Journal of Innovation & Knowledge 10 (2025) 100782

Contents lists available at ScienceDirect

Journal of Innovation & Knowledge

journal homepage: www.elsevier.com/locate/jik

Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European Countries

Nik Hynek [a], Beata Gavurova [b,*], Matus Kubak [c]

[a] Department of Security Studies, Faculty of Social Sciences, Charles University in Prague, Prague, Czechia
[b] Technical University of Košice, Faculty of Mining, Ecology, Process Control and Geotechnologies, Košice, Slovakia
[c] Technical University of Košice, Faculty of Economics, Košice, Slovakia

ARTICLE INFO

JEL classifications:
O32
O33
O35
O38

Keywords:
Generative artificial intelligence
Deepfakes
AI-Generated hyper-realistic audio-visual digital content
SWOT/Risks and benefits
Cross-National survey
Public opinion

ABSTRACT

This study provides an evidence-based integrated appraisal of artificial intelligence (AI)-generated deepfakes by integrating a cross-disciplinary literature synthesis with original opinion-poll evidence from seven European countries. A SWOT matrix distils convergent concerns—weaponised disinformation, privacy erosion, and the detection arms race—alongside under-explored opportunities in education, therapy, and creative industries. To test whether these scholarly themes resonate with citizens, a computer-assisted web survey (N = 7,083) measured perceived risks and benefits across 10 specific scenarios for each theme. Correspondence analysis and Bonferroni-adjusted means reveal a pronounced age gradient for benefits, whereas risk perceptions vary by country—younger cohorts are noticeably less alarmed only in Sweden, France, and Czechia. Geographically, Dutch, German, British, and Italian publics prove the most enthusiastic: the United Kingdom (UK) couples similar enthusiasm with markedly higher risk vigilance, whereas Czech and Swedish respondents remain consistently sceptical, underscoring a broad, though imperfect, west/south versus central/north divide. The Netherlands, Germany, the UK, and Italy value pro-social applications (i.e., realistic crisis drills, public-interest campaigns, and therapeutic 'mental-health' avatars), with the Netherlands topping four benefit items and Italy favouring commercial/entertainment uses such as virtual brand ambassadors. By contrast, Czech and Swedish respondents assign uniformly low benefit scores. Juxtaposed with risk perceptions, the UK and Czechia register the greatest vigilance, Sweden the most relaxed, and others intermediate. Divergence seems associated with digital literacy levels and regulatory maturity. The survey reveals a statistically and practically significant gap between perceived risks and benefits: across all seven countries, respondents, on average, rate risks higher than advantages. Regression estimates indicate that advancing age, lower household income, and gender (woman) enlarge this gap—primarily by undermining perceived benefits—whereas tertiary education and residence in certain western or southern European countries—notably, Germany and Italy—are associated with more balanced appraisals. This study concludes that layered governance, interoperable detection standards, and targeted literacy programmes are urgently required.

Introduction

Several critical factors drive the rapid proliferation of deepfakes and generative artificial intelligence (AI) technologies. First, the technological breakthroughs in generative adversarial networks (GANs) and deep learning have substantially lowered the impediments to creating hyper-realistic audio-visual forgeries, enabling even novice users to produce advanced deepfakes (Morris, 2024; Domenteanu et al., 2024). The abundance of data on social media platforms provides abundant material for these algorithms to learn and refine, further enhancing the generated content's perceived authenticity (Morris, 2024). Additionally, the broad accessibility of generative AI tools has democratised the production of deepfakes, allowing them to be employed for both benign and malicious ends, including misinformation, identity fraud, and political manipulation (Paterson, 2024; Singh & Dhiman, 2023). The societal ramifications of deepfakes are intensified by the swift dissemination mechanisms of social media, wherein fabricated videos can rapidly reach millions of users, thereby influencing public perception and eroding trust in information (The Rise of Deepfake Technology, 2023; Boté-Vericad & Vállez, 2022). The ethical and legal complications

* Corresponding author.
E-mail addresses: hynek@fsv.cuni.cz (N. Hynek), beata.gavurova@tuke.sk (B. Gavurova), matus.kubak@tuke.sk (M. Kubak).

https://doi.org/10.1016/j.jik.2025.100782
Received 4 May 2025; Accepted 7 August 2025
Available online 23 August 2025
2444-569X/© 2025 The Authors. Published by Elsevier España, S.L.U. on behalf of Journal of Innovation & Knowledge. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

posed by deepfakes are exacerbated by the challenge of detecting and regulating such material, as existing detection methods struggle to keep up with evolving deepfake techniques (Supriya et al., 2024; Hao, 2024). Moreover, the geopolitical climate—shaped by events such as the COVID-19 pandemic and Russia-Ukraine conflict—has accelerated the use of deepfakes for propaganda and misinformation, further fuelling their proliferation (Domenteanu et al., 2024). The convergence of these technological, social, and political elements emphasises the complex domain of deepfakes and generative AI, demanding a wide-ranging response involving technological innovation, regulatory measures, and public awareness to mitigate detrimental effects (Paterson, 2024; Ferrara, 2024).

Deepfakes—emerging from advanced AI and deep learning—endanger democracy, cybersecurity, social trust, and personal rights while generating multiple ethical and legal issues. Research suggests that these fabricated media forms undermine democratic processes by spreading misinformation and influencing public opinion, notably during election campaigns (Diakopoulos & Johnson, 2021), as observed in 11 countries in 2023 (Łabuz & Nehring, 2024). Their threat is amplified in developing regions, where slower technological progression and limited public knowledge heighten the harmful effects (Noor et al., 2024). Moreover, deepfakes compromise individuals by creating convincing yet misleading portrayals, including insertion into pornographic material (Fehring & Bonaci, 2023; Öhman, 2020). Their impact extends to cybersecurity, as they can alter public perception and endanger institutional stability, thereby spurring the demand for effective detection tools and legal frameworks (Fehring & Bonaci, 2023). The Political Deepfakes Incidents Database documents their prevalence and significance and offers valuable insights to policymakers and researchers (Walker et al., 2024). Although AI may enhance political discourse, manipulative deepfakes precipitate ethical dilemmas and cast doubt on the authenticity of democratic processes (Battista, 2024; Flattery & Miller, 2024; Pawelec, 2022). Consequently, mitigating these challenges requires an inclusive approach that integrates technological safeguards, public education, and legal interventions to uphold democratic values and social confidence (Noor et al., 2024; Fehring & Bonaci, 2023).

Deepfakes, as a communication innovation, carry meaningful advantages for society, creativity, and positive applications, despite their controversial aspects. They help generate lifelike synthetic media, which can serve multiple beneficial purposes. For example, deepfakes can foster improvements in creative industries by helping filmmakers and artists produce otherwise impractical or prohibitively expensive content, stimulating breakthroughs in media production (Whittaker et al., 2023). In education, deepfakes can create immersive and engaging experiences (e.g. historical re-enactments and language learning tools) that enhance learning outcomes and broaden access (Wazid et al., 2024). Additionally, they can safeguard intellectual property through systems such as FORGE, which can mislead intruders attempting to steal sensitive data, thereby bolstering cybersecurity (Alanazi et al., 2024). On a societal level, deepfakes can facilitate community-led initiatives and awareness campaigns to encourage unity and understanding (Machin-Mastromatteo, 2023). Moreover, digital health communities can employ them to tailor motivational content that inspires healthy behaviours and community participation (Nguyen et al., 2023). While they pose risks such as misinformation and privacy breaches, responsible frameworks and ethical standards can harness deepfakes' potential to produce various social benefits (Alanazi et al., 2024; Wazid et al., 2024; cf. Cavedon-Taylor, 2024). Consequently, despite inherent challenges, deepfakes' capacity to stimulate novel ideas and yield beneficial results underscores their relevance in modern society.

Scholarship on deepfakes remains divided across various fields, such as law, political science, media and communication studies, and computer science, creating a gap that calls for a unifying framework to synthesise these differing approaches. As existing empirical research on public perceptions of deepfakes is scarce, particularly on a multi-country European scale, this study aims to address three core research questions (RQs), while primarily aiming to construct an integrated framework that (a) consolidates disparate disciplinary insights on deepfakes and (b) assesses their real-world resonance through original, cross-national survey evidence, thereby supplying a coherent evidentiary foundation for scholarship and policy.

RQ1: How have deepfakes been studied across social sciences, humanities, and technical fields? By examining diverse approaches and focuses, we can establish a structured overview of the academic literature on deepfakes.

RQ2: What are the strengths, weaknesses, opportunities, and threats identified by different disciplines? Where do they converge or diverge? A comparative SWOT analysis can offer insight into each domain's major contributions and blind spots, revealing overlapping concerns and unique perspectives.

RQ3: What do European citizens believe about deepfakes and generative AI, and how do demographic, cultural, and literacy factors shape these perceptions? This empirical component highlights the importance of understanding not merely expert debates but also real-world beliefs and experiences. By integrating fragmented scholarship and presenting novel primary data in the form of empirical findings from our own statistically representative population sample in seven European countries, this study offers a cohesive, cross-disciplinary exploration of deepfake technologies, stressing the urgent need for informed policy decisions, ethical regulations, and public education initiatives that address the complexity of these evolving challenges.

This article's structure presents a strategic flow designed to link the literature review, comparative assessments, and empirical findings into a coherent whole. Section 1 offers a detailed literature review, highlighting foundational studies and disciplinary perspectives on deepfakes. Section 2 introduces a SWOT analysis that identifies central strengths, weaknesses, opportunities, and threats, drawing together the insights gathered from diverse academic fields. Section 3 presents the empirical findings derived from original opinion-poll data in seven European countries, offering an intricate view of how Europeans perceive and experience deepfakes and generative AI. Concluding remarks reflect on policy, ethical implications, and broader social consequences, thereby delivering practical guidance for stakeholders. Collectively, these sections situate deepfakes within a structured analytical framework, elucidate empirical realities, and propose strategies to inform future inquiry and decision-making within the European context.

Literature review and categorisation

This section systematically reviews and categorises existing scholarly research on deepfakes, emphasising each field's unique perspectives. Subsections highlight major debates, landmark studies, and thematic gaps.

The discourse surrounding deepfakes in Security Studies and International Relations features considerable debates, notable research, and unresolved questions, particularly regarding national security (Sayler & Harris, 2023), disinformation campaigns, and political manipulation (MacCarthaigh & McKeown, 2021). Deepfakes—synthetic videos created with advanced AI—threaten state stability by enabling misleading content that inflames social tensions or sways public sentiment (Navarro Martínez et al., 2024). Agarwal et al. (2004) caution that forging videos of world leaders could prompt constitutional crises or civil unrest. Likewise, Taylor (2021) highlights the anxiety among Western democracies regarding deepfakes' potential to compromise electoral integrity and undermine defence strategies. These synthetic media tools influence the balance of power by shaping deterrence, facilitating false-flag operations, and enabling espionage. Chang et al. (2022) emphasise how warped information can be harnessed to aid military leaders, illustrating how deepfakes might achieve similar disruptive outcomes. Meanwhile, Landon-Murray et al. (2019) address the ethical complexities of disinformation in U.S. foreign policy, noting how deepfakes intensify concerns in the 'post-truth' age. Adding to this

body of work, scholars have examined the role of deepfakes in the broader sphere of misinformation and political manipulation, underscoring that these threats have spread rapidly since deepfakes emerged publicly in 2017 (Carvajal & Iliadis, 2020; Frankovits & Mirsky, 2023). Despite these advances, gaps persist in clarifying the full effect on media producers and the measures required to limit these threats, especially in hybrid warfare and diplomatic contexts.

In parallel, research focusing on hybrid warfare and diplomacy points out the asymmetric risks posed by deepfakes, highlighting how adversaries exploit synthetic media to destabilise political climates and erode trust in digital networks (Veerasamy & Pieterse, 2022). Beretas (2020a; 2020b) frames deepfakes as a critical component of cyber hybrid warfare, stressing their capacity to undermine societal cohesion. Chemerys highlights the urgency of strengthening media literacy so that the public can recognise and reject manipulative content disseminated through deepfake channels (Chemerys, 2024). Building on the technological side, Katarya and Lal (2020) spotlight GANs as the driving force behind deepfakes, proposing SSTNet as a leading detection model while also noting that document and signature forgeries remain under-examined. Weikmann and Lecheler (2023) observe that while deepfakes are acknowledged as a looming challenge for fact-checkers, other forms of misleading material still pose immediate hurdles, suggesting that current countermeasures are not yet fully equipped for synthetic media. Patel et al. (2023) emphasise the need to deepen understanding of deepfake generation processes so that optimised detection tools can be deployed effectively. Collectively, these studies affirm that increased collaboration among policymakers, technologists, and civil society groups is essential for crafting robust detection methods and digital awareness initiatives. Consequently, stakeholders can more effectively address deepfakes' corrosive impact on security, diplomacy, and public trust (Ali et al., 2022).

The emergence of deepfakes has sparked significant debates within legal science, around the regulation of audio-visual manipulation and the challenges posed by advanced AI technologies. Deepfakes, which are hyper-realistic synthetic media created via deep learning, raise concerns over privacy rights, intellectual property, defamation, and the integrity of information in the digital age (Mustak et al., 2023; Milliere, 2022). Their legal implications are profound, as they can spread misinformation, influence elections, and enable non-consensual pornography, affecting social relationships, democracy, and the rule of law (Sloot & Wagensveld, 2022; Karasavva & Noorbhai, 2021). The European Union's (EU) evolving regulatory framework, including the AI Act and the Digital Service Act, aims to assign responsibilities to Very Large Online Platforms and establish AI standards (Sloot & Wagensveld, 2022). However, these measures are deemed insufficient, prompting calls for amendments to privacy and data laws, plus revised free speech policies to limit the harm of deepfakes (Sloot & Wagensveld, 2022). Landmark studies emphasise the necessity for robust detection and prevention methods, along with ethical guidelines for media and entertainment usage (Wang et al., 2022a; Lees et al., 2021; Lu & Yuan, 2024). Despite efforts, research gaps persist around legal definitions, enforcement, and support for victims (Karasavva & Noorbhai, 2021; Vizoso et al., 2021).

Regulating deepfakes is a complex and ongoing issue, prompting debates over definitions, enforcement, and other hurdles posed by AI-generated synthetic media. These ultra-realistic forgeries threaten multiple sectors, including criminal justice, where they can taint evidence and erode institutional trust (Sandoval et al., 2024). The absence of a universal legal definition hampers regulatory consistency, as shown by varying U.S. state laws that focus on factors such as AI and falsified media (Meneses, 2024). Similar shortcomings emerge in Canadian policy, which fails to address deepfake pornography effectively and calls for explicit language on non-consensual synthetic material (Karasavva & Noorbhai, 2021). Rapid dissemination through social media intensifies these threats, resulting in marketplace deceptions and personal harms such as identity theft (Mustak et al., 2023). Suggested solutions include privacy law reforms, constraints on certain expressions, and pre-emptive rules on deepfake technologies (Sloot & Wagensveld, 2022). Beyond these policy concerns, therapeutic contexts such as grief counselling present additional ethical and legal issues tied to reality distortion (Hoek et al., 2024). As calls for enhanced detection frameworks mount, researchers emphasise that these systems help curb misinformation and misuse (Rathoure et al., 2024). In sum, the legal regulation of deepfakes demands an approach that balances innovation with protective safeguards.

The EU's response to deepfakes intersects with privacy, intellectual property, defamation, and the wider regulatory landscape, including the GDPR, AI Act, and DSA. The AI Act attempts to define deepfakes and mandate transparency but faces critiques for overlooking certain scenarios, such as non-consensual pornography, leaving gaps in enforcement (Łabuz, 2024; 2023). Privacy breaches remain central, as the unauthorised use of personal data can lead to reputational harm and emotional distress. Existing rules, such as Article 1032 of the Civil Code and the Personal Information Protection Law, provide limited clarity, complicating protection efforts (Han, 2024). Moreover, deepfakes undermine democratic norms by diminishing trust; however, current EU measures are deemed inadequate (Karunian, 2024). Developer accountability has been proposed to limit harm, directing attention toward ethical and governance-based strategies (Pawelec, 2024; Rini & Cohen, 2022). While the AI Act represents a pioneering step, experts suggest stronger definitions and improved oversight to tackle AI-driven synthetic media (Hacker, 2023). Likewise, the DSA obligates Very Large Online Platforms to control harmful media, though its effectiveness remains uncertain (Birrer & Just, 2024). Subsequently, cooperation among regulators, technologists, and civil society is key to ensuring that EU laws evolve alongside fast-paced technological advancements.

From a Science and Technology Studies (STS) viewpoint, deepfakes offer varied opportunities and concerns, reflecting the dynamic relationship between technology and society. On the one hand, they can energise fields such as entertainment and education by enabling realistic simulations and re-enactments, enhancing creative expression and learning experiences (Stewart & Williams, 2000). On the other hand, their hyper-realistic form can be exploited to spread disinformation and propaganda, contributing to public confusion and undermining trust in institutions (Tahir et al., 2021). This mixed potential illustrates how social values shape—and are shaped by—new technologies. Limited public knowledge about deepfakes—often influenced by age, gender, and education—further amplifies the risk of misuse, indicating a growing need for digital literacy initiatives (Seibert et al., 2024; Tahir et al., 2021; Fallis, 2020; Habgood-Coote, 2023). In this context, an STS emphasis on co-evolution reminds us that technological development is non-deterministic, shaped by cultural, ethical, and political factors (Boerwinkel et al., 2014; Keulartz et al., 2004). By recognising deepfakes as more than mere technical artefacts, scholars encourage participatory mechanisms that include diverse stakeholders, thereby addressing gaps in current STS and ethical frameworks. Such measures can strengthen both societal understanding and regulatory approaches, reducing the risks while harnessing beneficial applications of deepfake technologies (Sabanovic, 2010).

Socio-technical imaginaries around deepfakes reflect both promising and troubling narratives, shaping how the public perceives their potential uses. For instance, human-computer interaction can benefit from the technology, as seen in design tasks where deepfake personas improve user engagement, though perceptual glitches still pose challenges (Kaate et al., 2024). Similarly, digital resurrection in campaigns such as 'Listen to My Voice' demonstrated the ability of deepfakes to evoke strong emotional reactions and raise social awareness, even if these efforts sparked ethical questions about autonomy and posthumous consent (Lowenstein et al., 2024). Balancing these positives, deepfakes remain closely linked to misinformation and the erosion of trust in digital content, prompting the development of countermeasures to minimise harm (Lyu, 2024). A general lack of familiarity with

deepfakes—especially among women—reinforces negative perceptions, calling attention to the influence of sociodemographic factors in shaping attitudes (Seibert et al., 2024). Emotional responses, such as fear or interest, play a pivotal role in these imaginaries, as demonstrated in the Irish anti-fracking context (Hughes, 2024). Deepfakes can also offer possibilities for activism or self-expression; however, concerns still arise over identity control and surveillance (Doğan Akkaya, 2024). These contrasting visions highlight the importance of ethical guidelines to steer deepfake innovation toward beneficial societal outcomes.

Deepfakes, emerging from deep learning and generative AI, present a dual scenario for the economy by introducing novel markets for AI-driven synthetic content and simultaneously posing significant challenges for creative industries (Akbar et al., 2023). On the positive side, this technology has the potential to revolutionise audio-visual production, lowering costs and enabling realistic media creation across entertainment, advertising, and education (Milliere, 2022; Kietzmann, Mills, & Plangger, 2021). Such possibilities could spark fresh economic growth, as new markets for AI-generated content evolve. However, the unauthorised use of creators' works for training data raises serious copyright and ownership issues, prompting lawsuits that may constrain generative AI to the public domain or specially licensed content (Samuelson, 2023). This tension could dampen innovation, limiting the economic benefits of emerging AI applications. Beyond intellectual property disputes, deepfakes can damage media credibility, undermining trust and leading to economic fallout in fields that rely on reputable content, such as journalism and advertising (Weikmann et al., 2024). Even when viewers do not fully accept these synthetic clips as real, they can harm individual and corporate reputations, resulting in financial losses (Harris, 2021). While certain studies remain optimistic about the creative potential of deepfakes, the overall economic balance hinges on carefully managing the risks and opportunities (Lu & Yuan, 2024; Patel et al., 2023).

Tackling deepfakes in business and cybersecurity introduces a blend of promise and peril, attesting to the need for robust detection and strategic planning. From a corporate perspective, these hyper-realistic media tools can elevate brand engagement through personalised marketing while simultaneously risking severe brand harm if misused (Mustak et al., 2023). In the cybersecurity realm, advanced methods such as SecureVision combine big data analytics and deep learning to identify and neutralise synthetic content (Kumar & Kundu, 2024). Complementary measures, such as perceptual-aware perturbations and decoy tactics, aim to thwart adversarial manipulation, demonstrating the growing sophistication of counter-technologies (Jointly Defending DeepFake, 2023; Wang et al., 2022b; Ding et al., 2023). Blockchain and distributed ledgers also feature in emerging strategies to validate content authenticity, showing that innovative solutions can help protect high-profile individuals from deepfake impersonations (Gambin et al., 2024; Boháček & Farid, 2022). The financial stakes are high: reputational damage and consumer mistrust may trigger significant losses, while advanced detection systems demand ongoing investment (Domenteanu et al., 2024). Against this backdrop, the public and private sectors weigh costs and benefits, assessing how deepfakes may be harnessed responsibly for digital forensics or creative content (Wang, 2023; Lyu, 2024). Ultimately, striking a balance between capitalising on deepfake innovations and safeguarding against their misuse is essential for sustaining economic growth and securing online environments.

From a sociological angle, deepfakes hold implications for social norms, collective trust, and the ways communities share and process information. In particular, they can amplify societal polarisation by disseminating false narratives that reinforce existing prejudices or sow discord (Verma, 2024; Wan, 2023). Although some argue that the credibility of a video hinges more on its source than its content (Harris, 2021), deepfakes may still disturb viewers psychologically, influencing memory formation and social interactions. This effect is seen when moviegoers find their shared viewing experiences altered by synthetic edits (Murphy et al., 2023). Complicating matters, deepfakes often trigger a third-person effect, where individuals believe they themselves are less likely to be misled than others (Ahmed, 2023a). Meanwhile, the entertainment value of these highly realistic fabrications can normalise deception, potentially dulling ethical concerns among the general public (Wan, 2023). Sociologists also point out that deepfakes undermine collective trust in audio-visual evidence, placing fresh burdens on institutions—such as news outlets—that rely on reliable video material (Shin & Lee, 2022; Whyte, 2020). Even if deepfakes are not always more convincing than other types of disinformation (Hameleers et al., 2022), their capacity to spread quickly on social media intensifies the danger they pose to communal beliefs and behaviours (Karpinska-Krakowiak & Eisend, 2024). Consequently, regulatory measures, technical defences, and digital literacy campaigns have emerged as essential avenues for mitigating these societal risks (McCosker, 2022; Mustak et al., 2023).

Media and communication studies examine deepfakes in terms of framing, virality, and moral panic (Brooks, 2021; Godulla, Hoffmann, & Seibert, 2021). Public discourse often paints them as a dire threat to democratic institutions, given their capacity to fabricate video or audio content that seems genuine (Fehring & Bonaci, 2023; Cover, 2022). This portrayal can heighten alarm, especially as society grapples with the blurred line between reality and fabrication (Broinowski, 2022; Momeni, 2024). Viral circulation on social platforms magnifies the impact, allowing manipulated clips to reach vast audiences swiftly, as observed during the 2020 U.S. Presidential Election (Prochaska et al., 2023; Sorell, 2023). Audiences typically struggle to identify altered material, which can produce public confusion and even incite real-world consequences (Momeni, 2024; Hameleers et al., 2023a). Despite this bleak narrative, some scholars call for a more measured view, noting that deepfakes may be harnessed for creative or educational ends (Cover, 2022; Broinowski, 2022). Ultimately, human agency shapes how these tools are used; however, fragmented research and limited theoretical models still make it difficult to fully grasp deepfakes' cultural impact (Vasist & Krishnan, 2022). Media literacy programmes, clear ethical guidelines, and tighter oversight represent potential remedies (Holzschuh, 2023; Beridze & Butcher, 2019). By balancing vigilance with an openness to positive applications, media specialists aim to address the complexities of deepfake technologies (Shahodat, 2022).

Psychological research focuses on the cognitive processes that affect individuals' detection and resistance to deepfake manipulation (Ask et al., 2023). Analytical thinkers tend to identify fabricated content more readily, assigning lower credibility to suspicious videos and images (Pehlivanoglu et al., 2024; Hameleers et al., 2023b). Similarly, cognitive flexibility—defined as the ability to adapt thinking processes—enables better detection accuracy and a more reliable judgement of one's own performance (The role of cognitive styles and cognitive flexibility, 2023). Familiarity with the subject matter also boosts detection rates: individuals are more adept at spotting distortions involving celebrities or topics they already know (Allen et al., 2023; 2022). Simultaneously, a 'truth bias' leads several people to assume that the incoming information is accurate, making them vulnerable to manipulation (Pehlivanoglu et al., 2024). Emotional reactions can heighten susceptibility; for instance, unsettling content featuring deceased personalities may trigger discomfort, further clouding judgement (Soto-Sanfiel & Wu, 2024). Real-world conditions such as low-quality footage or divided attention amplify the likelihood of missing deepfake cues, hinting that experimental data may underestimate genuine vulnerability (Josephs et al., 2023; Artifact magnification, 2023). Technological solutions, such as artefact magnification, can raise both detection accuracy and confidence, highlighting the combined importance of cognitive skills and supportive tools (Josephs et al., 2023).

Additional psychological considerations involve confirmation bias, motivated reasoning, and the emotional resonance of deepfakes. Individuals often seek information that aligns with their beliefs, ignoring contradictory or unsettling details (Dickinson, 2024). In the context of deepfakes, this bias can solidify harmful impressions, especially when

the fabricated content resonates with viewers' existing worldviews (Hameleers et al., 2023a, 2023b). Emotional investment heightens these effects, as strong reactions to ideologically charged or personally offensive materials can magnify confirmation bias (Dickinson, 2024). Consequently, political figures may be disaited more easily, and public trust in legitimate institutions can erode, especially if the manipulated content appears sufficiently realistic (Ruiter, 2021). Interestingly, some scholars argue that the threat posed by deepfakes is moderated by audiences' trust in the content's source: viewers may dismiss or doubt suspicious material if they have faith in its origin (Harris, 2021). However, third-person perception still emerges, with individuals believing that others are more susceptible than they are themselves (Ahmed, 2023b). Media literacy initiatives, focused on revealing how cheap and accessible deepfake technologies are, can help lower the veneer of authenticity and reduce undue influence (Shin & Lee, 2022). Overall, understanding these cognitive biases and emotional triggers highlights the urgency of interventions aimed at mitigating deepfakes' potentially disruptive psychological impact.

Deepfakes, driven by advanced AI methods such as GANs and diffusion models, have prompted significant debates in Computer Science and AI ethics because of their capacity to generate extremely convincing yet fabricated content (Laurier et al., 2024; Hao, 2024; Amerini et al., 2024; Chowdhury & Lubna, 2020). Detection techniques play a critical role in limiting the harm these creations can cause (Buo, 2020). Researchers have tested various strategies, including Convolutional Neural Networks (CNNs), to detect the fine-grained artefacts often hidden within deepfake media (Amerini et al., 2024; Guarnera et al., 2020). Optical flow-based CNNs assess motion inconsistencies in video sequences, effectively tackling cross-forgery scenarios (Amerini et al., 2019; Caldelli et al., 2021). Anomaly detection methods, such as self-adversarial variational autoencoders, help differentiate normal from anomalous latent variables, improving the recognition of manipulated material (Wang et al., 2020). Other innovations include dynamic prototype networks, which generate interpretable explanations for temporal anomalies in deepfake videos (Trinh et al., 2021). Model performance benefits further from data augmentation and transfer learning, yielding high precision and recall in real-world applications (Iqbal et al., 2023). However, the ongoing arms race between more sophisticated generation algorithms and detection tools still necessitates continued research to outpace malicious adaptation (Laurier et al., 2024; Hao, 2024).

Beyond detection, computer scientists have also focused on ethical frameworks to guide the responsible development and use of deepfake-generating AI. Such guidelines address a broad spectrum of ethical dilemmas, ranging from questions of privacy and authenticity to issues of fairness and broader societal impact. Ali and Aysan (2024) call for adaptive governance models that respond to the shifting ethical landscape of AI, highlighting the need for domain-specific oversight in education, healthcare, and industry. A separate scoping review by Hagendorff (2024) enumerates 378 normative concerns, emphasising the significance of fairness, content safety, and user privacy. In healthcare, Oniani et al. (2023) propose the 'GREAT PLEA' principles—governability, reliability, equity, accountability, traceability, privacy, lawfulness, empathy, and autonomy—to steer the ethical adoption of AI in clinical environments (Oniani et al., 2023). On the legislative front, the EU's AI Act aims to protect fundamental rights while setting international standards for AI governance (Gasser, 2023). Floridi's (2019) work on the Commission's guidelines for trustworthy AI adds further impetus to these endeavours, asserting that AI should advance human welfare and environmental stewardship. However, operationalising these frameworks remains challenging, as highlighted by Chen et al. (2023), who argue for interdisciplinary methods and inclusive user testing to embed morality into AI systems.

The listed outcomes of the research studies represent a foundation for further sub-operationalising the initial RQ3 into a set of six sub-RQs as follows:

RQ1: Does a consistent age gradient emerge across Europe in perceptions of the risks and benefits of deepfakes?

RQ2: Do European countries cluster in similar positions within the correspondence–analysis space based on their average risk and benefit assessments?

RQ3: Are there significant cross-national differences in perceptions of specific deepfake scenarios—political misinformation, legal evidence, and creative applications—and, if so, do countries rating a scenario as especially risky also view it as less beneficial?

RQ4: What thematic interdependencies can be identified among individual risk- and benefit-related perceptions of deepfake technologies as revealed by covariance analysis?

RQ5: Is there a statistically and practically significant difference between the perceived risks and perceived benefits of deepfake technologies among respondents?

RQ6: To what extent do demographic characteristics (age, gender, country of residence, educational attainment, and income level) explain variation in the difference between perceived risks and benefits of deepfake technologies?

Comparative SWOT analysis

This section presents a comparative SWOT analysis of deepfakes, drawing on insights from various disciplines such as Security Studies, International Relations, Law, Sociology, Media Studies, Psychology, and Computer Science. The aim is to reconcile different perspectives and highlight convergences, divergences, and gaps. Four criteria guide this comparative lens. First, each field emphasises primary concerns: national security, individual privacy, economic impact, or broader sociocultural issues. Second, proposed solutions range from regulatory approaches and technological innovations to educational and ethical frameworks. Third, conceptualisations of risk vary between immediate or long-term, targeted or global, and incremental or disruptive scenarios. Finally, forms of benefits cover creative industries, political engagement, activism, and novel art or entertainment applications. By applying these criteria, this analysis seeks to draw out the strengths and weaknesses of existing scholarship, pinpointing how each discipline addresses deepfake challenges and opportunities. The resulting synthesis uncovers areas of synergy, for instance, where legal strategies can inform technical solutions, as well as points of contention, such as differing views on whether social trust or national security should take precedence. Overall, this comparative approach highlights how deepfakes cut across multiple policy, cultural, and technological arenas, demanding a more unified understanding and a coordinated response.

Strengths across disciplines

Security Studies and International Relations highlight well-defined threat models that illuminate how deepfakes can manipulate public opinion, erode trust, and undermine global stability. They offer structured frameworks for analysing malicious uses, including disinformation campaigns and espionage. This clarity helps in prioritising immediate risks and informing governmental and intergovernmental strategies. Legal scholarship, meanwhile, excels in outlining potential avenues for regulation and liability, drawing on existing doctrines of defamation, privacy, and intellectual property. These fields also propose reforms to adapt current legislation and introduce responsibilities for content-hosting platforms. The synergy between these domains materialises in discussions of deterrence and enforcement: when legal guidelines reinforce security imperatives, states and multilateral bodies have clearer pathways to prosecute perpetrators or impose sanctions. Hence, the strengths in these areas lie in their focus on comprehensive risk assessments, potential litigation strategies, and the capacity to integrate measures across jurisdictions. They support the urgency of robust institutional frameworks—both to preserve national interests and to protect civil rights. Consequently, they point toward practical

approaches that address immediate threats without entirely stifling technological innovation.

Sociology and Media Studies excel at diagnosing how deepfakes shape social norms, collective trust, and media consumption practices. Their research highlights how false narratives become viral, how audiences may respond to manipulated content, and how moral panic can emerge. Such an understanding is crucial in designing public-awareness campaigns that speak directly to community values. Psychology brings a granular view of cognitive processes by revealing how biases—such as truth bias, motivated reasoning, and confirmation bias—amplify or mitigate susceptibility to deepfake content. This knowledge can guide interventions that improve individual resilience, such as targeted media literacy training for users who are especially prone to manipulation. Meanwhile, Computer Science research stands out for its robust detection methodologies, employing advanced algorithms and anomaly detection systems to spot synthetic content. This technical depth can bolster legal and policy measures by providing reliable means to identify illicit manipulations. Together, these fields demonstrate strong analytical capabilities and practical solutions. Sociology and Media Studies offer insights into public perception and cultural contexts, Psychology elucidates why some people are more affected than others, and Computer Science offers concrete techniques for thwarting malicious actors. These strengths collectively build a multifaceted skill set for understanding, detecting, and mitigating deepfake threats.

Weaknesses: fragmentation and limited scope

A recurring weakness across disciplines stems from their siloed nature. Security Studies might stress national-security related, geopolitical risks without fully capturing the human security aspect, that is, the personal harms experienced by individuals, such as non-consensual pornography or identity theft. Similarly, legal discussions can be narrowly focused on legislative loopholes and specific definitions, sometimes overlooking the everyday psychological toll of deepfake victimisation. Sociology and Media Studies risk underestimating the technical intricacies of generating and detecting synthetic content, limiting their capacity to propose concrete countermeasures. Furthermore, Computer Science research, while advanced in detection strategies, may overlook ethical and legal implications. These disciplinary blind spots testify to limited scope and fragmented research agendas that fail to connect the dots between, for instance, how psychological vulnerabilities can fuel security breaches or how legal frameworks might need to adapt to the evolving technical landscape. Interdisciplinary knowledge exchange is still sporadic, leaving several aspects of deepfakes unaddressed—especially regarding the experiences of marginalised communities, who may lack digital literacy in being exposed to complex manipulations. By operating in compartments, each discipline struggles to craft holistic approaches that can manage the broad range of threats posed by deepfakes.

Opportunities: innovative uses, collaborative governance, and enhanced public awareness

Despite significant concerns, deepfakes present opportunities to innovate in areas such as entertainment, education, and business. Sociologists and media scholars suggest that if used responsibly, these technologies could enrich interactive learning, enhance creative storytelling, and open new forms of artistic expression. Computer Science research indicates that advanced AI systems can accelerate scientific visualisation or training simulations, benefiting sectors ranging from healthcare to architecture. Meanwhile, economic analysis points to emerging markets for AI-driven content creation, stimulating job growth, and technological entrepreneurship. These positive potentials create compelling arguments for measured support of deepfake development, provided safeguards are in place. Moreover, opportunities manifest in the realm of social impact: community-driven campaigns leveraging deepfakes can highlight social issues, reach broader audiences, and provoke meaningful public discourse. Additionally, legal and policy frameworks might be fine-tuned to support ethical innovation by introducing flexible guidelines that protect against harm without stifling creativity or progress. In sum, deepfakes offer novel prospects for industry expansion, public engagement, and technological evolution—an outlook that can guide balanced policymaking focused on long-term benefits.

Another key opportunity across disciplines lies in forging collaborative governance models. International Relations research highlights the necessity of treaties or multinational agreements to prevent the misuse of deepfakes in propaganda and warfare. When combined with legal proposals for oversight, clear jurisdictional boundaries, and accountability for platforms, a global regulatory framework could emerge. These efforts would be strengthened by sociological and psychological studies identifying exactly where mis/disinformation thrives and how best to provide accurate counsel. Moreover, the synergy of computer scientists, social scientists, and policymakers can drive comprehensive public-awareness campaigns, teaching digital literacy through multifaceted interventions. Harnessing each discipline's expertise could also result in a standardised approach to labelling AI-generated content, enabling users to make informed decisions. An integrated coalition of academic researchers, industry stakeholders, government bodies, and civil society organisations may develop consistent guidelines and detection tools. This collaborative environment would not only curtail malicious actors but also nurture ethical, beneficial uses of deepfakes. By ensuring people understand the technology's full potential and pitfalls, it becomes easier to discourage exploitation while simultaneously encouraging positive innovation.

Threats: eroding trust and heightened risks, arms race, and ethical quagmires

From a broad perspective, deepfakes pose threats to democratic institutions, collective trust, and personal autonomy. Security analysts warn of state-sponsored disinformation that can destabilise regions, while legal experts caution that legislation is still insufficient or unevenly applied, leading to enforcement gaps. Sociologists point to the risk of moral panic, where the public may become overly suspicious of genuine media because of the fear of manipulation, further polarising societies. Media researchers suggest that virally spreading deepfakes can cause lasting reputational harm while simultaneously fuelling cynicism about all digital content. Technological arms races raise another concern: as deepfakes become more sophisticated, detection lags, compelling continuous research investment. There is also a risk of regulatory overreach, whereby policymakers enact broad restrictions that hamper legitimate creativity or stifle free expression. In a commercial sense, deepfake-related fraud undercuts consumer trust, potentially leading to significant economic ramifications. These multifaceted threats could converge into a perfect storm of social fragmentation, legal uncertainty, and inflated security costs. Thus, safeguarding against these hazards requires a careful balance of technological advancement, targeted regulation, and vigilant public education.

The accelerating arms race between deepfake generation tools and detection methodologies intensifies the threats faced by users and institutions alike. Every time a more advanced algorithm emerges, detection models must quickly adapt, and the cycle repeats itself. This constant cat-and-mouse dynamic forces stakeholders to invest heavily in research and development, potentially diverting resources from other critical areas. Ethical dilemmas add another layer of complexity. Even beyond malicious misuse, the sheer realism of deepfakes can blur the boundaries between consensual creations and exploitative manipulations, raising questions about identity rights and moral responsibility. If the technology becomes more widely accessible, unscrupulous individuals may produce deepfake content targeting personal relationships or private spheres, such as familial disputes, with potentially

devastating consequences. Furthermore, overdependence on automated detection can erode human vigilance, creating a false sense of security. If a detection system fails, it might be harder for humans to step in. Ultimately, these challenges underscore that threats are neither purely technical nor purely ethical; they encompass a broad range of dangers that compound across social, political, and individual domains, requiring cohesive, multi-layered responses.

Comparative analysis: similarities and differences

When synthesising insights across disciplines, clear similarities and differences emerge. A shared recognition is that deepfakes hold both disruptive and transformative potential, prompting calls for strong detection strategies and legal guidelines. However, each field prioritises differently: Security Studies focuses on national stability and disinformation, while media scholarship is more concerned with cultural impact, moral panic, and framing effects. Legal researchers examine concepts such as liability and regulatory gaps, whereas psychologists investigate cognitive susceptibility and emotional responses. Meanwhile, Computer Science concentrates on technological fixes, including detection models and anomaly identification. Some convergences include the universal emphasis on digital literacy, public awareness, and ethical guidelines as key elements of any solution. Divergences are seen in the scale of concerns—global for security experts, individual for psychologists—and in the pace of recommended interventions, with policymakers sometimes demanding immediate action while academic researchers advocate deeper empirical exploration. Overall, the synergy among disciplines rests on the acceptance that no single approach can fully address deepfakes. Only a tightly woven strategy, incorporating legal measures, technological breakthroughs, social education, and ethical considerations, can minimise the risks and harness potential benefits.

Empirical study: public perceptions in seven European countries

Data and research objectives

The study was conducted in partnership with Ipsos as part of an international survey spanning seven European countries. Its primary aim was to gauge public attitudes toward deepfake audio-visual content and, more broadly, AI. Data were collected through Computer-Assisted Web Interviewing with pre-recruited online panels; for the Czech sample, we used the Ipsos Populace.cz panel. Fieldwork was conducted from 10 to 17 February 2025 via a structured questionnaire that the respondents completed in roughly ten minutes. The final dataset comprises 7,083 respondents drawn from nationally representative samples in the Czech Republic (N = 1,005), the United Kingdom (UK) (N = 1,013), Germany (N = 1,000), France (N = 1,022), Italy (N = 1,014), Sweden (N = 1,011), and the Netherlands (N = 1,018). Each country's sample shows a balanced gender distribution: Czechia includes 493 men and 512 women, the UK 482 men and 531 women, Germany 493 men and 507 women, France 500 men and 522 women, Italy 497 men and 517 women, Sweden 513 men and 498 women, and the Netherlands 502 men and 516 women. The samples are likewise well balanced across age, education, monthly income, and settlement size.

Selecting Sweden, Italy, Germany, the Netherlands, France, Czechia, and the UK maximises structural variance within a compact European sample. The set spans the continent's cardinal points (Nordic Sweden, Mediterranean Italy, western Germany–the Netherlands–France, east-central Czechia, Atlantic UK), blends large (Germany, France, UK, Italy) and medium populations, contrasts a post-communist EU member with long-standing liberal democracies inside the EU and its leading external counterpart, and captures marked differences in economic models, technological capacity, educational outcomes, and dominant religious traditions. Such heterogeneity offers a rigorous basis for probing how institutional, socio-economic, and cultural contexts shape public attitudes while keeping the broader regional framework constant.

Methods and analysis

The first analytical step centres on the composite variable **Deepfakes Risk Perceptions**, derived from the respondents' evaluations of ten deepfake scenarios on a five-point Likert scale. Summing the item scores yields an index from 10 (lowest overall risk) to 50 (highest), providing a panoramic measure of concern about potential harms, ranging from misuse in media and politics to legal proceedings and social contexts.

Ten items on the questionnaire concerning risks (R) related to AI were as follows:

1. R1: A deepfake video depicts a prominent world leader announcing an unverified military action, potentially escalating global tensions.
2. R2: A deepfake is presented as evidence in court and may influence the outcome of a criminal or civil case.
3. R3: A deepfake campaign ad falsely portrays a candidate as supporting a controversial policy, influencing voter opinions.
4. R4: A deepfake video 'resurrecting' a deceased public figure is used for a public event without clear consent from their surviving relatives.
5. R5: A government or media corporation uses AI-generated 'news anchors' (deepfakes) to shape public opinion on certain policies.
6. R6: Companies deploy deepfake 'brand ambassadors' to personalise marketing, potentially distorting product claims.
7. R7: Risks of Deepfakes—a viral deepfake on social media accuses a celebrity or influencer of unethical behaviour, sparking public outrage.
8. R8: Deepfake content is deliberately designed to trigger emotional reactions (e.g., fear, anger) to manipulate public behaviour.
9. R9: Deepfake technologies are evolving faster than detection methods, fuelling an ongoing 'arms race'.
10. R10: A lack of regulation allows deepfake creators to freely publish deceptive content without disclosure.

The second analytical stage introduces the composite variable **Deepfakes Benefits Perceptions**, capturing the respondents' views on the potential advantages of deepfake technologies. This index aggregates ratings for ten benefit-oriented scenarios, each scored on a five-point Likert scale. Summing the item scores produces a scale from 10 (very low perceived benefit) to 50 (very high), enabling a clear quantification of how various social and demographic groups value technology's positive applications.

Ten items on the questionnaire towards benefits (B) related to AI were:

1. B1: Deepfakes are used in emergency preparedness exercises, creating realistic simulations for more effective training of military and diplomatic teams.
2. B2: Deepfake technologies are used to protect witnesses by altering their facial identity and voice, allowing them to testify anonymously without fear of retaliation.
3. B3: Deepfakes serve as interactive political-education tools, enabling students and citizens to virtually 'interview' AI-generated avatars of historical experts.
4. B4: A deepfake 'resurrection' is created for educational/historical exhibits, bringing long-deceased figures to life for museum visitors with a clear disclosure that it is AI-generated.
5. B5: Governments or NGOs conduct deepfake-based awareness campaigns on urgent social issues, using hyper-realistic scenarios to vividly illustrate potential future outcomes.

6. B6: Companies use deepfake 'virtual brand ambassadors' for small businesses that cannot afford celebrity endorsements, levelling the marketing playing field.
7. B7: Interactive media experiences feature deepfake hosts or performers who offer audiences highly personalised entertainment or educational content, enhancing social engagement.
8. B8: Therapeutic deepfakes are developed to allow patients to role-play conversations with AI-generated support avatars or practice exposure therapy in a controlled setting.
9. B9: Researchers develop open-source deepfake platforms for legitimate uses (e.g., film production, voice dubbing) with built-in ethical safeguards, supporting innovation while promoting responsible development.
10. B10: Clear policies on labelling deepfake content are enacted, increasing transparency and user trust, enabling creative uses of AI-generated media without misleading the public.

Statistical framework and analytical approach

This study applied a suite of quantitative techniques selected to match each research objective and to provide a reproducible account of how Europeans perceive the risks and benefits of deepfake technologies. We began with correspondence analysis, following the formulation of Hirschfeld (1935) and the later refinements of Benzécri (1973), to visualise the associations between countries and ordinal categories of risk and benefit. Prior chi-square tests confirmed that the contingency tables contained sufficient dependence for meaningful dimension reduction.

To compare nations on individual scenarios, we employed two-sided tests with Bonferroni-adjusted thresholds (Armstrong, 2014), limiting the probability that multiple comparisons would yield spurious significance. The overall balance of attitudes was examined with paired-samples t tests, which revealed that the respondents consistently rated risks higher than advantages. Covariance analysis then probed the thematic structure of the 20 questionnaire items; by retaining the unstandardised units, it exposed both tight within-cluster ties and lighter, yet interpretable, cross-cluster linkages—most noticeably, those involving regulation and transparency. Finally, an ordinary least-squares General Linear Model (GLM) assessed the demographic drivers of the risk–benefit differential. Advancing age, lower income, and gender (woman) each widened the gap, while higher educational attainment narrowed it; country effects were visible as well, with Czech and UK publics registering the greatest divergence.

All computations were performed in IBM SPSS Statistics 30, with supplementary tabulation and graphing in Microsoft Excel, ensuring full transparency and ease of replication.

Deepfakes risk perceptions

Fig. 1 charts perceived deepfake-related risk across seven European countries by age cohort. The y-axis denotes risk intensity, with higher values signifying greater concern. Perceptions differ both internationally and generationally. The respondents aged 16–24 generally register lower concern—most clearly in Sweden, France, and Czechia—whereas age effects are muted in Germany, Italy, and the UK. The UK shows the highest risk ratings across all cohorts, while Sweden and Italy record the lowest overall. Error bars illustrate within-group variability; uncertainty is greatest among the youngest and oldest participants. Collectively, the results indicate a generational divide in assessments of deepfake harm, plausibly linked to contrasts in digital literacy, media exposure, and trust in technology.

Fig. 2 positions countries and four Deepfakes Risk Perception categories in a correspondence map; shorter distances denote more similar response patterns. Dimension 1 (singular value = 0.086) accounts for

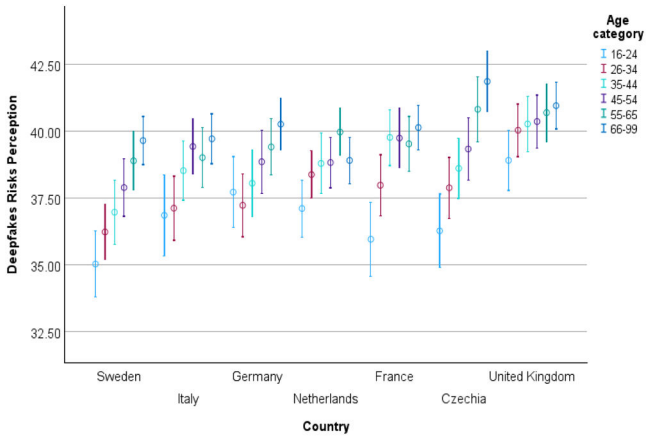

Fig. 1. Deepfakes Risk Perceptions by country and age group.

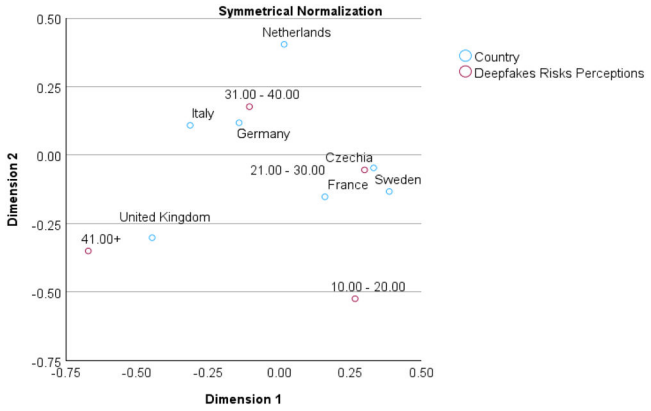

Fig. 2. Correspondence map – Deepfakes Risk Perceptions.

70.4 % of the total inertia, while Dimension 2 (singular value = 0.046) adds 20.4 %. Together, these axes capture about 91% of the variance, indicating that the two-dimensional display adequately summarises the data. Chi-square statistic of 72.271 with 48 degrees of freedom and a significance level of 0.013 confirms that the association between country and risk category is statistically meaningful.

Along Dimension 1, the chief axis of differentiation, Sweden (0.388), Czechia (0.332), and France (0.161) appear on the right-hand side, aligning with the lower-risk categories 11.00–20.00 and 21.00–30.00, which also carry positive scores. By contrast, the UK (-0.447), Italy (-0.313), and Germany (-1.141) plot to the left; the high-risk category 41.00+ (-0.672) sits nearby, indicating that the respondents in these countries are more likely to view deepfakes as highly risky. Dimension 2 captures less variance (≈ 20.4 %) but reveals a vertical split: the Netherlands registers a high positive loading (0.404), whereas France and Czechia fall below the horizontal axis. Among the risk groups, 11.00–20.00 is low on Dimension 2 (-0.524), whereas 31.00–40.00 shows a moderate positive value (0.177). The Netherlands, therefore, occupies a somewhat distinct position in the correspondence space.

Deepfakes benefits perceptions

Fig. 3 charts *Deepfakes Benefits Perceptions* by age cohort in seven European countries. A consistent age gradient emerged: the respondents aged 16–24 and 26–34 assigned the highest-benefit scores—around 33 in Germany and above 34 in the UK—whereas those aged 55–65 and 66–99 clustered between 28 and 31. Although the steepness of this decline varies, every country shows the same downward trajectory.

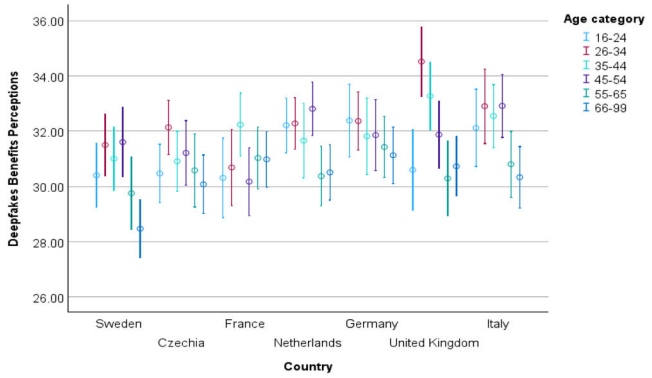

Fig. 3. Deepfakes Benefits Perceptions.

Although Czechia, France, and the Netherlands display a narrower range across cohorts, younger groups still lead. France's overlapping confidence intervals suggest modest cohort contrasts, whereas Sweden and the Netherlands reveal the sharpest drop, with the oldest Swedes registering the lowest mean. In sum, perceived benefits of deepfakes diminish steadily with age, highlighting a pervasive intergenerational divide across Europe.

We conducted a correspondence analysis for *Deepfakes Benefits Perceptions* (Fig. 4), mirroring the procedure used for the risk index. The first dimension has a singular value of 0.116 and explains 0.014 units of inertia—87 % of the total—thus, it captures the dominant pattern linking countries to benefit perceptions. The second dimension, with a singular value of 0.045 and inertia of 0.002, contributes an additional 13 %—a smaller share of the overall variance—raising the cumulative proportion of explained inertia to 99.9 %, meaning two dimensions are sufficient to summarise the data structure.

On Dimension 1, the UK and Czechia sit on the positive side near the **41.00+** benefits category, indicating that the respondents in these countries hold the most favourable views of deepfakes' creative and communicative potential; France also leans this way, though closer to the origin. Sweden, the Netherlands, and Italy fall on the negative side, alongside the **31.00–40.00** and **21.00–30.00** categories, reflecting a more cautious stance that likely emphasises ethical or social concerns. Germany, positioned close to zero on both axes, occupies a neutral midpoint between optimism and scepticism.

Taking up 13% of the inertia, Dimension 2 refines these distinctions: the Netherlands loads negatively on both axes, highlighting its distance from the highest-benefit category and accentuating its critical perspective, whereas Czechia loads positively on both, further confirming its optimistic outlook. Together, the coordinates revealed pronounced national contrasts in how the respondents assessed the value of deepfake technologies.

Difference between deepfakes-risk and deepfakes-benefit perceptions

Fig. 5 depicts the distribution of individual risk–benefit differentials for the full sample (N = 6,956). The x-axis records each respondent's composite risk score minus their composite benefit score; the y-axis represents frequency. Negative values—where benefits exceed risks—occur only sporadically, confirming their marginal presence. The histogram is positively skewed. Most observations clustered in the 0–10 range, indicating that the respondents generally judged deepfakes to be more hazardous than advantageous, though the margin was seldom extreme. The sample mean is 7.16, while the standard deviation of 9.38 reveals wide dispersion, signalling considerable diversity in individual assessments even as the aggregate leans toward heightened concern.

Correspondence Analysis of the Risk–Benefit Differential

We extended the correspondence–analysis procedure to the net difference between each respondent's risk and benefit indices. To aid interpretation, the continuous differential was discretised into six ordered categories as follows:

- ≤ **-1.00**: Respondents in this group actually perceive benefits as greater than risks – a small but noteworthy subset.
- **0.00–2.00**: Very balanced perception; risks are only slightly greater than benefits.
- **3.00–6.00**: Mild risk dominance; respondents see some imbalance, but not sharply.
- **7.00–10.00**: Moderate difference in favour of risks; growing concern.
- **11.00–17.00**: Substantial perceived risk dominance.
- **18.00+**: The most concerned group, respondents see deepfakes as far more risky than beneficial.

The model yields a total inertia of 0.016 and a statistically significant chi-square value ($\chi^2 = 113.368$, $p < 0.001$, $df = 36$), validating the relevance of the association between countries and risk–benefit categories. The first two dimensions together explain 92 % of the total inertia (Dimension 1 = 47.7 %, Dimension 2 = 44.3 %), confirming that a two-dimensional map captures the essential structure of the data.

Fig. 6 portrays a two-dimensional correspondence map in which the horizontal axis (47.7 % of inertia) represents a continuum from 'risk-dominant' perceptions on the left to 'benefit-sensitive' or balanced views on the right, while the vertical axis (44.3 %) captures the degree of internal dispersion within each national sample. On the right-hand side, Italy and Sweden occupy positive positions on Dimension 1, adjacent to the bins in which benefits either equal or exceed risks (≤ –1 and 0–2). Their publics, therefore, appear relatively open to the constructive potential of deepfakes or, at a minimum, unconvinced that hazards decisively prevail. Germany sits close to the map's centre, signalling a

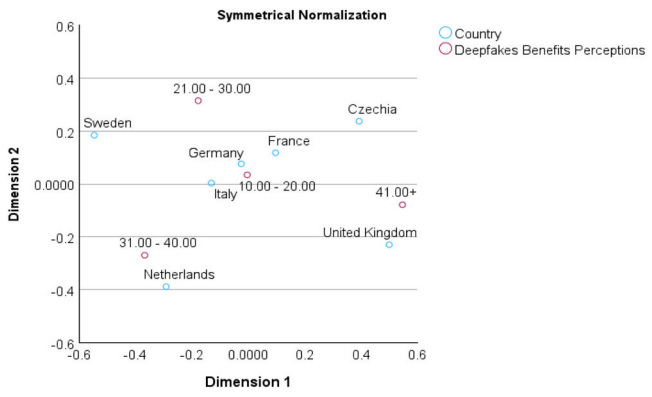

Fig. 4. Correspondence map – Deepfakes Benefits Perceptions.

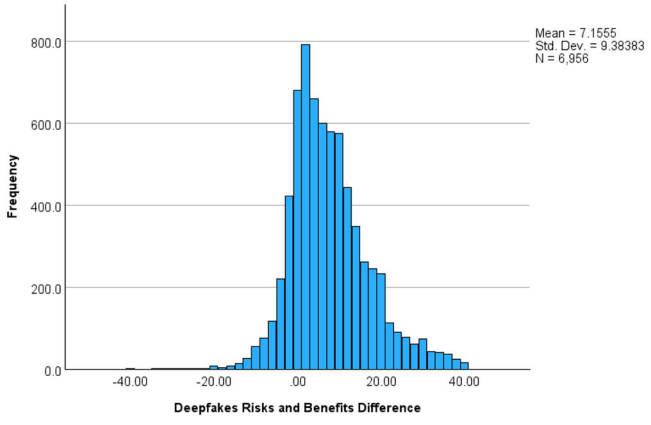

Fig. 5. Difference between Deepfakes-Risk and Deepfakes-Benefit Perceptions.

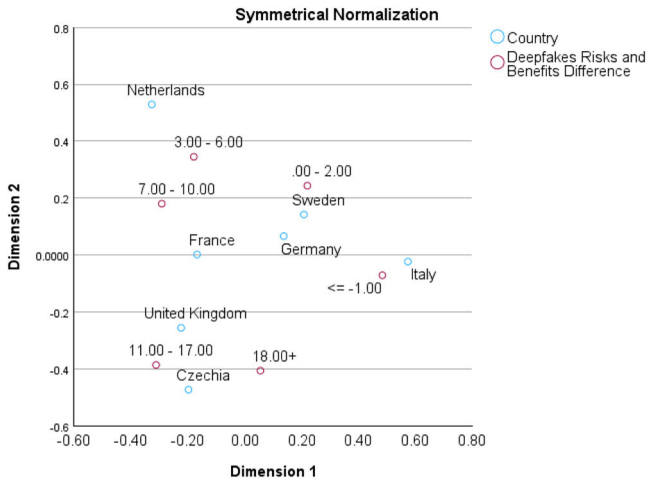

Fig. 6. Net risk–benefit differential: correspondence map.

middle-ground stance – neither markedly alarmist nor overtly optimistic. Subsequently, France, the Netherlands, the UK and Czechia gravitate toward bins denoting moderate to substantial risk dominance (7–10 and 11–17). However, the two clusters differ vertically. The Netherlands plots high on Dimension 2, indicating sizeable within-country variation: alongside a sizeable cautious segment, a counter-group registers either significantly milder or significantly steeper differentials, resulting in a dissemination of opinion. The UK and Czechia fall low on Dimension 2, revealing tighter consensus: both publics coalesce around the judgement that deepfakes pose clearly greater dangers than advantages, with the UK aligning most closely to the 11–17 category and Czechia extending toward the extreme 18+ group.

Collectively, the map depicts the following three patterns: (1) Italy and Sweden exhibit the most benefit-aligned or balanced orientations; (2) Germany and France hover near neutrality; and (3) the Netherlands, the UK, and Czechia adopt strongly risk-centred outlooks, though the Dutch respondents display greater internal disagreement than their Czech and British counterparts.

Cross-national differences

Subsequently, our analysis focuses on the perceived risks associated with deepfake technologies across countries in the study. The aim is to identify which countries consider specific deepfake scenarios to be most risky and whether there are statistically significant differences in their assessments. Using average scores of Likert scale responses (1 – Very low risk, 2 – Low risk, 3 – Moderate risk, 4 – High risk, and 5 – Very high risk) and two-sided tests with Bonferroni correction (Armstrong, 2014), the analysis highlights where perceptions diverge most clearly. Bonferroni correction is a statistical method used in multiple comparisons (e.g., when comparing multiple countries with each other) to reduce the risk of false-positive results (so-called 'Type I error', i.e., incorrect rejection of the null hypothesis). Table 1 presents the output of the analysis for risk perceptions. The letters next to the mean values indicate which countries have significantly lower ratings, highlighting where concerns are notably higher.

For the scenario in which a deepfake video shows a world leader announcing an unverified military action, the Czech respondents (M = 4.188) perceived significantly more risk than those in Germany (4.044), France (3.945), Italy (3.946), and Sweden (3.798). Citizens in the UK (4.161) also judged this threat more severe than participants in France, Italy, and Sweden, while Sweden registered the lowest concern overall. When deepfakes were presented as courtroom evidence, Czechia (3.958) and the UK (4.014) rated the danger markedly higher than Germany (3.618), France (3.569), Italy (3.718), Sweden (3.550), and the Netherlands (3.653). In campaign ads that falsely portray a candidate, the UK (4.183) showed the greatest alarm, scoring the risk significantly above France (4.051), Italy (3.995), Sweden (3.909), and the Netherlands (4.029). Germany (4.089), Czechia (4.056), and France (4.051) also rated this scenario higher than Sweden. For the 'digital resurrection' of a deceased public figure, Czechia (3.710) and the UK (3.692) expressed significantly more worry than Germany (3.427), France (3.455), Italy (3.478), Sweden (3.483), and the Netherlands (3.383), with the Dutch sample showing the lowest average risk. When assessing AI-generated news that anchors used to shape public opinion, Czechia (3.910) and the UK (3.905) rated the threat higher than Sweden (3.599) and the Netherlands (3.737); France (3.893) displayed the same pattern relative to these two countries.

No significant cross-national differences emerged for deepfake brand

Table 1
Cross-National Comparison of Deepfake Risk Perception.

Risks of Deepfakes	Country						
	Czechia (A)	United Kingdom (B)	Germany (C)	France (D)	Italy (E)	Sweden (F)	Netherlands (G)
World leader announces unverified military action	4.188 C D E F	4.161 D E F	4.044 F	3.945 F	3.946 F	3.798	4.067 F
Deepfake used as court evidence	3.958 C D E F G	4.014 C D E F G	3.618	3.569	3.718 F	3.550	3.653
Campaign ad misrepresents candidate	4.056 F	4.183 D E F G	4.089 F	4.051 F	3.995	3.909	4.029
"Resurrected" public figure used without consent	3.710 C D E F G	3.692 C D E F G	3.427	3.455	3.478	3.483	3.383
AI-generated "news anchors" shape opinion	3.910 F G	3.905 F G	3.809 F	3.893 F G	3.823 F	3.599	3.737
Deepfake "brand ambassadors" distort claims	3.676	3.786	3.663	3.772	3.745	3.720	3.681
Viral deepfake accuses celebrity	3.896	4.084 A F	4.013 F	4.114 A E F G	3.966	3.863	3.972
Emotion-triggering deepfake manipulates behaviour	4.145 E F	4.145 E F	4.087 F	4.063	4.016	3.940	4.040
Technology outpaces detection ("arms race")	3.978	4.103 C E F	3.951	4.056 E F	3.852	3.854	4.013 E F
Lack of regulation enables deception	4.023	4.234 A F	4.151 A F	4.222 A F	4.228 A F	4.013	4.188 A F

Results are based on two-sided tests assuming equal variances. For each significant pair, the key of the smaller category appears in the category with the larger mean.
Significance level for upper case letters (A, B, C): .05[1]

[1] Tests are adjusted for all pairwise comparisons within a row of each innermost subtable using the Bonferroni correction.

ambassadors in marketing; means clustered tightly between 3.663 (Germany) and 3.786 (UK). For viral deepfakes accusing celebrities of wrongdoing, France (4.114) reported significantly greater concern than Czechia (3.896), Italy (3.966), Sweden (3.863), and the Netherlands (3.972). The UK (4.084) was likewise higher than Czechia and Sweden. Regarding emotion-evoking deepfakes designed to manipulate behaviour, Czechia (4.145) and the UK (4.145) considered the risk significantly greater than Italy (4.016) and Sweden (3.940). The German (4.087), French (4.063), and Dutch (4.040) scores did not differ statistically from the leading pair. On the technological 'arms race', where deepfake methods outpace detection, the UK (4.103) voiced significantly higher anxiety than Germany (3.951), Italy (3.852), and Sweden (3.854); France (4.056) exceeded Italy and Sweden as well. The Netherlands (4.013) ranked among the more concerned publics, though its score did not differ significantly from that of the UK. Finally, for the absence of regulation, the participants in the UK (4.234), Italy (4.228), France (4.222), Germany (4.151), and the Netherlands (4.188) scored the risk significantly above Czechia (4.023) and Sweden (4.013). This suggests that Western and Southern European countries are particularly concerned about the regulatory vacuum around deepfake technologies.

Overall, Czechia and especially the UK register consistently high levels of concern across most scenarios, whereas Sweden repeatedly records the lowest. Bonferroni-adjusted comparisons confirm that these cross-national gaps are statistically robust.

We applied the same cross-national procedure to the benefit scenarios. Table 2 reports mean scores on a five-point scale (1 = Very low benefit, 2 = Low benefit, 3 = Moderate benefit, 4 = High benefit, and 5 = Very high benefit) and presents two-sided Bonferroni-adjusted tests that pinpoint where evaluations differ significantly.

For emergency-preparedness training, respondents in the Netherlands (M = 3.551), Germany (3.539), and the UK (3.526) judged the benefit significantly higher than their counterparts in Czechia (3.329) and Sweden (3.336). In the witness-protection scenario, average scores varied only marginally; no pairwise differences reached significance. When deepfakes served as interactive political-education tools, the UK (3.057) and the Netherlands (3.058) rated the benefit significantly above France (2.903). Germany's mean (3.010) exceeded the French figure numerically but not at a significant level. For the 'digital resurrection' scenario, the UK respondents (M = 3.237) rated the benefit significantly higher than those in Sweden (3.053). Perceptions in Czechia (3.180), Germany (3.111), France (3.163), Italy (3.184), and the Netherlands (3.117) did not differ significantly from either the UK or Sweden. In awareness campaigns on urgent social issues, the UK (3.186), Germany (3.065), France (3.139), and Italy (3.144) rated the benefit significantly higher than Czechia (2.912) and Sweden (2.899). The Netherlands (3.046) was directionally higher than these two countries, but the difference fell short of significance.

When respondents assessed virtual brand ambassadors, the benefit score in Italy (2.969) exceeded those in Czechia (2.808), the UK (2.808), Sweden (2.679), and the Netherlands (2.741); Germany (2.883) and France (2.826) also rated the scenario significantly higher than Sweden. For interactive-media hosts, Italy again stood out: its mean of 3.035 surpassed those of Czechia (2.806), the UK (2.851), France (2.828), Sweden (2.826), and the Netherlands (2.863). In the therapeutic setting, the Netherlands (3.264) scored significantly higher than Czechia (3.078), France (3.050), and Sweden (2.978); and Italy (3.181) did not differ significantly from the Dutch mean. Germany (3.143) also exceeded Sweden. When evaluating open-source deepfake platforms with ethical safeguards, Italy (3.118) and the Netherlands (3.120) rated the benefit significantly above France (2.961) and Sweden (2.939). Czechia (3.085) and Germany (3.087) were likewise higher than Sweden. Clear deepfakes labelling requirements drew the strongest endorsement: Germany (3.567) and the UK (3.521) rated this measure significantly above every other country in the comparison.

The benefit results extend the picture drawn from the risk analysis and point to a clear West–East gradient in public sentiment toward deepfake technologies. The respondents in the Netherlands, Germany, the UK, and Italy assigned comparatively high value to practical or prosocial applications—simulated crisis drills, public-interest campaigns, and mental-health therapies. This positive outlook is strongest in the Netherlands, which leads the field in four of the ten benefit items, and is shared by the German and British publics, whose support is bolstered when robust labelling rules are introduced. Italians, meanwhile, show remarkable enthusiasm for the commercial and entertainment potential of deepfakes, from virtual brand ambassadors to interactive-media hosts. By contrast, Czech and Swedish respondents consistently give the lowest benefit ratings, particularly in socially or culturally expressive scenarios.

When we place benefits alongside risks, a two-track picture appears: the UK and Czechia rate deepfake dangers highest, Sweden lowest, with

Table 2
Cross-National Comparison of Deepfake Benefits Perception.

Benefits of Deepfakes	Country						
	Czechia (A)	United Kingdom (B)	Germany (C)	France (D)	Italy (E)	Sweden (F)	Netherlands (G)
Emergency preparedness exercises	3.329	3.526	3.539	3.414	3.432	3.336	3.551
		A F	A F				A F
Witness protection through identity masking	3.471	3.479	3.405	3.366	3.374	3.386	3.427
Interactive political-education tools	2.932	3.057	3.010	2.903	2.965		3.058
		D				2.927	D
"Resurrection" exhibits of historical figures	3.180	3.237	3.111	3.163	3.184	3.053	3.117
		F					
Awareness campaigns on urgent social issues	2.912	3.186	3.065	3.139	3.144	2.899	3.046
		A F	A F	A F	A F		
Virtual brand ambassadors for small firms	2.808	2.808	2.883	2.826	2.969	2.679	2.741
			F		A B F G		
Interactive media hosts/performers	2.806	2.851	2.900	2.828	3.035	2.826	2.863
					A B D F G		
Therapeutic role-play or exposure therapy	3.078	3.124	3.143	3.050	3.181	2.978	3.264
			F		F		A D F
Open-source platforms with ethical safeguards	3.085	3.077	3.087	2.961	3.118	2.939	3.120
	F		F		D F		D F
Clear labelling of deepfake content	3.243	3.521	3.567	3.280	3.310	3.270	3.344
		A D E F G	A D E F G				

Results are based on two-sided tests assuming equal variances. For each significant pair, the key of the smaller category appears in the category with the larger mean.
Significance level for upper case letters (A, B, C): .05[1]

[1] Tests are adjusted for all pairwise comparisons within a row of each innermost subtable using the Bonferroni correction.

Germany, Italy, France, and the Netherlands in between; conversely, benefit optimism is strongest in the Netherlands, Germany, the UK, and Italy, whereas Czech and Swedish publics remain distinctly sceptical. These cross-currents likely reflect national gaps in digital literacy, the stringency of media regulation, and broader cultural attitudes toward emerging technologies. In settings where media-education programmes are well developed, and regulatory debates are advanced—as in the UK, Germany, and the Netherlands—citizens may feel confident enough to endorse constructive uses while still flagging potential harms. In countries with lower institutional visibility or public discussion of deepfakes, such as Czechia and Sweden, the technology elicits less optimism, either because its advantages are less salient or because trust in institutional safeguards is weaker.

The covariance analysis of deepfake technology perceptions

In the next section, we focus on the covariance analysis of individual questionnaire items. This analysis helps us understand how the statements about the risks and benefits of deepfakes are interrelated, specifically, whether and how the evaluation of one statement changes depending on the evaluation of others. Unlike correlation analysis, which measures the strength and direction of relationships, covariance provides an unstandardised indicator of shared variability, considering the degree of dispersion of individual variables.

Covariance analysis allows us to uncover thematic connections within groups of statements about deepfake technologies' risks and benefits, as well as cross-links between them.

Table 3 presents the covariances between individual questionnaire items. This analysis offers deeper insights into the structural dependencies within the dataset, enabling a more complex interpretation of how perceptions of deepfake risks and benefits evolve dynamically.

The covariance-matrix analysis of risk-related (R1–R10) and benefit-related (B1–B10) items reveals several notable regularities. The highest covariances lie within the risk and benefit clusters, showing strong thematic coherence in how respondents judge potential dangers and advantages of deepfakes. Cross-cluster covariances are weaker, though they point to a limited set of policy-relevant intersections: while deepfakes are largely viewed as risky, certain applications are simultaneously recognised as beneficial.

Regarding intra-risk covariances, a few links stand out. R1, the scenario of a fabricated video in which a world leader announces unverified military action, covaries strongly with R2 (0.69), indicating that political deception and the misuse of deepfakes as courtroom evidence are perceived as closely related threats. R4, which involves 'resurrecting' deceased public figures without family consent, correlates highly with R6 (0.53), flagging ethical concerns about identity exploitation both in entertainment and branding. Additionally, R7 through R10—viral deepfake scandals, emotional exploitation, lack of regulation, and the accelerating deepfake arms race—each show covariances above 0.44, reinforcing the view that misinformation and manipulation represent a single, interconnected risk complex.

Turning to intra-benefit covariances, the pattern is equally clear. B1 (emergency-preparedness training) covaries with B2 (anonymous witness protection) at 0.68, signalling that respondents treat realistic crisis simulations and protective law-enforcement tools as tandem security benefits. B3 (interactive historic education) clusters with B7 and B8 (both 0.61), suggesting that interactive learning and therapeutic avatars form a shared 'experiential' benefit. B4 (AI-generated historic exhibitions) aligns with B8 (0.62), indicating broad acceptance of AI personas when their synthetic nature is disclosed. Ethical deepfake development (B9) likewise aligns with B10, underlining preference for responsible AI innovation guided by clear, transparent labelling rules.

Although risks and benefits appear largely independent, a few cross-group covariances reveal perceived safeguards. The strongest such bridge is R10 (lack of regulation) with B10 (transparent labelling) at 0.16: the respondents most troubled by a regulatory vacuum also favour

Table 3
Covariance matrix.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
R1	1.08	0.69	0.56	0.46	0.56	0.40	0.46	0.47	0.47	0.46	0.09	0.14	0.06	0.05	0.08	0.01	-0.01	0.05	0.02	0.08
R2	0.69	1.45	0.51	0.55	0.61	0.43	0.40	0.42	0.45	0.40	0.03	0.12	0.06	0.02	0.06	0.05	0.00	0.05	0.01	0.02
R3	0.56	0.51	0.93	0.43	0.51	0.45	0.48	0.50	0.46	0.49	0.12	0.14	0.05	0.07	0.07	-0.02	-0.01	0.06	0.05	0.13
R4	0.46	0.55	0.43	1.33	0.45	0.53	0.44	0.39	0.40	0.36	-0.01	0.05	0.06	0.02	0.03	0.08	0.05	0.01	0.02	0.00
R5	0.56	0.61	0.51	0.45	1.14	0.49	0.40	0.44	0.44	0.42	0.05	0.08	0.02	0.02	0.04	0.02	-0.02	0.03	0.01	0.06
R6	0.40	0.43	0.45	0.53	0.49	1.00	0.46	0.43	0.43	0.42	0.01	0.04	0.01	0.00	0.02	0.00	-0.02	-0.01	-0.01	0.01
R7	0.46	0.40	0.48	0.44	0.40	0.46	0.94	0.49	0.44	0.46	0.09	0.10	0.04	0.07	0.06	0.00	0.00	0.03	0.03	0.10
R8	0.47	0.42	0.50	0.39	0.44	0.43	0.49	0.89	0.47	0.48	0.11	0.13	0.02	0.05	0.06	-0.05	-0.05	0.04	0.04	0.13
R9	0.47	0.45	0.46	0.40	0.44	0.43	0.44	0.47	0.90	0.47	0.10	0.11	0.02	0.02	0.06	-0.04	-0.04	0.04	0.02	0.11
R10	0.46	0.40	0.49	0.36	0.42	0.42	0.46	0.48	0.47	0.88	0.13	0.12	0.02	0.06	0.08	-0.05	-0.04	0.03	0.04	0.16
B1	0.09	0.03	0.12	-0.01	0.05	0.01	0.09	0.11	0.10	0.13	1.25	0.68	0.56	0.60	0.57	0.42	0.45	0.61	0.54	0.62
B2	0.14	0.12	0.14	0.05	0.08	0.04	0.10	0.13	0.11	0.12	0.68	1.38	0.53	0.58	0.55	0.43	0.45	0.57	0.52	0.62
B3	0.06	0.06	0.05	0.06	0.02	0.01	0.04	0.02	0.02	0.02	0.56	0.53	1.19	0.61	0.56	0.55	0.61	0.61	0.58	0.50
B4	0.05	0.02	0.07	0.02	0.02	0.00	0.07	0.05	0.02	0.06	0.60	0.58	0.61	1.30	0.52	0.47	0.55	0.62	0.58	0.57
B5	0.08	0.06	0.07	0.03	0.04	0.02	0.06	0.06	0.06	0.08	0.57	0.55	0.56	0.52	1.25	0.52	0.52	0.55	0.53	0.55
B6	0.01	0.05	-0.02	0.08	0.02	0.00	0.00	-0.05	-0.04	-0.05	0.42	0.43	0.55	0.47	0.52	1.21	0.63	0.50	0.52	0.39
B7	-0.01	0.00	-0.01	0.05	-0.02	-0.02	0.00	-0.05	-0.04	-0.04	0.45	0.45	0.61	0.55	0.52	0.63	1.13	0.56	0.58	0.45
B8	0.05	0.05	0.06	0.01	0.03	-0.01	0.03	0.04	0.04	0.03	0.61	0.57	0.61	0.62	0.55	0.50	0.56	1.21	0.57	0.55
B9	0.02	0.01	0.05	0.02	0.01	-0.01	0.03	0.04	0.02	0.04	0.54	0.52	0.58	0.58	0.53	0.52	0.58	0.57	1.14	0.54
B10	0.08	0.02	0.13	0.00	0.06	0.01	0.10	0.13	0.11	0.16	0.62	0.62	0.50	0.57	0.55	0.39	0.45	0.55	0.54	1.30

disclosure mandates. R3 (political manipulation) has a smaller but notable link to B10 (0.13), and R8 (emotional manipulation) shows modest associations with B10, B2, and B1 (0.13–0.11), suggesting that people differentiate between harmful emotional triggers and constructive, safeguard-supported applications such as crisis training or protected testimony.

Collectively, the covariance matrix confirms that the questionnaire items were thematically well structured. The respondents show consistent patterns within each risk and benefit group—legal–political deception versus security–educational–ethical use—while cross-group ties, though weaker, converge on regulatory solutions, especially transparent labelling, as the linchpin linking public concern to public optimism.

Inferential statistics for the risk–benefit gap

This stage of the analysis considers, with a paired-samples *t-test*, whether the respondents rate deepfake risks higher than benefits. The hypotheses are as follows:

Null Hypothesis **(H_0):** There is no significant difference between the mean scores of Deepfakes Risk Perceptions and Deepfakes Benefits Perceptions perceived by the respondents. Any observed difference is because of random variation rather than a systematic pattern in public perception.

Alternative Hypothesis **(H_1):** There is a statistically significant difference between the mean scores of Deepfakes Risk Perceptions and Deepfakes Benefits Perceptions, indicating that the respondents perceive deepfake technologies as more risky than beneficial.

Table 4 presents the Paired-samples t-test, Paired-Samples Correlations results, and Paired-Samples Effect Sizes for the entire sample and respective countries.

The dependent measures are the composite risk and benefit indices (possible range 10–50). Across the entire sample, the mean risk score is 39.01, and the mean benefit score is 31.27; their difference of 7.74 points produces $t(7\ 082) = 63.463$, $p < 0.001$. Cohen's $d = 0.754$ (Hedges-corrected), a *large* effect, which shows the gap is not only statistically reliable but also practically substantial. The correlation between the paired scores is $r = 0.074$ ($p < 0.001$), positive yet trivial, indicating that risk and benefit judgements are largely independent.

Country-specific tests replicate the pattern. The largest effects occur in Czechia (d = 0.828) and the UK (0.823), followed by France (0.808) and the Netherlands (0.800). Sweden (0.708) and Germany (0.685) show slightly smaller, though still strong, gaps, while Italy (0.651) records the least pronounced difference—risks remain dominant, but the margin is narrower.

Overall, every national sample rejects H_0 at $p < 0.001$, confirming that Europeans, irrespective of locale, regard deepfakes as appreciably more hazardous than beneficial.

Regression analysis

The section aims to identify which demographic variables most significantly influence the difference between perceived risks and benefits of deepfake technologies. While previous descriptive analyses revealed basic patterns of perception, regression analysis allows us to account for multiple factors simultaneously and examine their joint effects. The dependent variable is an index-based *difference* between each respondent's composite risk and benefit scores for ten deepfake scenarios. Positive values indicate that risk perception outweighs perceived benefit; negative values indicate the reverse. We estimated an ordinary-least-squares GLM with the predictors age, gender, country, education, and net monthly income:

The regression equation is specified as follows:

$$Difference = \beta_0 + \beta_1 Age + \beta_2 Gender + \sum_{j=1}^{J} \beta_3 Country_{ij} + \sum_{k=1}^{K} \beta_4 Education_{ik} \sum_{l=1}^{L} \beta_5 Income_{il} + \varepsilon_i$$

Where:

- $Difference_i$ is the perceived risk–benefit difference for respondent i;
- β_0 is the intercept (expected value for the reference group);
- β_1 to β_5 are regression coefficients;
- $Country_{ij}$, $Education_{ik}$, and $Income_{il}$ are dummy variables for country, education, and income, respectively; and
- ε_i is the random error term.

The reference category is a woman respondent from the Netherlands with a Master's or Doctorate degree and a net monthly income of over 4,000 EUR. Table 5 presents the regression analysis results.

The regression results provide information about how demographic factors shape perceptions of deepfake technologies. The reference group—women respondents in the Netherlands who hold a Master's or Doctorate and earn more than €4,000 per month—shows a baseline gap of 7.744 points, indicating that even the least-concerned cohort judges risks to outweigh benefits. Age is a decisive factor: each additional year adds 0.085 points to the gap, confirming that older individuals are more sceptical, perhaps owing to lower digital literacy or heightened caution. Gender matters as well; men score 1.258 points lower than women, implying that women see greater danger in deepfakes.

Country effects reveal cultural variation. Czechia (+1.160), the UK (+1.091), and France (+0.815) exhibit significantly larger risk–benefit differentials than the Dutch benchmark, whereas Germany, Sweden and—crucially—Italy (−0.783, p = 0.062) do not differ in a statistically reliable way. Educational attainment exerts a subtler influence: people with only upper-secondary credentials register an additional 0.837 points, while the bachelor's coefficient of +0.648 does not reach conventional significance (p = 0.055); lower levels show no effect. Income displays the steepest gradient: respondents earning under €1,000 add 2.720 points, those in the €1,000–2,000 band add 1.904, €2,000–3,000 band add 1.779, and €3,000–4,000 band add 0.740—each tier significantly above the top-income reference.

In sum, low income, advanced age, and gender (woman) most strongly amplify the perceived imbalance, while Italy's negative, non-significant coefficient tempers earlier impressions of uniformly heightened concern in southern Europe. Cultural context, educational ceiling effects, and economic security jointly explain why some publics, and

Table 4
T-test, Correlation, and Effect Size.

	Paired Samples Correlations		**Paired samples T test**				**Paired Samples Effect Sizes - Point Estimate**	
	Correlation	Two-Sided p	Mean	Std. Deviation	Std. Error Mean	Two-Sided p	Cohen's d	Hedges' correction
Czechia	.095	.002	8.696	10.505	.331	<.001	.828	.827
United Kingdom	.079	.012	8.439	10.251	.322	<.001	.823	.823
Germany	.073	.021	7.142	10.424	.330	<.001	.685	.685
France	.091	.004	8.211	10.162	.318	<.001	.808	.807
Italy	.040	.207	7.054	10.828	.340	<.001	.651	.651
Sweden	.031	.327	7.435	10.497	.330	<.001	.708	.708
Netherlands	.094	.003	7.231	9.033	.283	<.001	.800	.800
Whole sample	**.074**	**<.001**	**7.745**	**10.270**	**.122**	**<.001**	**.754**	**.754**

Table 5
Regression Analysis of the Difference Between Perceived Risks and Benefits of Deepfake Technologies.

Parameter	B	Std. Error	95% Wald Confidence Interval		Sig.
			Lower	Upper	
(Intercept)	7.744	.1575	7.435	8.053	<.001
Age	.085	.0065	.072	.098	<.001
Gender					
Male	-1.258	.2247	-1.699	-.818	<.001
Female					
Country					
Czechia (CZ)	1.160	.4130	.351	1.969	.005
United Kingdom (UK)	1.091	.4119	.283	1.898	.008
Germany (DE)	-.445	.4135	-1.255	.366	.282
France (FR)	.815	.4107	.010	1.620	.047
Italy (IT)	-.783	.4202	-1.607	.041	.062
Sweden (SE)	-.028	.4122	-.836	.780	.946
Netherlands (NL)	0[a]	.	.	.	.
Education					
Primary and lower secondary education	.151	.4735	-.777	1.079	.750
Upper secondary education	.837	.2922	.264	1.409	.004
Post-secondary, but non-tertiary education	.252	.4048	-.542	1.045	.534
Tertiary education – Bachelor's	.648	.3379	-.015	1.310	.055
Tertiary education – Master's and Doctorate	0[a]	.	.	.	.
Net monthly income					
Less than 1,000 EUR	2.720	.3845	1.967	3.474	<.001
1,000 to 2,000 EUR	1.904	.3864	1.147	2.662	<.001
2,000 to 3,000 EUR	1.779	.3409	1.111	2.447	<.001
3,000 to 4,000 EUR	.740	.3760	.003	1.477	.049
More than 4,000 EUR	0[a]	.	.	.	.

Dependent Variable: Deepfakes Risks and Benefits Difference
Model: (Intercept), Age, Country, Education, Net monthly income,
a. Set to zero because this parameter is redundant.

some segments within them, are markedly more anxious about deepfake risks than others.

Discussion

This study offers a thorough cross-national account of public perceptions concerning deepfake technologies in seven European countries, exposing notable discrepancies in perceived dangers and rewards. The outcomes highlight how national contexts, generational factors, and distinct deepfake uses shape public sentiment.

Concerning RQ1, a consistent age gap was identified in perceived benefits: younger people (aged 16–34) uniformly granted higher benefit scores than older participants (aged 55–99) in every nation. This generational split might stem from younger respondents' stronger familiarity with digital media and optimism regarding emerging technologies, along with their exposure to creative or entertainment-related uses of deepfakes. However, the age gradient for perceived risks was less straightforward. While younger respondents generally exhibited lower concern in Sweden, France, and Czechia, age-related effects were muted in the UK, Germany, and Italy. This suggests that risk perceptions of deepfakes may be influenced not just by age but also by factors such as local media environments and the extent of digital literacy required to detect harmful or misleading content.

Napshin et al. (2024) confirm that individual responsibility is negatively associated with individual interest, suggesting externalised accountability for deceptive deepfakes. Younger participants more often view deepfakes as humorous, presumably adjusting their sense of responsibility and showing a greater inclination toward malicious forms of the technology. These humorous aspects, along with the perception that self-accountability is lower, could decrease the sense that platforms should counter deepfake threats on a societal level (Cochran & Napshin, 2021). Despite considerable concern about deepfakes, perceptions differ among demographic segments depending on social media usage patterns.

Similarly, Li and Zhao (2024) note that the intention to watch or share deepfake videos varies by gender, age, and education level. Personal attitudes, subjective norms, and perceived control all appear to encourage the behavioural intention to engage with deepfakes. According to Bitton et al. (2025), age, gender, and educational attainment predict deepfake knowledge, though digital skills, individual innovativeness, and social media exposure also show a positive link to familiarity. Bitton et al. (2025) underline that age alone explains only a small fraction of how deepfakes are viewed, whereas women are more likely to exhibit limited knowledge and negative attitudes. Ahmed (2023a) contends that most people overestimate their own ability to identify synthetic clips relative to others, a self-perception that can complicate efforts to shield the public from deepfake hazards. Doss et al. (2023) also indicate that vulnerability to deepfakes grows with age and trust in information sources; it varies with political orientation and is higher among adults and teachers than among students, reinforcing the importance of contextual factors.

Many researchers emphasize stronger regulatory regimes and targeted media-literacy programs to enhance collective resilience and address the distinct ways deepfakes can harm different population groups (Alanazi et al., 2025). Those at greatest risk include women subjected to nonconsensual pornography, public figures, people with limited digital skills, minority communities, older adults, celebrities, and businesses. The negative outcomes can manifest as psychological, reputational, or emotional harm. Vaccari and Chadwick (2020) likewise warn that deepfakes intensify confusion on social media and threaten civic life and democratic processes.

Turning to RQ2, the correspondence analysis showed Sweden and, on several items, France reported comparatively lower risk, whereas the United Kingdom, Italy, and Germany indicated higher concern. On the benefits side, the Netherlands, Germany, the United Kingdom, and Italy showed comparatively positive views, whereas Czechia and Sweden adopted a more restrained stance. These patterns point to the influence of national contexts on broad attitudes toward deepfakes, possibly stemming from variations in cultural norms, media oversight, and public discourse about AI.

Regarding RQ3, significant cross-national differences emerged in perceptions of specific deepfake scenarios. The United Kingdom and Czechia consistently displayed higher worries about political misinformation, legal manipulation, and unauthorized use of AI-generated public figures, while Sweden repeatedly scored as least concerned. Concerning benefits, the Netherlands, Germany, the United Kingdom, and Italy put a higher value on functional or pro-social uses, whereas Czechia and Sweden were more skeptical. Notably, these findings show that a straightforward inverse relationship between perceived risks and benefits did not apply across all countries. For instance, the UK public judged deepfakes in political campaigning as risky but also held relatively favorable views of political-education applications. These outcomes underline the complexity of deepfake perceptions and the ways local contexts and each scenario's purpose can affect attitudes.

The divergence between the distinctly risk-averse United Kingdom (high concern, only moderate enthusiasm) and benefit-oriented countries like the Netherlands and Italy supports the conclusion that opinions on deepfakes vary considerably across Europe. Czechia's profile is more ambivalent – aggregate risk scores place it in the lower-risk cluster, yet scenario-level data reveal heightened sensitivity to specific high-salience threats. Sweden's consistently low assessments for both risks and benefits set it apart from the higher and more variable views among other European nations. Overall, the results suggest that engagement with deepfake technology is not uniform but rather rests on how local traditions, demographic tendencies, and media practices intersect with the emerging capabilities of advanced AI.

Zheng et al. (2025) observe that the United States, the European Union, and China have each initiated differing regulatory approaches, centered on specific use cases, the entire life cycle, or the core entities behind deepfakes. The US focuses on election interference and pornographic content; the EU applies layered regulation across multiple legal instruments; while China locates primary responsibility with the providers of deepfake services. Yet an examination of the digital ecosystem indicates that regulations often target content producers, platforms, and service providers, while the public's role remains comparatively overlooked.

Despite prevailing views of deepfakes as instruments of deception, some scholarship reveals their positive potential in spheres like education, healthcare, and entertainment (Navarro Martínez et al., 2024; Patterson, 2024). Nevertheless, the same advanced techniques that enable highly realistic dubbing in film (Chesney and Citron, 2019) or superior translation via lip synchronization (Suwajanakorn et al., 2017) can also open the door to unlawful or destructive uses. This intensifies the need to address wide-ranging social risks and examine unexplored areas of deepfake influence (Mirsky and Lee, 2021).

Existing work in Security Studies, International Relations, and Law emphasizes how deepfakes threaten global security, democratic foundations, and personal privacy (Agarwal et al., 2019; Sloot & Wagensveld, 2022; Havlík, 2023). Scholars call attention to espionage and hybrid-warfare risks, while underlining partial or inadequate legal structures at national and international levels (Łabuz, 2024; Meneses, 2024). The technical arms race between increasingly refined generative models and newly emerging detection techniques is often seen as an uphill battle (Laurier et al., 2024). Parallel contributions from Sociology, Media Studies, STS, and Psychology highlight how deepfakes can disrupt social norms and intensify biases. Audiences are vulnerable when lacking digital education or harboring strong partisan leanings (Hameleers et al., 2023a,b; Wan, 2023). Efforts to inform citizens and uphold institutional trust are often proposed, but interdisciplinary collaboration between legal and social-psychological approaches remains limited. Our empirical evidence supports these concerns: while certain countries display considerable apprehension aligned with expert views on misinformation, others appear comparatively unconcerned – suggesting the need to tailor responses based on public readiness.

This study's data both corroborate and extend existing debates. In certain nations, such as the UK and Czechia, heightened apprehension about political deception matches the legal and security scholarship's warnings. The moderate-to-high endorsement of potential advantages in the UK and Italy also resonates with more optimistic perspectives in STS or business literature, which propose that deepfakes may be adapted for educational or commercial purposes. Nevertheless, some countries – Sweden in particular – demonstrate subdued responses to deepfake threats and opportunities. This finding partially diverges from the academic consensus that sees deepfakes as a pressing concern worldwide. Factors such as national-level discussions, existing regulations, and local cultural norms may shape whether citizens perceive deepfake technology as a real and pressing issue. The patterns that emerged here demonstrate how interdisciplinary insights, combined with public feedback, are crucial for designing relevant policy measures.

Conclusion

The deepfake technologies represent significant challenges that can threaten trust in society, democratic processes, and harm privacy of individuals or even whole communities. This requires systematic creation of comprehensive strategies to mitigate deepfake risks and to support ethical innovation. The development of deepfakes technologies emphasises much more importance of creating media literacy for all the population groups, which can mitigate vulnerability to disinformation, improve individuals' abilities to critically evaluate information by raising ethical awareness, and to educate people in the digital ecosystem, who will resist the growing risks of development of digital technologies and take strictly ethical and responsible approaches to their use.

Political implications

Several critical policy implications emerge from this study. Foremost among these is the introduction of mandatory labeling for AI-generated content. As calls intensify for robust disclaimers clearly identifying synthetic visuals, videos, or audio, adopting consistent labeling standards could empower citizens to better distinguish authentic from manipulated media. Countries such as Germany have exhibited a positive reception to labeling initiatives, suggesting that increased clarity measures can significantly mitigate misinformation and improve public awareness. In tandem, AI literacy campaigns are essential. The data distinctly illustrate a generational divide in perceived benefits and risks associated with deepfake technologies, which vary substantially across national contexts. Coordinated educational campaigns spearheaded by civic organizations, educational institutions, psychologists, and media specialists are therefore recommended to bolster public awareness and enhance individuals' capacity to recognize manipulative media. Such interventions are particularly crucial for addressing vulnerabilities specific to older adults, populations with limited digital skills, or groups characterized by strong partisan alignment, who might otherwise face greater susceptibility.

Additionally, the observed heterogeneous patterns of public concern highlight the pressing need for flexible and multi-level regulatory coordination. Policymakers would benefit from supporting cross-border collaborations, reflecting the inherently global reach of deepfake technology. Multilateral cooperation, guided by insights from security experts, sociologists, and computer scientists, can establish shared standards, jointly develop advanced detection methods, and promote comprehensive data-sharing frameworks, thus strengthening collective resilience and ensuring consistent protection across Europe. Furthermore, industry and civil society partnerships are central to future regulatory effectiveness. Given the rapid pace of technological advancements, collaborative efforts involving private entities, civil society organizations, and government bodies are critical for driving innovation in detection methods and user-friendly interfaces that facilitate widespread adoption. Notably, respondents in countries indicating higher perceived risks also recognized potential value in pro-social deepfake applications, suggesting credible oversight can promote legitimate and beneficial uses while actively discouraging abuses.

From a theoretical perspective, this study demonstrates that public perceptions of deepfake technology depend upon factors extending well beyond technological dimensions alone. Cultural context, national regulatory approaches, media practices, and generational dynamics collectively shape attitudes toward deepfake applications. The complex interaction between risk and benefit perceptions across different scenarios highlights that acceptance of the technology is not necessarily inversely proportional to concern; societies may simultaneously appreciate educational or therapeutic potentials even while acknowledging significant apprehensions about misuse.

By revisiting our primary RQs, we can now specify how each has been addressed: RQ1 is answered through the systematic literature review, which shows that scholarship clusters around legal, security, sociotechnical, and computational lenses yet remains fragmented; RQ2 is resolved via the comparative SWOT, revealing consensus on threats and a relative silence on therapeutic and educational opportunities; and RQ3 is met by the seven-nation survey, which demonstrates a wide but patterned gap between hazards and advantages in European public opinion. The six sub-RQs derived from RQ3 are likewise fully satisfied – (1) a pronounced age gradient in benefit scores, (2) distinctive country clusters in correspondence space, (3) strong scenario-specific cross-national contrasts, (4) coherent intra-risk and intra-benefit covariance structures, (5) a statistically and practically large risk–benefit differential, and (6) significant demographic predictors led by income, age, and

gender. Collectively, these findings confirm that the analytical battery deployed here can bridge disciplinary insight with lay evaluation, thereby validating the integrated framework proposed at the outset.

The ongoing research in the examined field

Advancing future research calls for qualitative and mixed-method approaches, including targeted interviews and scenario-based experimental designs, to enrich insights into individual engagement with synthetic content. Investigating smaller or less-studied countries, as well as regions outside Europe, can further illuminate how differing regulatory frameworks and socio-economic contexts influence deepfake perceptions. Complementary research within the Global South – where digital literacy initiatives markedly differ – may reveal unique vulnerabilities and innovative strategies for positively leveraging deepfake technology. Longitudinal panel studies would be especially valuable for tracing whether today's measured enthusiasm among younger Europeans persists as the technology matures, or whether risk sensitivities converge across age cohorts once exposure and personal experience accumulate. Equally, field experiments that test the effectiveness of disclosure labels, authenticity watermarks, or AI-assisted detection prompts could provide actionable evidence for regulators and platform designers.

Interdisciplinary collaboration remains critical for addressing legal uncertainties, refining detection technologies, and developing comprehensive best practices. Enhanced cooperation across disciplines such as Law, Security Studies, Sociology, Psychology, Media, and Science and Technology Studies (STS) can clarify grey areas in accountability, bridging theoretical concerns with practical solutions. Policymakers should adapt legal frameworks dynamically to the shifting technological landscape, informed by comprehensive cross-disciplinary expertise. In summary, deepfake technologies represent a rapidly evolving phenomenon with significant potential to reshape social, economic, and political domains. The comparative SWOT reveals the necessity for integrated responses that draw on disciplinary strengths, address existing gaps, seize opportunities, and mitigate looming threats. This cross-national analysis provides a robust empirical foundation for refining policy, legal frameworks, educational interventions, and industry collaboration. By building upon this comprehensive empirical investigation into diverse European contexts, stakeholders – from legislators and regulators to educators and media organizations – can effectively guide the responsible utilization of deepfake technology, capitalizing on its positive potentials while safeguarding against associated risks.

Funding

This research was funded by the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences as a part of the research project VEGA No. 1/0554/24 and VEGA No. 1/0700/25. This research was supported by the Institutional support by Charles University, Program Cooperatio (IPS FSV).

Ethics

All subjects were informed about the study, and all provided informed consent. The study procedures were carried out in accordance with the Declaration of Helsinki and was approved through Ethics Consent (CZDEMOS4AI) by Ethics Committee FSV - School of Social Sciences, Charles University in Prague.

CRediT authorship contribution statement

Nik Hynek: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Conceptualization. **Beata Gavurova:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Matus Kubak:** Validation, Software, Formal analysis, Data curation, Resources.

Declaration of competing interest

Nik Hynek declares he has no conflict of interest.
Beata Gavurova declares she has no conflict of interest.
Matus Kubak declares he has no conflict of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jik.2025.100782.

Data availability

The data used in this study are uploaded as supplementary material.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deep fakes. *Computer Vision and Pattern Recognition*, 38–45.

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

Ahmed, S. (2023b). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society, 25*(5), 1108–1129. https://doi.org/10.1177/14614448211019198

Akbar, M., Suaib, M., & Hussain, M. S. (2023). The rise of deepfake technology. *Advances in Human and Social Aspects of Technology Book Series*, 178–201. https://doi.org/10.4018/978-1-6684-8133-2.ch010

Alanazi, S., Asif, S., & Moulitsas, I. (2024). Examining the societal impact and legislative requirements of deepfake technology: A comprehensive study. *International Journal of Social Science and Humanity*. https://doi.org/10.18178/ijssh.2024.14.2.1194

Alanazi, S., Asif, S., Caird-daley, A., & Moulitsas, I. (2025). Unmasking deepfakes: A multidisciplinary examination of social impacts and regulatory responses. *Human-Intelligent Systems Integration*, 1–23. https://doi.org/10.1007/s42454-025-00060-4

Ali, A. B. M., Ghouri, K. F. K., Naseem, H., Soomro, T. R., Mansoor, W., & Momani, A. (2022). Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security. In *International Conference Control and Robots* (pp. 1–5). https://doi.org/10.1109/ICCR56254.2022.9995821

Ali, H., & Aysan, A. F. (2024). Ethical dimensions of generative AI: A cross-domain analysis using machine learning structural topic modeling. *International Journal of Ethics and Systems, 41*(1), 3–34. https://doi.org/10.1108/ijoes-04-2024-0112

Allen, C., Payne, B. R., Abegaz, T., & Robertson, C. (2022). What you see is not what you know: Deepfake image manipulation. *KSU Proceedings on Cybersecurity Education, Research and Practice*. https://doi.org/10.32727/28.2023.1

Allen, C., Payne, B., Abegaz, T. T., & Robertson, C. (2023). What you see is not what you know: Studying deception in deepfake video manipulation. *Journal of Cybersecurity Education, Research & Practice, 2024*(1). https://doi.org/10.32727/8.2023.25

Amerini, I., Bartolini, F., Battiato, S., et al. (2024). Deepfake media forensics: State of the art and challenges ahead. *arXiv*.. https://doi.org/10.48550/arxiv.2408.00388

Amerini, I., Galteri, L., Caldelli, R., & Bimbo, A. D. (2019). Deepfake video detection through optical flow based CNN. In *International Conference on Computer Vision* (pp. 1205–1207). https://doi.org/10.1109/ICCVW.2019.00152

Armstrong, RA. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502–508. https://doi.org/10.1111/opo.12131

Ask, T. F., Lugo, R., fritsch, J., Veng, K., Eck, J., Özmen, M.-T., Bärreiter, B., Knox, B. J., & Sütterlin, S. (2023). Cognitive flexibility but not cognitive styles influence deepfake detection skills and metacognitive accuracy. *Center for Open Science*. https://doi.org/10.31234/osf.io/a9dwe

Battista, D. (2024). Political communication in the age of artificial intelligence: An overview of deepfakes and their implications. *Society Register, 8*(2), 7–24. https://doi.org/10.14746/sr.2024.8.2.01

Beretas, C. P. (2020a). Cyber hybrid warfare: Asymmetric threat. *Biomedical Journal of Scientific and Technical Research, 27*(4). https://doi.org/10.26717/BJSTR.2020.27.004524

Beretas, C. P. (2020b). Cyber hybrid warfare: Asymmetric threat. *Research & Development in Material Science,, 13*(2). https://doi.org/10.31031/RDMS.2020.13.000808

Benzécri, J. P. (1973). L'Analyse des Données:. In *L'Analyse des Correspondances, 2* Paris: Dunod.

Beridze, I., & Butcher, J. (2019). When seeing is no longer believing. *Nature Machine Intelligence, 1*(8), 332–334. https://doi.org/10.1038/S42256-019-0085-5

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

Bitton, D. B., Hoffmann, C. P., & Godulla, A. (2025). Deepfakes in the context of AI inequalities: Analysing disparities in knowledge and attitudes. *Information,*

Communication & Society, 28(2), 295–315. https://doi.org/10.1080/1369118x.2024.2420037
Boerwinkel, D. J., Boerwinkel, D. J., Swierstra, T., Waarlo, A. J., & Waarlo, A. J. (2014). Reframing and articulating socio-scientific classroom discourses on genetic testing from an STS perspective. *Science & Education, 23*(2), 485–507. https://doi.org/10.1007/S11191-012-9528-7
Boháček, M., & Farid, H. (2022). Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences of the United States of America, 119*(48). https://doi.org/10.1073/pnas.2216035119
Boté-Vericad, J.-J., & Vállez, M. (2022). Image and video manipulation: The generation of deepfakes. In: Freixa, P.; Codina, L.; Pérez-Montoro, M.; Guallar, J. (Ed.). Visualisations and narratives in digital media. *Methods and current trends, (pp.* 116-127*)*. https://doi.org/10.3145/indocs.2022.8.
Broinowski, A. (2022). Deepfake nightmares, synthetic dreams: A review of dystopian and utopian discourses around deepfakes, and why the collapse of reality may not be imminent - yet. *Journal of Asia-Pacific Pop Culture, 7*(1), 109–139. https://doi.org/10.5325/jasiapacipopcult.7.1.0109
Brooks, C. F. (2021). Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 159–163. https://doi.org/10.1089/CYBER.2020.0183
Buo, S. A. (2020). The emerging threats of deepfake attacks and countermeasures. *arXiv: Cryptography and Security*. https://doi.org/10.13140/RG.2.2.23089.81762
Caldelli, R., Galteri, L., Amerini, I., & Bimbo, A. D. (2021). Optical flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters, 146*, 31–37. https://doi.org/10.1016/J.PATREC.2021.03.005
Carvajal, L., & Iliadis, A. (2020). Deepfakes: a preliminary systematic review of the literature. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/SPIR.V2020I0.11190
Cavedon-Taylor, D. (2024). Deepfakes: a survey and introduction to the topical collection. In *Synthese, 204*. https://doi.org/10.1007/s11229-024-04634-8
Cochran, J. D., & Napshin, S. A. (2021). Deepfakes: awareness, concerns, and platform accountability. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 164–172. https://doi.org/10.1089/cyber.2020.0100
Chang, Y., Keblis, M. F., Li, R., Iakovou, E., & White, C. C. (2022). Misinformation and disinformation in modern warfare. *Operations Research, 70*(3), 1577–1597. https://doi.org/10.1287/opre.2021.2253
Chemerys, H. (2024). Deepfakes as a problem of modernity: a brief overview and current state. *Scientific journal of Khortytsia National Academy*, 162–173. https://doi.org/10.51706/2707-3076-2023-10-17
Chen, F., Zhou, J. H., Holzinger, A., Fleischmann, K. R., & Stumpf, S. (2023). Artificial intelligence ethics and trust: From principles to practice. *IEEE Intelligent Systems, 38*(6), 5–8. https://doi.org/10.1109/mis.2023.3324470
Chowdhury, S. M. A. K., & Lubna, J. I. (2020). Review on deep fake: A looming technological threat. In *International Conference on Computing, Communication and Networking Technologies* (pp. 1–7). https://doi.org/10.1109/ICCCNT49239.2020.9225630
Cover, R. (2022). Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. *Continuum: Journal of Media & Cultural Studies, 36*(4), 609–621. https://doi.org/10.1080/10304312.2022.2084039
Diakopoulos, N., & Johnson, D. G. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society, 23*(7). https://doi.org/10.1177/1461444820925811. s. 2072–2098.
Dickinson, D. L. (2024). Political ideology, emotion response, and confirmation bias. *Economic Inquiry, 23*(7), 2072–2098. https://doi.org/10.1111/ecin.13253
Ding, F., Fan, B., Shen, Z., Yu, K., Srivastava, G., Dev, K., & Wan, S. (2023). Securing facial bioinformation by eliminating adversarial perturbations. *IEEE Transactions on Industrial Informatics, 19*(5), 6682–6691. https://doi.org/10.1109/TII.2022.3201572
Doğan Akkaya, F. (2024). Deepfake dilemmas: Imagine tomorrow's surveillance society through three scenarios. *Journal of Economy Culture and Society, 70*, 121–134. https://doi.org/10.26650/jecs2024-1462119
Domenteanu, A., Tătaru, G.-C., Crăciun, L., Molănescu, A.-G., Cotfas, L., & Delcea, C. (2024). Living in the age of deepfakes: A bibliometric exploration of trends, challenges, and detection approaches. *Information, 15*(9), 525. https://doi.org/10.3390/info15090525
Doss, N., Wu, Y., Yang, P., & Zhou, H. H. (2023). Optimal estimation of high-dimensional Gaussian location mixtures. *The Annals of Statistics, 51*(1), 62–95. https://doi.org/10.1214/22-aos2207
Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology, 34*(4), 623–643. https://doi.org/10.1007/S13347-020-00419-2
Fehring, J. A., & Bonaci, T. (2023). It looks like me, but it isn't me: On the societal implications of deepfakes. *IEEE Potentials, 42*(5), 33–38. https://doi.org/10.1109/mpot.2022.3229823
Ferrara, E. (2024). GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science, 7*(1), 549–569. https://doi.org/10.1007/s42001-024-00250-1
Flattery, T., & Miller, C. (2024). Deepfakes and Dishonesty. *Philosophy & Technology, 37*(4). https://doi.org/10.1007/s13347-024-00812-1
Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262. https://doi.org/10.1038/S42256-019-0055-Y
Frankovits, G., & Mirsky, Y. (2023). Discussion paper: The threat of real time deepfakes. *arXiv.*. https://doi.org/10.48550/arXiv.2306.02487
Gambin, A. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review, 57*(3). https://doi.org/10.1007/s10462-023-10679-x
Gasser, U. (2023). An EU landmark for AI governance. *Science, 380*(6651), 1203. https://doi.org/10.1126/science.adj1627. –1203.
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
Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. *IEEE Access, 8*, 165085–165098. https://doi.org/10.1109/ACCESS.2020.3023037
Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese, 201*(3). https://doi.org/10.1007/s11229-023-04097-3
Hacker, P. (2023). AI Regulation in Europe: From the AI act to future regulatory challenges. *arXiv.*. https://doi.org/10.48550/arxiv.2310.04072
Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A comprehensive scoping review. *Minds and Machines, 34*(4). https://doi.org/10.1007/s11023-024-09694-w
Meer, T. G. van der Hameleers, M., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media and Society, 8*(3). https://doi.org/10.1177/20563051221116346.
van der Hameleers, M., Meer, T. G., & Dobber, T. (2023a). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, Article 108096. https://doi.org/10.1016/j.chb.2023.108096.
van der Hameleers, M., Meer, T. G., & Dobber, T. (2023b). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39*(1), 56–70. https://doi.org/10.1177/02673231231184703.
Han, M. (2024). The infringement of deepfake technology on personal privacy and legal protection: A discussion based on article 1032 of the civil code. *Journal of Education, Humanities and Social Sciences, 41*, 188–197. https://doi.org/10.54097/s0a47e08
Hao, J. (2024). Advances in deepfake generation and detection technologies: Challenges and opportunities. *Transactions on Computer Science and Intelligent Systems Research, 6*, 13–21. https://doi.org/10.62051/m3s6nc42
Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese, 199*(5–6), 13373–13391. https://doi.org/10.1007/S11229-021-03379-Y
Havlík, M. (2023). Deepfake as an advanced manipulative technique for spreading propaganda. *Vojenské Rozhledy, 32*(1), 3–17. https://doi.org/10.3849/2336-2995.32.2023.01.003-017
Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society, 31*(4), 520–524. https://doi.org/10.1017/s0305004100013517
Hoek, S., Metselaar, S., Ploem, C., & Bak, M. (2024). Promising for patients or deeply disturbing? The ethical and legal aspects of deepfake therapy. *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2024-109985
Holzschuh, N. (2023). Doctor Faustus Redux: Deepfake technology, sociological disruption of true knowledge, and the moral crisis facing our generation. *International Journal for Multidisciplinary Research, 5*(3). https://doi.org/10.36948/ijfmr.2023.v05i03.3325
Hughes, S. (2024). Hearts and minds: The technopolitical role of affect in sociotechnical imaginaries. *Social Studies of Science, 54*(6), 907–930. https://doi.org/10.1177/03063127241257489
Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*, 147. Available at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war.
Iqbal, F., Abbasi, A., Javed, A. R., Almadhor, A., Jalil, Z., Anwar, S., & Rida, I. (2023). Data Augmentation-based novel deep learning method for deepfaked images detection. *ACM Transactions on Multimedia Computing, Communications, and Applications, 20*(11), 1–15. https://doi.org/10.1145/3592615
Jointly Defending DeepFake. (2023). Manipulation and adversarial attack using decoy mechanism. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(8), 9922–9931. https://doi.org/10.1109/tpami.2023.3253390
Josephs, E., Fosco, C., & Oliva, A. (2023). Artifact magnification on deepfake videos increases human detection and subjective confidence. *arXiv.*. https://doi.org/10.48550/arXiv.2304.04733
Kaate, I., Salminen, J., Santos, J. M., Jung, S., Almerekhi, H., & Jansen, B. J. (2024). There is something rotten in Denmark": Investigating the Deepfake persona perceptions and their Implications for human-centered AI. *Computers in Human Behavior. Artificial Humans, 2*(1), Article 100031. https://doi.org/10.1016/j.chbah.2023.100031
Karasavva, V., & Noorbhai, A. (2021). The real threat of deepfake pornography: A review of Canadian policy. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 203–209. https://doi.org/10.1089/CYBER.2020.0272
Mags Karpinska-Krakowiak, M., & Eisend, M. (2024). Realistic portrayals of untrue information: The effects of deepfaked ads and different types of disclosures. *Journal of Advertising*, 1–11. https://doi.org/10.1080/00913367.2024.2306415.
Karunian, A. Y. (2024). The imitation game: Examining regulatory challenges of political deepfakes in the European Union. *Center for Open Science*. https://doi.org/10.31235/osf.io/59qjw
Katarya, R., & Lal, A. (2020). A study on combating emerging threat of deepfake weaponization. In *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)* (pp. 485–490). https://doi.org/10.1109/I-SMAC49090.2020.9243588
Keulartz, J., Schermer, M., Korthals, M., & Swierstra, T. (2004). Ethics in technological culture: A programmatic proposal for a pragmatist approach. *Science, Technology, & Human Values, 29*(1), 3–29. https://doi.org/10.1177/0162243903259188
Kietzmann, J., Mills, A. J., & Plangger, K. (2021). Deepfakes: Perspectives on the future "reality" of advertising and branding. *International Journal of Advertising, 40*(3), 473–485. https://doi.org/10.1080/02650487.2020.1834211

Kumar, N., & Kundu, A. (2024). SecureVision: Advanced cybersecurity deepfake detection with big data analytics. *Sensors, 24*(19), 6300. https://doi.org/10.3390/s24196300

Li, W., & Zhao, H. (2024). It's Up to Me Whether I Do – Or Don't – Watch Deepfakes": Deepfakes and behavioral intention. *SAGE Open, 14*(4), Article 21582440241302282. https://doi.org/10.1177/21582440241302282

Łabuz, M. (2023). Regulating deep fakes in the artificial intelligence act. *Applied Cybersecurity & Internet Governance, 2*(1), 1–42. https://doi.org/10.60097/acig/162856

Łabuz, M. (2024). Deep fakes and the Artificial Intelligence Act – An important signal or a missed opportunity? *Policy & Internet, 16*(4), 783–800. https://doi.org/10.1002/poi3.406

Łabuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science, 23*(4), 454–473. https://doi.org/10.1057/s41304-024-00482-9

Landon-Murray, M., Mujkic, E., & Nussbaum, B. (2019). Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of Fake News, Social Media, and Artificial Intelligence. *Public Integrity, 1*(5), 512–522. https://doi.org/10.1080/10999922.2019.1613832

Laurier, L., Giulietta, A., Octavia, A., & Cleti, M. (2024). The cat and mouse game: The ongoing arms race between diffusion models and detection methods. *arXiv*. https://doi.org/10.48550/arxiv.2410.18866

Lees, D., Bashford-Rogers, T., & Keppel-Palmer, M. (2021). The digital resurrection of Margaret Thatcher: Creative, technological and legal dilemmas in the use of deepfakes in screen drama. *Convergence, 27*(4), 954–973. https://doi.org/10.1177/13548565211030452

Lowenstein, H., Steinfeld, N., & Rosenberg, H. (2024). The ethical implications of prosocial synthetic resuscitation: Analysing user comments to a deepfake campaign addressing intimate partner violence. *Journal of Creative Communications, 20*(1), 23–40. https://doi.org/10.1177/09732586241276984

Lu, H., & Yuan, S. (2024). I know it's a deepfake": the role of AI disclaimers and comprehension in the processing of deepfake parodies. *Journal of Communication, 20*(1), 23–40. https://doi.org/10.1093/joc/jqae022

Lyu, S. (2024). DeepFake the menace: Mitigating the negative impacts of AI-generated content. *Organizational Cybersecurity Journal, 4*(1), 1–18. https://doi.org/10.1108/ocj-08-2022-0014

MacCarthaigh, M., & McKeown, C. (2021). A political science perspective on fake news. *The Information Retrieval Series* (pp. 233–243). https://doi.org/10.1007/978-3-030-62696-9_11

Machin-Mastromatteo, J. D. (2023). Community-driven and social initiatives. *Information Development, 39*(3), 393–401. https://doi.org/10.1177/02666669231197243

McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society, 26*(5), 2786–2803. https://doi.org/10.1177/14614448221093943

Meneses, J.-P. (2024). Seeking to define deepfakes from U.S. state laws. *Communication & Society*, 219–235. https://doi.org/10.15581/003.37.3.219-235

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

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR), 54*(1), 1–41. https://doi.org/10.48550/arXiv.2004.11138

Momeni, M. (2024). Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *Journal of Creative Communications, 20*(1), 41–56. https://doi.org/10.1177/09732586241277335

Morris, K. W. (2024). Deepfake Sockpuppets: The toxic "realities" of a weaponised internet. *Palgrave Gothic* (pp. 61–79). https://doi.org/10.1007/978-3-031-43852-3_5

Murphy, G., Twomey, J. E., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *PLOS ONE, 8*(7), Article e0287503. https://doi.org/10.1371/journal.pone.0287503

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, Article 113368. https://doi.org/10.1016/j.jbusres.2022.113368

Napshin, S., Paul, J., & Cochran, J. (2024). Individual responsibility around deepfakes: It's no laughing matter. *Cyberpsychology, Behavior, and Social Networking, 27*(2), 105–110. https://doi.org/10.1089/cyber.2023.0274

Navarro Martínez, O., Fernández-García, D., Cuartero Monteagudo, N., & Forero-Rincón, O. (2024). Possible health benefits and risks of deepfake videos: A qualitative study in nursing students. *Nursing Reports, 14*(4), 2746–2757. https://doi.org/10.3390/nursrep14040203

Nguyen, L., Ngwenyama, O. K., Bandyopadhyay, A., & Nallaperuma, K. (2023). Realising the potential of digital health communities: A study of the role of social factors in community engagement. *European Journal of Information Systems, 33*(6), 1033–1068. https://doi.org/10.1080/0960085x.2023.2252390

Noor, L., Malahat, I., & Noor, H. (2024). The socio-political implications of deepfakes in developing countries. https://doi.org/10.20944/preprints202409.1654.v1.

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology, 22*(2), 133–140. https://doi.org/10.1007/S10676-019-09522-1

Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G., & Wang, Y. (2023). Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *Npj Digital Medicine, 6*(1). https://doi.org/10.1038/s41746-023-00965-x

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access, 1*, 143296–143323. https://doi.org/10.1109/access.2023.3342107

Paterson, J. M. (2024). AI Deepfakes on the Web: The "Wicked" Challenges for AI Ethics, Law and Technology. In *Proceedings of the ACM Web Conference 2024*. https://doi.org/10.1145/3589334.3649116

Pawelec, M. (2022). Deepfakes and democracy (Theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society, 1*(2). https://doi.org/10.1007/s44206-022-00010-6

Pawelec, M. (2024). Decent deepfakes?. *Professional deepfake developers' ethical considerations and their governance potential. AI and Ethics*. https://doi.org/10.1007/s43681-024-00542-2

Pehlivanoglu, D., Zhu, M., Zhen, J., Gagnon-Roberge, A. A. C. M., Kern, R., Woodard, D. L., Cahill, B. S., & Ebner, N. C. (2024). Is this real? Susceptibility to deepfakes in machines and humans. *Center for Open Science*. https://doi.org/10.31219/osf.io/etxzw

Prochaska, S. L., Duskin, K. R., Kharazian, Z., Blucker, S., West, J. D., & Starbird, K. (2023). Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 U.S. Presidential Election. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–39. https://doi.org/10.1145/3579616

Rathoure, N., Pateriya, R. K., Bharot, N., & Verma, P. (2024). Combating deepfakes: A comprehensive multilayer deepfake video detection framework. *Multimedia Tools and Applications, 83*(38). https://doi.org/10.1007/s11042-024-20012-5. s. 85619–85636.

Rini, R., & Cohen, L. H. (2022). Deepfakes, deep harms. *Journal of Ethics & Social Philosophy, 22*(2). https://doi.org/10.26556/jesp.v22i2.1628

de Ruiter, A. D. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology, 34*(4), 1311–1332. https://doi.org/10.1007/S13347-021-00459-2

Sabanovic, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics, 2*(4), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Samuelson, P. (2023). Generative AI meets copyright. *Science, 381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

Sandoval, M. P., Vau, M., de, A., Solaas, J., & Rodrigues, L. F. D. (2024). Threat of deepfakes to the criminal justice system: A systematic review. *Crime Science, 13*(1). https://doi.org/10.1186/s40163-024-00239-1

Sayler, K.M., & Harris, L.A. (2023). Deep fakes and national security. *CRS Report F11333 (Version 7, Updated)*..

Seibert, D., Hoffmann, C. P., & Godulla, A. (2024). Deepfakes in the context of AI inequalities: Analysing disparities in knowledge and attitudes. *Information, Communication & Society, 28*(2), 295–315. https://doi.org/10.1080/1369118x.2024.2420037

Shahodat, B. (2022). Deepfakes, intellectual cynics, and the cultivation of digital sensibility. *Royal Institute of Philosophy Supplement, 92*, 67–85. https://doi.org/10.1017/s1358246122000224

Shin, S. Y., & Lee, J. (2022). The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes. *Digital Journalism, 10*(3), 412–432. https://doi.org/10.1080/21670811.2022.2026797

Singh, P., & Dhiman, B. (2023). Exploding AI-generated deepfakes and misinformation: A threat to global concern in the 21st century. *Institute of Electrical and Electronics Engineers (IEEE)*.. https://doi.org/10.36227/techrxiv.24715605.v1

Sloot, B. V. D., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. *Computer Law & Security Review, 46*, Article 105716. https://doi.org/10.1016/j.clsr.2022.105716

Sorell, T. (2023). Deepfakes and political misinformation in U.S. elections. *Techné: Research in Philosophy and Technology, 27*(3), 363–386. https://doi.org/10.5840/techne20231110190

Soto-Sanfiel, M. T., & Wu, Q. (2024). Individual attitudes toward deepfakes of deceased artists. *Center for Open Science*. https://doi.org/10.31235/osf.io/7gve2

Stewart, J., & Williams, R. (2000). The co-evolution of society and multimedia technology. *Social Dimensions of Information Technology* (pp. 46–62). https://doi.org/10.4018/978-1-878289-86-5.CH004

Supriya, M., Shree, S., Arya, R., & Roy, S. K. (2024). Investigating the evolving landscape of deepfake technology: Generative AI's role in it's generation and detection. *International Research Journal on Advanced Engineering Hub, 2*(05), 1489–1511. https://doi.org/10.47392/irjaeh.2024.0206

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG), 36*(4), 1–13. https://doi.org/10.1145/3072959.3073640

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 174*, 1–16. https://doi.org/10.1145/3411764.3445699

Taylor, B. C. (2021). Defending the state from digital Deceit: the reflexive securitization of deepfake. *Critical Studies in Media Communication, 38*(1), 1–17. https://doi.org/10.1080/15295036.2020.1833058

Trinh, L., Tsang, M., Rambhatla, S., & Liu, Y. (2021). Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. In *IEEE Winter Conference on Applications of Computer Vision* (pp. 1972–1982). https://doi.org/10.1109/WACV48630.2021.00202

Vasist, 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*, 590–636. https://doi.org/10.17705/1cais.05126

Navarro Martínez, O., Fernández-García, D., Cuartero Monteagudo, N., & Forero-Rincón, O. (2024). Possible Health Benefits and Risks of DeepFake Videos: A qualitative study in nursing students. *Nursing Reports, 14*(4), 2746–2757. https://doi.org/10.3390/nursrep14040203

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

Veerasamy, N., & Pieterse, H. (2022). Rising Above Misinformation and Deepfakes. *Proceedings of the International Conference on Information Warfare and Security, 17*(1), 340–348. https://doi.org/10.34190/iccws.17.1.25

Verma, N. (2024). One Video Could Start a War": A qualitative interview study of public perceptions of deepfake technology. *Proceedings of the Association for Information Science and Technology, 61*(1), 374–385. https://doi.org/10.1002/pra2.1035

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication, 9*(1), 291–300. https://doi.org/10.17645/MAC.V9I1.3494

Walker, C., Schiff, D., & Schiff, K. J. (2024). Merging AI incidents research with political misinformation research: Introducing the political deepfakes incidents database. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23053–23058. https://doi.org/10.1609/aaai.v38i21.30349

Wan, Y. (2023). The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research, 33*(5), 1750–1773. https://doi.org/10.1108/intr-07-2022-0561

Wang, L., Zhou, L., Yang, W., & Yu, R. (2022a). Deepfakes: A new threat to image fabrication in scientific publications? *Patterns, 3*(5), Article 100509. https://doi.org/10.1016/j.patter.2022.100509

Wang, R., Huang, Z.-S., Chen, Z., Liu, L., & Wang, L. (2022b). Anti-Forgery: Towards a Stealthy and Robust DeepFake disruption attack via adversarial perceptual-aware perturbations. In *International Joint Conference on Artificial Intelligence*. https://doi.org/10.48550/arXiv.2206.00477

Wang, X., Du, Y., Lin, S., Cui, P., Shen, Y., & Yang, Y. (2020). adVAE: A self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection. *Knowledge Based Systems, 190*, Article 105187. https://doi.org/10.1016/J.KNOSYS.2019.105187

Wang, Y. (2023). Synthetic Realities in the Digital Age: Navigating the opportunities and challenges of AI-Generated Content. *Institute of Electrical and Electronics Engineers (IEEE)*.. https://doi.org/10.36227/techrxiv.23968311

Wazid, M., Mishra, A. K., Mohd, N., & Das, A. K. (2024). A secure Deepfake mitigation framework: Architecture, issues, challenges, and societal impact. *Cyber Security and Applications, 2*, Article 100040. https://doi.org/10.1016/j.csa.2024.100040

Weikmann, T. E., & Lecheler, S. (2023). 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

Weikmann, T. E., Greber, H., & Nikolaou, A. (2024). After deception: How falling for a Deepfake affects the way we see, hear, and experience media. *The International Journal of Press/Politics, 30*(1), 187–210. https://doi.org/10.1177/19401612241233539

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation,, 125*, Article 102784. https://doi.org/10.1016/j.technovation.2023.102784

Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy, 5*(2), 199–217. https://doi.org/10.1080/23738871.2020.1797135

Zheng, G., Shu, J., & Li, K. (2025). Regulating deepfakes between Lex Lata and Lex ferenda – a comparative analysis of regulatory approaches in the US, the EU and China. *Crime, Law and Social Change, 83*(1), 1–23. https://doi.org/10.1007/s10611-024-10197-z