The Roots and Routes of Deepfakes: Towards an ontological, typological and sociotechnical perspective

Johnny Tang-Ear
University of Sydney
johnny.tang-ear@sydney.edu.au

Catherine Hardy
University of Sydney
catherine.hardy@sydney.edu.au

Mike Seymour
University of Sydney
mike.seymour@sydney.edu.au

Abstract

This paper offers a critical review of the evolving concept of deepfakes, tracing their roots in machine learning-based face-swapping technologies to their broader sociotechnical implications in contemporary society. While early research focused on technical detection and generation, the term 'deepfake' now broadly refers to identity deception. We analyze how the barriers to deepfake creation have lowered due to accessible tools, and how public understanding has shifted from technical specificity to generalized concern. Drawing on literature from information systems (IS), media studies, and elsewhere, we develop a typology that categorizes deepfakes by intent, realism, technological accessibility, and sociotechnical impact. We also propose an initial ontology to clarify conceptual boundaries, distinguishing deepfakes from related terms such as "cheap fakes" and "synthetic media." This interdisciplinary perspective contributes to IS discourse by framing deepfakes as dynamic artifacts embedded within complex sociotechnical systems.

Key words: sociotechnical, AI, Deepfakes, ontology, hermeneutic, critical review.

1. Introduction

Deepfake phenomenon is now omnipresent, attracting attention at the highest levels of government, commented upon widely in the media and experienced by people in their everyday lives (Atlam et al., 2025). Since the term deepfake was created in 2017 (Seymour et al, 2023), it has been used abundantly in the scholarly and practitioner literature. Yet, definitional issues remain (Altuncu et al., 2024; Birrer & Just, 2024; Rancourt-Raymond & Smaili, 2023; Whittaker et al., 2023). This paper traces the emergence of the term 'deepfake' and how its meaning has evolved alongside the phenomena it describes. We address this aim in the following research question: *How is the Deepfake term represented in the scholarly and practitioner literature?*

The research serves two key objectives. Firstly, we provide definitional clarity as to what the term encompasses, developing a typology and ontology, for guiding research (Vasist & Krishnan, 2022). In the critical review we focus on how the term "deepfake" has evolved to assist with identifying what it is and where it came from (its roots) as well as what it does and how it is used (the routes). Secondly, several reviews of deepfakes have been conducted that have largely adopted a systematic (e.g. Whittaker et al., 2023) and integrative (e.g. Vasist & Krishnan, 2022) approach to synthesize the literature and identify research gaps with the exception of the critical reviews by Vasist and Krishnan (2023) and Twomey et al. (2025). We build upon these contributions, engaging with a more hermeneutical approach to critically scrutinize how the deepfake phenomenon is constructed and its underlying assumptions, or what Alvesson and Sandberg, (2020) refer to as a problematizing review. In doing so we reveal contestations and inconsistencies in how the term is deployed in scholarly and public discourse, uncovering the dynamics of social and technical aspects and opening up possibilities for novel ways of thinking about the deepfake phenomenon through a sociotechnical framing (Sarker et al., 2019).

The paper is organized as follows. Firstly, we provide a conceptual background of existing studies and understandings of deepfakes. We then introduce our methodology. The findings of the critical review follow. Finally, we propose a typology and an initial ontology of the term, highlight our contributions, discuss limitations and future research opportunities and provide a conclusion.

URI: https://hdl.handle.net/10125/111932
978-0-9981331-9-5

2. Background

We first describe the core ideas of deepfakes and their application. Then, we summarize previous literature reviews related to deepfakes and justify the need for this review.

2.1 An introduction to deepfakes

The origins of the term deepfake trace back to an anonymous user named "deepfakes" on the Reddit social media platform in 2017 creating sexually explicit videos of famous celebrities using face swapping technologies to produce non-consensual content (Seymour et al, 2023, 2019; Kietzmann, Mills, et al., 2020; Kirchengast, 2020; Lyu, 2020; Meskys et al., 2019). The subsequent release of an application for generating deepfakes in the Reddit community led to an explosion of fake content and is commonly recognized as the beginning of the deepfake phenomenon (Hawkins, 2018 cited in Delfino, 2019; Kietzmann, Lee, et al., 2020).

After the term 'deepfake' was coined, early attention focused on technical aspects, mainly training and testing machine learning (ML) models (Afchar et al., 2018; Guera & Delp, 2018; Li et al., 2018; Maras & Alexandrou, 2018). In these early stages, deepfakes were predominantly associated with face-swapping technologies requiring significant technical proficiency in areas such as coding, machine learning, and algorithmic modeling, including autoencoders. Tools like DeepFaceLab began to emerge, marking a shift toward greater accessibility, although still requiring users to download and compile source code locally.

Over time the term has come to be used widely in the scholarly and practitioner literature. Yet there is currently no generally agreed upon definition. There is some general consensus about deepfakes as a form of media and content and a technology that generates media and content through AI and machine learning techniques (Altuncu et al., 2024; Birrer & Just, 2024; Rancourt-Raymond & Smaili, 2023; Whittaker et al., 2023). The deepfake term itself is often viewed as harmful and pejorative (Millière, 2022). It is has been characterized as "creepy artifacts" of "toxic geek masculinity incubating in the dark underbelly of the internet," misinformation on a "catastrophic scale" and associated with apocalyptic visions of the "gold standard" of evidentiary truth diminishing, presenting an "epistemic threat to Western democracy" (Broinowski, 2022). Further, advances in artificial neural network (ANN) based technologies have made tools more accessible to non-technical specialists for creating deepfakes and what is generated more realistic and believable (Broinowski, 2022; Rana et al., 2022), amplifying fears of the further weaponization of deepfakes by "resource-rich bad actors" (Wahl-Jorgensen & Carlson, 2021 cited in Twomey et al., 2025).

The harmful impacts arising from the use of deepfake technology are not just dystopian predictions and unevidenced. There have been multiple cases of malicious and threatening behaviors witnessed over time. For example, recent reported cases of deepfakes posted in Australia about the India-Pakistan conflict (ABC News, 2025), Australian public servants creating sexual deepfakes of colleagues (Braue, 2025) and teenagers selling deepfake schoolgirl porn images to build their "DIY porn hubs" (Panagopoulos & Bita, 2025). Cases, such as these, have been sufficiently harmful for an array of government and industry efforts to be initiated across multiple jurisdictions (Broinowski, 2022) to address ethical, policy and legal issues related to deepfakes (see for e.g. "Criminal Code Amendment (Deepfake Sexual Material) Bill," 2024) and artificial intelligence (AI) more broadly (see for e.g. Smuha, 2025). However, some argue that there is an "asymmetrical coverage" of deepfakes in the news media focusing attention on potential harms rather than their current uses (Twomey et al., 2025) and with limited cross-geographic coverage (Vasist & Krishnan, 2022). Hence a disparity exists in the reporting of deepfake technology that has been applied beneficially, such as, cloning voices to assist those who have degenerative illnesses to continue to verbally communicate (Langa, 2021; Meskys et al., 2019) or the "Malaria Must Die" campaign, in which David Beckham's likeness was synthetically altered to deliver a multilingual awareness message (Meskys et al., 2019). Further, dystopian and utopian narratives surrounding deepfake technology have arguably replicated "hysterias" that have been witnessed in the past with the emergence of 'new' technologies (Twomey et al., 2025).

Despite the ubiquity of the deepfake phenomenon it remains somewhat of an enigma. Definitions and meanings continue to be the subject of contest, impacting research progress and raising questions about what is being governed when governing deepfakes. We elaborate upon this in the next section and argue that an ontology is needed that describes the technology and its generating tools as well as the application of this technology within different settings to advance understanding.

2.2 A review of previous literature reviews

Previous deepfake literature reviews are examined to illustrate how this study contributes to this stream of research. Table 1 presents an overview of prior reviews, that have adopted a systematic (Rana et al., 2022; Stroebel et al., 2023; Whittaker et al., 2023) or integrative (Vasist & Krishnan, 2022) approach to synthesize the literature. Vasist and Krishnan (2023) and Twomey et al. (2025) are exceptions providing a critical review, using a multidisciplinary and qualitative approach. The reviews were based on academic publications except for Westerlund (2019) who analyzed public news articles sourced from USA based news websites. The review discussion is divided into two sections, namely definitional issues and theoretical framing.

Table 1: Overview of prior deepfake literature reviews

Author(s)	Deepfake(s) “is/are ...”	Review period & sources
Westerlund (2019)	“Deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never happened”	2018-2019 84 public news articles from 11 USA news companies’ websites
Vasist and Krishnan (2022)	Definitional issues relating to three aspects: (1) not a new phenomenon as there is a long history of video alterations although technological advancements have enabled more sophisticated manipulations; (2) unclear what is included as multimedia content (e.g. only face swapping technology or is text also included) (3) the “term ‘fake’ is etymologically erroneous”	2001-2022 68 peer-reviewed journal articles (includes discussion and opinion papers). Excludes technical aspects of deepfake production or detection.
Rana et al. (2022)	“The term ‘Deepfake’ is derived from ‘Deep Learning (DL) and ‘Fake’ and it describes specific photo-realistic video or image contents created with DL’s support.”	January 2018-December 2020 112 research articles and reviews (including conference, workshops, journals and archives) on deepfake detection.
Vasist and Krishnan (2023)	“Deepfakes are a relatively new phenomenon that has been extensively debated for both their positive and bad applications.” Need to “shed light on the socio-technical aspects of deepfake engagement.”	2012-2022 Meta-synthesis of 16 research studies using a qualitative research design only.
Stroebel et al. (2023)	“a name coined from Deep Learning and Fake which exploits the powers of deep learning to create fake, realistic video and image content.”	January 2021 – August -2022 83 papers of deepfake detection methods.
Whittaker et al. (2023)	“Deepfakes are synthetic media generated using artificial intelligence and deep learning technology which produce realistic yet fake representations of people undertaking actions or saying words in the form of video, image, or audio content.”	January 1, 2017 – June 9, 2021 80 peer-reviewed journals (includes articles, editorials & commentaries). Excludes deepfake technical development
Twomey et al. (2025)	“Deepfakes are inconsistently defined but there is some general consensus as to what constitutes a deepfake.” For example, “their use of deep-learning technology, a subset of machine learning and artificial intelligence.”	January 1, 2018 – January 11, 2023 102 academic papers

2.2.1 Definitional issues. While the origin of the deepfake term is widely agreed upon, more recent reviews conducted by Vasist and Krishnan (2022) and Twomey et al. (2025) highlight ongoing definitional issues. Twomey et al. (2025) found that efforts to synthesize definitions in previous reviews and theoretical work had “added to the confusion,” summarized here as including three key aspects. Firstly, previous reviews had not critically reflected on different disciplinary perspectives of the nature and impacts of deepfakes. While Twomey et al. (2025) found similar themes to the definition of Whittaker et al. (2023), their critical analysis highlighted the importance of *mimicry* and *identity*. Mimicry was considered by Twomey et al. (2025) to have been understated in previous reviews of deepfake definitions, emphasizing the generation of the spoken word rather than how deepfakes mimic “people ‘saying’ or doing things which they never actually ‘said’ or ‘did’” Further, Twomey et al. (2025) argue that the role of manipulated identity is fundamental to the harmful impacts of deepfakes due to the “capacity of the technology to control or appropriate the identity of its target and to erase or disguise the identity of its source".

Secondly, previous reviews had not interrogated in sufficient detail how deepfake technology is novel or different to technologies preceding them. Comparisons of deepfakes with other media technologies have tended to be represented in the *deepfake-shallowfake* dichotomy, the *cheapfake-deepfake* spectrum and as part of a taxonomy of synthetic media (Twomey et al., 2025). Shallowfakes are distinguished from deepfakes based on whether media is manipulated solely through human intervention with respect to the former or using deep learning (DL) technology with respect to the latter (Twomey et al., 2025). Cheapfakes are distinguished from deepfakes along a spectrum based on the sophistication of technology used to create fake media, with 'lower-quality deepfakes' or 'cheap fakes' associated with 'consumer grade or 'simple video-editing techniques' (Twomey et al., 2025). These two terms and comparisons also interacted with the ideas of deepfakes as a technology being more accessible and as more believable form. However, Twomey et al. (2025) also argued that these terms 'muddied the waters' as to what constituted a deepfake as: traditional editing can be complex and expensive; realistic deepfakes may still require some specialized technical skills; and the forms of image manipulation referred to have already been defined for decades. Twomey et al. (2025) viewed characterizations of deepfake technology alongside synthetic media as not only fundamental to understanding how deepfakes are different to other modes of media but also useful in determining what constitutes a deepfake highlighting the work of Millière (2022); who interestingly focused on the notion of *Deep Learning – based synthetic audiovisual media* (DLSAM) given the lack of a single clear definition of deepfakes. Millière (2022) proposed a broad taxonomy of audiovisual media, situating DLSAM alongside traditional media. Notwithstanding this research classifications of modes of media remain blurred. Kietzmann, Lee, et al. (2020), for example, classified four types of deepfakes, photo, audio, video as well as audio and video. Vasist and Krishnan (2022) called for further investigation into whether modified text should be included as a form of content or whether photographs that had been altered without the use of AI related technology should be viewed as deepfakes. Millière (2022) argued that boundaries between the categories of DLSAM and the generic label of 'deepfake' required greater scrutiny and that ML-based approaches had started to blur the lines between partial and totally synthetic media.

Thirdly, applications of deepfake technology across multiple fields have revealed malicious and beneficial aspects that are both gendered and exist at multiple levels (individual, organization, society) (Dagar & Vishwakarma, 2022). Twomey et al. (2025) classified three key types of harm: (1) "personal harms of defamatory pornographic deepfakes, particularly harassment and objectification" (2) "societal harms of defamatory political deepfakes," (e.g. elections and disinformation); and (3) epistemic harms of deepfakes to video (harms to truth and evidence)." Positive uses of deepfakes, across different disciplines, were focused on artistic, educational and self-expressive uses such as digitally resurrecting deceased actors in films, also noting the potential ethical issues of appropriating the identity of dead actors (Twomey et al., 2025). There are divergent perspectives on the use of deepfakes within and across several fields. For example, in the marketing field deepfakes are described as an emerging communication innovation (Whittaker et al., 2023) and eroding consumer trust in marketing (Kietzmann, Mills, et al., 2020) or in the political domain corrupting democratic processes and enabling regional dialects during the 2019 elections in India in the political domain (Vasist & Krishnan, 2022). Categorizations of deepfakes uses need reconsideration as they are not purely positive and negative, nor necessarily bound within individual organizational or societal domains. For example, deepfakes used to target specific groups such as celebrities, activists and politicians may not only cause defamatory harm but also themselves are political as they are designed to silence and harass targets that has individual, gender and societal effects (Twomey et al., 2025).

2.2.2 Theoretical framing. Three of the six reviews proposed a theoretical framework (Vasist & Krishnan, 2022; Whittaker et al., 2023) not limited to detection techniques only (see Rana et al., 2022; Stroebel et al., 2023). These three more holistic frameworks are reviewed in further detail.

Underpinning the framework developed by Vasist and Krishnan (2022) is a dyadic theoretical perspective grounded in the elaboration likelihood model (ELM) and the theory of reasoned action (TRA). These are described by Vasist and Krishnan (2022) as persuasion theories. The theoretical framework consists of five aspects, namely: 1) Motivations of deepfake creation (positive, negative peripheral). 2) Deepfake viewer responses (deepfake beliefs and detection). 3) Antecedents of deepfake dissemination (e.g. social influence). 4) Deepfake sharing mechanisms (intentional & unintentional) and 5) outcomes of deepfake dissemination (micro-individual, meso- organizational, macro – societal).

The framework developed by Whittaker et al. (2023) conceptualizes deepfakes as an innovation and utilizes a consumption values theory and ecosystem lens to develop understanding of the potential creative and destructive impacts of deepfakes. The use of deepfakes is examined at a micro (individual) and meso (organizational) level. Key actors are represented as: (1) creators (e.g. individual and commercial content creators); and (2) groups depicted (e.g. politicians, deceased). Potential value outcomes (epistemic, functional, emotional, social) of deepfakes for different ecosystem actors are categorized into two groups (1) value creation; and (2) value destruction.

Vasist and Krishnan (2023) viewed deepfakes as a socio-technical phenomenon and based their integrated conceptual framework on the social shaping of technology theory. Embedded in the framework are eight components: (1) motivations (malevolent and not malevolent); (2) dichotomous assertions (pro-deepfake and anti-deepfake); (3) digital platform enablers (technological expertise, networked spaces, marketplace ecosystem); (4) deepfake genres (malicious fabrications and innocuous fabrications); (5) dissemination (intentional sharing and inadvertent sharing); (6) challenges posed by digital platforms (measurement ambiguities, vagueness of platform standards, lack of control and moderating effects); (7) governing interventions (human-oriented mechanisms, technological infrastructure, legal instruments); and (8) consequences (social, organizational, individual).

3. Toward a socio-technical perspective

The need for a more comprehensive understanding of deepfakes is clearly recognized in the literature reviews discussed in Section 2. There is a complexity of ideas, meanings and technologies embedded in the deepfake term that has become a shorthand label for several phenomena. Since deepfake technology emerged, it has attracted significant attention across multiple disciplinary domains including information systems, computer science, social sciences, law and politics. The term encompasses multiple applications, a range of technologies, breadth of activities and characterizations of harm and benefits. From its original, technically specific meaning, namely, machine learning-based face-swapping it has evolved to a broader category encompassing various forms of identity deception. This semantic shift reflects a growing public awareness of the social and ethical risks associated with identity manipulation, often without a corresponding understanding of the underlying technologies. The "deepfake" term has become a catch-all label for technologically mediated identity deception, regardless of whether it involves machine learning, artificial intelligence, or other digital techniques. As a result, the term has expanded in both scope and usage, while simultaneously becoming less precise in describing the specific methods by which such deceptions are created.

Further, the need for a more comprehensive understanding of deepfakes is clearly recognized in these literature reviews. However, based on Sarker et al. (2019) categorization of sociotechnical research, we found that the reviews tended to focus predominately on technical or social aspects. That is, some prioritized technical elements in the use of deepfake technology, focusing on deepfake detection techniques (Rana et al., 2022; Stroebel et al., 2023). Other reviews emphasized primarily social elements. Two of the three theoretical frameworks focused on behaviors associated with beliefs or values associated with the use of deepfake technology (Vasist & Krishnan, 2022; Whittaker et al., 2023). Whilst Vasist and Krishnan (2023) conceptualization of deepfake engagement did take into account social and technical aspects together the relationship was represented in a unidirectional way, from social to technical. Further, the dichotomous classification of positive or malicious deepfakes was highlighted by Twomey et al. (2025) as a limitation. A deeper examination of what brings the social and technical aspects together is needed as such interactions pertain to the conditions of business models, organizational forms and governance arrangements (Sarker et al., 2019).

Ambiguity remains about the form and content of deepfakes, the technology that produces it and social impacts. Building on these existing studies, we undertake a critical review to bring further clarification as to what a deepfake is and is not. Ambiguity remains about the form and content of deepfakes, the technology that produces it and social impacts. We argue that this is a necessary step to assist with progressing the conceptualization of deepfakes as a sociotechnical phenomenon.

4. Review method

This research adopts a qualitative approach, involving a combination of a critical literature review (CLR) as well as a hermeneutic approach to provide structure through a thorough analysis of the current state (Williams et al., 2019 cited in Boell & Cecez-

Kecmanovic, 2015) of deepfakes and problematize the underlying challenges associated with deepfake technology (Boell & Cecez-Kecmanovic, 2014). Furthermore, a thematic analysis is undertaken using NVivo, a qualitative data analysis software to identify themes and concepts over time as well as what is qualitatively different about deepfakes and their impacts from previous technologies.

Table 2: Keywords and filters

Keywords:	**Filtered by:**
(“deepfake”) OR (“deep” AND “fake”) OR (“deep” AND “fakes”) OR (“deepfake” AND “technology”) OR (“deep” AND “fake” AND “technology”) OR (“deepfakes” AND “organisation”) OR (“deep” AND “fake” AND “organisations”) OR (“deepfake” AND “technology” AND “Australian organisation”) OR (“deepfakes”) OR (“AI” AND “Ethics”)	• Publication Years (2018 – 2024) • Document Types (Review Article/Peer -reviewed journals)

4.1 Critical Literature Review (CLR)

The search period and terms used for the critical review are set out in Table 2. The keyword analysis allowed for a broad search into the academic research on deepfake technology and narrow the scope to focus upon definitions of deepfakes over time. The journal articles selected in the CLR fall between the social and technical spectrum which exclude journals that purely discuss the technical aspects of deepfakes.

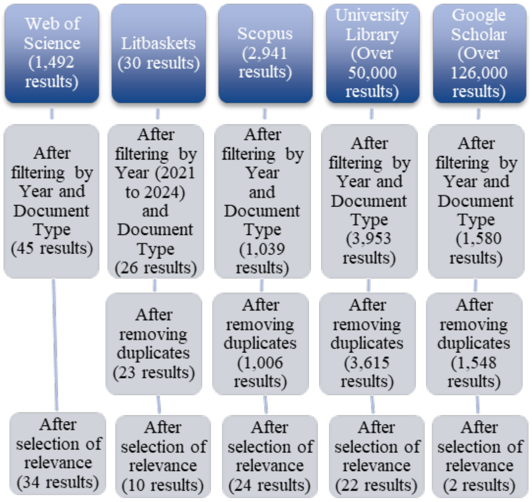
Web of Science (1,492 results)
Litbaskets (30 results)
Scopus (2,941 results)
University Library (Over 50,000 results)
Google Scholar (Over 126,000 results)
After filtering by Year and Document Type (45 results)
After filtering by Year (2021 to 2024) and Document Type (26 results)
After filtering by Year and Document Type (1,039 results)
After filtering by Year and Document Type (3,953 results)
After filtering by Year and Document Type (1,580 results)
After removing duplicates (23 results)
After removing duplicates (1,006 results)
After removing duplicates (3,615 results)
After removing duplicates (1,548 results)
After selection of relevance (34 results)
After selection of relevance (10 results)
After selection of relevance (24 results)
After selection of relevance (22 results)
After selection of relevance (2 results)

Figure 1: Data sources of literature review

A combination of data sources was used, consisting of academic and practitioner publications, as shown in Figure 1.

4.2 Hermeneutic approach

A hermeneutic approach involves an exploration and interpretation of text to uncover the expression of the different contexts within the literature in relation to cultural, historical and subjective factors. The approach also incorporates the intent of the author, historical background and the reader’s own perspective involving an iterative process (Hammersley, 2022).

The hermeneutic approach aligns with a constructivist epistemology, recognizing that meanings of deepfake are socially constructed and context dependent. This perspective acknowledges that not only how deepfakes are interpreted but also how they are governed, contested, and legitimated in sociotechnical systems.

5. Findings: Toward a typology

A thematic analysis was undertaken to classify how the deepfake term was represented in the 92 articles. Twenty descriptive codes were identified, using the NVivo software package (Lumivero, 2023), as listed in Table 3.

Table 3: Themes of deepfakes

Replacement	Tool	Trust	Disinformation
Superimpose	Deceptive	Misuse	Misinformation
Manipulation	Authentic	Harm	Malicious Intent
Improve Experiences	Fake News	Synthetic	Facial Reenactment
Face Swap	Realistic	Forgery	Technologies

A further classification of these themes was performed, using cluster analysis, revealing four conceptual dimensions as shown in Figure 2 informing the development of the typology; adopting Bailey’s (1994) definition of classification. These include consequences, detection, manipulation and generation.

Consequences: The use of deepfakes has intended and unintended consequences. *Harm* is a potential negative consequence, or a risk that has been materialized, such as the creation of deepfake pornography. Furthermore, *trust* in the use of deepfakes has significant consequences that can be positive (e.g. improved communication) or negative (e.g. loss of confidence) (Gambín et al., 2024; Habbal et al., 2024; Jacobsen & Simpson, 2024; Ray, 2021). Consequences are far-reaching impacting individuals, institutions and society more broadly, such as reputational damage (Abbas & Taeihagh, 2024; Broinowski, 2022; Habbal et al., 2024).

Detection: Tools are central to designing technology for deepfake detection. However, advances in AI have made this harder, as highly realistic synthetic faces can now be easily generated and animated. Ultimately, the purpose of detection is to differentiate deepfake usages with *malicious intent* from those that are *improving experiences.* (Brooks et al., 2022; Jacobsen & Simpson, 2024; Rana et al., 2022; Rao et al., 2021).

Manipulation: Facial reenactment and replacement are manipulation techniques (Ascott, 2020; Li et al., 2020), involving privacy and data security considerations (Siau & Wang, 2020).

Generation: The process of *generating* deepfakes, the technologies used, and types of media and content created (synthetic, fake news). Advances in AI and open-source tools now enable realistic manipulated content with little specialist expertise (Gong & Li, 2024; Tolosana, 2020 cited in Khoo et al., 2021). Deepfake technologies can result in harmful consequences, including the spread of misinformation, reputational damage to individuals, and a broader erosion of public trust (Malik et al., 2024; Renier et al., 2024). They can be weaponized (Karasavva & Noorbhai, 2021) for cybercriminal activities such as employee impersonation and document forgery (Dagar & Vishwakarma, 2022; Rana et al., 2022) and have the potential to exacerbate existing challenges like the proliferation of fake news (Siau & Wang, 2020).

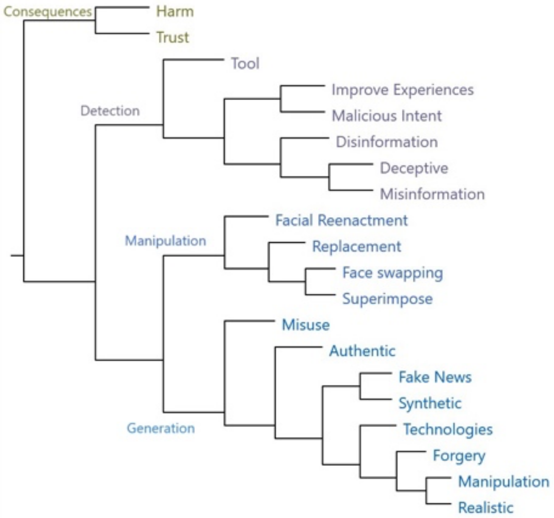

Figure 2: Cluster analysis from NVivo

The analysis revealed that in the academic literature, the focus tends to be more on video-based deepfakes than audio. Concurrently, the term "fake" is becoming increasingly problematic. As synthetic content becomes more realistic, distinguishing genuine from manipulated media has grown more difficult. Moreover, not all synthetic or altered content is intended to deceive or cause harm, although the term "deepfake" remains heavily associated with negative intent.

These findings align with earlier definitional challenges and reflect a growing trend toward describing such content as synthetic media. Originally, "deepfake" referred specifically to AI-driven face-swapping using ML technologies. Over time, however, the term has broadened to describe a wider range of identity-altering or generative media, particularly those perceived as harmful. Across multiple disciplines there also appears to be a shift towards using the term synthetic media in encapsulating deepfakes (Brooks et al., 2022; Firc et al., 2023; Khoo et al., 2021; Lees, 2024; McCosker, 2022; Rancourt-Raymond & Smaili, 2023; Rao et al., 2021; Sandoval et al., 2024; Sloot & Wagensveld, 2022; Twomey et al., 2025). Synthetic media then refers to AI-generated or AI-manipulated video, audio, or images that fabricate or mimic real-world content. Unlike traditional media creation or editing, synthetic media relies on generative models such as GANs or diffusion networks. Importantly, this category includes, but is not limited to, deepfakes.

The distinction lies in intent and framing. Deepfake implies deception and harm, while synthetic media captures a broader spectrum of uses, including creative, assistive, and artistic applications without malicious purpose, such as virtual avatars or multilingual public health campaigns such as the David Beckham "Malaria Must Die" initiative (Meskys et al., 2019). As Vasist and Krishnan (2022) argue, "fake" is both etymologically and conceptually misleading. Thus, while all deepfakes are a form of synthetic media, not all synthetic media should be classified as deepfakes. This shift in terminology reflects a deeper need to reframe how such technologies are understood and categorized within scholarly and public discourse.

The thematic analysis presented here in Section 5 provides the empirical grounding for the ontology developed in Section 6. The NVivo-coded clusters revealed recurring themes around intent (e.g., deception vs. creative expression), accessibility of tools, realism of outputs, and the varying levels of impact. By moving from descriptive themes to conceptual categories, we translate this analysis into an ontological framework.

6. Toward an ontology of *deepfake*

While the typology captures how deepfakes are described and interpreted across the literature, our ontology translates the recurring patterns of intent, accessibility, realism and impact into a more formalized conceptual structure. An ontology refers to an examination of the relationship through conceptualizing a computational artefact (Guarino et al., 2009). This shift from descriptive categorization to ontological modelling allows us to capture the complex interplay between technological generation and sociocultural context, positioning deepfakes not just as artifacts but as dynamic sociotechnical phenomena shaped by purpose, perception, and infrastructure.

Originally, *deepfake* referred to a narrow class of ML technologies, however, this definition has expanded considerably. The ontology must therefore begin with a technical core, encompassing methods of generation (e.g., face replacement, voice cloning, motion synthesis) and modalities or media types (video, audio, image). Deepfakes have also transitioned from isolated technological approaches to sociotechnical artifacts, where meaning is contingent not only on the tools used but also on user intent, context, and societal impact.

We propose a four-dimensional ontology that conceptualizes deepfakes along a sociotechnical continuum, enabling a more nuanced understanding than binary classifications of "real" or "fake." The first dimension: *Intent and use*, distinguishes between malicious applications such as political sabotage or defamation, and benign or beneficial uses such as education, or entertainment. The second dimension: *perceived authenticity*, captures the technical spectrum from indistinguishable from reality to artificial, shaping how audiences interpret and respond to the content. Third: *accessibility and tooling* which considers the ease with which such content can be produced, ranging from expert-only systems to mass-accessible consumer applications. Finally: *context impact* or the level at which the deepfake exerts influence on the individual or society. These dimensions are not independent; they interact dynamically to express the degree of "deepfakeness" of any given artifact. For example, a deepfake created with publicly available tools and deployed with malicious intent in a societal context, such as to disrupt an election, represents one end of the spectrum. In contrast, a moderately realistic, consent-based synthetic video of a public figure delivering multilingual health messages reflects positive intent, constructive impact, and ethical application, despite relying on similar technologies.

What distinguishes this ontology from prior frameworks is its explicit integration of both technical and social dimensions within a unified sociotechnical entanglement. While earlier studies have proposed taxonomies or conceptual models focused on specific aspects, such as motivations, innovation potential or disciplinary framings, these efforts often treat social and technical factors as separate influences. In contrast, our ontology treats deepfakes as dynamic artifacts whose meaning and impact emerge from the interaction of generation methods, user intent, perceived authenticity, accessibility, and contextual consequences. By embedding deepfakes within a continuum that spans individual, organizational, and societal levels, our ontology offers a more multidimensional structure for classification and analysis. This makes it not only more adaptable to emerging use cases (e.g., personalized AI educators), but also more useful for governance, detection, and policy-making efforts that must grapple with deepfakes' dual nature as both technical outputs and socially embedded phenomena.

Once a precise term for ML-generated face-swaps, 'deepfake' now broadly also includes low-tech and non-AI identity manipulation. Our ontology therefore defines a criterion that distinguishes deepfakes from adjacent terms like "cheap fakes" (simple edits) or "synthetic media" (a much broader umbrella). Anchoring it in the sociotechnical continuum (Sarker et al., 2019) allows researchers to differentiate artifacts not only by how they are made, but also by their performative role in social systems, trust infrastructures, and media ecologies.

7. Contribution and Future Research

Our research reveals definitional inconsistencies and disciplinary fragmentation while analyzing the evolving concept of "deepfakes" through a thematic and cluster analysis to derive a typology with themes of generation, detection, manipulation and consequences. Furthermore, a hermeneutic and critical approach evaluates different dimensions – *intent and use*, *accessibility and tooling*, *perceived authenticity* and *context impact* to enhance the clarity of "deepfake" and captures the dynamic interwoven dimensions with an ontology.

Once limited to AI-driven face-swapping, deepfakes now encompass a broader set of generative practices, prompting the need for clearer conceptual boundaries. In response, we proposed a multidimensional ontology that captures the complexity of deepfakes as sociotechnical artifacts, shaped not only by their method of production but also by intent, perceived authenticity, accessibility, and societal impact. The ontology offers a structured and flexible framework to differentiate between types of synthetic media and their potential implications. Unlike earlier taxonomies that treat technical and social factors separately, our approach foregrounds their interaction. This sociotechnical framing enables a more nuanced assessment of deepfakes—one that is sensitive to both technical sophistication and contextual deployment.

A critical literature review is limited to a variety of sources which in this case was very broad. Our boundaries were established to narrow the scope of the research across different fields that could be publicly sourced from different databases. Furthermore, with the rapid development of tools to create and manipulate deepfakes, the critical hermeneutic review provides only a historical summary of the literature and not a prediction of the future for such phenomenon.

Future research could explore how this ontology can support decision-making in areas such as regulation, governance, and platform moderation. Additionally, as generative AI technologies continue to evolve, further refinement of the framework will be needed to account for new modalities, cultural interpretations, and application domains. We argue that deepfakes should not be treated as a monolithic threat, but rather as a diverse class of sociotechnical phenomena requiring context-aware responses.

8. References

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems With Applications*, *252*, 1 - 38.

ABC News. (2025). VIDEO: Deepfakes posted as India-Pakistan conflict. *ABC News*.

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: a Compact Facial Video Forgery Detection Network* 2018 IEEE International Workshop on Information Forensics and Security (WIFS), China.

Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. *Frontiers in Big Data*, *7*, 1 - 31.

Alvesson, M., & Sandberg, J. (2020). The Problematizing Review: A Counterpoint to Elsbach and Van Knippenberg's Argument for Integrative Reviews. *Journal of Management Studies*, *57*(6), 1290 - 1304.

Ascott, T. (2020). Microfake: How small-scale deepfakes can undermine society. *Journal of Digital Media & Policy*, *11*(2), 215 - 222.

Atlam, E.-S., Almaliki, M., Elmarhomy, G., Almars, A. M., Elsiddieg, A. M. A., & ElAgamy, R. (2025). SLM-DFS: A systematic literature map of deepfake spread on social media. *Alexandria Engineering Journal*, *111*, 446-455.

Bailey, K. D. (1994). *Typologies and Taxonomies*. SAGE Publications Inc.

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*, 1 - 20.

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches. *Communications of the Association for Information Systems*, *34*, 258 - 286.

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'Systematic' in Literature Reviews in IS. *Journal of Information Technology*, *30*(2), 161 – 173.

Braue, D. (2025). Public servant creates sexual deepfakes of colleagues. *Information Age*.

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.

Brooks, R., Yuan, Y., Liu, Y., & Chen, H. (2022). DeepFake and its Enabling Techniques - A Review. *APSIPA Transactions on Signal and Information Processing*, *11*(2), 1 - 31.

Criminal Code Amendment (Deepfake Sexual Material) Bill 2024, Commonwealth Australia (2024).

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International Journal of Multimedia Information Retrieval*, *11*, 219 – 289.

Delfino, R. A. (2019). Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. *Fordham Law Review*, *88*(3), 887 - 938.

Firc, A., Malinka, K., & Han´aˇcek, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools & attack vectors. *Heliyon*, *9*(4), 1- 33.

Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: current and future trends. *Artificial Intelligence Review*, *57*(64), 1 - 32.

Gong, L. Y., & Li, X. J. (2024). A Contemporary Survey on Deepfake Detection - Datasets, Algorithms, and Challenges. *Electronics*, *13*(3), 1 - 22.

Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (2 ed., pp. 1 - 20). Springer.

Guera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), New Zealand.

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems With Applications*, *240*, 1 - 14.

Hammersley, M. (2022). Emergent Design. In U. Flick & Editor (Eds.), *The SAGE Handbook of Qualitative Research Design*. SAGE Publications, Limited.

Jacobsen, B. N., & Simpson, J. (2024). The tensions of deepfakes. *Information, Communication & Society*, *27*(6), 1095 – 1109.

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.

Khoo, B., Phan, R. C.-W., & Lim, C.-H. (2021). Deepfake attribution: On the source identification of artificially generated images. *WIREs Data Mining and Knowledge Discovery*, *12*(3), 1 - 21.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135 - 146.

Kietzmann, J., Mills, A. J., & Plangger, K. (2020). Deepfakes: perspectives on the future "reality" of advertising and branding. *International Journal of Advertising*, *40*(3), 473 – 485.

Kirchengast, T. (2020). Deepfakes and image manipulation: criminalisation and control. *Information & Communications Technology Law*, *29*(3), 308 - 323.

Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation to combat threats posed by deepfakes. *Boston University Law Review*, *101*(2), 761 - 802.

Lees, D. (2024). Deepfakes in documentary film production: images of deception in the representation of the real. *Studies in Documentary Film*, *18*(2).

Li, Y., Chang, M.-C., & Lyu, S. (2018). *In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking* 2018 IEEE International Workshop on Information Forensics and Security (WIFS), China.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics* 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), USA.

Lumivero. (2023). *NVivo*. In (Version 14)

Lyu, S. (2020). *Deepfake detection: Current challenges and next steps* 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, United Kingdom.

Malik, S., Surbhi, A., & Roy, D. (2024). Blurring boundaries between truth and illusion: Analysis of human rights and regulatory concerns arising from abuse of deepfake technology. International Conference Series on ICT, entertainment technologies, and intelligent information management in education and industry, Japan.

Maras, M.-H., & Alexandrou, A. (2018). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, *23*(3).

McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *new media & society*, *26*(5), 2786 – 2803.

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2019). Regulating Deep-Fakes: Legal and Ethical Considerations. *Journal of Intellectual Property Law & Practice*, *15*(1), 24 - 31.

Millière, R. (2022). Deep learning and synthetic media. *Synthese*, *200*(231), 1 - 27.

Panagopoulos, J., & Bita, N. (2025, 24/05/2025). Teens sell deepfake schoolgirl porn pics. *The Australian*.

Rana, M. s., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature_Review. *IEEE Access*, *10*, 25494 - 25513.

Rancourt-Raymond, A. d., & Smaili, N. (2023). The unethical use of deepfakes. *Journal of Financial Crime*, *30*(4), 1066 - 1077.

Rao, S., Verma, A. K., & Bhatia, T. (2021). A review on social spam detection: Challenges, open issues, & future directions. *Expert Systems With Applications*, *186*, 1 - 31.

Ray, A. (2021). Disinformation, deepfakes and democracies: The need for legislative reform. *UNSW Law Journal*, *44*(3), 983 - 1013.

Renier, L. A., Shubham, K., Vijay, R. S., Mishra, S. S., Kleinlogel, E. P., Jayagopi, D. B., & Mast, M. S. (2024). A deepfake-based study on facial expressiveness and social outcomes. *Scientific Reports (Nature Publisher Group)*, *14*(1), 1 - 13.

Sandoval, M. P., Vau, M. d. A., Solaas, J., & Rodrigues, L. (2024). Threat of deepfakes to the criminal justice system. *Crime Science*, *13*(41), 1 - 16.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The Sociotechnical Axis of Cohesion for the IS Discipline: Its historical legacy and its continued relevance. *MIS Quarterly*, *43*(3), 695 - 720.

Seymour, M., Riemer, K., Yuan, L., & Dennis, A. R. (2023). Beyond deep fakes. Communications of the ACM, 66(10), 56–67.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, *31*(2), 74 - 87.

Sloot, B. v. d., & Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review*, *46*, 1 - 15.

Smuha, N. A. (2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence. Cambridge: Cambridge University Press.* Cambridge University Press.

Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, *7*(2), 83 - 113.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2025). What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Transactions on Technology & Society*, *6*(1).

Vasist, P. N., & Krishnan, S. (2022). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *CAIS 51*(1), 590 – 636.

Vasist, P. N., & Krishnan, S. (2023). Engaging with deepfakes: a meta-synthesis from the perspective of social shaping of technology theory. *Internet Research*, *33*(5), 1670 - 1726.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 39 - 52.

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review. *Technovation*, *125*, 1- 17.