothesizing a potential union of data and media literacy. In contrast to this, Carmi et al., (2020) state that the sets of literacies reflect on the political and technological context of their development, leading to newer literacies such as data or digital literacies encompassing older forms of media or information literacy. Yet, Twidale et al. (2013) claim that despite the conceptual overlap, literacies should be distinguished depending on the scale, genre, and usage. However, especially when turning towards a critical literacy perspective, it becomes obvious that the content and data dimensions are closely interconnected (Musi et al., 2022). Mcdougall (2019) argues for acknowledging the intricacies of "dynamic literacies", blending or transcending the boundaries between different spaces and roles. As such, we believe that to combat online disinformation, a close examination of the interplay between media and data literacy competencies is crucial for educators, learners, and tool developers alike. To do so, we reflect in Table 12 on the CML dimensions of awareness, reflection, and empowerment proposed by Schmitt et al. (2018).

Dimension		Description
CML 1	**Awareness**	Awareness, in this case, means becoming aware of the existence of disinformation and possibly encountering it (J. B. Schmitt et al., 2018): ▪ Knowledge of various forms of disinformation and manipulation (e.g., rhetorical resources, distorted articles, and pseudo-media outlets) ▪ Deeper understanding of how media and online media, including algorithms, operate Awareness can trigger subsequent activities such as reflection.
CML 2	**Reflection**	Reflection in the context of CML is about applying analytical criteria to media content and determining whether or not it is deceptive (J. B. Schmitt et al., 2018): ▪ Conscious consideration and thorough thinking before an article is liked or shared, or a headline is taken at face value ▪ Utilizes an individual's knowledge, abilities, and attitudes to critically evaluate (media-communicated) information based on specific criteria including credibility, source, and quality
CML 3	**Empowerment**	Individuals' confidence in their ability to detect manipulative messages, participate in social discourses, and actively position themselves against disinformation is cultivated through empowerment strategies and methods: ▪ A certain form of behavior that encompasses a person's ability to recognize and state doubts about specific content as well as express their own thoughts. Empowerment relies on individuals' knowledge (awareness) and analytical thinking (reflection) regarding messages conveyed through the media. Moreover, it could also be a factor that anticipates increased awareness.

Table 12. Dimensions of the CML framework by Schmitt et al. (2018).

Awareness, reflection, and empowerment are considered intertwined dimensions. We show in Table 13 how they can be enriched by the three domains of data citizenship – data thinking, doing, and participation – which in turn subsume different competencies of DL (Carmi et al., 2020; Yates et al., 2020).

Dimension		Description
DL 1	**Data Thinking**	Data thinking revolves around aspects of critical data understanding when citizens view or analyze situations from the angle of data (Yates et al., 2020). First, it involves attitudes and knowledge, such as understanding aspects of data collection or the data economy (Carmi et al., 2020; Yates et al., 2020): ▪ Data and data products are increasingly disseminated and contextualized both in the field of professional journalism and in social media (Schüller et al., 2019) ▪ They can be misused to serve vested interests (Pullinger, 2021) such as recruiting unsuspecting adolescents for far-right groups (Liang & Cross, 2020) ▪ Platform design and business models can influence user behavior (Carmi et al., 2020) Thus, promoting data thinking could be a valuable extension to the dimension of awareness in a datafied environment. Additionally, data thinking also includes aspects of critical usage of data, such as the ability to critically consider and discuss data analysis and communication (Yates et al., 2021), which makes it relevant for the dimensions of reflection and empowerment and overlaps with the aspects of data doing and data participation.
DL 2	**Data Doing**	In a data-driven debate, critical reflection on and active positioning in relation to content can necessitate thinking about and engaging with the underlying data. Along with data thinking, literacy skills in actively engaging with data (data doing) (Carmi et al., 2020; Yates et al., 2020) may be essential to the dimension of reflection and empowerment. Data doing revolves around aspects of data engagement on a day-to-day basis (Yates et al., 2020): ▪ Everyday data engagement may be necessary when reflecting on content. On social media, for instance, users might need to identify and assess a data source in a post or interpret different formats the data is presented in (Yates et al., 2020) ▪ Aspects of data doing such as data creation or citation in a blog, on social media, or in other contexts (Yates et al., 2021) might play a role when envisioning empowered citizens who actively participate and position themselves in the public sphere ▪ On an individual level, skills of data literacy can support empowerment when citizens are enabled to utilize data in their local context (Bhargava et al., 2015)
DL 3	**Data Participation**	Data participation describes the ability to engage proactively with and about data, going from an individual to a network perspective, focusing on the "collective and interconnected nature of data society" (Yates et al., 2020, p. 10) and could thus enrich the dimension of empowerment: ▪ Highlights how DL enables citizens to actively shape the community by getting involved in disinformation debates, utilizing data for civic action, or supporting others in their literacy journey (Yates et al., 2020) ▪ Goes beyond an individual literacy level toward ways of mutual and collective enablement. It seeks to counteract disparities of power and feelings of disempowerment in datafied environments (Yates et al., 2021)

Table 13. Dimensions of the data citizenship framework and their relation to CML.

As awareness of disinformation grows, so does the ability to reflect critically upon it. Critical reflection on deceptive content, in turn, necessitates knowledge regarding the presence of such content on the internet. Reflection on deceptive content affects the feasibility of proactively opposing such content (empowerment) and may increase awareness of the contributions of those who have already stood up against disinformation on the internet (Schmitt et al., 2018). Therefore, essentially, CML claims to promote both critical consumers and creators of media (Allen et al., 2022). Concepts of DL, from data thinking to data participation, can support dimensions of awareness and reflection while enriching aspects of empowerment in particular.

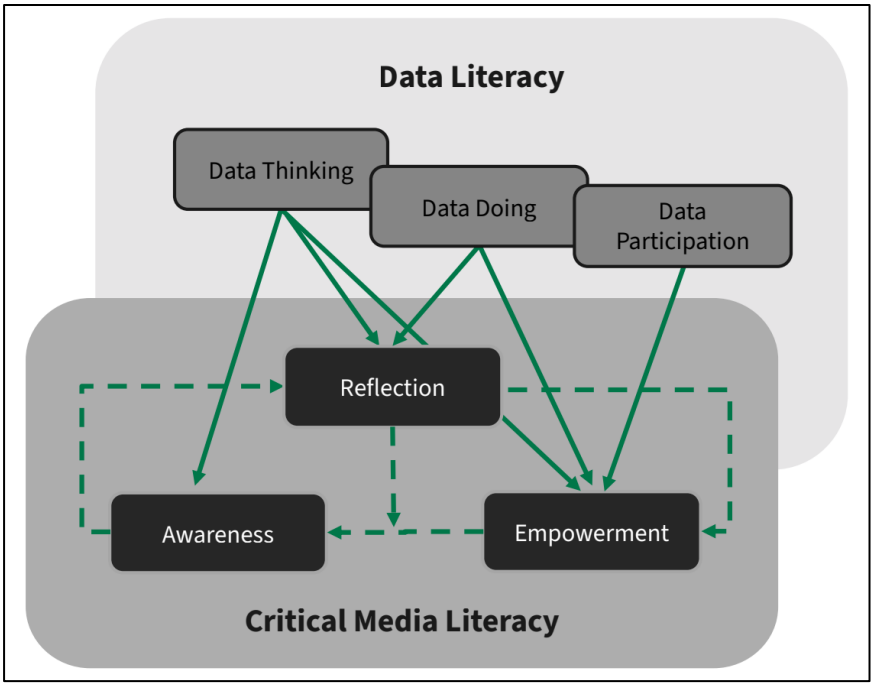

Figure 41. Synergistic Literacy Model Against Disinformation.

From our perspective, these two concepts can be combined in our Synergistic Literacy Model Against Disinformation (Figure 41) by referring to the elements of data thinking, doing, and participation, which are subsumed under the umbrella term DL, as elements that can enhance the development of a broader critical mindset on the individual level along the three dimensions of CML (awareness, reflection, empowerment). Accordingly, we propose to allow for their interplay in our new theoretical concept and in the groundwork for developing practical media educational formats and methods.

10.5 Proposing Learning Opportunities and Digital Infrastructures for Democratic Resilience

Having established the theoretical foundation for the amalgamation of CML and DL (see Table 12 and Table 13), we shift our focus to the tangible benefits that emerge, particularly in practical application and influence. Many fundamental issues in dealing with media and information in different areas of society are not new but have to be reclassified for the digitization context (Peissl & Sedlaczek, 2022). Competent and critical media action is thus becoming a central social challenge in the digital age. A prime example of this can be seen in the ongoing debate on approaches to combat disinformation (Diepeveen & Pinet, 2022). Individual interventions can highlight the individual's responsibility to develop necessary literacies, while structural interventions can invoke platform design or tools (Diepeveen & Pinet, 2022). The latter might include interventions focusing on facilitating media or data handling or supporting educational goals (Twidale et al., 2013). In this section, we therefore explore the need to create learning opportunities for individuals to build literacies and investigate the role that technological interventions can play. Furthermore, we argue that this conceptual linkage can play a significant role in terms of (1) a more adequate and target-group-specific conceptualization of digital learning settings, (2) a more accurate development of digital structures and usable tools, and (3) positive impacts on a societal level.

10.5.1 Using Emerging Learning Opportunities

When it comes to learning opportunities, it is crucial to regard the learner as a person with several modes of perception and, therefore, offer a setting that attracts different senses and modes of learning (Pritchard, 2017; Schunk, 2012). Practically, educators must consider this context while fostering a critical data mindset. They can employ diverse media like text, videos, podcasts, and images to present content. Additionally, a mix of activities such as reading, researching, manipulating, and creating data should be integrated to provide a comprehensive data-handling experience. This approach enables learners to grasp, interpret, and apply data within novel contexts, aligning with their unique learning preferences. To make a learning experience more realistic and relatable to everyday life, it has been shown that digital learning should not focus on individual competencies but rather address a set of similar and connected skills (Fischer et al., 2020; Moser, 2020; Soßdorf & Gallach, 2022). Therefore, we suggest reflection on complex problems related to data, based on real cases with multiple dimensions. This enables individuals to learn by dealing with actual data problems in our world but also to have a learning opportunity that shows how important data skills are interconnected and interdependent.

When talking about digital skills, the focus should lie on the development and establishment of a certain digital mindset, which means being open to new digital techniques and methods, being self-confident in navigating the digital sphere and finding individual solutions, and being aware of the permanent digital condition (Soßdorf, 2023; Stalder, 2016) that affects our lives. It is important to be aware of the fact that digital platforms and tools are always available for our convenient use, but under the condition that data is scraped and monetized while we are navigating the digital sphere. Keeping these conditions in mind, reflecting on them, and being able to find as well as choose individual paths in everyday life – with or without the use of digital tools or settings – is what we refer to as a digital mindset. It can thus be regarded as an overarching, general skill (Soßdorf, 2023) since it is not specifically bound to certain tools or platforms but addresses a way of living and coping with the challenges of digital life.

10.5.2 Leveraging Digital Infrastructures and Tools

The idea of promoting digital learning brings with it certain requirements for infrastructure and digital platforms. Digital tools and infrastructures to support the handling of data and information already exist on a large scale (D'Ignazio & Bhargava, 2016; Musi et al., 2022): they include powerful tools such as R, Python, and Excel, which need training to be used effectively, but also simpler tools that facilitate individual tasks (D'Ignazio & Bhargava, 2016). The latter requires system designers to anticipate the needs of their users as learners and to focus on learning processes (D'Ignazio & Bhargava, 2016). For the specific context of strengthening literacies for combating dis- and misinformation, Musi et al. (2022) identified 22 gamified tools that are designed to enable learning. However, they point out that although the tools are intended to be educational, they require comprehensive assessment to improve their effectiveness. In this context, we propose that theoretical literacy models could help both guide the design of tools and enable their structured evaluation. Applying our structural model may help to recognize the functionalities of tools that can target different educational outcomes, from raising awareness to supporting reflection to empowering users. Likewise, the tool's focus can be on different activities, from supporting critical thinking to enabling active doing to social participation. Through the debunking tool *New-Wise*, for instance, the user's ability to judge the truth in headlines is assessed through a direct debunk, promoting awareness and reflection on deceptive news content. The user is invited to think about the information contained in the headline but does not require active doing in terms of checking sources or searching for additional data. Debunking tools contain mainly gamification elements, whereas prebunking tools encourage engagement in addition to awareness and critical reflection through sophisticated forms of explicit gameful design, such as simulations and serious games. In the *Vaccination News* chatbot, for instance, users are guided through a sequence

of critical inquiries that highlight possibly flawed arguments, cautioning them to question the credibility of a news piece. Through this gamified pre-bunking tool, users are encouraged to think about the underlying data, but also to actively handle data on their own, for example, by accessing and evaluating other sources. The game spurs users to critically review content (reflection) and enables them to actively express their doubts (empowerment) (Musi et al., 2022). While the latter is an important prerequisite to data participation, the game does not further animate the users to utilize their skills for participation.

For existing tools, the application of our theoretical model allows us to systematically describe a tool's focus which could in turn help to assess the technological landscape and identify gaps. Moreover, new technologies could be developed alongside all or a subset of our three dimensions (awareness, reflection, empowerment) to assist individuals in developing the needed skills. Likewise, the dimensions can be utilized to systematically evaluate the effectiveness of tools. As such, our framework could be useful to system designers in both a conceptual and operational phase of tool production.

Evaluation of tools is especially important as they come with certain limitations and can potentially produce side effects, depending on the usage scenario. The use of simulation tools like *Bad News*, *GoViral!*, and *Fake It To Make It* places players in the role of a disinformation website editor, helping them to understand the mechanisms behind creating and spreading fraudulent content. This implies the risk that players may become more sympathetic toward the creators of disinformation, especially if the playful element of the application is the focus of the specific usage scenario. For instance, empirical evidence from studies on video game design suggests that players might develop empathy toward their in-game characters and see them as role models for future behavior (Konijn et al., 2007). To address this bias, *GoViral!* includes face-threatening outcomes, where players receive messages from disappointed friends about their behavior. Similarly, *Harmony Square* visually portrays the harm caused by disinformation by showing the game's neighborhood going downhill. Despite these efforts, fictional goals like earning money for personal needs may make the decision to spread disinformation more relatable and justifiable to players (Musi et al., 2022). Furthermore, such gamified tools may only reach a very limited target group: As these tools have a clear educational purpose, they attract individuals who are already interested in learning about how disinformation spreads. However, to effectively reach people who are vulnerable to disinformation or have authoritarian right-wing tendencies, and thus strengthen their democratic resilience, it is essential to include educational content about disinformation in games that have a broader scope and appeal to a wider audience (Musi et al., 2022).

Eventually, in the development of educational tools and platforms, it is essential to include instruction on argument-checking in addition to proper fact-checking (Brave et al.,

2022). Argument-checking means evaluating the overall argumentation for its acceptability, relevance, and sufficiency. This approach not only empowers individuals to distinguish factual from deceptive information but also equips them with the skills to become better content producers. Existing platforms and digital systems can easily be addressed as topics of learning sessions to critically analyze their mechanisms with the aim of proposing necessary regulations and developing appropriate policies. Moreover, complex problems that occur on platforms can be addressed, not only on the individual level but also on a societal level, which requires regulation and responsibility on the part of digital organizations and corporations as well as policymakers. It is crucial that individual learners as well as society at large have the opportunity to reflect on platforms' strategies and procedures and have a chance to exert influence. When addressing right-wing extremist movements, it is vital to note their robust digital organization and embedding of various nationalistic characteristics. To effectively counteract their influence, democratic societies must grasp media dynamics and influencing tactics. This allows the development of tandem literacies: critical analysis of content disseminated by such groups, and data skills in comprehending platform operations and data leveraging to disseminate ideas.

10.5.3 Society and Democracy

As developing the aforementioned skills is a collective societal endeavor, resources are required on the individual (micro) level as well as on the educational (meso) and political (macro) levels. The overarching goal is to enable learners to be(come) active citizens, to recognize their interests, opportunities, and responsibilities arising from digitization, and to make well-informed decisions about their media actions (Peissl et al., 2018). In this sense, the *Synergistic Literacy Model* applies not only to the individual level but also to the societal system as a whole, which must develop appropriate competencies to stabilize and strengthen itself from within. Therefore, we would like to emphasize that it is not only the individual who is responsible for acquiring appropriate skills to participate in an increasingly complex and digital social space: Representative institutions must also work to create appropriate framework conditions so that the necessary competencies can be learned. The transitions between the roles of the individual, institutions, and society as a whole are fluid. In the first sense, literacy interventions benefit the individual, but in the second sense, they ideally enable the individual to initiate and support the learning process of other people. Accordingly, individuals become literate not only for their own needs and purposes, which are described in the *Synergistic Literacy Model*: In the stage of empowerment, individuals may feel encouraged to actively support other people in building their own literacy skills. The critical skills this requires involve agency, as learners and educators become co-creators of their own knowledge and competencies (Wright

et al., 2023). The model can thus also be adapted on the societal level, where individuals can benefit from each other's skills and knowledge.

Disinformation poses an existential threat to democracy as without access to accurate information, individuals can be prevented from realizing their own societal visions. Ultimately, the manipulation of media hinders meaningful participation in shaping society (Higdon, 2020). Those who would like to participate in the media discourse must be capable of critically analyzing and assessing the social dynamics and significance of this discourse (Peissl & Sedlaczek, 2022). According to research conducted by Pennycook and Rand (2018), the primary factor behind vulnerability to disinformation is inadequate critical thinking rather than other factors such as partisan bias. Therefore, to effectively combat the dissemination of deceptive information, users need to develop a higher level of critical media competence. An examination of critical competencies should enable individuals to expand their ability to act in a democratic society, to form opinions independently, to constructively shape media content themselves, and to participate in political life (Peissl & Sedlaczek, 2022).

10.6 Conclusion

In this paper, we develop a new model to combat (right-wing extremist) disinformation online. In our *Synergistic Literacy Model*, we combine two digital competencies, critical media literacy and data literacy, and argue that in combination, they can function as a theoretical foundation for a digital learning environment. Our theoretical starting point is the so-called "digital condition" (Stalder, 2018), which describes today's reality as a permanent digital environment in which our digital and analog ecosystems are in flux. From this perspective, people are both users and creators of digital content and culture and, therefore, need certain competencies as individuals but also as members of a democratic society. After decades of discourse on the necessary skills for a digitalized world, two competencies have been identified as distinct but at the same time intertwined: information literacy and media literacy.

Our paper takes up this emerging discourse and explores an interpretation resulting in a *Synergistic Literacy Model* to combat disinformation, especially in the context of right-wing extremism. This model suggests combining the two literacies CML and DL. While CML refers to the skill of critically reflecting on media content, digital ecosystems, and the impact of digital exposure and usage, DL describes the skill of being able to understand, interpret, use, and evaluate data and data-driven products. We show that the three central dimensions of CML – awareness, reflection, and empowerment – can be partially connected to the DL concept in order to create stronger synergies. DL, on the other hand, can provide meaningful extensions to CML with its elements of data thinking, data doing,

and data participation. Through these linkages in our model, we show how aspects of CML and DL can enrich each other and therefore create a helpful blueprint for the design of digital learning settings as well as digital platforms.

Concerning the implications for digital learning settings and digital platforms, we set forth several considerations for a concrete conceptualization of learning opportunities. First, we argue for a focus on the learners' perspective, where educational setups, methods, and materials are composed in such a way that learners with diverse backgrounds and requirements can take part equally. Second, we suggest that digital skills should be regarded as connected abilities to navigate the digital sphere and that educational settings must, therefore, be interlinked and focused on realistic cases and examples. As a third proposition, we assert that cultivating a digital mindset is essential for confidently navigating, identifying, and resolving digital challenges in everyday life.

In addition to the broader learning context, we contemplate the function of digital infrastructure in fostering literacy development, emphasizing the relevance of our new integrated model. Through an exploration of well-designed digital interventions, we illustrate how they can be methodically aligned with our model, aiding the identification of right-wing organizations and technological gaps. In assessing the constraints of technological solutions, we advocate for a critical evaluation of their efficacy and deliberation on individual versus structural accountabilities. Beyond the individual reasoning, we also argue that having the abilities and knowledge in the use of digital skills can enhance democracy. Detecting disinformation, engaging in discussions to counter disinformation, and collaborating with others are vital in safeguarding democratic integrity and might thereby become a potent countermeasure against right-wing extremism.

Finally, our model can serve as a foundation for assessing the efficacy of digital literacy interventions and inspiring the creation of new literacy combinations. We encourage the scientific community to seek additional synergies among theoretical concepts and frameworks for digital skills. Given the complex challenges we face, conceptual connections may generate fresh insights into prevailing (harmful) frameworks. It is worth noting that the subject matter discussed here represents just one focal point and that the model can be seamlessly adapted to other critical digital contexts, such as climate communication or cybercrime.

In this paper, we focus on the dynamics of right-wing extremism in a digitally connected world and assert that it is imperative to disrupt these dynamics to strengthen our democratic culture. This proposal extends to the academic community, urging continuous vigilance, identification of emerging threats, and exploration of future research directions. As Twidale et al. (2013) argue for the case of fostering DL, we need a "sociotechnical ecology where data, information, people and technology co-evolve" (p. 250). We believe

this remains true when extending the argument to fostering literacies against disinformation. Rather than adhering to one literacy curriculum or intervention to combat disinformation, we must develop multiple frameworks and approaches to fit the changing shape of our digitized society.

Part V

Conclusion

11 Conclusion and Outlook

As this dissertation draws to a close, it is essential to synthesize the diverse theoretical, empirical, and design-oriented contributions presented across the preceding chapters. The conclusion chapter revisits the central research questions, reflecting on how the findings collectively advance our understanding of online disinformation and the multifaceted challenges it poses to digital societies. By critically examining the conceptual frameworks, methodological approaches, and practical interventions developed throughout this work, the chapter aims to situate these insights within the broader landscape of Information Systems research and ongoing debates about trust, explainability, and resilience in the face of information manipulation. This final synthesis not only highlights the dissertation's key achievements but also delineates the boundaries of its explanatory reach, setting the stage for future inquiry and practical innovation.

11.1 Contributions

This dissertation addresses the phenomenon of online disinformation by responding to three overarching research questions. The individual chapters contribute distinct theoretical, empirical, and design-oriented insights to these questions. The following synthesis presents how each research question is addressed across the dissertation.

Research Question 1*: How can online disinformation be characterized and differentiated based on conceptually grounded characteristics?*

This research question is addressed through a multi-dimensional approach that spans the analytical landscape review in Chapter 3, the development of a formal conceptual model in Chapter 4, and its operationalization and empirical validation in Chapter 5. Together, these chapters provide a comprehensive and theoretically grounded answer to the challenge of characterizing and differentiating online disinformation. Chapter 3 lays the groundwork by critically examining the current state of publicly funded research on false information and hate speech. Through a systematic mapping of German and European research projects, the chapter identifies significant gaps in the Information Systems discipline's engagement with the broader disinformation landscape. While the field is notably oriented towards technological solutions, particularly in the form of machine learning and digital tools, it lacks substantive contributions to theoretical, policy-oriented, and qualitative research. This limited perspective constrains a deeper understanding of the multifaceted nature of disinformation. The chapter argues for a broader conceptual lens that moves beyond purely technical paradigms and incorporates dimensions such as rhe-

torical strategy, semantic ambiguity, and the sociotechnical contexts in which disinformation operates. By identifying these limitations, Chapter 3 underscores the need for structured, theory-informed frameworks to more effectively characterize and differentiate disinformation; a need that is directly addressed in the subsequent chapters. Building on this foundation, Chapter 4 introduces TAXODIS, a structured, SKOS-based taxonomy designed to provide a systematic classification of disinformation. TAXODIS encapsulates key linguistic, semantic, and pragmatic features of manipulative content, including intent, veracity, rhetorical strategy, emotional framing, and target audience. As the first openly accessible taxonomy of this kind, TAXODIS offers a conceptually grounded vocabulary that enables the consistent annotation and differentiation of disinformation across research contexts. Its alignment with semantic web standards ensures interoperability and fosters reuse in both academic and applied settings. By formalizing the conceptual dimensions highlighted as lacking in Chapter 3, TAXODIS provides a shared analytical language that supports more nuanced and theory-driven approaches to disinformation detection. Chapter 5 operationalizes this taxonomy through the creation of the DeFaktS dataset, a large-scale, span-level annotated corpus of German-language political discourse on Twitter (now X). Unlike binary fact-checking datasets, DeFaktS applies the TAXODIS framework to capture the multifaceted nature of disinformation and polarized rhetoric. This granularity enables the training of more sophisticated machine learning classifiers capable of identifying subtle manipulative patterns in digital texts. The empirical evaluation of DeFaktS demonstrates the practical applicability of conceptually grounded features in real-world detection tasks. In doing so, it closes the loop between theoretical modeling and computational implementation.

In sum, Chapters 3, 4, and 5 collectively offer a robust answer to Research Question 1. Chapter 3 frames the conceptual problem space and identifies unmet needs in current research; Chapter 4 translates these needs into a formal, interoperable taxonomy; and Chapter 5 validates the taxonomy through practical application and empirical analysis. This integrated approach not only advances theoretical understanding of disinformation but also provides the tools and data necessary for its differentiated detection in computational and non-computational contexts. By bridging conceptual, methodological, and empirical domains, the thesis contributes a comprehensive and actionable framework for the study of online disinformation.

***Research Question 2**: How does an XAI component for disinformation detection have to be designed to help users trust the algorithm's assessment?*

This research question is addressed through a Design Science Research (DSR) approach, which unfolds across a systematic literature review (Chapter 6), a qualitative user study (Chapter 7), and a quantitative online experiment (Chapter 8). Collectively, these chapters

provide a comprehensive and theoretically grounded response to the challenge of designing trustworthy XAI systems for disinformation detection by investigating the interrelations among explainability, user trust, perceived usability, and system comprehension. Chapter 6 establishes the theoretical foundation by systematically reviewing existing XAI literature, focusing on commonly employed explanation techniques, such as saliency maps, counterfactual explanations, and uncertainty indicators. Drawing on these insights, it proposes a series of low-fidelity mock-ups that synthesize the most promising explanatory patterns.

Building upon this foundation, Chapter 7 undertakes an in-depth qualitative investigation to examine how users engage with these prototypes. Through focus group discussions, the study uncovers nuanced use perceptions which help refine the design space to three interactive prototypes. Chapter 8 subjects these prototypes to rigorous empirical validation via a large-scale online experiment involving 344 participants. This quantitative phase evaluates the impact of different explanation formats on key outcome variables, including trust, usability, and understandability, and finds that transparency through explanation can indeed impair the overall user experience instead of improving it. Taken together, these chapters problematize the assumption that more transparency invariably leads to greater trust. From a theoretical standpoint, the research contributes a nuanced perspective that integrates cognitive load theory and user-centered design principles. From a practical angle, it culminates in a set of design guidelines recommending progressive disclosure, optional detail, and the integration of trust-enhancing mechanisms such as user feedback loops.

In conclusion, Chapters 6, 7, and 8 provide a coherent and evidence-based answer to Research Question 2. Chapter 6 establishes the foundation through a literature-driven design framework; Chapter 7 sharpens and contextualizes this framework through qualitative insight; and Chapter 8 validates it through scalable empirical testing. The resulting framework deepens our understanding of the interplay between explainability and trust and offers actionable design principles for the development of user-sensitive XAI components in disinformation detection systems.

Research Question 3*: How can the key challenges in detecting information manipulation be effectively addressed through practical tools and strategies?*

The third research question is addressed in Chapters 9 and 10, which focus on applied, user-centered responses to the evolving threat of information manipulation, especially pertaining to deepfakes and disinformation in the context of right-wing extremism. Chapter 9 presents an empirical study exploring practitioners' perspectives across various domains regarding the risks posed by deepfakes and their expectations for detection tools.

The findings emphasize a growing concern over the increasing sophistication and potential impact of deepfakes on public trust, political discourse, and institutional integrity. Participants highlight the necessity of detection tools that are user-friendly, transparent, and legally compliant, while also capable of handling multimodal content in context-aware ways. The preference for open-source or publicly governed systems over proprietary black-box solutions underscores the demand for trustworthiness not just in system functionality but also in system governance. These results directly inform design principles for developing practical tools that address real-world constraints and expectations. Chapter 10 complements this applied focus with a conceptual contribution in the form of the Synergistic Literacy Model. This model integrates critical media literacy (CML) and data literacy (DL) into a unified framework that supports both individual and societal resilience against disinformation. It argues for the development of digital learning environments that promote not only technical competence but also reflective and participatory engagement with digital content. By linking cognitive and critical literacies, the model addresses the structural and educational roots of vulnerability to manipulation, particularly in the context of right-wing extremist narratives. It further highlights the need for inclusive, realistic, and participatory educational settings that empower users to navigate, evaluate, and counter digital disinformation effectively.

Together, Chapters 9 and 10 advance Research Question 3 by providing empirically informed design requirements and a normative, literacy-based model for practical intervention. These contributions emphasize the importance of socio-technical alignment, participatory design, and cross-sector collaboration in developing resilient responses to the challenges of information manipulation in digital democracies. The dual focus on practitioner-informed tool development and a theoretically grounded literacy model offers a comprehensive response to the operational and educational dimensions of disinformation resilience. By bridging empirical insight with conceptual innovation, this line of research strengthens the foundation for sustainable and user-responsive strategies to counter information manipulation in an increasingly complex digital ecosystem.

In sum, the contributions presented in this dissertation collectively advance our understanding of online disinformation along conceptual, technical, and socio-educational dimensions. By addressing the three research questions through a multi-method and interdisciplinary lens, this work not only enriches the theoretical discourse within the Information Systems field but also provides actionable insights for practitioners, policymakers, and system designers. The findings underscore the importance of integrating conceptual clarity, user-centered system design, and digital literacies in the ongoing effort to mitigate disinformation and safeguard democratic integrity. These insights lay the foundation for future research and design endeavors, particularly those aimed at enhancing

the resilience, inclusiveness, and accountability of digital platforms in an increasingly complex information landscape.

11.2 Limitations and Discussion

Building on the cumulative insights and contributions outlined in the previous chapters, this chapter offers a critical synthesis of the study's key conceptual, methodological, and empirical boundaries. Beyond simply cataloguing limitations, the chapter aims to interrogate the foundational assumptions, interpretive frameworks, and design choices that have shaped the research trajectory. In doing so, it seeks to contextualize the findings within a broader epistemological landscape and articulate the contingent nature of the knowledge claims advanced throughout the dissertation. While earlier chapters have identified specific constraints tied to individual components of the research, this discussion extends beyond those localized reflections to consider more systemic and structural limitations, those that arise not only from methodological trade-offs or data availability but also from the inherent complexity of modeling trust in AI-driven disinformation detection systems. These considerations are essential for clarifying the boundaries of the study's explanatory reach and the scope of its normative claims.

At the same time, the chapter takes a discursive turn by exploring how these limitations illuminate deeper tensions, unresolved questions, and theoretical ambiguities in the field. Rather than positioning limitations as mere shortcomings, the discussion reframes them as productive constraints – points of friction that invite critical engagement with current assumptions about AI explainability, epistemic trust, and the evolving nature of digital deception. Through this lens, the chapter contributes to a more reflexive and layered understanding of the research, while delineating how its insights might inform, challenge, or refine ongoing scholarly and practical debates.

11.2.1 The Role of Trust in AI-Driven Disinformation Detection Systems

The construct of trust in AI, particularly within the domain of disinformation detection, is inherently complex and multidimensional. It is imperative for researchers to acknowledge that trust does not constitute a binary or static attribute but rather manifests as a dynamic, context-contingent phenomenon. Its formation and evolution are influenced by a constellation of factors, including system reliability, perceived transparency, and the overall user experience (Fulmer & Gelfand, 2013; Huang & Bashir, 2017). Empirical findings from this dissertation suggest that the interplay between transparency and trust is nuanced, thereby necessitating a more critical and layered conceptualization of trust within AI-mediated environments. Within such systems, trust operates as a mediating mechanism that influences the extent to which users are willing to rely on algorithmic

outputs (Cabiddu et al., 2022; Lee, 2018). In the context of disinformation detection, this mediating role becomes particularly salient, as users must continually assess whether to accept or reject content that the system has flagged. Crucially, trust should not be misconstrued as a monolithic or absolute metric. Instead, it must be understood as a calibrated equilibrium between confidence in the system's functional efficacy and a degree of epistemic vigilance that facilitates critical user engagement (Jalava, 2006; Ting et al., 2021; Yan & Holtmanns, 2008). The study presented in Chapter 8 underscores that comprehensive transparency is not necessarily a prerequisite for cultivating trust; on the contrary, excessive information disclosure may paradoxically diminish user trust by inducing cognitive overload or interpretive ambiguity. Hence, the objective is not the maximization of trust per se, but the cultivation of an optimal trust state that supports informed, yet cautious, interaction with the system (Wicks et al., 1999).

Trust is also inextricably linked to perceptions of system reliability and predictive validity (Chavaillaz et al., 2016; Desai et al., 2012). Particularly in disinformation detection tasks, outcome-based trust emerges as a critical dimension: users' confidence in the system tends to grow when its outputs consistently align with observable truths or expert consensus (Nourani et al., 2019). However, this relationship is inherently non-linear. Trust is not instantaneously conferred but rather accrues incrementally, contingent on the system's sustained and demonstrable performance across varied contexts (Schaefer et al., 2016). In this light, trust must be conceptualized as an emergent property of longitudinal system-user interactions, rather than as a static variable. Its development is governed by a feedback loop wherein correct predictions enhance trust, which in turn increases user reliance, so long as the system continues to meet performance expectations. Initial trust may be informed by extrinsic cues, such as institutional reputation or third-party endorsements, but its persistence is predicated on continued reliability and intelligible system behavior (Nilsson & Mattes, 2015). Consequently, the cultivation of trust should not aim for uniformity across user populations but rather support individualized pathways through which trust is incrementally constructed. However, it is important to acknowledge that the empirical studies presented in Part III of this dissertation did not explicitly account for the longitudinal nature of trust formation. The study design primarily captured users' immediate trust responses, without tracking how trust might evolve over time through extended interaction with the system. This constitutes a limitation, as trust in AI systems, particularly in high-stakes domains such as disinformation detection, is likely to develop gradually, influenced by accumulated experiences and iterative system evaluations. Future research would benefit from longitudinal methodologies that observe trust dynamics over extended periods, thereby yielding deeper insights into the temporal dimensions of trust calibration and maintenance.

Of particular concern is the phenomenon of blind trust – an uncritical deference to algorithmic authority that can have deleterious implications, especially within the high-stakes arena of disinformation detection (Schmitt et al., 2024). Such unquestioning acceptance of AI outputs risks perpetuating algorithmic biases, legitimizing false positives or negatives, and amplifying the societal harms associated with disinformation (Bansal et al., 2021; Grissinger, 2019). To counteract this, it is imperative that AI systems are designed not only to foster trust but also to enable robust user interrogation of system decisions. Critical transparency, the strategic communication of rationale and uncertainty, emerges here as a vital design principle. This entails furnishing users with sufficient explanatory scaffolding to evaluate outputs meaningfully, without overwhelming them with technical minutiae (Bansal et al., 2021). In this framework, trust is not the absence of doubt but the presence of a well-calibrated disposition toward engaged scrutiny (Norris, 2022). From an ethical standpoint, the responsibility for cultivating trustworthy AI rests with system developers and researchers, who must eschew manipulative design practices aimed at artificially inflating user trust. Oversimplified explanations or interface features designed to obscure system fallibility may yield short-term compliance but ultimately undermine user autonomy and informed consent (Bennett et al., 2023; Friedman, 1998). Ethical system design must therefore prioritize honest disclosure regarding the system's limitations and the probabilistic nature of its outputs. Encouraging a stance of cautious optimism, wherein users are invited to consider AI recommendations without surrendering critical agency, supports more resilient and ethically sound trust relationships (Spector & Ma, 2019). Achieving such resilient trust necessitates a long-term, sustainability-oriented approach to AI deployment. Trust should be reconceptualized not as a terminal state but as an iterative process contingent on continuous system performance and user learning (Siau & Wang, 2018). The cultivation of such trust hinges on several interrelated pillars: empirical reliability, epistemically appropriate transparency, the facilitation of critical engagement, and adherence to ethical design principles. Collectively, these dimensions form the foundation for a trust architecture that resists both undue skepticism and naive acceptance.

In conclusion, the role of trust in AI-driven disinformation detection is not to be construed in terms of maximalist objectives, but rather as the construction of a dynamic, context-sensitive relationship between users and technology. Trust must remain flexible, critically informed, and responsive to both system performance and user cognition. By emphasizing healthy skepticism over blind acceptance, AI systems can be leveraged not only as tools of computational efficiency but also as catalysts for more discerning and autonomous user engagement. In this light, the aims of critical media literacy must also evolve: if such literacy is to empower individuals as critical consumers and producers of media, it should likewise encompass the capacity to engage with AI-powered tools that moderate digital content. This includes not only the use of such systems but the development of

skills to interrogate their classifications, evaluate their assumptions, and critically assess their outputs. Integrating these competencies into the framework of digital literacies can foster a more reflective and informed public capable of navigating the epistemic uncertainties of algorithmically governed information environments. Eventually, this approach holds the potential to enhance both the efficacy and the integrity of AI applications in sensitive and socially consequential domains.

11.2.2 Predominantly Unimodal Focus in a Rapidly Multimodal Disinformation Landscape

While this dissertation advances the conceptualization, detection, and mitigation of digital disinformation with a focus on text-based content, it must be acknowledged that such a unimodal approach constitutes a methodological limitation. The vast majority of chapters in this dissertation center on the textual modality, examining linguistic patterns, conceptual taxonomies, and XAI tools tailored for text analysis. Although Chapter 9 extends this scope by addressing the growing phenomenon of deepfakes and analyzing the technical and organizational requirements for multimodal detection systems, this remains an exception rather than the rule. This emphasis reflects both pragmatic and epistemological decisions: textual data is comparatively more accessible, structured, and conducive to current explainability frameworks (Fankhauser et al., 2014; Ford et al., 2016; Xu et al., 2024). Moreover, much of the disinformation circulating in early digital environments was predominantly text-based (Alam et al., 2022), making such a focus historically appropriate. However, the current and emerging information ecosystem is increasingly defined by multimodal content, where text, image, audio, and video are fused in complex ways to deceive, manipulate, and emotionally engage users (Hameleers et al., 2020; Qi et al., 2021; Tanwar & Sharma, 2021).

The shift toward multimodality is not merely a trend but a structural transformation of the digital public sphere. Advances in generative AI – particularly models capable of producing synthetic faces, voices, and full-motion videos – are fundamentally altering how disinformation is created, distributed, and perceived (Bontcheva et al., 2024; Mirsky & Lee, 2021). Deepfakes and other multimodal fabrications present unique epistemic challenges: they exploit visual and auditory trust heuristics more powerfully than text alone, often bypassing traditional critical evaluation processes (Kietzmann et al., 2020). As disinformation becomes more immersive and sensorially rich, its psychological and emotional impact also intensifies, reinforcing ideological echo chambers and reducing users' ability to discern authenticity across modalities (Weikmann & Lecheler, 2023).

The relative absence of multimodal analysis throughout the thesis, therefore, represents a gap in fully capturing the contemporary dynamics of manipulated information. By focusing primarily on unimodal textual data, the research risks underestimating both the complexity and the reach of disinformation in its current forms. Multimodal disinformation not only demands new detection techniques but also necessitates different interpretive paradigms; ones that can integrate cross-modal coherence, temporal sequencing, and affective impact into the analytic framework. Nevertheless, the foundational insights developed in this dissertation provide a vital basis for such future work. The conceptual, technical, and methodological foundations developed here offer an important springboard for future multimodal work. The classification schemes, transparency principles, and human-centered design approaches established in this research may be extended to support detection and mitigation strategies across modalities. As such, while the scope of this dissertation is predominantly unimodal, its insights remain adaptable to the more complex, sensorily rich forms of disinformation that are already reshaping the digital landscape. A deeper engagement with multi-modal disinformation, both in terms of content analysis and detection systems, represents a critical next step in this research trajectory.

11.2.3 The Blurring Boundary Between AI-Generated and Human-Produced Content

A further challenge, closely aligned with the concerns raised in the World Economic Forum's Global Risks Report (2024), is the increasing indistinguishability between AI-generated and human-produced content. This issue represents not only a technical hurdle but also an epistemological dilemma: as the fidelity of synthetic content approaches and sometimes surpasses that of authentic human communication, the capacity to identify and assess manipulated information becomes fundamentally destabilized (Rana et al., 2022). This dissertation acknowledges the growing prevalence of generative AI in disinformation ecosystems, particularly through discussions of deepfakes and larger language models. However, most detection efforts explored here still rest on the assumption that manipulated content, even when sophisticated, retains some identifiable anomalies – semantic, structural, or contextual – that can be flagged by human or algorithmic systems. Yet this assumption is increasingly tenuous (Somoray & Miller, 2023). As noted by the World Economic Forum, detection mechanisms are struggling to keep pace with the sophistication of generative models, and the disparity in funding between foundational AI technologies and the tools designed to detect their misuse further exacerbates this gap.

Moreover, the epistemic problem is not merely one of accuracy but of visibility and interpretation. Even when synthetic content is labeled, via watermarks, metadata, or platform warnings, digital labels may not persist when content is shared across platforms or stripped out during download (Hameleers, 2023; Krafft & Donovan, 2020). In the course

of this, the emotive power of AI-generated content can override rational processing, allowing fabricated videos or narratives to shape opinion and behavior even when explicitly marked as AI-generated (Bakir et al., 2024; Ienca, 2023). This raises deeper questions about the boundary between malignant and benign uses of generative AI. A political campaign video created with synthetic voice and imagery may comply with legal standards, yet still manipulate emotional responses and reinforce ideological bias (Haq et al., 2024; Reveland, 2025). In such cases, the ethical stakes lie not only in identifying content as AI-generated but in evaluating its intent, impact, and context – tasks that add layers of complexity to binary classification. The methodological frameworks in this dissertation, while robust in their treatment of textual content and explainability, are not yet fully equipped to address this ambiguity. The underlying models are designed to detect deception or manipulation in fairly well-defined formats, but they do not yet engage with the more ambiguous and indeterminate terrain where realism, intention, and audience interpretation intersect.

In light of this, advancing research must go beyond detection to grapple with the interpretability of authenticity itself. Future work will need to explore how systems can better communicate uncertainty, provenance, and intent, and how users interpret such signals in high-stakes environments. It will also require a stronger interdisciplinary focus on the effective and cognitive dimensions of how synthetic content is perceived and acted upon. This limitation underscores an urgent need not only for technical innovation but for deeper societal conversations about the epistemic authority of AI-generated content in democratic discourse.

11.2.4 The Epistemic Instability of Static Annotation Schemes in Generative Contexts

Building upon these concerns around multimodality and the ontological ambiguity introduced by generative systems, a further methodological limitation of this dissertation lies in its reliance on fixed taxonomies and static annotation guidelines for the fine-grained analysis of disinformation. Central to the TAXODIS framework (Chapter 4) and its application within the DeFaktS dataset (Chapter 5) is the assumption that deceptive communication can be systematically decomposed into a set of recurring linguistic cues, such as emotional appeals, logical inconsistencies, or semantic manipulations, that can be reliably identified and labeled by human annotators. While this assumption holds considerable value in structuring analytical insight and supporting transparency in disinformation detection, it becomes increasingly unstable in an environment where the very boundaries of deception are algorithmically reconfigurable (Knight, 2021; Lyons, 2020).

The proliferation of generative AI, particularly large language models capable of producing highly coherent, grammatically impeccable, and contextually plausible text, complicates the viability of static labeling schemes (Schuster et al., 2020). The outputs of such systems often do not exhibit the kinds of overt disinformation cues codified in TAXODIS. Instead, they may rely on more subtle, emergent, or hybrid forms of manipulation: confident misstatements framed with epistemic modesty, rhetorical devices borrowed from legitimate discourse communities, or context-sensitive omissions that reshape narrative meaning without introducing explicit falsehoods (Shoaib et al., 2023). These characteristics resist taxonomic capture, exposing the limitations of cue-based annotation when deception is latent, distributed, or semiotic rather than explicitly linguistic. Moreover, the epistemic volatility introduced by generative systems is not merely a matter of sophistication but of adaptability. Unlike human-crafted disinformation, which tends to follow historically or ideologically entrenched patterns, AI-generated content can mutate in response to detection frameworks, introducing a form of adversarial co-evolution (Beyer, 2023; Shoaib et al., 2023; World Economic Forum, 2024). This dynamic makes any fixed taxonomy perishable by design: what constitutes a salient deception cue today may become obsolete tomorrow, repurposed or masked by the next generation of generative adversaries.

This challenge is further compounded by the epistemological load placed on annotators. Human annotators working with TAXODIS and DeFaktS are tasked with identifying deception at a granular level. Yet, as generative content becomes more ambiguous and less tethered to conventional forms of manipulation (H. Zhao et al., 2021), the cognitive and interpretive burden on annotators intensifies. Without the support of adaptive frameworks or machine-in-the-loop guidance, annotation risks devolving into an exercise in subjective inference, eroding inter-annotator reliability, and diminishing the reproducibility of results. From a methodological standpoint, this raises a critical tension between the desire for explainable, interpretable classification schemes, which taxonomies like TAXODIS facilitate, and the increasingly indeterminate, fluid nature of deceptive content in the generative era. If the goal is to maintain analytic clarity and system transparency, the framework must become more dynamic. Static taxonomies must evolve into adaptive infrastructures capable of incorporating novel cues, contextual shifts, and annotator feedback in real time. Such an evolution would require integrating techniques from unsupervised clustering (Hosseinimotlagh & Papalexakis, 2018), anomaly detection (Tam et al., 2019), and active learning (Sahan et al., 2021) – methods that allow the taxonomy to grow alongside the threat landscape it seeks to map.

Thus, while the TAXODIS framework and its instantiation in DeFaktS have laid essential groundwork for fine-grained disinformation analysis, it must be understood as foundational but incomplete. Its utility lies not in their permanence but in their adaptability

(Nickerson et al., 2013); the degree to which they can inform the development of more responsive, reflexive, and epistemologically robust systems. As with the broader limitations discussed above, the future of disinformation detection will depend not only on technical sophistication but on methodological humility: the recognition that in an era of generative content, no classificatory scheme can remain static for long.

11.3 Propositions for Future Research

Building on the findings and limitations identified in this dissertation, this chapter outlines a series of targeted propositions for future research. These suggestions are designed to extend the theoretical frameworks, methodological tools, and practical implementations introduced throughout this study. In particular, this chapter highlights key gaps in our current understanding of AI-driven disinformation detection systems and proposes new directions that can address the evolving complexity of this domain.

11.3.1 Expanding into Multimodal Disinformation Detection

While this dissertation advances the conceptualization, detection, and mitigation of digital disinformation with a focus on text-based content, it must be acknowledged that such a unimodal approach constitutes a methodological limitation. Most chapters center on the textual modality, examining linguistic patterns, conceptual taxonomies, and XAI tools tailored for text analysis. Although Chapter 9 extends this scope by exploring the rise of deepfakes and outlining technical and organizational requirements for multimodal detection systems, this remains the exception rather than the rule. These emphases reflect pragmatic considerations: textual data are comparatively more accessible, structured, and compatible with current explainability frameworks (Fankhauser et al., 2014; Ford et al., 2016; Xu et al., 2024), and much early online disinformation was indeed text-dominant (Alam et al., 2022). Yet, the contemporary information ecosystem is now defined by multimodality, where text, image, audio, and video are fused in increasingly sophisticated ways to deceive, manipulate, and emotionally engage users (Hameleers, 2023; Qi et al., 2021; Tanwar & Sharma, 2021). These multimodal fabrications exploit visual and auditory heuristics far more powerfully than text alone and often bypass traditional critical-evaluation processes (Kietzmann et al., 2020).

Future research should therefore build on the conceptual, technical, and methodological foundations developed in this dissertation to rigorously address the challenges of multimodal disinformation. One important direction involves the development of cross-modal coherence models that can assess semantic and temporal consistency across modalities; for example, detecting lip-sync mismatches between audio and video, or evaluating the

alignment between captions and images. Such architectures can help surface discrepancies that are likely to be missed by unimodal detection approaches. In parallel, explainability techniques must evolve to accommodate multimodal inputs. Extending current XAI methods to mixed-media content might involve pairing salience heatmaps, attention rollouts, or timeline overlays with textual rationales, thereby enabling users to interrogate why a system has flagged a specific frame, track, or phrase as deceptive. These developments would benefit not only expert users but also lay audiences who may require intuitive visual explanations to build trust in detection outcomes. Equally important are human-centered usability studies that examine how different user groups engage with AI-generated explanations when interacting with multimodal content. Such research should account for varying levels of media literacy, including the needs of visual learners and lower-literacy populations, who may interpret and scrutinize multimedia information in distinct ways. This is especially relevant given that sensory-rich disinformation often bypasses critical scrutiny more effectively than text, leveraging affective cues to manipulate perception. Finally, future work should consider how the DeFaktS architecture can be adapted to support multimodal pipelines. This could involve integrating joint embedding spaces or late-fusion ensembles that allow for unified analysis across text, imagery, and audiovisual materials. The goal would be to create a cohesive dashboard in which fact-checkers and end-users alike can assess heterogeneous media inputs side-by-side, with consistent standards of transparency and interpretability. By pursuing this research agenda, future scholarship can more fully capture the breadth and sophistication of contemporary disinformation tactics while extending the transparency principles, classification schemes, and human-centered design ethos established in this dissertation. A deeper engagement with multimodal content is thus not merely a promising direction but a necessary one for sustaining users' discernment and safeguarding democratic discourse in an increasingly immersive media environment.

11.3.2 Developing Dynamic and Reflexive Annotation Frameworks

The TAXODIS taxonomy and the DeFaktS annotation guidelines introduced in this dissertation constitute an important step toward the structured analysis of disinformation. Yet, as emphasized in Section 1.2.4, their fixed, cue-based design is increasingly vulnerable in a media environment shaped by generative AI. Large language and image models can now craft persuasive content that evades the overt linguistic or psychological markers codified in TAXODIS, continually mutating in response to detection efforts and thereby rendering any static label set perishable (Knight, 2021; Lyons, 2020; Schuster et al., 2020). In such a context, the epistemic burden on annotators rises sharply: without adaptive support, fine-grained labelling risks devolving into subjective inference, eroding inter-annotator reliability and diminishing reproducibility (Zhao et al., 2021).

To address this instability, future work should reconceptualize annotation as a dynamic, reflexive process rather than a one-off coding exercise. Adaptive annotation pipelines can be designed to evolve in real-time, integrating techniques such as active learning to surface ambiguous or novel instances for rapid human adjudication (Sahan et al., 2021). Community-driven revision mechanisms, akin to version-controlled knowledge bases, would allow the taxonomy itself to grow alongside emerging threat patterns, capturing latent or hybrid manipulations that elude existing labels. Hybrid frameworks that combine supervised cues with unsupervised clustering, anomaly detection, and prompt-based zero-shot methods (Hosseinimotlagh & Papalexakis, 2018; Tam et al., 2019) can further expand coverage, automatically flagging content that diverges from known patterns. Crucially, the annotation interface must foreground reflexivity: annotators should be able to contest labels, propose new categories, and document uncertainty. Such metadata, when fed back into the learning loop, can calibrate model confidence and guide the prioritization of future annotation tasks. In turn, the DeFaktS system can leverage these dynamic signals to maintain a continually updated understanding of deception strategies, ensuring that explanation modules reflect the most current threat landscape. By shifting from static to adaptive annotation infrastructures, researchers can preserve the transparency and interpretability benefits of taxonomic analysis while accommodating the fluid, adversarial nature of generative disinformation. The methodological humility advocated in the limitations chapter thus becomes operational: annotating frameworks are treated not as immutable artefacts but as living instruments that learn, adapt, and iterate in concert with the evolving information battlefield.

11.3.3 Investigating the Temporal Dynamics of Trust

While this dissertation emphasized the multifaceted and contingent nature of trust in AI-driven disinformation detection systems, a key methodological limitation lies in its focus on immediate user responses. The experimental design primarily captured snapshot assessments of system trustworthiness without accounting for the longitudinal processes through which trust is cultivated, challenged, or eroded over time. However, as prior research underscores, trust is not instantaneously conferred; rather, it emerges incrementally through repeated interactions, informed by users' accumulated experiences with system performance, error management, and transparency practices (Nilsson & Mattes, 2015; Schaefer et al., 2016). The empirical framework presented in Chapter 8 thus offers only a partial view, omitting the temporal dynamics that shape real-world trust formation in iterative, adaptive environments.

To address this gap, future research should adopt longitudinal methodologies capable of capturing how trust evolves across extended engagements. Diary studies, field deployments, and repeated-measures experiments may enable a more granular understanding of

how trust is recalibrated in response to ongoing system use, interface updates, or shifting error profiles (Desai et al., 2012; Fulmer & Gelfand, 2013). In particular, the concept of *trust calibration*, the degree to which user confidence aligns with actual system capabilities, requires sustained observation to determine whether users become overreliant or unduly skeptical over time (Lee, 2018; Ting et al., 2021). Furthermore, comparative analysis across distinct demographic or occupational groups may reveal divergent trust trajectories shaped by domain expertise, prior exposure to algorithmic systems, or differing epistemic expectations (Chavaillaz et al., 2016; Nourani et al., 2019). Given that this dissertation aimed to support the development of a broadly applicable disinformation detection system, future studies should consider the trust dynamics or more narrowly defined user populations. Adolescents, older adults, and professionals in high-stakes domains such as journalism, education, or public health may each exhibit unique forms of cognitive engagement, affective response, and epistemic vigilance. Tailoring DeFaktS to reflect these differential needs would enhance its inclusivity, foster more sustainable forms of trust, and mitigate the risks of both under-reliance and blind deference. Ultimately, understanding trust as an emergent and temporally situated property of human-AI interaction will be essential for designing systems that are not only functionally robust but also capable of sustaining trust and reliability in social interactions.

11.3.4 Addressing the Risks of Blind Trust and Algorithmic Deference

A critical challenge identified in this dissertation is the risk of blind trust – uncritical acceptance of AI outputs that may arise from perceived authority, interface polish, or algorithmic confidence cues. As discussed in Section 1.2.1, such trust can be epistemically corrosive, especially in disinformation contexts where system errors may lead to the misclassification of legitimate content or the uncritical endorsement of manipulated narratives (Bansal et al., 2021; Schmitt et al., 2024). Blind trust undermines the normative goal of AI-assisted information vetting: not merely to automate judgment, but to support users in exercising critical discernment within complex and adversarial media environments.

To counteract this phenomenon, future development of systems such as DeFaktS should integrate design strategies that foster *engaged skepticism*, a mode of interaction in which users remain attentive to system limitations while still benefiting from algorithmic support (Friedman, 1998; Norris, 2022). This entails moving beyond simplistic notions of transparency as a disclosure and instead operationalizing *critical transparency,* the selective and strategic communication of rationale, uncertainty, and possible alternatives (Bansal et al., 2021; D. Bennett et al., 2023). For example, interface features that highlight model confidence intervals, generate counterfactual explanations, or juxtapose competing

interpretations may help maintain user autonomy and epistemic vigilance without inducing cognitive overload. Educational interventions also play a key role in shaping the conditions under which trust is responsibly exercised. Integrating media literacy frameworks that include algorithmic literacy components may equip users with the conceptual tools necessary to interrogate system outputs, recognize fallibility, and contextualize algorithmic recommendations (Huang & Bashir, 2017; Spector & Ma, 2019). Tutorials, use-case simulations, and guided reflection modules could further encourage users to question rather than defer to AI decisions. Such interventions are essential in establishing a healthy equilibrium, one in which users are neither paralyzed by skepticism nor seduced by unwarranted trust. In high-stakes applications like disinformation detection, trust must not be maximized at all costs; it must be carefully calibrated, ethically grounded, and epistemically earned (Li et al., 2025).

11.3.5 Establishing Ethical and Normative Design Principles for Trust

As articulated in Section 11.2.1, trust in AI systems, particularly those tasked with identifying and interpreting disinformation, cannot be treated solely as a functional variable. It is also an ethical relation, shaped by sociotechnical infrastructures, institutional norms, and communicative practices. This dissertation emphasizes that trust must be understood not as an artifact of engineering but as a construct that is co-produced by system behavior, user interpretation, and the values embedded in design choices (Cabiddu et al., 2022; Wicks et al., 1999). In this light, cultivating trust requires a deliberate commitment to normative design principles that prioritize user agency, transparency, and public accountability.

Future work should thus focus on developing ethical frameworks that guide the construction of trustworthy AI systems from the ground up. This involves embedding co-design practices that engage marginalized or underrepresented groups – users whose perspectives, media experiences, and risk exposures may differ markedly from those of majority populations (Friedman, 1998). Involving these groups in the iterative refinement of systems like DeFaktS not only enhances inclusivity but also ensures that the resulting tools are sensitive to a broader range of informational harms and trust deficits. Additionally, standards for transparency must be reimagined to avoid manipulation of performative disclosure. Explanatory interfaces should provide actionable insights that facilitate user understanding without overpromising interpretability or concealing probabilistic uncertainty (D. Bennett et al., 2023; Nourani et al., 2019). Governance mechanisms, such as third-party audits and democratic oversight structures, are also essential for aligning trustworthiness with institutional legitimacy and civic responsibility (Siau & Wang, 2018). By foregrounding these ethical commitments, future systems can ensure that trust is not merely behavioral compliance but a reflection of reciprocal accountability between users,

developers, and the public sphere. Ultimately, the challenge is to move beyond metrics of short-term user satisfaction and toward a model of trust that is sustainable, reflexive, and ethically coherent. In the context of AI-mediated disinformation detection, such a model demands that technical excellence be coupled with principled governance and normative integrity. Only by embedding these values into the design and deployment of systems like DeFaktS can we hope to cultivate trust relationships that are both epistemically sound and democratically legitimate.

11.3.6 Grappling with the Ambiguity of AI-Generated versus Human Content

As AI-generated content increasingly mirrors the complexity and fluency of human expression, conventional markers of source credibility are being steadily eroded, complicating efforts to assess authenticity and intent. Section 1.2.3 emphasizes that the convergence between synthetic and human-authored media has introduced profound epistemological challenges, particularly in disinformation detection contexts (Rana et al., 2022; World Economic Forum, 2024). The assumption that manipulated content inevitably contains detectable anomalies, semantic, structural, or contextual, is no longer tenable given the rapid advances in generative model sophistication (Somoray & Miller, 2023). Even when AI-generated content is labeled with watermarks or platform disclosures, such cues can be stripped or ignored, and users often prioritize affective resonance over metadata when evaluating credibility (Hameleers, 2023; Krafft & Donovan, 2020). These developments highlight the limitations of binary classification frameworks and underscore the need for more nuanced approaches that incorporate contextual and interpretive dimensions (Bakir et al., 2024; Ienca, 2023).

Future research should therefore move beyond traditional classifications to develop interpretive frameworks that acknowledge uncertainty and nuance. First, studies must investigate how users construe authenticity when traditional source cues, such as voice timbre, writing style, and production quality, can be algorithmically replicated or invented. Methodologies such as mixed-methods perception studies or experimental designs could clarify the relative influence of surface realism versus contextual indicators (e.g., platform reputation, topical familiarity) on credibility judgments. In addition, researchers should examine how contextual framing, such as comment threads or platform affordances, modulates users' interpretations of content authenticity and intent. These ambient signals often shape trust judgments more powerfully than explicit labels, particularly in fast-moving or emotionally charged information environments (Jiao et al., 2022; Waddell, 2018). Finally, rather than promising definite authenticity verdicts – an increasingly untenable goal – researchers and developers of AI systems should aim to cultivate epistemic resilience: the capacity of users to navigate information environments marked by ambiguity

with critical poise. This entails interfaces that surface uncertainty without inducing paralysis, encourage cross-source triangulation, and facilitate reflection on intent and impact rather than mere factuality. By embedding such interpretive supports, detection tools can shift from gatekeeping truth claims to scaffolding discernment, helping users retain confidence and agency (Soßdorf et al., 2024) even when categorical answers are unavailable. In doing so, future scholarship can address the deeper stakes identified in this dissertation: sustaining democratic deliberation in a media ecosystem where the very notion of "authentic" content is perpetually in transition.

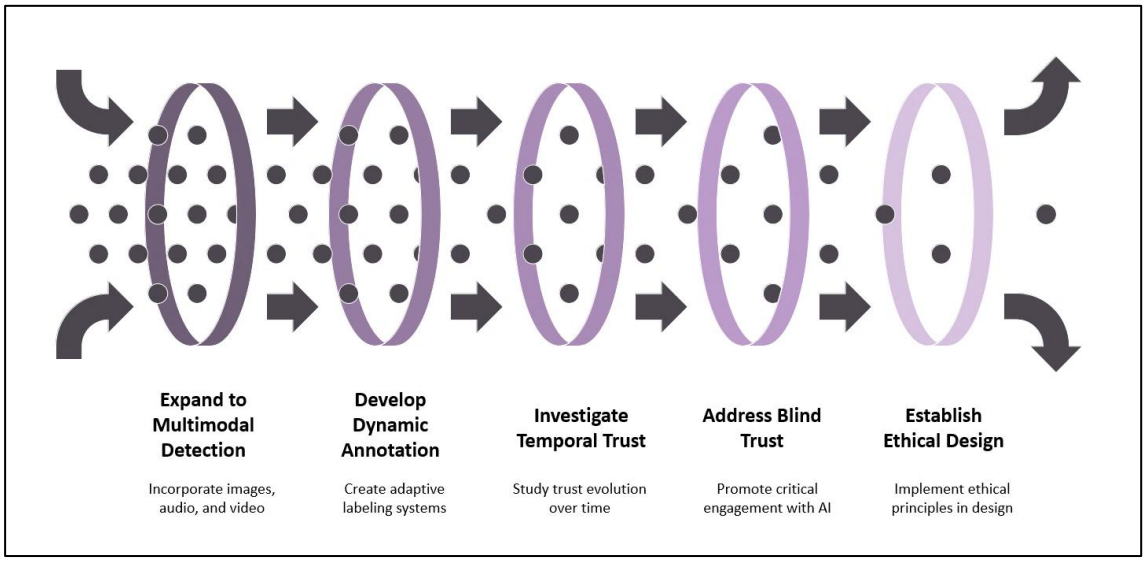

Figure 42. Propositions for future research.

This chapter has proposed several directions for future research that build upon the theoretical, technical, and methodological contributions of this dissertation. Figure 42 presents a summary of these directions and provides a suggestion of how they may build upon each other. The visual density of the dots reflects the relative complexity and scope of each research direction, with earlier stages requiring more comprehensive approaches to address numerous interconnected challenges, while later stages represent increasingly focused and refined methodologies built upon the more foundational work. By addressing key challenges such as the evolving nature of disinformation, the limitations of current detection frameworks, and the complexities of user interaction with AI systems, these propositions aim to strengthen the effectiveness and adaptability of AI-based approaches to disinformation detection. Future work may benefit from examining the long-term dynamics of system use, expanding into multimodal content, refining annotation practices, and ensuring that detection tools are transparent, ethical, and responsive to the fast-changing digital landscape. Through these avenues, researchers and practitioners can contribute to more robust, context-aware, and socially responsible systems capable of countering the

increasingly sophisticated forms of disinformation circulating in today's information ecosystems.

11.4 Concluding Remarks

This dissertation was developed during a time of significant political, technological, and societal flux. As the digital sphere continues to evolve in both form and function, so do the challenges and opportunities associated with democratic engagement. The studies presented here are situated against the backdrop of rising global concerns about disinformation, growing mistrust in institutions, and the increasing influence of platform economies on public discourse. These developments underscore the urgent need to rethink how participation, deliberation, and trust are structured in digital environments. In many ways, the tensions explored in this dissertation, between control and openness, automation and judgment, structure and freedom, mirror broader anxieties about the role of technology in democratic life – such as concerns over the erosion of civic agency and apprehensions regarding the concentration of power in technological infrastructures. The work's central focus on how system design can either enable or inhibit critical public engagement reflects a wider societal struggle to reconcile the promises of technological innovation with the need to uphold core democratic values. In this light, the design and evaluation of systems like DeFaktS are not merely technical or methodological contributions, but normative interventions into debates about the future of digital public spheres.

At the time of writing, the world continues to grapple with the consequences of war, displacement, and resurgent authoritarianism. These developments have rendered democracy more fragile – and more essential – than ever. While these forces manifest dramatically in the form of territorial aggression or coordinated disinformation campaigns, the subtler erosion of democratic culture often occurs through everyday processes: through the gradual normalization of information that prioritizes emotion over facticity, the algorithmic shaping of public attention, or the weakening of civic trust in digital spaces. In his book *Demokratie: Eine gefährdete Lebensform,* Till Van Rahden (2019) argues that democracy must be practiced as a lived experience, embedded in the rhythms and interactions of daily life. He calls for the cultivation of democratic experiential spaces, settings in which individuals are invited not only to vote or comment, but to participate meaningfully in shaping collective life. This dissertation aligns with and extends that vision by examining how such spaces might be reimagined within the architectures of digital platforms. It suggests that design is never neutral and that every interface, every algorithm, carries with it assumptions about who participates, how, and to what end.

As these dynamics continue to evolve, new technological developments further complicate the relationship between truth, perception, and participation. One particularly pressing development is the rise of fully synthetic content within online social networks. Media reporting on recent developments at TikTok and Meta (Westfall, 2025) has highlighted how platforms are not only contending with AI-generated content but are increasingly involved in its production, whether through synthetic accounts, generative tools, or automated media creation. This evolution complicates existing models of disinformation detection and challenges conventional notions of authenticity, trust, and participation in digital publics. Further, these developments shift the terrain in which truth claims are made and challenged. In this context, the contributions of this dissertation, particularly the emphasis on linguistic and psychological cues in the design of explainable systems like DeFaktS and modular frameworks like TAXODIS, acquire renewed relevance. While not originally conceived for purely synthetic content, these approaches provide a foundation for future research that aims to adapt detection and deliberation tools to emerging platform realities. This work positions itself as a starting point: a set of conceptual and practical tools that can be taken up, reinterpreted, and further developed by researchers, practitioners, and platform actors. Different environments will yield different types of knowledge and needs, and it is precisely this openness that gives the work its relevance. The hope is that these results invite further inquiry, and perhaps experimentation, into how systems supporting digital democratic life can be made more transparent, responsive, and inclusive.

More broadly, this research responds to the fading optimism that once accompanied the early digital age. The euphoric belief in the internet as an inherently democratizing force has given way to more sober assessments of its vulnerabilities. And yet, even within its climate of skepticism, there remains a critical space for creative, responsible, and reflexive innovation. This dissertation does not offer a solution to the structural challenges facing democracy, but it contributes to the ongoing effort to understand how digital technologies might be shaped in ways that serve democratic purposes rather than undermine them. As democratic institutions face pressure both from without and within, Information Systems research must take seriously the task of supporting their renewal, not only by diagnosing failure, but by imagining alternatives. By foregrounding the role of judgment, transparency, and human agency in the design of content moderation and disinformation detection systems, this dissertation offers one such alternative. It demonstrates that automated systems can, and should, be built to reflect social complexity rather than flatten it. The hope is that this work will serve as a foundation for further inquiry into how digital infrastructures can be made accountable, inclusive, and responsive to the public they claim to serve. At a time when the stakes for digital democracy have never been higher, such inquiry is not merely academic. It is, as Van Rahden reminds us, a contribution to the fragile and ongoing project of living democracy.

Appendix

Bibliography & Supplementary Information

Bibliography

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, *252*, 124260. https://doi.org/10.1016/j.eswa.2024.124260

Abd Rahim, N. H., & Basri, M. S. H. (2022). Malcov: Covid-19 fake news dataset in the malay language. *2022 International Visualization, Informatics and Technology Conference (IVIT)*, 239–244. https://ieeexplore.ieee.org/abstract/document/10033374/?casa_token=UnzgL0jI9lAAAAAA:DFbkFxjNOC26QM8pWMp5RhGEZDxYeqfry_8P4BjGSf4BJNmEmulzHxCtHMU0iYmZ3nEvIK6S0Q

Abdullah All Tanvir, Mahir, E. M., Akhter, S., & Huq, M. R. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1–5. https://doi.org/10.1109/ICSCC.2019.8843612

Abdullah, S. M., Cheruvu, A., Kanchi, S., Chung, T., Gao, P., Jadliwala, M., & Viswanath, B. (2024). An analysis of recent advances in deepfake image detection in an evolving threat landscape. *2024 IEEE Symposium on Security and Privacy (SP)*, 91–109. https://ieeexplore.ieee.org/abstract/document/10646853/?casa_token=FsmEyqljfh8AAAAA:CFjXZ4TuWSrNjKjcr4ms1d3l_PiPpmWCH8BoYlatrZ-VT0ULmC36pTtlCo-NHQu9hv63AB3q

Abonizio, H. Q., de Morais, J. I., Tavares, G. M., & Barbon Junior, S. (2020). Language-independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet*, *12*(5), 87.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138–52160.

Adam, M. T., Gregor, S., Hevner, A., & Morana, S. (2021). Design science research modes in human-computer interaction projects. *AIS Transactions on Human-Computer Interaction*, *13*(1), 1–11.

Agrawal, C., Pandey, A., & Goyal, S. (2021). A survey on role of machine learning and nlp in fake news detection on social media. *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 1–7. https://ieeexplore.ieee.org/abstract/document/9573875/?casa_token=rVfeXpQ1YPMAAAAA:PPjvxPkaTIWEa62Aak5XoVSixvO5vPffDZeMdthc6vPJbXeQ6

ztnWS4_eHPFs42NKDkuws26gg

Ahuja, N., & Kumar, S. (2023). Mul-FaD: Attention based detection of multiLingual fake news. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 2481–2491. https://doi.org/10.1007/s12652-022-04499-0

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Akert, R., Aronson, E., & Wilson, T. (2008). Sozialpsychologie. 6., aktual. *Aufl. München [u. a.]: Pearson Studium*.

Akhtar, Z. (2023). Deepfakes generation and detection: A short survey. *Journal of Imaging*, *9*(1), 18.

Akinyemi, B., Adewusi, O., & Oyebade, A. (2020). An Improved Classification Model for Fake News Detection in Social Media. *international journal of Information Technology and Computer Science (IJITCS)*, *12*(1), 34–43.

Akinyemi, J.-P., Chew, S., Geeling, S., Heuer, M., WANG, Q., Hassan, N., & Kude, T. (2024). Seeing Isn't Believing: AI Disclosure Labels and Sharing Behavior in the Era of Deepfakes. *ICIS 2024 Proceedings*. https://aisel.aisnet.org/icis2024/paperathon/paperathon/3

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., Shaar, S., Firooz, H., & Nakov, P. (2022). *A Survey on Multimodal Disinformation Detection* (arXiv:2103.12541). arXiv. https://doi.org/10.48550/arXiv.2103.12541

Alanazi, S., & Asif, S. (2024). Exploring deepfake technology: Creation, consequences and countermeasures. *Human-Intelligent Systems Integration*, *6*(1), 49–60. https://doi.org/10.1007/s42454-024-00054-8

Al-Asadi, M., & Tasdemir, S. (2022). *Using Artificial Intelligence Against the Phenomenon of Fake News: A Systematic Literature Review* (S. 39–54). https://doi.org/10.1007/978-3-030-90087-8_2

Alassad, M., Spann, B., & Agarwal, N. (2021). Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations. *Information Processing & Management*, *58*(1), 102385.

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, *97*(22), 3242–3250.

Alexander, J. M., & Smith, J. M. (2010). Disinformation: A Taxonomy. *Department of Computer & Information Science*, *Technical Reports (CIS)*.

Ali, S. B., Kechaou, Z., & Wali, A. (2022). Arabic fake news detection in social media Based on AraBERT. *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 214–220. https://ieeexplore.ieee.org/abstract/document/10101635/

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election.

Journal of economic perspectives, *31*(2), 211–236.

Allen, J. K., Griffin, R. A., & Mindrila, D. (2022). Discerning (Dis) information: Teacher perceptions of critical media literacy. *Journal of Media Literacy Education*, *14*(3), Article 3.

Almars, A. M. (2021). Deepfakes Detection Techniques Using Deep Learning: A Survey. *Journal of Computer and Communications*, *9*(5), Article 5. https://doi.org/10.4236/jcc.2021.95003

Alsaidi, H., & Etaiwi, W. (2022). Empirical Evaluation of Machine Learning Classification Algorithms for Detecting COVID-19 Fake News. *Int. J. Advance Soft Compu. Appl*, *14*(1).

Alt, R. (2021). Electronic Markets on digital platforms and AI. *Electronic Markets*, *31*(2), 233–241. https://doi.org/10.1007/s12525-021-00489-w

Amadeu Antonio Foundation. (2020). QAnon in Deutschland. *de: hate report*, *1*.

Andres, J., Wolf, C. T., Cabrero Barros, S., Oduor, E., Nair, R., Kjærum, A., Tharsgaard, A. B., & Madsen, B. S. (2020). Scenario-based XAI for Humanitarian Aid Forecasting. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3382903

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights*, *1*(2), 100052.

Appel, M. (2020). Die Psychologie des Postfaktischen–Einleitung und Überblick. In *Die Psychologie des Postfaktischen: Über Fake News,,Lügenpresse ", Clickbait & Co.* (S. 1–7). Springer.

Appel, M., & Doser, N. (2020). Fake News. In *Die Psychologie des Postfaktischen: Über Fake News,,Lügenpresse ", Clickbait & Co.* (S. 9–20). Springer.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Ashraf, S., Bezzaoui, I., Andone, I., Markowetz, A., Fegert, J., & Flek, L. (2024). DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4580–4591. https://aclanthology.org/2024.lrec-main.409/

Aufderheide, P. (2018). Media literacy: From a report of the national leadership conference on media literacy. In *Media literacy in the information age* (S. 79–

86). Routledge.
Aufenanger, S. (2001). Multimedia und Medienkompetenz—Forderungen an das Bildungssystem. *Jahrbuch Medienpädagogik 1*, 109–122.
Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.
Auger, J. H. (2014). Living With Robots: A Speculative Design Approach. *Journal of Human-Robot Interaction*, *3*(1), 20. https://doi.org/10.5898/JHRI.3.1.Auger
Avgerou, C. (2010). Discourses on ICT and development. *Information technologies and international development*, *6*(3), 1–18.
Azevedo, L., d'Aquin, M., Davis, B., & Zarrouk, M. (2021). *LUX (Linguistic aspects Under eXamination): Discourse Analysis for Automatic Fake News Classification*. 41–56.
Baacke, D. (1999). „Medienkompetenz ": Theoretisch erschließend und praktisch folgenreich. *medien und erziehung*, *43*(1), Article 1.
Bachmann, S., Putter, D., & Duczynski, G. (2023). Hybrid warfare and disinformation: A Ukraine war perspective. *Global Policy*, *14*(5), 858–869. https://doi.org/10.1111/1758-5899.13257
Bada, M., Sasse, A. M., & Nurse, J. R. C. (2019). *Cyber Security Awareness Campaigns: Why do they fail to change behaviour?* (arXiv:1901.02672). arXiv. https://doi.org/10.48550/arXiv.1901.02672
Badaskar, S., Agarwal, S., & Arora, S. (2008). *Identifying real or fake articles: Towards better language modeling*. Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.
Bailer, W., Thallinger, G., Backfried, G., & Thomas-Aniola, D. (2021). Challenges for Automatic Detection of Fake News Related to Migration. *2021 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 133–138. https://ieeexplore.ieee.org/abstract/document/9475929/?casa_token=DHAwdTiotPcAAAAA:ljGZ-HNfmo_zWKk3RQhU66sc3C5y_HSx_devcu8qzF-8ZH4NenZkKR3I06gz09umiyCS_0Kf7w
Bakir, V., Laffer, A., McStay, A., Miranda, D., & Urquhart, L. (2024). On manipulation by emotional AI: UK adults' views and governance implications. *Frontiers in Sociology*, *9*, 1339834.
Bąkowicz, K. (2019). Introduction to the Definition and Classification of the Fake News. *Media Studies*, *78*(3).
Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423–443.
Banas, J. A., & Rains, S. A. (2010). A Meta-Analysis of Research on Inoculation Theory. *Communication Monographs*, *77*(3), 281–311.

https://doi.org/10.1080/03637751003758193

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717

Barrutia-Barreto, I., Seminario Córdova, R., & Chero-Arana, B. (2022). *Fake News Detection in Internet Using Deep Learning: A Review* (S. 55–67). https://doi.org/10.1007/978-3-030-90087-8_3

Basil, V. R., & Turner, A. J. (1975). Iterative enhancement: A practical technique for software development. *IEEE Transactions on Software Engineering*, *4*, 390–396.

Bäumler, J., Kaufhold, M.-A., Voronin, G., & Reuter, C. (2024). Towards an Online Hate Speech Classification Scheme for German Law Enforcement and Reporting Centers: Insights from Research and Practice. . . *September*.

Bell, P. (2024, Juni 24). Public Trust in Government: 1958-2024. *Pew Research Center*. https://www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/

Ben Aissa, F., Hamdi, M., Zaied, M., & Mejdoub, M. (2024). An overview of GAN-DeepFakes detection: Proposal, improvement, and evaluation. *Multimedia Tools and Applications*, *83*(11), 32343–32365. https://doi.org/10.1007/s11042-023-16761-4

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, *11*(3), 369–386. https://doi.org/10.2307/248684

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, *6*(3), 4.

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). *Complexity and Information Systems Research in the Emerging Digital World* (SSRN Scholarly Paper 3539079). Social Science Research Network. https://papers.ssrn.com/abstract=3539079

Benkler, Y. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press. https://books.google.de/books?hl=de&lr=&id=MVRuDwAAQBAJ&oi=fnd&pg=PP1&dq=Benkler,+Y.,+Faris,+R.,+%26+Roberts,+H.+(2018).+Network+propaganda:+Manipulation,+disinformation,+and+Radicalization+in+American+politics.+New+York,+US:+Oxford+University+Press.&ots=W9gswoHCqg&sig=yoQee61ujLoOlFi5WFKCtvsyIgw

Bennet, L. W., & Livingston, S. (2018). The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions. *European Journal*

of Communication, *Vol. 33 (2)*, 122–139.

Bennett, D., Metatla, O., Roudaut, A., & Mekler, E. (2023). How does HCI Understand Human Autonomy and Agency? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3580651

Bennett, W. L., & Pfetsch, B. (2018). Rethinking Political Communication in a Time of Disrupted Public Spheres. *Journal of Communication*, *68*(2), 243–253. https://doi.org/10.1093/joc/jqx017

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*, 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Berger, C., Freihse, C., & Otto, M. zu S. (2024). *Effektiver Umgang mit Desinformation*. *Reinhard Mohn*(Upgrade Democracy).

Berinsky, A. J. (2017). Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science*, *47*(2), 241–262. https://doi.org/10.1017/S0007123415000186

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First monday*, *21*(11–7).

Beyer, J. N. (2023, Juni 10). *The race to detect AI can be won*. POLITICO. https://www.politico.eu/article/artificial-intelligence-ai-detection-race-can-be-won/

Bezzaoui, I. (2022). *Distinguishing Between Truth and Fake Using Explainable AI to Understand and Combat Online Disinformation*.

Bezzaoui, I., Fegert, J., & Weinhardt, C. (2022a). Distinguishing Between Truth and Fake: Using Explainable AI to Understand and Combat Online Disinformation. *The 16th International Conference on Digital Society*. 16th International Conference on Digital Society.

Bezzaoui, I., Fegert, J., & Weinhardt, C. (2022b). Truth or Fake? Developing a Taxonomical Framework for the Textual Detection of Online Disinformation. *International journal on advances in internet technology*, *15*(3/4), Article 3/4.

Bezzaoui, I., Nikolajevic, N., & Fegert, J. (2023). *Demokratiegefährdende Plattform-Mechanismen – Erkennen, Verstehen, Bekämpfen*. PolKomm.

Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). *Beyond data literacy: Reinventing community engagement and empowerment in the age of data*. https://dspace.mit.edu/bitstream/handle/1721.1/123471/Beyond%20Data%20Literacy%202015.pdf

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Binder, M., Heinrich, B., Hopf, M., & Schiller, A. (2022). Global reconstruction of

language models with linguistic rules – Explainable AI for online consumer reviews. *Electronic Markets*, *32*(4), 2123–2138. https://doi.org/10.1007/s12525-022-00612-5

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016). *„ 8 amazing secrets for getting more clicks“: Detecting clickbaits in news streams using article informality*. Thirtieth AAAI conference on artificial intelligence.

Blackman, R., & Ammanath, B. (2022). When—And Why—You Should Explain How Your AI Works. *Harvard Business Review*, *31*.

Boghardt, T. (2009). Operation Infektion. *Soviet Bloc Intelligence and Its AIDS Disinformation Campaign. Stud Intell*, *53*, 1–24.

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., Teyssou, D., Severio Nucci, F., Spangenberg, J., & Srba, I. (2024). *Generative AI and disinformation: Recent advances, challenges, and opportunities*. https://lirias.kuleuven.be/retrieve/758830

Bontcheva, K., & Posetti, J. (2020). *Balancing act: Countering digital disinformation while respecting freedom of expression: Broadband Commission research report on ‘Freedom of Expression and Addressing Disinformation on the Internet’—UNESCO Digital Library* (UNESCO Broadband Commission Report). https://unesdoc.unesco.org/ark:/48223/pf0000379015

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2021). Contextualising local explanations for non-expert users: An XAI pricing interface for insurance. *Joint Proceedings of the ACM IUI 2021 Workshops*, *2903*. https://hal.science/hal-03844389

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, *10*(1), 7.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*.

Bozarth, L., & Budak, C. (2020). *Toward a better performance evaluation framework for fake news classification*. *14*, 60–71.

Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Computational propaganda research project*. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. https://digitalcommons.unl.edu/scholcom/207/

Branley-Bell, D., Whitworth, R., & Coventry, L. (2020). User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. In M. Kurosu (Hrsg.), *Human-Computer Interaction. Human Values and Quality of Life* (S. 382–399). Springer International Publishing.

https://doi.org/10.1007/978-3-030-49065-2_27

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), 26. https://doi.org/10.1007/s12525-023-00644-5

Braun, J. A., & Eklund, J. L. (2019). Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism. *Digital Journalism*, *7*(1), 1–21. https://doi.org/10.1080/21670811.2018.1556314

Brave, R., Russo, F., & Wagemans, J. (2022). *Argument-Checking: A Critical Pedagogy Approach to Digital Literacy* (SSRN Scholarly Paper 4016734). Social Science Research Network. https://papers.ssrn.com/abstract=4016734

Brendel, A. B., Greve, M., Diederich, S., Bührke, J., & Kolbe, L. M. (2020). You are an Idiot!-How Conversational Agent Communication Patterns Influence Frustration and Harassment. *AMCIS*. https://core.ac.uk/download/pdf/326836248.pdf

Brennen, A. (2020). What Do People Really Want When They Say They Want „Explainable AI?“ We Asked 60 Stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3383047

Brinda, T., Brüggen, N., Diethelm, I., Knaus, T., Kommer, S., Kopf, C., Missomelius, P., Leschke, R., Tilemann, F., & Weich, A. (2020). *Frankfurt-Dreieck zur Bildung in der digital vernetzten Welt. Ein interdisziplinäres Modell*.

Bronstein, M., Pennycook, G., Bear, A., & Cannon, T. (2018). Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. *Journal of Applied Research in Memory and Cognition*, *8*. https://doi.org/10.1016/j.jarmac.2018.09.005

Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). Truth, lies, and automation. *Center for Security and Emerging technology*, *1*(1), 2.

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Bussone, A., Stumpf, S., & O’Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, *40*(5), 685–706. https://doi.org/10.1016/j.emj.2022.06.001

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262.

https://doi.org/10.1145/3301275.3302289

Cai, Z., Ghosh, S., Dhall, A., Gedeon, T., Stefanov, K., & Hayat, M. (2023). *Glitch in the matrix*: A large scale benchmark for content driven audio–visual forgery detection and localization. *Computer Vision and Image Understanding*, *236*, 103818. https://doi.org/10.1016/j.cviu.2023.103818

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, *9*(2), Article 2.

Carr, N. (2020). *The shallows: What the Internet is doing to our brains*. WW Norton & Company. https://books.google.de/books?hl=de&lr=&id=-HuqDwAAQBAJ&oi=fnd&pg=PP6&dq=the+shallows+nicholas+carr&ots=1N7NA8ztTg&sig=GJNSaeKRbIqkZk98Gs5kPLZ0uYA

Carrella, F., Miani, A., & Lewandowsky, S. (2023). IRMA: The 335-million-word Italian coRpus for studying MisinformAtion. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, *2023*, 2339. https://pmc.ncbi.nlm.nih.gov/articles/PMC7615326/

Carretero, S., Vuorikari, R., & Punie, Y. (2017). *DigComp 2.1: The digital competence framework for citizens*.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*, 20–23.

Catering, C. (2018). *Internet Research Agency Indictment*. United States Department of Justice. https://www.justice.gov/d9/fieldable-panel-panes/basic-panes/attachments/2018/02/16/internet_research_agency_indictment.pdf

Cellan-Jones, R. (2021, Januar 8). *Tech Tent: Did social media inspire Congress riot?* https://www.bbc.com/news/technology-55592752

Center for Countering Digital Hate. (2024). *Rated Not Helpful: How X's Community Notes system falls short on misleading election claims*. Center for Countering Digital Hate | CCDH. https://counterhate.com/research/rated-not-helpful-x-community-notes/

Chaka, C. (2022). Digital marginalization, data marginalization, and algorithmic exclusions: A critical southern decolonial approach to datafication, algorithms, and digital citizenship from the Souths. *Journal of E-Learning and Knowledge Society*, *18*(3), 83–95. https://doi.org/10.20368/1971-8829/1135678

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, *52*, 333–342.

Chen, B., & Tan, S. (2021). FeatureTransfer: Unsupervised Domain Adaptation for Cross-Domain Deepfake Detection. *Security and Communication Networks*, *2021*. https://doi.org/10.1155/2021/9942754

Cheng, F., Ming, Y., & Qu, H. (2020). Dece: Decision explorer with counterfactual

explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1438–1447.

Chersoni, E., Santus, E., Huang, C.-R., & Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, *47*(3), 663–698.

Chien, S.-Y., Yang, C.-J., & Yu, F. (2022). XFlag: Explainable Fake News Detection Model on Social Media. *International Journal of Human–Computer Interaction*, *38*(18–20), 1808–1827. https://doi.org/10.1080/10447318.2022.2062113

Choudhary, M., Jha, S., Saxena, D., & Singh, A. K. (2021). A review of fake news detection methods using machine learning. *2021 2nd international conference for emerging technology (INCET)*, 1–5. https://ieeexplore.ieee.org/abstract/document/9456299/

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational behavior and human decision processes*, *48*(1), 147–168.

Chromik, M. (2021). Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Hrsg.), *Human-Computer Interaction – INTERACT 2021* (Bd. 12933, S. 641–651). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_37

Chu, D., & Lee, A. Y. (2014). Media education initiatives by media organizations: The uses of media literacy in Hong Kong media. *Journalism & mass communication educator*, *69*(2), Article 2.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In M. P. Zanna (Hrsg.), *Advances in Experimental Social Psychology* (Bd. 24, S. 201–234). Academic Press. https://doi.org/10.1016/S0065-2601(08)60330-5

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Ciora, R. A., & Cioca, A. L. (2022). RoCoFake-A Romanian Covid-19 Fake News Dataset. *2022 E-Health and Bioengineering Conference (EHB)*, 1–4. https://ieeexplore.ieee.org/abstract/document/9991411/?casa_token=p5Ln-Gb2IwYAAAAA:uyQFSOQM95F5n6iVfxjiYi59tZUWyS5cTlwSTJC0sCE_DZkuxjmWoLznJ8yNvvDwTTNRnpR6bQ

Cirqueira, D., Nedbal, D., Helfert, M., & Bezbradica, M. (2020). Scenario-Based Requirements Elicitation for User-Centric Explainable AI: A Case in Fraud Detection. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Hrsg.), *Machine Learning and Knowledge Extraction* (Bd. 12279, S. 321–341). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-

8_18
Clarke, R. (2016). Big data, big risks. *Information Systems Journal*, *26*(1), Article 1.
Colomina, C., Margalef, H. S., Youngs, R., & Jones, K. (2021). The impact of disinformation on democratic processes and human rights in the world. *Brussels: European Parliament*, 1–19.
Compton, J. (2013). Inoculation theory. In *The SAGE handbook of persuasion: Developments in theory and practice, 2nd ed* (S. 220–236). Sage Publications, Inc.
Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(1), e1391.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. https://doi.org/10.48550/arXiv.1911.02116
Conway, M., Scrivens, R., & McNair, L. (2019). *Right-wing extremists' persistent online presence: History and contemporary trends*.
Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3
Das, D., Nishimura, Y., Vivek, R. P., Takeda, N., Fish, S. T., Plötz, T., & Chernova, S. (2023). Explainable Activity Recognition for Smart Home Systems. *ACM Transactions on Interactive Intelligent Systems*, *13*(2), 1–39. https://doi.org/10.1145/3561533
Davis, F. D., & Grani, A. (1989). *The Technology Acceptance Model*.
De Blasio, E., & Selva, D. (2021). Who Is Responsible for Disinformation? European Approaches to Social Platforms' Accountability in the Post-Truth Era. *American Behavioral Scientist*, *65*, 000276422198978. https://doi.org/10.1177/0002764221989784
Debruyne, C., Kearns, A., O'Neill, C., Colclough, M., Grehan, L., & O'Sullivan, D. (2021). *DALIDA: Data Literacy Discussion Workshops for Adults*. 23–25.
Deepak, P., Chakraborty, T., Long, C., & Kumar, G. S. (2021). *Data science for fake news: Surveys and perspectives*. https://pure.qub.ac.uk/en/publications/data-science-for-fake-news-surveys-and-perspectives
Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, *113*(3), 554–559.
Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the Seventh Annual*

ACM/IEEE International Conference on Human-Robot Interaction, 73–80. https://doi.org/10.1145/2157689.2157702

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186. https://aclanthology.org/N19-1423/?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-_m9bbH_7ECE1h3lZ3D61TYg52rKpifVNjL4fvJ85uqggrXsWDBTB7YooFLJeNXHWqhvOyC

Dhar, A., Mukherjee, H., Dash, N. S., & Roy, K. (2021). Text categorization: Past and present. *Artificial Intelligence Review*, *54*(4), 3007–3054.

Diaz Ruiz, C. (2023). Disinformation on digital media platforms: A market-shaping approach. *New Media & Society*, 14614448231207644. https://doi.org/10.1177/14614448231207644

Diepeveen, S., & Pinet, M. (2022). User perspectives on digital literacy as a response to misinformation. *Development Policy Review*, *40*(S2), e12671. https://doi.org/10.1111/dpr.12671

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*, *23*(1), Article 1.

D'Ignazio, C., & Bhargava, R. (2016). *DataBasic: Design principles, tools and activities for data literacy learners*.

DISARM. (2023). *DISARM Framework Explorer* [Dataset]. https://disarmframework.herokuapp.com

Dolensky, A., Laube, S., & Gorbacheva, E. (2015). How Can Information Systems Help to Make Policymaking Be More Sensitive to Global Long-Term Perspectives? In J. vom Brocke, A. Stein, S. Hofmann, & S. Tumbas (Hrsg.), *Grand Societal Challenges in Information Systems Research and Education: Ideas from the ERCIS Virtual Seminar Series* (S. 21–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-15027-7_3

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608). arXiv. http://arxiv.org/abs/1702.08608

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports*, *13*(1), 13429. https://doi.org/10.1038/s41598-023-39944-3

Dowse, A., & Bachmann, S. D. (2022). Information warfare: Methods to counter disinformation. *Defense & Security Analysis*, *38*(4), 453–469. https://doi.org/10.1080/14751798.2022.2117285

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in

climate change preference formation. *Nature Climate Change*, *9*(2), 111–119.

Durães, D., Freitas, P. M., & Novais, P. (2023). The relevance of deepfakes in the administration of criminal justice. In *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (S. 351–369). Springer International Publishing Cham. https://library.oapen.org/bitstream/handle/20.500.12657/86900/1/978-3-031-41264-6.pdf#page=355

Ecker, U. K. H., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, *49*(4), 631–644. https://doi.org/10.3758/s13421-020-01129-y

Eckert, S., Metzger-Riftkin, J., Kolhoff, S., & O'Shay-Wallace, S. (2021). A hyper differential counterpublic: Muslim social media users and Islamophobia during the 2016 US presidential election. *New Media & Society*, *23*(1), 78–98. https://doi.org/10.1177/1461444819892283

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3411764.3445188

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312787

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic perspectives*, *30*(3), 133–140.

Epstein, Z., Foppiani, N., Hilgard, S., Sharma, S., Glassman, E., & Rand, D. (2022). Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 183–193. https://doi.org/10.1609/icwsm.v16i1.19283

Equality Labs. (2019). *Facebook India Report 2019—Equality Labs*. http://archive.org/details/facebook-india-report-2019-equality-labs

European Commission. (2018). *A Multi-Dimensional Approach to Disinformation. Report of the Independent High Level Group on Fake News and Online Disinformation*. Publications Office of the European Union.

European Court of Auditors. (2020, Juni 11). *Sonderbericht 09/2021: Desinformation und ihre Auswirkungen auf die EU: Problem erkannt, aber nicht gebannt*. European Court of Auditors. http://www.eca.europa.eu/de/Pages/Report.aspx?did=58682&TermStoreId=8935807f-8495-4a93-a302-f4b76776d8ea&TermSetId=172d3e3c-ae5e-4a25-82c7-

8d37334fcfe2&TermId=5f6589a2-5a2e-4ae8-8cc6-e05bf544b71f

European Union. (2024). *Germany – EU country profile | European Union*. https://european-union.europa.eu/principles-countries-history/eu-countries/germany_en

Fabuyi, J., Olaniyi, O. O., Olateju, O., Aideyan, N. T., Selesi-Aina, O., & Olaniyi, F. G. (2024). Deepfake Regulations and Their Impact on Content Creation in the Entertainment Industry. *Archives of Current Research International*, *24*(12), 10–9734.

Fafalios, P., Iosifidis, V., Ntoutsi, E., & Dietze, S. (2018). Tweetskb: A public and large-scale rdf corpus of annotated tweets. *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, 177–190.

Fallis, D. (2015). What Is Disinformation? *Library Trends*, *63*(3), 401–426. https://doi.org/10.1353/lib.2015.0014

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2020). Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research*, *294*(1), 567–592. https://doi.org/10.1007/s10479-018-2818-y

Fankhauser, P., Knappen, J., & Teich, E. (2014). Exploring and visualizing variation in language resources. *LREC*, 4125–4128. https://www.researchgate.net/profile/Peter-Fankhauser/publication/265865192_Exploring_and_Visualizing_Variation_in_Language_Resources/links/542002330cf241a65a1af4f0/Exploring-and-Visualizing-Variation-in-Language-Resources.pdf

Faragó, L., Krekó, P., & Orosz, G. (2023). Hungarian, lazy, and biased: The role of analytic thinking and partisanship in fake news discernment on a Hungarian representative sample. *Scientific Reports*, *13*(1), 178. https://doi.org/10.1038/s41598-022-26724-8

Farhoudinia, B., Ozturkcan, S., & Kasap, N. (2024). Emotions unveiled: Detecting COVID-19 fake news on social media. *Humanities and Social Sciences Communications*, *11*(1), 640. https://doi.org/10.1057/s41599-024-03083-5

Fatima, S. A., Zafar, A., & Malik, K. M. (2023). YouFake: A Novel Multi-Modal Dataset for Fake News Classification. *2023 3rd International Conference on Artificial Intelligence (ICAI)*, 148–152. https://ieeexplore.ieee.org/abstract/document/10136667/?casa_token=y1K90EfKF24AAAAA:g-EruPp_s6cgKRyqbtzW9E9g7uy9mkz8r7nwsJ2xKw9gEib6DeS9ZVGlQ13EN-d-vpESBX5Pow

Fausset, R., & Bogel-Burroughs, N. (2021). *8 Dead in Atlanta Spa Shootings, With Fears of Anti-Asian Bias—The New York Times*.

https://www.nytimes.com/live/2021/03/17/us/shooting-atlanta-acworth

Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. *Soft Computing*, 1–9.

Ferguson, A. N., Franklin, M., & Lagnado, D. (2022). Explanations that backfire: Explainable artificial intelligence can cause information overload. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/3d97g0n3

Fernandez, A. C. T. (2019). *Computing the Linguistic-Based Cues of Credible and Not Credible News in the Philippines Towards Fake News Detection*.

Fernández Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, *57*(3), 64. https://doi.org/10.1007/s10462-023-10679-x

Ferreira, V. C., Kundu, S., & França, F. M. (2022). *Analysis of Fake News Classification for Insight into the Roles of Different Data Types*. 75–82.

Fetzer, J. (2004). Disinformation: The Use of False Information. *Minds and Machines*, *14*, 231–240. https://doi.org/10.1023/B:MIND.0000021683.28604.5b

Fiorenza, C. E., Kashyap, A. S., Chauhan, K., Mokaria, K., & Chandra, A. (2018). *Fake Product Review Monitoring and Removal for Genuine Online Reviews*. *8*(4).

Fischer, G., Lundin, J., & Lindberg, J. O. (2020). Rethinking and reinventing learning, education and collaboration in the digital age—From creating technologies to transforming cultures. *The International Journal of Information and Learning Technology*, *37*(5), Article 5.

Flores-Yeffal, N. Y., Vidales, G., & Martinez, G. (2019). #WakeUpAmerica, #IllegalsAreCriminals: The role of the cyber public sphere in the perpetuation of the Latino cyber-moral panic in the US. *Information, Communication & Society*, *22*(3), 402–419. https://doi.org/10.1080/1369118X.2017.1388428

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1007–1015.

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, *25/26*, 56. https://doi.org/10.2307/466240

Frauenberger, C. (2019). Entanglement HCI The Next Wave? *ACM Trans. Comput.-Hum. Interact.*, *27*(1), 2:1-2:27. https://doi.org/10.1145/3364998

Freelon, D., & Wells, C. (2020). Disinformation as Political Communication. *Political Communication*, *37 (2)*, Article 37 (2).

French, A., Storey, V. C., & Wallace, L. (2024). A typology of disinformation intentionality and impact. *Information Systems Journal*, *34*(4), 1324–1354. https://doi.org/10.1111/isj.12495

Friedman, B. (1998). User Autonomy: Who Should Control What and When? A CHI 96

workshop. *ACM SIGCHI Bulletin*, *30*(1), 26–29. https://doi.org/10.1145/280571.280583

Frischlich, L., & Humprecht, E. (2021). *Trust, democratic resilience, and the infodemic*. https://www.zora.uzh.ch/id/eprint/202660/

Fulmer, C. A., & Gelfand, M. J. (2013). How Do I Trust Thee? Dynamic Trust Patterns and Their Individual and Social Contextual Determinants. In K. Sycara, M. Gelfand, & A. Abbe (Hrsg.), *Models for Intercultural Collaboration and Negotiation* (S. 97–131). Springer Netherlands. https://doi.org/10.1007/978-94-007-5574-1_5

Galli, A., Masciari, E., Moscato, V., & Sperlí, G. (2022). A comprehensive Benchmark for fake news detection. *Journal of Intelligent Information Systems*, *59*(1), 237–261. https://doi.org/10.1007/s10844-021-00646-9

Gangopadhyay, S., Boland, K., Dessí, D., Dietze, S., Fafalios, P., Tchechmedjiev, A., Todorov, K., & Jabeen, H. (2023). Truth or dare: Investigating claims truthfulness with claimskg. *D2R2 2023-2nd International Workshop on Linked Data-driven Resilience Research*, *3401*.

Gangopadhyay, S., Schellhammer, S., Hafid, S., Dessi, D., Koß, C., Todorov, K., Dietze, S., & Jabeen, H. (2024). Investigating Characteristics, Biases and Evolution of Fact-Checked Claims on the Web. *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, 246–258.

Ganguin, S., & Sander, U. (2015). Zur Entwicklung von Medienkritik. In *Medienpädagogik-ein Überblick*.

Garaialde, D., Bowers, C. P., Pinder, C., Shah, P., Parashar, S., Clark, L., & Cowan, B. R. (2020). Quantifying the impact of making and breaking interface habits. *International Journal of Human-Computer Studies*, *142*, 102461.

Gasquet, M., Brechtel, D., Zloch, M., Tchechmedjiev, A., Boland, K., Fafalios, P., Dietze, S., & Todorov, K. (2019). Exploring Fact-checked Claims and their Descriptive Statistics. *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26-30, 2019*.

Gerlach, J., Hoppe, P., Jagels, S., Licker, L., & Breitner, M. H. (2022). Decision support for efficient XAI services—A morphological analysis, business model archetypes, and a decision tree. *Electronic Markets*, *32*(4), 2139–2158. https://doi.org/10.1007/s12525-022-00603-6

Germani, F., Spitale, G., & Biller-Andorno, N. (2024). The Dual Nature of AI in Information Dissemination: Ethical Considerations. *Jmir Ai*, *3*. https://doi.org/10.2196/53505

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining explanations: An overview of interpretability of machine learning*.

80–89.

Gkeredakis, M., Lifshitz-Assaf, H., & Barrett, M. (2021). Crisis as opportunity, disruption and exposure: Exploring emergent responses to crisis through digital technology. In *Information and Organization* (Bd. 31, Nummer 1, S. 100344). Elsevier. https://www.sciencedirect.com/science/article/pii/S1471772721000105?casa_token=sbbOnv9V-9cAAAAA:c9WPut58z73o_38vCpRxo7qqHI1VQC0KAj8L_H0ahu9TaarGs-IF3TMZZzoo9SuLEelSMRJIs1k

Glaser, S., Pfeiffer, T., & Yavuz, C. (2017). Hassimnetz: Frei–sozial–multimedial. Entwicklungslinien rechtsextremer Online-Präsenzen. *Erlebniswelt Rechtsextremismus. Modern–subversiv–hasserfüllt: Hintergründe und Methoden für die Praxis der Prävention*, *5*, 104–117.

Gnewuch, U., & Maedche, A. (2022). Toward a Method for Reviewing Software Artifacts from Practice. In A. Drechsler, A. Gerber, & A. Hevner (Hrsg.), *The Transdisciplinary Reach of Design Science Research* (S. 337–350). Springer International Publishing. https://doi.org/10.1007/978-3-031-06516-3_25

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

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

Golovchenko, Y., Hartmann, M., & Adler-Nissen, R. (2018). State, media and civil society in the information warfare over Ukraine: Citizen curators of digital disinformation. *International affairs*, *94*(5), 975–994.

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2024). A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, *7*(1), 587–623. https://doi.org/10.1007/s42001-024-00248-9

Górski, Ł., & Ramakrishna, S. (2021). Explainable artificial intelligence, lawyer's perspective. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 60–68. https://doi.org/10.1145/3462757.3466145

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794.

https://doi.org/10.1177/2053951719897945
Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics*, *43*, 11–22.
Granik, M., & Mesyura, V. (2017). *Fake news detection using naive Bayes classifier* (S. 903). https://doi.org/10.1109/UKRCON.2017.8100379
Gregor, S., & Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*, 337–356. https://doi.org/10.25300/MISQ/2013/37.2.01
Grissinger, M. (2019). Understanding human over-reliance on technology. *Pharmacy and Therapeutics*, *44*(6), 320.
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), e2110013119. https://doi.org/10.1073/pnas.2110013119
Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, *20*(9), 1389–1407.
Gruppi, M., Horne, B. D., & Adali, S. (2018). An exploration of unreliable news classification in Brazil and the US. *arXiv preprint arXiv:1806.02875*.
Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), Article 27.
Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, *10*.
Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature human behaviour*, *4*(5), 472–480.
Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850
Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.
Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). *Building trust in interactive machine learning via user contributed interpretable rules*. 537–548.
Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Combating fake news: Stakeholder interventions and potential solutions. *Ieee Access*, *10*, 78268–78289.
Gurzick, D., & Lutters, W. G. (2009). Towards a design theory for online communities. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 1.

https://doi.org/10.1145/1555619.1555634

Habermas, J. (1962). *The Structural Transformation of the Public Sphere*. MIT Press. https://mitpress.mit.edu/9780262581080/the-structural-transformation-of-the-public-sphere/

Habermas, J. (2022). *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik: Platz 1 der Sachbuchbestenliste der WELT* (3. Aufl.). Suhrkamp Verlag.

Hacker, K. L., & van Dijk, J. (2000). *Digital democracy: Issues of theory and practice*. Sage.

Haigh, M., Haigh, T., & Kozak, N. (2017). Stopping Fake News: The work practices of peer-to-peer counter propaganda. *Journalism Studies*, *19*, 1–26. https://doi.org/10.1080/1461670X.2017.1316681

Hakak, S., Khan, W. Z., Bhattacharya, S., Reddy, G. T., & Choo, K.-K. R. (2020). Propagation of Fake News on Social Media: Challenges and Opportunities. In S. Chellappan, K.-K. R. Choo, & N. Phan (Hrsg.), *Computational Data and Social Networks* (Bd. 12575, S. 345–353). Springer International Publishing. https://doi.org/10.1007/978-3-030-66046-8_28

Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and Beyond. *arXiv preprint arXiv:2301.08801*.

Hamed, S. Kh., Ab Aziz, M. J., & Yaakub, M. R. (2023). Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. *Sensors (Basel, Switzerland)*, *23*(4), 1748. https://doi.org/10.3390/s23041748

Hameleers, M. (2023). Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, *33*(1), 1–10. https://doi.org/10.1093/ct/qtac021

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

Hamm, L. (2020). *The Few Faces of Disinformation*. EU DisinfoLab. https://www.disinfo.eu/wp-content/uploads/2020/05/20200512_The-Few-Faces-of-Disinformation.pdf

Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets*, *33*(1), 17. https://doi.org/10.1007/s12525-023-00640-9

Hangloo, S., & Arora, B. (2022). Combating multimodal fake news on social media: Methods, datasets, and future perspective. *Multimedia Systems*, *28*(6), 2391–2422. https://doi.org/10.1007/s00530-022-00966-y

Hanley, H. W. A., & Durumeric, Z. (2023). *Machine-Made Media: Monitoring the*

Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites (arXiv:2305.09820; Version 1). arXiv. https://doi.org/10.48550/arXiv.2305.09820

Haq, E.-U., Zhu, Y., Hui, P., & Tyson, G. (2024). History in Making: Political Campaigns in the Era of Artificial Intelligence-Generated Content. *Companion Proceedings of the ACM Web Conference 2024*, 1115–1118. https://doi.org/10.1145/3589335.3652000

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Hassanian-Moghaddam, H., Zamani, N., Kolahi, A.-A., McDonald, R., & Hovda, K. E. (2020). Double trouble: Methanol outbreak in the wake of the COVID-19 pandemic in Iran—a cross-sectional assessment. *Critical Care*, *24*(1), 402. https://doi.org/10.1186/s13054-020-03140-w

Havrlant, L., & Kreinovich, V. (2017). A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, *46*(1), 27–36. https://doi.org/10.1080/03081079.2017.1291635

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery*, *14*(2), e1520. https://doi.org/10.1002/widm.1520

Henig, D., & Knight, D. M. (2023). Polycrisis: Prompts for an emerging worldview. *Anthropology Today*, *39*(2), 3–6. https://doi.org/10.1111/1467-8322.12793

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, *292*, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Hepenstal, S., Zhang, L., Kodagoda, N., & Wong, B. L. W. (2021). Developing Conversational Agents for Use in Criminal Investigations. *ACM Transactions on Interactive Intelligent Systems*, *11*(3–4), 1–35. https://doi.org/10.1145/3444369

Herm, L.-V., Steinbach, T., Wanner, J., & Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets*, *32*(4), 2185–2205. https://doi.org/10.1007/s12525-022-00606-3

Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75–105.

Higdon, N. (2020). What is fake news? A foundational question for developing effective critical news literacy education. *Democratic Communiqué*, *29*(1),

Article 1.

Hobbs, R. (2017). Measuring the digital and media literacy competencies of children and teens. In *Cognitive development in digital contexts* (S. 253–274). Elsevier.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 68:1-68:26. https://doi.org/10.1145/3392878

Hooghe, M. (2018). Trust and Elections. In E. M. Uslaner (Hrsg.), *The Oxford Handbook of Social and Political Trust* (S. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.17

Hopmann, D. N., Shehata, A., & Strömbäck, J. (2015). Contagious Media Effects: How Media Use and Exposure to Game-Framed News Influence Media Trust. *Mass Communication and Society*. https://www.tandfonline.com/doi/abs/10.1080/15205436.2015.1022190

Hornbostel, S. (2001). Third party funding of German universities. An indicator of research activity? *Scientometrics*, *50*(3), 523–537. https://doi.org/10.1023/A:1010566916697

Horne, B., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the international AAAI conference on web and social media*, *11*(1), 759–766.

Hosseinimotlagh, S., & Papalexakis, E. E. (2018). Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. https://www.cs.ucr.edu/~epapalex/papers/wsdm18-mis2-fakenews.pdf

Howard, P. N., & Hussain, M. M. (2013). *Democracy's fourth wave?: Digital media and the Arab Spring*. Oxford University Press. https://books.google.de/books?hl=de&lr=&id=ayHOyrmmT8kC&oi=fnd&pg=PP2&dq=Howard,+P.+N.,+%26+Hussain,+M.+M.+(2013).+Democracy%E2%80%99s+fourth+wave%3F:+Digital+media+and+the+Arab+spring.+Oxford:+Oxford+University+Press.&ots=mNxADu0kmd&sig=5C_tqtbguH16QUA4bVD_mLyrr7Y

Hoxtell, W. (2023). *Umgang mit Desinformation in Europa. Bertelsmann Stiftung*. https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/Umgang_mit_Desinformation_in_Europa._Herausforderungen_und_Gelegenheiten_fuer_zivilgesellschaftliche_Organisationen_und_Privatsektor.pdf

Huang, H.-Y., & Bashir, M. (2017). Personal Influences on Dynamic Trust Formation in Human-Agent Interaction. *Proceedings of the 5th International Conference on Human Agent Interaction*, 233–243.

https://doi.org/10.1145/3125739.3125749

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. B. Randolph, & G. Müller-Putz (Hrsg.), *Information Systems and Neuroscience* (S. 237–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-88900-5_27

Hug, T. (2011). Von der Medienkompetenz-Diskussion zu den «neuen Literalitäten» – Kritische Reflexionen in einer pluralen Diskurslandschaft. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *20*, 159–174. https://doi.org/10.21240/mpaed/20/2011.09.18.X

Hugger, K.-U. (2008). Medienkompetenz. In U. Sander, F. von Gross, & K.-U. Hugger (Hrsg.), *Handbuch Medienpädagogik* (S. 93–99). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91158-8_10

Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3347–3356. https://doi.org/10.1109/WACV48630.2021.00339

Ienca, M. (2023). On Artificial Intelligence and Manipulation. *Topoi*, *42*(3), 833–842. https://doi.org/10.1007/s11245-023-09940-3

IFLA. (2017). *How To Spot Fake News*. https://repository.ifla.org/handle/20.500.14598/167

Islam, A. K. M. N., Laato, S., Talukder, S., & Sutinen, E. (2020). Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, *159*, 120201. https://doi.org/10.1016/j.techfore.2020.120201

Jackson, S. J., & Kreiss, D. (2023). Recentering power: Conceptualizing counterpublics and defensive publics. *Communication Theory*, *33*(2–3), 102–111. https://doi.org/10.1093/ct/qtad004

Jacobs, S., Dawson, E. J., & Brashers, D. (1996). Information manipulation theory: A replication and assessment. *Communications Monographs*. https://doi.org/10.1080/03637759609376375

Jalava, J. (2006). *Trust as a decision: The problems and functions of trust in Luhmannian systems theory* [PhD Thesis, Helsingin yliopisto]. https://helda.helsinki.fi/bitstreams/9cd5c7ee-bb29-429f-bca6-7c23bd1eebf6/download

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.

Jeronimo, C. L. M., Marinho, L. B., Campelo, C. E., Veloso, A., & da Costa Melo, A.

S. (2019). *Fake news classification based on subjective language*. 15–24.

Jiao, Z., Li, C., & Chen, J. (2022). Knowledge platform affordances and knowledge collaboration performance: The mediating effect of user engagement. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1041767

Jörissen, B. (2011). «Medienbildung» – Begriffsverständnisse und Reichweiten. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *20*, 211–235. https://doi.org/10.21240/mpaed/20/2011.09.20.X

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, Л., & Liu, Y. (2021). *Countering Malicious DeepFakes: Survey, Battleground, and Horizon*. https://doi.org/10.48550/arxiv.2103.00218

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kachelmann, M., & Reiners, W. (2023). The European Union's Governance Approach to Tackling Disinformation–protection of democracy, foreign influence, and the quest for digital sovereignty. *L'Europe en Formation*, *396*(1), 11–36.

Kahan, D. M. (2015). The politically motivated reasoning paradigm. *Emerging Trends in Social & Behavioral Sciences, Forthcoming*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2703011

Kahan, D. M. (2017). *Misconceptions, misinformation, and the logic of identity-protective cognition*.

Kaiser, R. (2014). *Qualitative Experteninterviews. Konzeptionelle Grundlagen und praktische Durchführung.* Springer VS.

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A Systematic Review on Hate Speech among Children and Adolescents: Definitions, Prevalence, and Overlap with Related Phenomena. *Trauma, Violence & Abuse*, *24*(4), 2598–2615. https://doi.org/10.1177/15248380221108070

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A Systematic Literature Review on Disinformation: Toward a Uified Taxonomical Framework. *New Media & Society*, *23*(5), 1301–1326.

Kaplan, A., & Mazurek, G. (2018). Social Media. In *Handbook of Media Management and Economics* (2. Aufl.). Routledge.

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*, *57*(6), 159. https://doi.org/10.1007/s10462-024-10810-6

Kebede, A. Y., Ali, A. C., & Moges, M. A. (2022). Examining journalists organizational trust pursuant to predictive variables in the Ethiopian media industry: The case study of Amhara Media Corporation. *Cogent Social Sciences*, *8*(1), 2068271. https://doi.org/10.1080/23311886.2022.2068271

Keller, C. I., Freihse, C., & Berger, C. (2024). State actions against disinformation.

Research Series.

Kellner, D., & Share, J. (2005). Toward Critical Media Literacy: Core concepts, debates, organizations, and policy. *Discourse: studies in the cultural politics of education*, *26*(3), 369–386.

Kellner, D., & Share, J. (2007). Critical Media Literacy, Democracy, and the Reconstruction of Education. In D. Macedo & S. R. Steinberg, *Media Literacy: A Reader* (S. 3–23). Peter Lang Publishing.

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition*, *47*(1), 33–46. https://doi.org/10.3758/s13421-018-0848-y

Kerres, M. (2020). Bildung in der digitalen Welt: Über Wirkungsannahmen und die soziale Konstruktion des Digitalen. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 1–32.

Khaled, W. (2022). *Disinformation is the next big cybersecurity threat | Security Magazine*. https://www.securitymagazine.com/articles/97990-disinformation-is-the-next-big-cybersecurity-threat

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135–146.

Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., Heinrich, R., Baragwanath, R., & Raskutti, G. (2018). The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook. *Political Communication*, *35*(4), 515–541. https://doi.org/10.1080/10584609.2018.1476425

Kleemann, A. (2024). *Wie man erfolgreich Desinformation bekämpft. Reaktive Ansätze–Potentiale und Grenzen*. https://policycommons.net/artifacts/18031306/wie-man-erfolgreich-desinformation-bekampft/18930572/

Knight, W. (2021). AI Can write disinformation now—And dupe human readers. *Wired, May*.

Kohlberg, L. (1994). Stage and sequence: The cognitive-developmental approach to socialization. *The first half of the chapter is a revision of a paper prepared for the Social Science Research Council, Committee on Socialization and Social Structure, Conference on Moral Development, Arden House, Nov 1963.* https://psycnet.apa.org/record/1995-97182-001

Koltay, T. (2022). Speculations on a Union Between Media Literacy and Data Literacy. *Media Education Research Journal*, *11*(2), Article 2.

Konijn, E. A., Nije Bijvank, M., & Bushman, B. J. (2007). I wish I were a warrior: The role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental Psychology*, *43*(4), 1038–1044. https://doi.org/10.1037/0012-1649.43.4.1038

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J.,

Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. *Political Communication*, *37*(2), 194–214. https://doi.org/10.1080/10584609.2019.1686094

Kreiss, D. (2021). Polarization Isn't America's Biggest Problem—Or Facebook's. *Wired*. https://www.wired.com/story/polarization-isnt-americas-biggest-problem-or-facebooks/

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung*. Beltz Juventa.

Kuckartz, U., & Rädiker, S. (2019). Datenaufbereitung und Datenbereinigung in der qualitativen Sozialforschung. In *Handbuch Methoden der empirischen Sozialforschung* (S. 441–456). Springer.

Kumar, A., Burtscher, C., & Eckhardt, A. (2024). Unmasking the Phantom: Discovering Exogenous Cues to Detect Political Deepfake Images. *ICIS 2024 Proceedings*. https://aisel.aisnet.org/icis2024/soc_impactIS/soc_impactIS/25

Kumar, D. (2006). Media, War, and Propaganda: Strategies of Information Management During the 2003 Iraq War. *Communication and Critical/Cultural Studies*, *3*(1), 48–69. https://doi.org/10.1080/14791420500505650

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Kumar, S., West, R., & Leskovec, J. (2016). *Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes*. 591–602.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An update for taxonomy designers. *Business & Information Systems Engineering*, *64*(4), 421–439.

Kuo, R., & Marwick, A. (2021). Critical disinformation studies: History, power, and politics. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-76

Ladd, J. M. (2011). Why Americans hate the media and how it matters. *Why Americans Hate the Media and How It Matters*, 1–270.

Lange, B., & Lechterman, T. (2021). *Combating disinformation with AI: Epistemic and ethical challenges* (S. 5). https://doi.org/10.1109/ISTAS52410.2021.9629122

Larkin, B. (2013). The politics and poetics of infrastructure. *Annual review of anthropology*, *42*(2013), 327–343.

Lasotte, Y., Garba, E., Malgwi, Y., & Buhari, M. (2022). An Ensemble Machine

Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier. *European Journal of Electrical Engineering and Computer Science*, *6*(2), 1–7.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). *Explaining Model Confidence Using Counterfactuals* (arXiv:2303.05729). arXiv. http://arxiv.org/abs/2303.05729

Ledford, H. (2023). Researchers scramble as Twitter plans to end free data access. *Nature*, *614*(7949), 602–603. https://doi.org/10.1038/d41586-023-00460-z

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684

Lehrer, C., Wieneke, A., Brocke, J. vom, Jung, R., & Seidel, S. (2018). How Big Data Analytics Enables Service Innovation: Materiality, Affordance, and the Individualization of Service. *Journal of Management Information Systems*, *35*(2), 424–460. https://doi.org/10.1080/07421222.2018.1451953

Lemieux, V., & Smith, T. D. (2018). *Leveraging archival theory to develop a taxonomy of online disinformation*. 4420–4426.

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, *32*. https://doi.org/10.1080/10463283.2021.1876983

Lewandowsky, S., & Yesilada, M. (2021). Inoculating against the spread of Islamophobic and radical-Islamist disinformation. *Cognitive Research: Principles and Implications*, *6*, 1–15.

Li, J., & Su, M.-H. (2020). Real Talk About Fake News: Identity Language and Disconnected Networks of the US Public's "Fake News" Discourse on Twitter. *Social Media + Society*, *6*(2), 2056305120916841. https://doi.org/10.1177/2056305120916841

Li, J., Yang, Y., Liao, Q. V., Zhang, J., & Lee, Y.-C. (2025). *As Confidence Aligns: Exploring the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making* (arXiv:2501.12868). arXiv. https://doi.org/10.48550/arXiv.2501.12868

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176. https://doi.org/10.1145/3461702.3462531

Liang, C. S., & Cross, M. J. (2020). White crusade: How to prevent right-wing

extremists from exploiting the internet. *Geneva Centre for Security Policy*, *11*, 1–27.

Lim, S. S., & Tan, K. R. (2020). Front liners fighting fake news: Global perspectives on mobilising young people as media literacy advocates. *Journal of Children and Media*, *14*(4), Article 4.

Lin, Y., Hu, Z., Alias, H., & Wong, L. P. (2020). Influence of Mass and Social Media on Psychobehavioral Responses Among Medical Students During the Downward Trend of COVID-19 in Fujian, China: Cross-Sectional Study. *Journal of Medical Internet Research*, *22*(7), e19982. https://doi.org/10.2196/19982

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Linder, R., Mohseni, S., Yang, F., Pentyala, S. K., Ragan, E. D., & Hu, X. B. (2021). How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, *2*(4), e49. https://doi.org/10.1002/ail2.49

Lindner, R., & Aichholzer, G. (2020). E-democracy: Conceptual foundations and recent trends. *European e-democracy in practice*, 11–45.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Liu, H., Lai, V., & Tan, C. (2021). Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2), 408:1-408:45. https://doi.org/10.1145/3479552

Livingstone, S. (2004). What is media literacy? *Intermedia*, *32*(3), Article 3.

Livingstone, S., Papaioannou, T., Pérez, M. del M. G., & Wijnen, C. W. (2012). Critical insights in European media literacy research and policy. *Media Studies*, *3*(6), Article 6.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301

Lukito, J. (2020). Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017. *Political Communication*, *37*(2), 238–255. https://doi.org/10.1080/10584609.2019.1661889

Lukyanenko, R., Parsons, J., Wiersma, Y., Wachinger, G., Huber, B., & Meldt, R. (2017). Representing Crowd Knowledge: Guidelines for Conceptual Modeling

of User-generated Content. *Journal of the Association for Information Systems*, *18*, 1–50. https://doi.org/10.17705/1jais.00456

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Lutz, B., Adam, M., Feuerriegel, S., Pröllochs, N., & Neumann, D. (2024). Which Linguistic Cues Make People Fall for Fake News? A Comparison of Cognitive and Affective Processing. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–22. https://doi.org/10.1145/3641030

Lyons, K. (2020). A college student used GPT-3 to write fake blog posts and ended up at the top of hacker news. *The Verge, August*.

Lyu, S. (2024). DeepFake the menace: Mitigating the negative impacts of AI-generated content. *Organizational Cybersecurity Journal: Practice, Process and People*, *4*(1). https://doi.org/10.1108/ocj-08-2022-0014

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology*, *36*(4), 427–445. https://doi.org/10.1177/0268396220915917

MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. K. H. (2021). Refuting spurious COVID-19 treatment claims reduces demand and misinformation sharing. *Journal of Applied Research in Memory and Cognition*, *10*(2), 248–258. https://doi.org/10.1037/h0101793

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th australasian conference on information systems*, *53*, 6–8. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b8eda9593fbcb63b7ced1866853d9622737533a2

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535–544. https://doi.org/10.1007/s12599-019-00600-8

Mahyoob, M., Al-Garaady, J., & Alrahaili, M. (2020). Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics*, *11*(1), Article 1.

Majchrzak, A., & Markus, M. L. (2012). Technology affordances and constraints in management information systems (MIS). *Encyclopedia of Management Theory,(Ed: E. Kessler), Sage Publications, Forthcoming*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2192196

Maliaroudakis, E., Boland, K., Dietze, S., Todorov, K., Tzitzikas, Y., & Fafalios, P. (2021). ClaimLinker: Linking text to a knowledge graph of fact-checked claims. *Companion Proceedings of the Web Conference 2021*, 669–672.

Malkiel, I., Ginzburg, D., Barkan, O., Caciularu, A., Weill, J., & Koenigstein, N. (2022). Interpreting BERT-based text similarity via activation and saliency

maps. *Proceedings of the ACM Web Conference 2022*, 3259–3268.
Maloy, R., Butler, A., & Goodman, L. (2022). Critical media literacy in teacher education: Discerning truth amidst a crisis of misinformation and disinformation. *Journal of Technology and Teacher Education*, *30*(2), Article 2.
Manning, M. J., & Romerstein, H. (2004). *Historical Dictionary of American Propaganda*. 1–448.
March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2
Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one*, *9*(8), e105937.
Marten, H. (2010). Evaluating media literacy education: Concepts, theories and future direction. *Journal of Media Literacy Education*, *2*(1), Article 1.
Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2022). NLP &IRUNED at CheckThat! 2022: Ensemble of classifiers for fake news detection. *Working Notes of CLEF*.
Marwick, A. E., & Lewis, R. (2017). *Media manipulation and disinformation online*.
Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, *53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z
Matasick, C., Alfonsi, C., & Bellantoni, A. (2020). Governance responses to disinformation: How open government principles can inform policy options. *OECD Working Papers on Public Governance*, *39*, 0_1-45.
Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets Hate: A Temporal Study of Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 92:1-92:24. https://doi.org/10.1145/3415163
Mattern, J., Qiao, Y., Kerz, E., Wiechmann, D., & Strohmaier, M. (2021). FANG-COVID: A new large-scale benchmark dataset for fake news detection in German. *Proceedings of the fourth workshop on fact extraction and verification (fever)*, 78–91. https://aclanthology.org/2021.fever-1.9/
Mayring, P. (2015). Qualitative Content Analysis: Theoretical Background and Procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Hrsg.), *Approaches to Qualitative Research in Mathematics Education: Examples of Methodology and Methods* (S. 365–380). Springer Netherlands. https://doi.org/10.1007/978-94-017-9181-6_13
McCornack, S. (2015). Information Manipulation Theory. In *The International Encyclopedia of Interpersonal Communication* (S. 1–7). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118540190.wbeic072
McCornack, S. A. (1992). Information manipulation theory. *Communications*

Monographs. https://doi.org/10.1080/03637759209376245

Mcdougall, J. (2019). Media literacy versus fake news: Critical thinking, resilience and civic engagement. *Media studies*, *10*(19), Article 19.

McGUIRE, W. J., & PAPAGEORGIS, D. (1962). EFFECTIVENESS OF FOREWARNING IN DEVELOPING RESISTANCE TO PERSUASION*. *Public Opinion Quarterly*, *26*(1), 24–34. https://doi.org/10.1086/267068

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

McMahon, L., Kleinman, Z., & Subramanian, C. (2025, Januar 7). *Meta to replace „biased" fact-checkers with moderation by users*. https://www.bbc.com/news/articles/cly74mpy8klo

McQuail, D. (1993). Media Performance: Mass Communication and the Public Interest. *Canadian Journal of Communication*, *18*(4). https://doi.org/10.22230/cjc.1993v18n4a783

Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, *153*, 112986.

Mehta, R., Sahni, J., & Khanna, K. (2018). Internet of Things: Vision, Applications and Challenges. *Procedia Computer Science*, *132*, 1263–1269. https://doi.org/10.1016/j.procs.2018.05.042

Mensah, G. B. (2023). Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI Systems. *Preprint, November*, *10*.

Menz, B. D., Modi, N. D., Sorich, M. J., & Hopkins, A. M. (2024). Health Disinformation Use Case Highlighting the Urgent Need for Artificial Intelligence Vigilance. *JAMA Internal Medicine*, *184*(1). https://doi.org/10.1001/jamainternmed.2023.5947

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Meske, C., & Bunde, E. (2020). *Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support*. 54–69.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics*, *59*, 210–220.

Miller, L. M. S., & Bell, R. A. (2012). Online Health Information Seeking: The Influence of Age, Information Trustworthiness, and Search Challenges. *Journal of Aging and Health*, *24*(3), 525–541. https://doi.org/10.1177/0898264311428167

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social

sciences. *Artificial intelligence*, *267*, 1–38.
Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 1–66.
Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, *54*(1), 7:1-7:41. https://doi.org/10.1145/3425780
Mirza, S., Begum, L., Niu, L., Pardo, S., Abouzied, A., Papotti, P., & Pöpper, C. (2023). Tactics, Threats & Targets: Modeling Disinformation and its Mitigation. *Proceedings 2023 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA, USA. https://doi.org/10.14722/ndss.2023.23657
Modgil, S., Singh, R. K., Gupta, S., & Dennehy, D. (2021). A Confirmation Bias View on Social Media Induced Polarisation During Covid-19. *Information Systems Frontiers*, *26*(2), 417–441. https://doi.org/10.1007/s10796-021-10222-9
Mody, S. (2020, Dezember 30). Managing The Risk of Fake News. *IRM India Affiliate*. https://www.theirmindia.org/blog/managing-the-risk-of-fake-news/
Mohseni, S., Ragan, E., & Hu, X. (2019). Open issues in combating fake news: Interpretability as an opportunity. *arXiv preprint arXiv:1904.03016*. https://arxiv.org/abs/1904.03016
Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., & Ragan, E. (2021). Machine Learning Explanations to Prevent Overtrust in Fake News Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 421–431. https://doi.org/10.1609/icwsm.v15i1.18072
Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3–4), 1–45. https://doi.org/10.1145/3387166
Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, *65*(2), 180–212.
Moscadelli, A., Albora, G., Biamonte, M. A., Giorgetti, D., Innocenzio, M., Paoli, S., Lorini, C., Bonanni, P., & Bonaccorsi, G. (2020). Fake News and Covid-19 in Italy: Results of a Quantitative Observational Study. *International Journal of Environmental Research and Public Health*, *17*(16), Article 16. https://doi.org/10.3390/ijerph17165850
Moser, H. (2020). Überlegungen zum Lernen mit und über Medien im Zeitalter der Digitalisierung. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 709–732.
Mridha, M. F., Keya, A. J., Hamid, M. A., Monowar, M. M., & Rahman, M. S. (2021). A comprehensive review on fake news detection with deep learning. *IEEE access*, *9*, 156151–156170.
Mullarkey, M. T., & and Hevner, A. R. (2019). An elaborated action design research

process model. *European Journal of Information Systems*, *28*(1), 6–20. https://doi.org/10.1080/0960085X.2018.1451811

Mulvey, K., Shulman, S., Anderson, D., Cole, N., Piepenburg, J., & Sideris, J. (2015). *The climate deception dossiers: Internal fossil fuel industry memos reveal decades of corporate disinformation* [Report]. Union of Concerned Scientists. https://apo.org.au/node/55839

Munn, L. (2020). Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), 1–11.

Murayama, T., Hisada, S., Uehara, M., Wakamiya, S., & Aramaki, E. (2022). *Annotation-Scheme Reconstruction for „Fake News“ and Japanese Fake News Dataset* (arXiv:2204.02718). arXiv. https://doi.org/10.48550/arXiv.2204.02718

Murphy, H. (2023, Oktober 16). *Israel conflict lets loose a deluge of falsehoods on social media*. https://www.ft.com/content/01650afb-dab4-4668-b16a-6add6ade0c04

Musi, E., Federico, L., & Riotta, G. (2022). Human–computer interaction tools with gameful design for critical thinking the media ecosystem: A classification framework. *AI & society*, 1–13.

Myers, M. D. (2019). *Qualitative research in business and management*. https://www.torrossa.com/gs/resourceProxy?an=5018482&publisher=FZ7200

Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information and Libraries Journal*, *38*(2), 143–149. https://doi.org/10.1111/hir.12320

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text annotation tool for human. *Software available from https://github. com/doccano/doccano*, *34*.

Nelson-Field, K., Riebe, E., & Newstead, K. (2013). The emotions that drive viral video. *Australasian Marketing Journal*, *21*(4), 205–211.

Ngueajio, M., Aryal, S., Atemkeng, M., Washington, G., & Rawat, D. (2025). Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *ACM Computing Surveys*, 3711123. https://doi.org/10.1145/3711123

Nguyen, A., Kharosekar, A., Lease, M., & Wallace, B. (2018). An Interpretable Joint Graphical Model for Fact-Checking From Crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.11487

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A Method for Taxonomy Development and its Application in Information Systems. *European Journal of Information Systems*, *22*, 336–359.

Niebert, K., & Gropengiesser, H. (2013). *Leitfadengestützte Interviews* (S. 121–132). https://doi.org/10.1007/978-3-642-37827-0_10

Nikolov, D., Flammini, A., & Menczer, F. (2021). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *Harvard Kennedy School*

Misinformation Review. https://doi.org/10.37016/mr-2020-55

Nilsson, M., & Mattes, J. (2015). The spatiality of trust: Factors influencing the creation of trust and the role of face-to-face contacts. *European Management Journal*, *33*(4), 230–244.

Noreen, I., Muneer, M. S., & Gillani, S. (2022). Deepfake attack prevention using steganography GANs. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/peerj-cs.1125

Norris, P. (2022). *In praise of skepticism: Trust but verify*. Oxford University Press. https://books.google.de/books?hl=de&lr=&id=J-t5EAAAQBAJ&oi=fnd&pg=PP1&dq=In+praise+of+skepticism:+Trust+but+verify&ots=0UpWSy_q7L&sig=hEBjhE4XfJMT7gayyc-3b-gCadU

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 97–105. https://ojs.aaai.org/index.php/HCOMP/article/view/5284

Nyow, N. X., & Chua, H. N. (2019). Detecting fake news with tweets' properties. *2019 IEEE conference on application, information and network security (AINS)*, 24–29. https://ieeexplore.ieee.org/abstract/document/8968706/?casa_token=kUOnnewJDLIAAAAA:3r77zECzTZbbgb0k2Al9xHiTq7_MjOyhamToBNDU1zY3YKIG9T1wKGy7Ejd7Tsml2ivPGlfnNr4

Odeh, A. (2024). Unmasking Deepfakes: Advances in Fake Video Detection. *Revue d'Intelligence Artificielle*, *38*(4), 1119.

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposureis linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-024

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174223

Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science*, *11*(4), 404–428. https://doi.org/10.1287/orsc.11.4.404.14600

O'Sullivan, D., Yurieff, K., & Bourdet, K. (2021). *Misinformation watch on the 2020 election | CNN Business*. https://edition.cnn.com/business/live-news/election-2020-misinformation/index.html

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration. *Media and Communication*, *9*(1), 171–180.

https://doi.org/10.17645/mac.v9i1.3399

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Päivärinta, T., & Sæbø, Ø. (2006). Models of E-Democracy. *Communications of the Association for Information Systems*, *17*. https://doi.org/10.17705/1CAIS.01737

Pamment, J. (2020). *The EU's Role in Fighting Disinformation: Crafting A Disinformation Framework | Carnegie Endowment for International Peace* (2; Carnegie Endowment for International Peace Future Threats Future Solutions series). https://carnegieendowment.org/research/2020/09/the-eus-role-in-fighting-disinformation-crafting-a-disinformation-framework?lang=en

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). *How model accuracy and explanation fidelity influence user trust* (arXiv:1907.12652). arXiv. http://arxiv.org/abs/1907.12652

Parikh, S. B., & Atrey, P. K. (2018). *Media-rich fake news detection: A survey*. 436–441.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021). Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445304

Park, J., Park, L. H., Ahn, H. E., & Kwon, T. (2024). Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge. *IEEE Access*, *12*. https://doi.org/10.1109/access.2024.3353785

Parkinson, H. J. (2016, November 14). Click and elect: How fake news helped Donald Trump win a real election. *The Guardian*. https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies

Parthiban, G., & Peter, S. (2022). *Review of Fake News Detection in Social Media using Machine Learning Techniques* (S. 501). https://doi.org/10.1109/ICAISS55157.2022.10010796

Payne, G., & Payne, J. (2004). *Key concepts in social research*. https://www.torrossa.com/gs/resourceProxy?an=4912774&publisher=FZ7200

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design Science Research Evaluation. In K. Peffers, M. Rothenberger, & B. Kuechler (Hrsg.), *Design Science Research in Information Systems. Advances in Theory and Practice* (Bd. 7286, S. 398–410). Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-642-29863-9_29

Peffers, K., Tuunanen, T., Rothenbergre, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Peissl, H., & Sedlaczek, A. (2022). Kritische Medienkompetenz vor dem Hintergrund der Digitalisierung. Media and Information Literacy (MIL) und Critical Media Literacy (CML) im Vergleich. *Magazin erwachsenenbildung. at*, *44/45*, Article 44/45.

Peissl, H., Sedlaczek, A., Eppensteiner, B., & Stenitzer, C. (2018). *Kritische Medienkompetenz und Community Medien*. CONEDU – Verein für Bildungsforschung und -medien. https://doi.org/10.25656/01:16869

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2018a). Lazy, not biased: Suceptibility to partisan news is better explained by lack of reasinong than by motivated reasoning. *Cognition*, 1–12.

Pennycook, G., & Rand, D. G. (2018b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 1–12. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388–402.

Pereira, S., & Moura, P. (2019). Assessing media literacy competences: A study with Portuguese young people. *European Journal of Communication*, *34*(1), 20–37.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic Detection of Fake News. In E. M. Bender, L. Derczynski, & P. Isabelle (Hrsg.), *Proceedings of the 27th International Conference on Computational Linguistics* (S. 3391–3401). Association for Computational Linguistics. https://aclanthology.org/C18-1287

Peters, E., Diefenbach, M. A., Hess, T. M., & Västfjäll, D. (2008). Age Differences in Dual Information-Processing Modes: Implications for Cancer Decision Making. *Cancer*, *113*(12 Suppl), 3556–3567. https://doi.org/10.1002/cncr.23944

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, *64*(6), 763–774.

Pfeiffer, J., Lachenmaier, J. F., Hinz, O., & Van Der Aalst, W. (2024). New Laws and Regulation: Opportunities for BISE Research. *Business & Information Systems Engineering*, *66*(6), 653–666. https://doi.org/10.1007/s12599-024-00902-6

Pfiffner, M., Sterel, S., & Hassler, D. (2021). 4K und digitale Kompetenzen. *Chancen und Herausforderungen*, *1*.

Phang, C. W., & Kankanhalli, A. (2008). A framework of ICT exploitation for e-participation initiatives. *Communications of the ACM*, *51*(12), 128–132.

https://doi.org/10.1145/1409360.1409385

Phillips, W., & Milner, R. M. (2021). *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape*. MIT Press.

Pires, T., Schlinger, E., & Garrette, D. (2019). *How multilingual is Multilingual BERT?* (arXiv:1906.01502). arXiv. https://doi.org/10.48550/arXiv.1906.01502

Plepi, J., Sakketou, F., Geiss, H. J., & Flek, L. (2022). Temporal graph analysis of misinformation spreaders in social media. *Proceedings of textgraphs-16: graph-based methods for natural language processing*, 89–104. https://aclanthology.org/2022.textgraphs-1.10/

Pohle, J., & Thiel, T. (2020). Digital sovereignty. *Pohle, J. & Thiel*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4081180

Posetti, J., & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. *International Center for Journalists*.

Potter, W. J. (2010). The state of media literacy. *Journal of broadcasting & electronic media*, *54*(4), Article 4.

Potter, W. J. (2013). Review of literature on media literacy. *Sociology Compass*, *7*(6), 417–435.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). *Manipulating and Measuring Model Interpretability* (arXiv:1802.07810; Nummer arXiv:1802.07810). arXiv. http://arxiv.org/abs/1802.07810

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Pritchard, A. (2017). *Ways of learning: Learning theories for the classroom*. Routledge.

Pullinger, J. (2021). Misuse of statistics: Time to speak out. *Statistical Journal of the IAOS*, *37*(1), Article 1.

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. *2019 international engineering conference (IEC)*, 200–204. https://ieeexplore.ieee.org/abstract/document/8950616/?casa_token=fg4s8iPN_HAAAAAA:BKjNCkuMAhVxIcsL8CJIbLVzXc_aFjIo4xN9PSoHItBiPJ51K66X1g9HkOwzokhYK3A9dKjyGw

Qi, P., Cao, J., Li, X., Liu, H., Sheng, Q., Mi, X., He, Q., Lv, Y., Guo, C., & Yu, Y. (2021). Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. *Proceedings of the 29th ACM International Conference on Multimedia*, 1212–1220.

https://doi.org/10.1145/3474085.3481548

Qureshi, I., Bhatt, B., Gupta, S., & Tiwari, A. (2021). *Causes, Symptoms and Consequences of Social Media Induced Polarization (SMIP)*.

Qureshi, J., & Khan, S. (2024). Deciphering Deception–the Impact of AI Deepfakes on Human Cognition and Emotion. *Journal of Advances in Artificial Intelligence*, *2*(1), 1–101.

Qureshi, K. A., Malick, R. A. S., Sabih, M., & Cherifi, H. (2021). Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter. *IEEE Access*, *9*, 139636–139656.

Radwan, E., Radwan, A., & Radwan, W. (2020). The role of social media in spreading panic among primary and secondary school students during the COVID-19 pandemic: An online questionnaire study from the Gaza Strip, Palestine. *Heliyon*, *6*(12). https://doi.org/10.1016/j.heliyon.2020.e05807

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, *10*, 25494–25513.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). *Truth of varying shades: Analyzing language in fake news and political fact-checking*. 2931–2937.

Rasi, P., Vuojärvi, H., & Ruokamo, H. (2019). Media literacy education for all ages. *Journal of Media Literacy Education*, *11*(2), Article 2.

Rauh, C., & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. *Harvard Dataverse*, *1*, 1.

Reboot Foundation. (2022, September 5). *Science Fictions: Low Science Knowledge, Poor Critical Thinking Linked to Conspiracy Beliefs | REBOOT FOUNDATION*. https://reboot-foundation.org/research/science-fictions-low-science-knowledge-poor-critical-thinking-linked-to-conspiracy-beliefs/

Reddy, P., Sharma, B., & Chaudhary, K. (2020). Digital literacy: A review of literature. *International Journal of Technoethics (IJT)*, *11*(2), Article 2.

Reeder, S., Jensen, J., & Ball, R. (2023). Evaluating Explainable AI (XAI) in Terms of User Gender and Educational Background. In H. Degen & S. Ntoa (Hrsg.), *Artificial Intelligence in HCI* (S. 286–304). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_18

Reveland, C. (2025). *KI-Influencer im Wahlkampf: Rechts, weiblich, Fake*. tagesschau.de. https://www.tagesschau.de/faktenfinder/kontext/rechte-ki-influencer-100.html

Ribeiro Bezerra, J. F. (2021). Content-based fake news classification through modified voting ensemble. *Journal of Information and Telecommunication*, *5*(4), 499–

513.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *„ Why should i trust you?“ Explaining the predictions of any classifier*. 1135–1144.

Rieger, D., Ernst, J., Schmitt, J. B., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Propaganda und Alternativen im Internet–Medienpädagogische Implikationen. *Medien+ Erziehung: Merz*, *3*, Article 3.

Rjoob, K., Bond, R., Finlay, D., McGilligan, V., Leslie, S. J., Rababah, A., Iftikhar, A., Guldenring, D., Knoery, C., McShane, A., & Peace, A. (2021). Towards Explainable Artificial Intelligence and Explanation User Interfaces to Open the 'Black Box' of Automated ECG Interpretation. In T. Reis, M. X. Bornschlegl, M. Angelini, & M. L. Hemmje (Hrsg.), *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications* (Bd. 12585, S. 96–108). Springer International Publishing. https://doi.org/10.1007/978-3-030-68007-7_6

Rocha, Y. M., De Moura, G. A., Desidério, G. A., De Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolete, L. D. (2021). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health*, 1–10.

Rodríguez-Fernández, L. (2019). *Disinformation and organisational communication: A study of the impact of fake news* (74. Aufl.). Revista Latina de Comunicación Social. https://doi.org/10.4185/RLCS-2019-1406en

Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., Hong, W.-C., & Sharma, R. (2022). A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects. *IEEE Access*.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Rosińska, K. A. (2021). Disinformation in Poland: Thematic classification based on content analysis of fake news from 2019. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *15*(4).

Rosso, P., Chulvi, B., Korenčić, D., Taulé, M., Casals, X. B., Camacho, D., Panizo, A., Arroyo, D., Gómez, J., & Rangel, F. (2024). *XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics*. https://ceur-ws.org/Vol-3729/p14_rev.pdf

Rowan, W., & Pears, N. (2022). *The Effectiveness of Temporal Dependency in Deepfake Video Detection*. https://doi.org/10.48550/arxiv.2205.06684

Rushkoff, D. (2016). *Throwing rocks at the Google bus: How growth became the enemy of prosperity*. Penguin. https://books.google.de/books?hl=de&lr=&id=3g_VCQAAQBAJ&oi=fnd&pg=PR9&dq=Throwing+Rocks+at+the+Google+Bus:+How+Growth+Became+the+Enemy+of+Prosperity+(2016).&ots=A1zqyW2cQU&sig=qGVz3typCTXfqdykn

hhPDKb1Ulg

Sahan, M., Smidl, V., & Marik, R. (2021). *Active Learning for Text Classification and Fake News Detection*. 87–94.

Sakketou, F., Plepi, J., Cervero, R., Geiss, H.-J., Rosso, P., & Flek, L. (2022). *FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias* (arXiv:2205.06181). arXiv. https://doi.org/10.48550/arXiv.2205.06181

Salthouse, T. A. (1992). Influence of processing speed on adult age differences in working memory. *Acta psychologica*, *79*(2), 155–170.

Salthouse, T. A. (1994). The nature of the influence of speed on adult age differences in cognition. *Developmental Psychology*, *30*(2), 240–259. https://doi.org/10.1037/0012-1649.30.2.240

Salthouse, T. A., McGuthry, K. E., & Hambrick, D. Z. (1999). A Framework for Analyzing and Interpreting Differential Aging Patterns: Application to Three Measures of Implicit Learning. *Aging, Neuropsychology, and Cognition*, *6*(1), 1–18. https://doi.org/10.1076/anec.6.1.1.789

Sample, C., Jensen, M. J., Scott, K., McAlaney, J., Fitchpatrick, S., Brockinton, A., Ormrod, D., & Ormrod, A. (2020). Interdisciplinary Lessons Learned While Researching Fake News. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.537612

Sander, I. (2020). What is critical big data literacy and how can it be implemented? *Internet Policy Review*, *9*(2), Article 2.

Sandoval, M.-P., Vau, M., Solaas, J., & Rodrigues, L. (2024). Threat of deepfakes to the criminal justice system: A systematic review. *Crime Science*, *13*. https://doi.org/10.1186/s40163-024-00239-1

Sanneman, L., & Shah, J. A. (2022). The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction*, *38*(18–20), 1772–1788. https://doi.org/10.1080/10447318.2022.2081282

Sarker, S., Chatterjee, S., & Xiao, X. (2013). How" Sociotechnical" is our IS Research?: An Assessment and Possible Ways Forward. *Proceedings of the 34th International Conference on Information Systems. ICIS 2013*, 1185. https://research.cbs.dk/en/publications/how-sociotechnical-is-our-is-research-an-assessment-and-possible-

Schacht, S., Morana, S., & Maedche, A. (2015). The Evolution of Design Principles Enabling Knowledge Reuse for Projects: An Action Design Research Project. *JITTA: Journal of Information Technology Theory and Application*, *16*(3), 5.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400.

https://doi.org/10.1177/0018720816634228
Schäfer, M. S. (2015). Digital public sphere. *The international encyclopedia of political communication*, *15*, 1–7.
Schank, H., & McGuinness, T. D. (2021). *Power to the Public: The Promise of Public Interest Technology*. Princeton University Press. https://doi.org/10.1515/9780691216638
Schemmer, M. (2022). *A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making*.
Schlichtkrull, M., Ousidhoum, N., & Vlachos, A. (2023). The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who. In H. Bouamor, J. Pino, & K. Bali (Hrsg.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (S. 8618–8642). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.577
Schmidt, T., Biessmann, F., & Teubner, T. (2020). Transparency and Trust in Artificial Intelligence Systems. *Journal of Decision Systems*, *29*(4), 260–278.
Schmitt, J. B., Ernst, J., Rieger, D., & Roth, H.-J. (2020). Die Förderung von Medienkritikfähigkeit zur Prävention der Wirkung extremistischer Online-Propaganda. In J. B. Schmitt, J. Ernst, D. Rieger, & H.-J. Roth (Hrsg.), *Propaganda und Prävention: Forschungsergebnisse, didaktische Ansätze, interdisziplinäre Perspektiven zur pädagogischen Arbeit zu extremistischer Internetpropaganda* (S. 29–44). Springer Fachmedien. https://doi.org/10.1007/978-3-658-28538-8_2
Schmitt, J. B., Rieger, D., Ernst, J., & Roth, H.-J. (2018). Critical Media Literacy and Islamist Online Propaganda: The Feasibility, Applicability and Impact of Three Learning Arrangements. *International Journal of Conflict and Violence*, *12*, 1–19.
Schmitt, V., Villa-Arenas, L.-F., Feldhus, N., Meyer, J., Spang, R. P., & Möller, S. (2024). The Role of Explainability in Collaborative Human-AI Disinformation Detection. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2157–2174. https://doi.org/10.1145/3630106.3659031
Schorb, B., Hartung-Griemberg, A., & Dallmann, C. (Hrsg.). (2017). *Grundbegriffe Medienpädagogik* (6., neu verfasste Auflage). kopaed.
Schreiber, D., Picus, C., Fischinger, D., & Boyer, M. (2021). The defalsif-AI project: Protecting critical infrastructures against disinformation and fake news/Das Projekt defalsif-AI: Schutz kritischer Infrastrukturen vor Desinformation und Fake News. *Elektrotechnik und Informationstechnik*, *Vol. 138 (7)*, Article Vol. 138 (7).
Schüller, K., Busch, P., & Hindinger, C. (2019). Future skills: Ein framework für data literacy. *Hochschulforum Digitalisierung*, *46*, 1–128.
Schunk, D. H. (2012). *Learning theories an educational perspective*. Pearson

Education, Inc.

Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*, *46*(2), 499–510. https://doi.org/10.1162/coli_a_00380

Schwerter, F., & Zimmermann, F. (2020). Determinants of trust: The role of personal experiences. *Games and Economic Behavior*, *122*, 413–425. https://doi.org/10.1016/j.geb.2020.05.002

Secosan, I., Virga, D., Crainiceanu, Z. P., Bratu, L. M., & Bratu, T. (2020). Infodemia: Another Enemy for Romanian Frontline Healthcare Workers to Fight during the COVID-19 Outbreak. *Medicina*, *56*(12), Article 12. https://doi.org/10.3390/medicina56120679

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly*, *35*(1), 37–56. https://doi.org/10.2307/23043488

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. 618–626.

Shae, Z., & Tsai, J. (2019). AI blockchain platform for trusting news. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 1610–1619. https://ieeexplore.ieee.org/abstract/document/8884985/?casa_token=bheCt5gA6bwAAAAA:VygMMQMfRO3M_c1Ru6KIVEjvmJtrNVSj_qsk29r5Ps70HVcVnDX-bXOrH4TpLqsw0po3pkxItw

Shahi, G. K., Jaiswal, A. K., & Mandl, T. (2024). FakeClaim: A Multiple Platform-Driven Dataset for Identification of Fake News on 2023 Israel-Hamas War. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Hrsg.), *Advances in Information Retrieval* (S. 66–74). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56069-9_5

Shahid, W., Jamshidi, B., Hakak, S., Isah, H., Khan, W. Z., Khan, M. K., & Choo, K.-K. R. (2022). Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities. *IEEE Transactions on Computational Social Systems*. https://ieeexplore.ieee.org/abstract/document/9789171/?casa_token=A4nwVoG_HbgAAAAA:kHhH2Sj1gzOYj-m_BJrsEt7vR-aQaUwKifS4ugkHsHMgpNVG2yyKziC9OIIyqcbRWuLS-U_LNA

Shane, S. (2017, September 7). The Fake Americans Russia Created to Influence the Election. *The New York Times*. https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques.

ACM Transactions on Intelligent Systems and Technology, *10*(3), 1–42. https://doi.org/10.1145/3305260

Sharma, K., Zhang, Y., & Liu, Y. (2021). *COVID-19 vaccines: Characterizing misinformation campaigns and vaccine hesitancy on twitter*.

Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19906-1

Shelke, M., Ranjan, N. M., Kharade, A., Gaikwad, P., Arakh, S., & Kalore, A. (2023). Combining Computer Vision Techniques and Intraframe Noise Methods to Detect a Deepfake. *Data Science and Intelligent Computing Techniques*. https://doi.org/10.56155/978-81-955020-2-8-43

Shin, D. (2021). The Effects of Explainability and Causability on Perception, Trust and Acceptance: Implications for Explainable AI. *International Journal of Human-Computer Studies*, *146*.

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond User Experience: What Constitutes Algorithmic Experiences? *International Journal of Information Management*, *52*, 1–11.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). *Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models* (arXiv:2311.17394). arXiv. https://doi.org/10.48550/arXiv.2311.17394

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020a). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), Article 6.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020b). Combating Disinformation in a Social Media Age. *WIREs Data Mining and Knowledge Discovery*, *10*, 1–23.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, *31*(2), 47–53.

Silverman, C. (2016, November 16). *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook*. BuzzFeed News. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Simonofski, A., Zuiderwijk, A., Clarinval, A., & Hammedi, W. (2022). Tailoring open government data portals for lay citizens: A gamification theory approach. *International journal of information management*, *65*, 102511.

Sirlin, N., Epstein, Z., Arechar, A. A., & Rand, D. G. (2021). *Digital literacy is*

associated with more discerning accuracy judgments but not sharing intentions.

Snow, N., & Taylor, P. M. (2006). The Revival of the Propaganda State: US Propaganda at Home and Abroad since 9/11. *International Communication Gazette*, *68*(5–6), 389–407. https://doi.org/10.1177/1748048506068718

Sohrawardi, S. J., Seng, S., Chintha, A., Ptucha, R., Wright, M., & Hickerson, A. (2020). DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection. *USENIX Symposium on Usable Privacy and Security (SOUPS)*.

Solon, O. (2018, März 12). Tim Berners-Lee: We must regulate tech firms to prevent „weaponised" web. *The Guardian*. https://www.theguardian.com/technology/2018/mar/11/tim-berners-lee-tech-companies-regulations

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*, 107917. https://doi.org/10.1016/j.chb.2023.107917

Sonnenberg, C., & vom Brocke, J. (2012). Evaluation Patterns for Design Science Research Artefacts. In M. Helfert & B. Donnellan (Hrsg.), *Practical Aspects of Design Science* (S. 71–83). Springer. https://doi.org/10.1007/978-3-642-33681-2_7

Soral, W., Liu, J., & Bilewicz, M. (2020). Media of Contempt: Social Media Consumption Predicts Normative Acceptance of Anti-Muslim Hate Speech and Islamoprejudice. *International Journal of Conflict and Violence (IJCV)*, *14*, 1–13. https://doi.org/10.4119/ijcv-3774

Soßdorf, A. (2023). Kompetenzen in einer Kultur der Digitalität: Brauchen wir generalistische Kompetenzmodelle? *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 257–280.

Soßdorf, A., & Gallach, L. (2022). Menü statt à la carte – Warum wir digitale, politische und ethische Bildung gemeinsam denken müssen. In G. Marci-Boehncke, M. Rath, M. Delere, & H. Höfer (Hrsg.), *Medien – Demokratie – Bildung: Normative Vermittlungsprozesse und Diversität in mediatisierten Gesellschaften* (S. 135–151). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36446-5_9

Soßdorf, A., Stein, C., Bezzaoui, I., & Fegert, J. (2024). Literacies against Fake News: Examining the Role of Data Literacy and Critical Media Literacy to Counteract Disinformation. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *59*, 55–76.

Spanhel, D. (2011). Medienkompetenz oder Medienbildung? Begriffliche Grundlagen für eine Theorie der Medienpädagogik. *Medienbildung und Medienkompetenz: Beiträge zu Schlüsselbegriffen der Medienpädagogik*, 95–120.

Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: From artificial intelligence back to human intelligence. *Smart Learning Environments*, *6*(1), 8, s40561-019-0088-z.

https://doi.org/10.1186/s40561-019-0088-z

Speith, T., & Langer, M. (2023). A New Perspective on Evaluation Methods for Explainable Artificial Intelligence (XAI). *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 325–331. https://doi.org/10.1109/REW57809.2023.00061

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26). https://doi.org/10.1126/sciadv.adh1850

Stalder, F. (2016). *Kultur der Digitalität*. Suhrkamp Verlag.

Stalder, F. (2018). *The digital condition*. John wiley & sons. https://books.google.de/books?hl=de&lr=&id=MIFSDwAAQBAJ&oi=fnd&pg=PP2&dq=Stalder,+Felix.+2018.+The+digital+condition.+John+Wiley+and+Sons.&ots=vZXh7ZFxcF&sig=jBlLVHdSjbR3nJ4W_uI5QDb4nPI

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Starbird, K., & Wilson, T. (2020). Cross-Platform Disinformation Campaigns: Lessons Learned and Next Steps. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-002

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining*, *3*(4), 1277–1291. https://doi.org/10.1007/s13278-012-0079-3

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Stitini, O., Kaloun, S., & Bencharef, O. (2022). Towards the detection of fake news on social networks contributing to the improvement of trust and transparency in recommendation systems: Trends and challenges. *Information*, *13*(3), 128.

Stray, J. (2019). Institutional Counter-disinformation Strategies in a Networked Democracy. *Companion Proceedings of The 2019 World Wide Web Conference*, 1020–1025. https://doi.org/10.1145/3308560.3316740

Strömbäck, J. (2005). In Search of a Standard: Four models of democracy and their normative implications for journalism. *Journalism Studies*, *6*(3), 331–345. https://doi.org/10.1080/14616700500131950

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press. https://books.google.de/books?hl=de&lr=&id=AJ_eBJtHxmsC&oi=fnd&pg=PR7&dq=Suchman,+L.+A.+(1987).+Plans+and+Situated+Actions:+The+Problem+of+Human-Machine+Communication.+Cambridge+University+Press.&ots=KtKpjMLLHN

&sig=3kVqe-fMGWufkTbzoxYaUOj3-F0

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy Theories: Causes and Cures. *Journal of Political Philosophy*, *17*(2), 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Suryavardan, S., Mishra, S., Patwa, P., Chakraborty, M., Rani, A., Reganti, A., Chadha, A., Das, A., Sheth, A., Chinnakotla, M., Ekbal, A., & Kumar, S. (2023). *Factify 2: A Multimodal Fake News and Satire News Dataset* (arXiv:2304.03897). arXiv. https://doi.org/10.48550/arXiv.2304.03897

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.

Swapna, H., & Soniya, B. (2022). A review on news-content based fake news detection approaches. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6. https://ieeexplore.ieee.org/abstract/document/9885447/?casa_token=WF9Jk2pw37oAAAAA:0Fod-CU0Ig3vlOvNRzuWf_0PKOORWHyMkZ_IoNAr6cs_P4Hb_aaGyfxkiqbNTZ0aQSUmGG4J-Q

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1948–1961. https://doi.org/10.1037/xlm0000422

Szakacs, J., & Bognar, E. (2021). The impact of disinformation campaigns about migrants and minority groups in the EU. *Policy Department for External Relations Directorate General for External Policies of the Union. https://www. europarl. europa. eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN. pdf*.

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific reports*, *11*(1), 23705.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, *50*(3), 755–769.

Tam, N. T., Weidlich, M., Zheng, B., Yin, H., Hung, N. Q. V., & Stantic, B. (2019). From anomaly detection to rumour detection using data streams of social platforms. *Proc. VLDB Endow.*, *12*(9), 1016–1029. https://doi.org/10.14778/3329772.3329778

Tambini, D. (2017). *Fake news: Public policy responses*.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

https://doi.org/10.1080/21670811.2017.1360143

Tanwar, V., & Sharma, K. (2021). A Review on Enhanced Techniques for Multimodal Fake News Detection. In P. K. Singh, Y. Singh, M. H. Kolekar, A. K. Kar, J. K. Chhabra, & A. Sen (Hrsg.), *Recent Innovations in Computing* (Bd. 701, S. 767–777). Springer Singapore. https://doi.org/10.1007/978-981-15-8297-4_61

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2022). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, *113*(3), 591–607. https://doi.org/10.1111/bjop.12551

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., & Todorov, K. (2019). ClaimsKG: A Knowledge Graph of Fact-Checked Claims. *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings*, 309–324.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence. *Journal of Marketing*, *83*(4), 1–20. https://doi.org/10.1177/0022242919841034

Thaler, R., & Sunstein, C. (2008). Nudge: Improving decisions about health, wealth and happiness. *Amsterdam Law Forum; HeinOnline: Online*, 89. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/amslawf1&section=49

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Thornton, B. (2000). The Moon Hoax: Debates About Ethics in 1835 New York Newspapers. *Journal of Mass Media Ethics*, *15*(2), 89–100. https://doi.org/10.1207/S15327728JMME1502_3

Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of design science research questions. *Communications of the Association for Information Systems*, *44*(1), 20.

Ting, H. L. J., Kang, X., Li, T., Wang, H., & Chu, C.-K. (2021). On the trust and trust modeling for the future fully-connected digital world: A comprehensive study. *IEEE Access*, *9*, 106743–106783.

Torgheh, F., Keyvanpour, M. R., Masoumi, B., & Shojaedini, S. V. (2021). A Novel Method for Detecting Fake news: Deep Learning Based on Propagation Path Concept. *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 1–5. https://ieeexplore.ieee.org/abstract/document/9420601/

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). Focus Groups for Artifact Refinement and Evaluation in Design Research. *Communications of the*

Association for Information Systems, *26*. https://doi.org/10.17705/1CAIS.02627

Trinh, L., Tsang, M., Rambhatla, S., & Liu, Y. (2021). Interpretable and trustworthy deepfake detection via dynamic prototypes. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1973–1983. http://openaccess.thecvf.com/content/WACV2021/html/Trinh_Interpretable_and_Trustworthy_Deepfake_Detection_via_Dynamic_Prototypes_WACV_2021_paper.html

Trültzsch-Wijnen, C. W. (2020). *Medienhandeln zwischen Kompetenz, Performanz und Literacy*. Springer.

Truong, B. T., Lou, X., Flammini, A., & Menczer, F. (2024). Quantifying the vulnerabilities of the online public square to adversarial manipulation tactics. *PNAS Nexus*, *3*(7), pgae258. https://doi.org/10.1093/pnasnexus/pgae258

Tsai, C.-H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445101

Tulodziecki, G. (1998). Entwicklung von Medienkompetenz als Erziehungs-und Bildungsaufgabe. *Pädagogische Rundschau*, *52*(6), Article 6.

Tulodziecki, G. (2011). Zur Entstehung und Entwicklung zentraler Begriffe bei der pädagogischen Auseinandersetzung mit Medien. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *20*, 11–39. https://doi.org/10.21240/mpaed/20/2011.09.11.X

Tulodziecki, G. (2015). Dimensionen von Medienbildung: Ein konzeptioneller Rahmen für medienpädagogisches Handeln. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 31–49.

Tuters, M., & Hagen, S. (2020). *(((They))) rule: Memetic antagonism and nebulous othering on 4chan—Marc Tuters, Sal Hagen, 2020*. https://journals.sagepub.com/doi/10.1177/1461444819888746

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases*. *185*.

Twidale, M. B., Blake, C., & Gant, J. P. (2013). *Towards a data literate citizenry*.

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), e0291668. https://doi.org/10.1371/journal.pone.0291668

Tymann, K. M., Lutz, M., Palsbröker, P., & Gips, C. (2019). *GerVADER-A German adaptation of the VADER sentiment analysis tool for social media texts*. https://www.hsbi.de/publikationsserver/record/3170

United Nations. (2022). *Countering disinformation for the promotion and protection of human rights and fundamental freedoms.* [Report of the Secretary-General.].

N2245924.pdf (un.org)

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

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press. https://books.google.de/books?hl=de&lr=&id=h05WDwAAQBAJ&oi=fnd&pg=PP1&dq=Vaidhyanathan,+S.+(2018).+Antisocial+Media:+How+Facebook+Disconnects+Us+and+Undermines+Democracy.+Oxford+University+Press.&ots=WhmD11pptB&sig=-DZhPdWtDfvwOQ68A-rjnaHAlv0

Vaishnavi, K., Bindu, L. H., Sathvika, M., Lakshmi, K. U., Harini, M., & Ashok, N. C. (2023). Deep learning approaches for robust deep fake detection. *World Journal of Advanced Research and Reviews*, *21*(3). https://doi.org/10.30574/wjarr.2024.21.3.0889

Valtonen, T., Tedre, M., Mäkitalo, K., & Vartiainen, H. (2019). Media Literacy Education in the Age of Machine Learning. *Journal of Media Literacy Education*, *11*(2), 20–36.

van der Linden, S. (2019). Countering science denial. *Nature Human Behaviour*, *3*(9), 889–890. https://doi.org/10.1038/s41562-019-0631-5

Van Rahden, T. (2019). *Demokratie: Eine gefährdete Lebensform*. Campus Verlag. https://books.google.de/books?hl=de&lr=&id=dztrEAAAQBAJ&oi=fnd&pg=PA7&dq=demokratie+eine+gef%C3%A4hrdete+lebensform&ots=W9Ecf9w8Qh&sig=NKsaZFYG03DAfNZmQXyETqoNdyg

Vasist, P. N., Chatterjee, D., & Krishnan, S. (2024). The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers*, *26*(2), 663–688. https://doi.org/10.1007/s10796-023-10390-w

Vasist, P. N., & Krishnan, S. (2022). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Communications of the Association for Information Systems*, *51*(1). https://doi.org/10.17705/1CAIS.05126

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186–204.

Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: Word embedding over linguistic features for fake news detection. *IEEE Transactions*

on Computational Social Systems, *8*(4), Article 4.

Vicario, M. D., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and Fake News: Early Warning of Potential Misinformation Targets. *ACM Transactions on the Web*, *13*(2), 1–22. https://doi.org/10.1145/3316809

Vogel, I., & Jiang, P. (2019). Fake News Detection with the New German Dataset "GermanFakeNC". In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Hrsg.), *Digital Libraries for Open Knowledge* (Bd. 11799, S. 288–295). Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_25

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, *10*(3152676), 10–5555.

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In J. Vom Brocke, A. Hevner, & A. Maedche (Hrsg.), *Design Science Research*. Cases, Cham.

Vom Brocke, J., Stein, A., Hofmann, S., & Tumbas, S. (2015). *Grand Societal Challenges in Information Systems Research and Education: Ideas from the ERCIS Virtual Seminar Series*. Springer International Publishing. https://doi.org/10.1007/978-3-319-15027-7

Vraga, E. K., & Bode, L. (2020). Correction as a Solution for Health Misinformation on Social Media. *American Journal of Public Health*, *110*(S3), S278–S280. https://doi.org/10.2105/AJPH.2020.305916

Waddell, T. F. (2018). What does the crowd think? How online comments and popularity metrics affect news credibility and issue importance. *New Media & Society*, *20*(8), 3068–3083. https://doi.org/10.1177/1461444817742905

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, *22*(11), 1531–1543. https://doi.org/10.1080/1369118X.2019.1648536

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Information Systems Research*, *3*(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wang, L. (2005). *Support vector machines: Theory and applications* (Bd. 177). Springer Science & Business Media. https://books.google.de/books?hl=de&lr=&id=uTzMPJjVjsMC&oi=fnd&pg=PA1&dq=Lipo+Wang.+2005.+Support+vector+machines:+theory+and+applications,+volume+177.+Springer+Science+%26+Business+Media.&ots=GGDF5wWGnc&sig=tY3x3CvNfA5gE2ziWcoz0iiD0s4

Wang, L., Wang, Y., de Melo, G., & Weikum, G. (2019). Understanding archetypes of fake news via fine-grained classification. *Social Network Analysis and Mining*, *9*(1), 1–17.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-Generated

Images Are Surprisingly Easy to Spot… for Now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8692–8701. https://doi.org/10.1109/CVPR42600.2020.00872

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets*, *32*(4), 2079–2102. https://doi.org/10.1007/s12525-022-00593-5

Wardle, C. (2019). A new world disorder. *Scientific American*, *321*(3), 88–95.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Bd. 27). Council of Europe Strasbourg.

Warren, G., Shklovski, I., & Augenstein, I. (2025). Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. *arXiv preprint arXiv:2502.09083*.

Washington, A. L., & Kuo, R. (2020). Whose side are ethics codes on? Power, responsibility and the social good. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 230–240. https://doi.org/10.1145/3351095.3372844

Washington, J. (2023). *Combating Misinformation and Fake News: The Potential of AI and Media Literacy Education* (SSRN Scholarly Paper 4580385). https://doi.org/10.2139/ssrn.4580385

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129–140.

Weber, S., & Knorr, E. (2020). Kognitive Verzerrungen und die Irrationalität des Denkens. *Die Psychologie des Postfaktischen: Über Fake News,„Lügenpresse “, Clickbait & Co.*, 103–115.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Wiriting a Literature Review. *MIS Quarterly*, *26*(2), 13–23.

Wei, X., Zhang, Z., Zhang, M., Chen, W., & Zeng, D. D. (2019). Combining Crowd and Machine Intelligence to Detect False News on Social Media. *MIS Quarterly*.

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

Weikmann, T., & Lecheler, S. (2024). Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network. *Digital Journalism*, *12*(10), 1505–1522. https://doi.org/10.1080/21670811.2023.2194665

Weinhardt, C., Fegert, J., Hinz, O., & van der Aalst, W. M. P. (2024). Digital Democracy: A Wake-Up Call. *Business & Information Systems Engineering*, *66*(2), 127–134. https://doi.org/10.1007/s12599-024-00862-x

Weismueller, J., Gruner, R. L., Harrigan, P., Coussement, K., & Wang, S. (2024).

Information sharing and political polarisation on social media: The role of falsehood and partisanship. *Information Systems Journal*, *34*(3), 854–893. https://doi.org/10.1111/isj.12453

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). *„ Do you trust me?“ Increasing user-trust by integrating virtual agents in explainable AI interaction design*. 7–9.

Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.550030

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, *9*(11).

Westfall, C. (2025). *Meta Opens Floodgates For AI-Generated Accounts On Facebook, Instagram*. Forbes. https://www.forbes.com/sites/chriswestfall/2025/01/02/meta-opens-floodgates-on-ai-generated-accounts-on-facebook-instagram/

Wicks, A. C., Berman, S. L., & Jones, T. M. (1999). The Structure of Optimal Trust: Moral and Strategic Implications. *The Academy of Management Review*, *24*(1), 99. https://doi.org/10.2307/259039

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). *Large language models can consistently generate high-quality content for election disinformation operations* (arXiv:2408.06731). arXiv. https://doi.org/10.48550/arXiv.2408.06731

Wills, M. (2017, November 7). *How the Sun Conned the World With „The Great Moon Hoax“*. JSTOR Daily. https://daily.jstor.org/how-the-sun-conned-the-world-with-the-moon-hoax/

Wölker, A., & Powell, T. E. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, *22*(1), 86–103. https://doi.org/10.1177/1464884918757072

Woolley, S. C., & Howard, P. N. (2016). Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents—Introduction. *International Journal of Communication*, *10*(0), Article 0.

World Economic Forum. (2024). *Global Risks Report 2024*. World Economic Forum. https://www.weforum.org/publications/global-risks-report-2024/

Wright, R. R., Sandlin, J. A., & Burdick, J. (2023). What is critical media literacy in an age of disinformation? *New Directions for Adult and Continuing Education*, *2023*(178), Article 178.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explor. Newsl.*, *21*(2), 80–90. https://doi.org/10.1145/3373464.3373475

Xu, Q., Feng, Z., Gong, C., Wu, X., Zhao, H., Ye, Z., Li, Z., & Wei, C. (2024). Applications of explainable AI in natural language processing. *Global Academic*

Frontiers, *2*(3), 51–64.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). *Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia* (arXiv:1812.06280). arXiv. https://doi.org/10.48550/arXiv.1812.06280

Yan, Z., & Holtmanns, S. (2008). Trust modeling and management: From social trust to digital trust. In *Computer security, privacy and politics: Current issues, challenges and solutions* (S. 290–323). IGI Global Scientific Publishing. https://www.igi-global.com/chapter/trust-modeling-management/6870

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., & Hu, X. (Ben). (2019). XFake: Explainable Fake News Detector with Visualizations. *The World Wide Web Conference*, 3600–3604. https://doi.org/10.1145/3308558.3314119

Yates, S., Carmi, E., Lockley, E., Wessels, B., & Pawluczuk, A. (2021). *Me and My Big Data: Understanding Citizens Data Literacies Research Report* [Report]. University of Liverpool. https://livrepository.liverpool.ac.uk/3180002

Yates, S., Carmi, E., Pawluczuk, A., Lockley, E., Wessels, B., & Gangneux, J. (2020). *Understanding citizens data literacy: Thinking, doing & participating with our data* (Me and My Big Data). University of Liverpool.

Yu, S., & Lo, D. (2020). Disinformation Detection using Passive Aggressive Algorithms. *ACM Southeast Conference, Session 4*, 324f.

Zahodne, L. B., Glymour, M. M., Sparks, C., Bontempo, D., Dixon, R. A., MacDonald, S. W. S., & Manly, J. J. (2011). Education Does Not Slow Cognitive Decline with Aging: 12-Year Evidence from the Victoria Longitudinal Study. *Journal of the International Neuropsychological Society*, *17*(6), 1039–1046. https://doi.org/10.1017/S1355617711001044

Zhang, D., Li, C., Lin, F., Zeng, D., & Ge, S. (2021). Detecting Deepfake Videos with Temporal Dropout 3DCNN. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2021/178

Zhang, Y., Lukito, J., Su, M.-H., Suk, J., Xia, Y., Kim, S. J., Doroshenko, L., & Wells, C. (2021). Assembling the Networks and Audiences of Disinformation: How Successful Russian IRA Twitter Accounts Built Their Followings, 2015–2017. *Journal of Communication*, *71*(2), 305–331. https://doi.org/10.1093/joc/jqaa042

Zhang, Y., Lukito, J., Suk, J., & McGrady, R. (2024). Trump, Twitter, and Truth Social: How Trump used both mainstream and alt-tech social media to drive news media attention View supplementary material. *Journal of Information Technology & Politics*. https://doi.org/10.1080/19331681.2024.2328156

Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021). Multi-attentional Deepfake Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2185–2194.

https://doi.org/10.1109/CVPR46437.2021.00222

Zhao, W., Rachwalski, A., Berndt, M., & Jin, Y. (2025). An Examination of Management of AI-Triggered Organisational Threats From Communication Practitioners' Perspective. *Journal of Contingencies and Crisis Management*, *33*(1). https://doi.org/10.1111/1468-5973.70031

Zhou, C., Li, K., & Lu, Y. (2021). Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Information Processing & Management*, *58*(6), 102679.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, *13*(1), 81–106.

Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2019). Fake news early detection: An interdisciplinary study. *arXiv preprint arXiv:1904.11679*.

Zimmermann, F., & Kohring, M. (2018). „Fake News" als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs. *Medien & Kommunikationswissenschaft*, *66*(4), 526–541. https://doi.org/10.5771/1615-634X-2018-4-526

Zorn, I. (2011). Zur Notwendigkeit der Bestimmung einer auf Digitale Medien fokussierten Medienkompetenz und Medienbildung. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, *20*, 175–209. https://doi.org/10.21240/mpaed/20/2011.09.19.X

List of Abbreviations

AC	Attention Check
ADR	Action Design Research
AI	Artificial Intelligence
API	Application Programming Interface
BMBF	German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung)
CMC	Computer-Mediated Communication
CML	Critical Media Literacy
CMS	Content Management System
CNN	Convolutional Neural Network
DCMT	Dublin Core Metadata Terms
DFG	German Research Foundation (Deutsche Forschungsgemeinschaft)
DL	Deep Learning
DL (Chapter 10)	Data Literacy
DP	Design Principle
DSA	Digital Services Act
DSR	Design Science Research
EU	European Union
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HCI	Human-Computer Interaction
IAA	Inter-Annotator Agreement
ICT	Information and Communication Technology
IMC	Instructional Manipulation Check
IMT	Information Manipulation Theory
IRA	Internet Research Agency
IS	Information Systems
IT	Information Technology

LIME	Local Interpretable Model-Agnostic Explanations
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
MR	Meta-Requirement
NGO	Non-Governmental Organization
NLP	Natural Language Processing
OSN	Online Social Network
PR	Public Relations
RDF	Resource Description Framework
RNN	Recurrent Neural Network
RQ	Research Question
SHAP	Shapley Additive Explanations
SKOS	Simple Knowledge Organization System
SVM	Support Vector Machine
U.S.	United States
XAI	Explainable Artificial Intelligence

List of Figures

List of Tables