Making Sense of Deepfakes

Epistemic Harms and the EU Policy Response

Lucas Christian Staab

Combined Thesis Project

Philosophy of Science, Technology and Society

Public Administration

Supervisors:

Dr. Adam Henschke (PSTS)

Dr. Ringo Ossewaarde (PA)

Dr. Michel Bourban (PSTS)

Wordcount: 28.578

Abstract

Recent advances in generative artificial intelligence have enabled the production of realistic pieces of audiovisual media footage, so-called deepfakes. Among other kinds of harm linked to the technology, deepfakes have raised concerns regarding their impact on public discourse and politics, as well as their epistemic consequences more broadly. Generally, three different kinds of epistemic harm from deepfakes are distinguished in the philosophical literature: deception, jeopardizing evidence and erosion of trust. However, little work has been done with respect to how deepfakes are interpreted by recipients and how they may or may adjust their beliefs in the process. In this thesis, I argue that previous conceptions of epistemic harm from deepfakes hinge on their ability to deceive; to be mistaken for authentic recordings. Drawing on Peircean Semiotics and Epistemic Vigilance, I further argue that the evaluation of deepfakes partially depends pre-existing beliefs, interests and the perception of benevolence. This introduces possibilities for non-deception-based harms. I then turn to the European policy discourse around deepfakes. Using Quantitative Content Analysis (QCA), I uncover the understanding of epistemic harm and the measures that target them in relevant policy documents. The analysis shows that deepfakes are subsumed under the broader phenomenon of disinformation. Policy-makers are primarily occupied with harms to epistemic goods which result from deception and manipulation. They seek to address this primarily through providing authoritative information recipients encounter online. However, as this fails to account for how deepfakes may cause non-deception-based harm, this policy response is incomplete.

Table of Contents

Acknowledgements

My Mom for always being patient. I swear I will move houses less now.

Andi, Freya and Hannah for teaching me more than I can tell.

Andi, Janna and Sam for making a second home.

Benny and Aylin for caring.

Alexis, Dami, Gemma, Guillaume, Juan, Pietro and Vishal for always making me laugh.

Celina, Hannah and Johanna for making me like shared houses again.

Dan Sperber for a chance encounter that changed this thesis.

Adam, Michel and Ringo for guiding me through and getting me to (mostly) enjoy the process.

1. Introduction

Since the initial invasion of eastern Ukraine by Russia in 2014 and accelerated by the success of the Donald J. Trump presidential campaign, online disinformation has become a prominent concern in the European Union (Datzer, Lonardo 2022). Disinformation in the policy discourse in the European Union is generally defined as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm" (Wardle, Derkahshan 2017: 20). False information online, be it spread out of bad faith or ignorance, probably dates back to the beginnings of the public internet (Floridi 1996). Since then, motivated parties or hobbyists have gained a variety of means to manufacture and spread their own flavour of disinformation. Recent advances in artificial intelligence have added to these means by enabling the generation audiovisual media content – audio, images and video – that may seem authentic, but is partially or entirely fabricated.

These pieces of synthetic audiovisual media generated with machine-learning systems are commonly called 'deepfakes', a term that combines 'deep learning' and 'fake' (Tolosana et al. 2020). While deepfake technology has a wide range of possible uses, it is often deployed to generate media that portrays the likeness of a person doing or saying something that never happened (*see* 2.1). In this capacity, deepfakes pose a series of problems as they may be used for harmful purposes in various ways, including disinformation.

The currently most prevalent form of harmful deepfakes, however, is non-consensual deepfake pornography. According to a report by Deeptrace, 96% of deepfakes on the internet in 2019 were pornographic (Ajder et al. 2019: 1). This is unsurprising, as the very term 'deepfake' emerged on a reddit forum dedicated to producing non-consensual deepfake pornography of famous women (Millière 2022: 2). Usually, these videos layer the faces of women who had pictures of them scraped from the internet over the faces of women in pre-existing pornographic videos without the consent of either party (Hunter 2023). While arguably demeaning in itself (de Ruiter 2021: 1314; Öhman 2022), this form of highly gendered abuse can of course be made public – there are many sites dedicated to deepfake pornography (Ajder et al. 2019: 6) – or to inflict harm in other ways, e.g. extortion (Chesney, Citron 2019: 1772f). Deepfakes may further be used for impersonation to conduct phishing or social engineering attacks in order to defraud individuals or companies (ibid.).

Nonetheless, in this thesis I am concerned with the epistemic harms deepfakes cause in digital information environments.[1] Deepfakes have raised concerns about harmful impacts on politics and public discourse, including forms of political and diplomatic sabotage, from the beginning of critical engagement with the technology (*e.g.* Bateman 2020; Chesney, Citron 2019; Smith, Mansted 2020). The technology emerged and matures in a time of entrenched social tensions, and prolific online dis- and misinformation. Many expect that deepfake technology will be exploited to accelerate both dynamics (ibid.).[2] Given the relevance of online dis- and misinformation for public discourse (Brown 2018) and the societal problems that are associated with it, e.g. harms to public health and safety (Dabbous et al. 2022; WTO 2022), investigating the epistemic effects of deepfakes in digital information environments contributes to the overarching goal of safeguarding public discourse as a key democratic institution (Anderson 2006).

Deepfaked content has already featured in the realm of politics in a variety of contexts. In 2018, a deepfaked video of Barak Obama speaking about the dangers of deepfakes was arguably the first instance of a deepfake gaining broad public attention (Mack 2018). Since then, more cases have emerged that bolster the political significance of deepfakes and illustrate how deepfakes are of concern to the EU. Two recent examples in the European context stand out, both of which are related to Russia's renewed invasion of Ukraine. On March 2nd 2022, news channel Ukraine24 shared a video on Twitter showing Ukrainian president Volodymyr Zelenskyy urging Ukrainian soldiers to surrender (Telegraph 2022). Soon after posting the video, Ukraine24 announced their account had been hacked and that the video was a deepfake. Various social media platforms then stopped the spread of the video (Simonite 2022). The harm caused by this deepfake was limited by its relatively poor quality, members of the public drawing attention to this lack of quality, and the quick response by social media platforms. Prior to this event, strategies to respond to this specific kind of threat have already been developed and discussed with relevant actors (Charlet, Citron 2019). Later the same year, another

[1] Briefly put, I understand deepfakes as pieces as pieces of audiovisual media showing someone doing or saying something, epistemic harms as a setback to people's interest of having accurate beliefs, and digital information environments as platforms that allow users to publicly post audiovisual content, and others to react to this content in some capacity (*see* 2.1).

[2] Such harm from deepfakes is also not limited to actual instances of deepfaked media. The mere possibility of footage being a deepfake enables people who have been caught on tape in a scandalous act to pass it of as fake. Alternatively, opposing parties may call the authenticity of recordings into question. This is referred to as the 'liar's dividend' (Chesney, Citron 2019; Fallis 2020; Harris 2020; Rini 2020).

suspected case of malicious deepfake use emerged. In June 2022, the mayors of Berlin, Vienna and Madrid all spoke to someone impersonating Vitali Klitschko, the mayor of Kyiv, presumably using deepfake technology to look and sound like him in a video conference (Oltermann 2022). Andrea Giffey, mayor of Berlin, grew suspicious when the supposed Ukrainian official began bringing up Ukrainian refugees defrauding the German welfare state and confirmed with the Ukrainian embassy that she was talking to an imposter (ibid.).[3]

In the both of these above cases, negative consequences were largely avoided. However, they show the potential severity of harm deepfakes may cause in the public sphere. In light of their potentially harmful impacts, deepfakes have gathered substantial scholarly attention. These works concern the various harms deepfakes may cause (*e.g.* Chesney, Citron 2019; Smith, Mansted 2020; Hwang 2020; Rini, Cohen 2022), non-consensual deepfake pornography as a gendered form of abuse (*e.g.* Dunn 2020; Maddocks 2020; Öhman 2020; Kerner, Risse 2020; Viola, Voto 2023), and the epistemic impact of deepfakes (*e.g.* Rini 2020; Fallis 2021; Harris 2021; Atencia-Linares, Artiga 2022; Matthews 2023).

The existing philosophical literature has firmly established a need for action on the harmful reality and potential of deepfakes. This need to act has also been recognized by some governing bodies, including the EU (*e.g.* van Huijstee et al. 2021; Kugler, Pace 2021; Congressional Research Service 2022; Geng 2023). In various policy documents in the areas of disinformation and artificial intelligence policy, EU bodies have proposed policy measures that apply to deepfakes (*see* 4.1, 4.2). Through large-scale regulatory efforts like the General Data Protection Regulation (GDPR)(Regulation (EU) 2016/679), Digital Services Act (DSA)(Regulation (EU) 2022/2065) and the proposed Artificial Intelligence Act (AIA)(EC 2021a) the EU has acquired a reputation as a global frontrunner in digital policy. The Union has been highly influential on global digital policy, among other domains, by leveraging common-market access (*see* Bradford 2012, 2020; Brattberg et al., 2020). The stance the EU takes on deepfakes may therefore have wider implications on global deepfake governance, rendering EU deepfake policy a relevant research object.

The potential impact of deepfakes on public discourse has been a primary concern, both in scholarly literature and among policymakers (*e.g.* van Huijstee et al. 2021). However, in the epistemological literature, this impact is largely attributed to the capacity of deepfakes to

[3] It is not entirely certain that the imposter utilized deepfake technology. Nonetheless, the example shows a potential use-case with high-stakes consequences.

deceive, leading to a diminished role of recordings as evidence, and undermining public trust (*see* 3.1). While this may certainly be the case, literature on the interpretation of media (Bode 2021; Chandler 2023) and the acquisition of beliefs (Sperber 1997; Sperber et al. 2010; Mercier 2017, 2019) suggests that deception may not be the only or even most crucial vector for epistemic harm. At the same time, how recipients – those who encounter footage online – make sense of deepfakes and, as a result, form beliefs on their basis is not thoroughly explored. This poses a gap in the philosophical literature and raises questions how the impact of deepfakes are conceived of and handled in EU policy.

Given this background, I seek to answer the following question in this thesis: *Does EU policy address the epistemic harms caused by deepfakes*? In answering this question, I aim to achieve three things. First, to expand on the existing philosophical literature on epistemic harms caused by deepfakes by tackling the literature gap identified above. In this effort, I provide a framework that accounts for how recipients make sense of media footage (*see* Ch. 3). Second, uncovering how the epistemic harms of deepfakes are understood in EU policy and, resulting from this understanding, which measures are proposed to address those harms. As mentioned above, EU policy on deepfakes stands to be influential on a global scale, making the understanding EU institutions have of the epistemic harms from deepfakes and their policy responses especially significant. Lastly, I aim to see whether current EU policy addresses the epistemic harms of deepfakes comprehensively. Importantly, I do not attempt to evaluate whether EU policy on deepfakes is effective. Rather, my interest here is whether the epistemic harms I identify in this thesis are addressed by EU policy at all. The evaluation of those policies will have to come at a later point.

Corresponding to those aims, the overall research question of this thesis can be divided into three sub-questions:

1. *How do deepfakes interact with the beliefs of their recipients and which epistemic harms may arise from this?*
2. *How is the epistemic harm of deepfakes understood and addressed in relevant EU policy documents*?
3. *Are the identified kinds of epistemic harm caused by deepfakes addressed by EU policy?*

Sub-question one presents the theoretical research interest of this thesis, whereas sub-questions two is investigated empirically. Sub-question three presents a synthesis of the theoretical and the empirical component. Answering the questions above will proceed as follows.

Chapter 2 provides the basis for the analysis in the following chapters. Section 2.1 defines digital information environments, epistemic harm and deepfakes, 2.2 explains the technological foundation of deepfakes and 2.3 shows how the epistemic significance of deepfakes is commonly understood in the literature. Section 2.4 describes how deepfake applications enable producers to communicate meanings. I close the chapter with a summary of key insights.

In Chapter 3, I answer sub-question one. Section 3.1 summarizes the current understanding of epistemic harms from deepfakes in the philosophical literature. So far, this literature has not provided a conceptual mechanism for how recipients make sense of footage and form beliefs as a result. In sections 3.2, I provide such an account by drawing on Peircean semiotics and epistemic vigilance. Semiotics helps to understand how footage is interpreted, whereas epistemic vigilance establishes a suite of mechanisms through which recipients evaluate the believability of new information. In section 3.3, I apply these frameworks to deepfakes and argue that, considering how recipients make sense of footage, deepfakes may cause epistemic harms that have so far been underappreciated. Section 3.4 offers a summary of the key insights of the chapter as well as an answer to sub-question one.

Chapter 4 outlines the approach to the empirical component. I provide a case description situation deepfakes within European disinformation and artificial intelligence policy (4.1), a justification for the EU policy documents selected for analysis (4.2), and description of the method chosen for this analysis: qualitative content analysis (QCA; 4.3). The specific implementation of QCA in this thesis partially utilizes deductive categories based on the discussion of deepfakes in previous sections. How these categories are developed is described in section 4.4. The research design is summarized in section 4.5.

Chapter 5 presents the results of the empirical analysis. As the understanding of epistemic harm partially emerges from how it is addressed, I begin with an overview of the measures proposed in the analysed documents (5.1) before detailing how epistemic harm is understood in EU policy (5.2). In section 5.3, I summarize the overarching results of the empirical component and answer sub-question two.

Having answered the research sub-questions of the theoretical and empirical component of this thesis, I contrast both answers and answer the final sub-question in section 6. I conclude with an overview of my argument and an answer to the overarching research question. I further situate my research in the broader literature on deepfakes and disinformation and offer some limitations as well as possible directions for future research. Lastly, I provide some policy recommendations in light of my findings (7).

2. Understanding Deepfakes

In this chapter, I define key terms and introduce key concepts necessary for understanding deepfakes and the epistemic harms they cause. In doing so, this chapter builds a basis for the discussion of those harms and the policies which tackle them in later chapters. Aside of providing necessary background, I argue that the common attribution of epistemic significance of deepfakes is mainly rooted in their ability to imitate their training data (2.2) in a way that allows them to achieve realism that is otherwise only available to recordings (2.3). Further, deepfake applications (DFAs) – system for the production of deepfakes – afford a great degree of freedom to the producers of deepfakes, which in turn can make use of this to communicate meanings to their audience (2.4).

Before building this basis, a clarification of the scope of this thesis is needed. In what follows, deepfakes will predominately be discussed under reference to images and videos of human subjects. This is not to diminish the significance of audio-deepfakes or deepfakes of non-human subjects, e.g. vehicles or buildings. However, prominent instances of the phenomenon tend to be images or videos of public figures. Nonetheless, the arguments made in this thesis should principally also apply to audio-deepfakes and deepfakes of non-human subjects. Further, I am concerned with political deepfakes, broadly speaking. These are all deepfakes that portray events that are politically significant, for example because the portray a politician or tap into existing socio-political tensions. This also means that I am not concerned with deepfakes that are only produced for private consumption, for example to illustrate personal fantasies without ever sharing them (Öhman 2020).

2.1 Defining Digital Information Environments, Epistemic Harm and Deepfakes

The scholarly literature can be inconsistent in how they approach deepfakes and how they understand the epistemic harms they cause (Vasist, Krishnan 2022; *see* 3.1). Further, the policies analysed in the empirical component also differ in how they differentiate the services they apply to. In this section, I set out to define digital information environments, epistemic harm and deepfakes in a way that is able to unify these diverging concepts and set out a clear scope for my analysis.

Digital Information Environments

For the purposes of this thesis, I understand digital information environments as platforms that allow users to publicly post audiovisual content, and others to react to this content in some capacity (*see* Bode 2021). The paradigm cases are social media platforms like Facebook, TikTok or Youtube. However, online forums like Reddit or 8chan also fall under this definition. The way recipients interact with each other and the content they encounter in digital information environments depends on technological affordances (like-buttons, resharing, algorithmic recommendation systems, etc.), rules of the platform (terms of service, content moderation practices) and the "cultural, ethical, and aesthetic spoken or unspoken rules" (ibid.: 922) that emerge between recipients. As such, these environments are characterized by technological, institutional and emergent social modalities (ibid.).

Epistemic Harm

I consider an epistemic harm as *a violation of an agent's right for others to abstain from obstructing that agent's epistemic success in a relevant area where such success does not infringe on the rights of others*. This definition is an amalgam of concept of harm, epistemic value and normative limitations to knowledge and can be broken down into four components which are explained below:

1. A negative right of a prospective knower,
2. the notion of obstructing epistemic success,
3. a criterion for relevance, and
4. the potential for conflicting rights.

That an agent has a negative right means that they are entitlement that others refrain from doing something (*see* Wenar 2020: 2.1.8), in this case that they do not obstruct an agent's epistemic success. Broadly construed, epistemic success consists in the realization of epistemic value, e.g. an agent having or acquiring true beliefs (Steup, Neta 2020). There is a considerable degree of debate on the entities that can have epistemic success (ibid.: 1.5).[4] For the purposes here, I am concerned with the epistemic successes of participants of public discourse in digital information environments. I consider obstructions of their epistemic success to be those actions that negatively impact the individual epistemic success of individuals as well as practices, groups, or (social) systems that are conducive of such epistemic successes (Fleisher, Šešelja 2023: 8; *see also* Carey 2023). Practices, groups and (social) systems conducive of epistemic success can be described as *epistemic goods.* Well-functioning democratic institutions may, for example, present such an epistemic good (Anderson 2006).

Of course, this poses the question what epistemic value consists in. Again, there is substantial debate regarding this (Pritchard et al. 2018; Steup, Neta 2020). Khalifa and Millson (2020) offer a useful account. Whereas some see epistemic value solely rooted in the possession of true beliefs (ibid.: 87ff), the authors argue that not all areas one may have true beliefs about are equally valuable. Some areas of epistemic success are more relevant for an agent than others. The value of information, according to Khalifa and Millson, depends on the context of an agent, meaning their "personal interests, social roles, and background assumptions" (ibid.: 91). I will summarize this as an agent's '*interest*' in certain kinds of information. This interest includes practical ends, e.g. information about the location of a key, but is not limited by them. An agent may also be interested in information for non-practical reasons, e.g. satisfying their curiosity (ibid.) or because they have a social role that requires them to have accurate information in a given area (ibid., *citing* Hart 1969: 212). For example, doctors have a social responsibility to have and provide accurate medical information to their patients. Lastly, information may be of interest because it relates to background assumptions that links it to other relevant areas (Khalifa, Millson 2020: 91f). For example, if a doctor assumes it is medically relevant to eat one apple each day they have a role-based interest in knowing whether their patient eats an apple each day, whether that assumption is accurate or not. Epistemic value is realized when an agent obtains true information in a relevant area (ibid.: 102). This allows to dismiss instances of obstructing epistemic success in irrelevant areas as harms.

[4] The authors refer to what I describe here as *cognitive* rather than *epistemic* success. I chose to simplify this for my purposes here, as the notion of '*epistemic success*' makes it more legible how the harm I define is epistemic.

Despite this useful characteristic, there is an issue with this account. Millson and Khalifa (2020) formulate their argument based on scientific inquiry. Whereas it can reasonably be presumed that researchers are interested in information they ought to be interested in given their social role, this is not necessarily the case for participants in public discourse. Beyond what may de facto interest them, a criterion for epistemic harm in the realm of public discourse needs to specify which information ought to be of interest for involved agents qua their role of participating in public discourse. While this may be a vary wide range of different kinds of information, for the purpose of this thesis I will take the criterion of relevance to be met insofar as pieces of information are politically relevant (*see* 2). Epistemic harm, as understood here, is caused if an agent obstructs the epistemic success of another in the area of politically relevant information.

However, I hold that there are limitations to when such an obstruction constitutes epistemic harm. This is the case in instances where there are legitimate limitations to which inquiries an agent may justifiably undertake. This is the case where it infringes upon the rights of other agents if the inquiring agent has access to certain information. This may be the case e.g. because the information in question is protected by privacy rights (Marmor 2015). Consider the following example. A parent is interested in what their child has written in their diary yesterday. They know that their child hides their diary under their pillow. Should the child's sibling decide to hide that diary somewhere else so that the parent cannot access the private information in the diary, I do not hold that the sibling inflicts epistemic harm on the parent.

In summary, I operationalize the above definition in this thesis as follows: *Epistemic harm is caused if an agent obstructs, without legitimizing reason, the epistemic success of another in the area of politically relevant information*. I do not seek to argue that this is the only reasonable concept of epistemic harm, even when it comes to political discourse.[5] However, this definition provides some advantages for my present purposes. First, stipulating epistemic harm as the

[5] Other concepts of epistemic harm may for example be more concerned with agents as prospective knowns. This plays a more prominent role e.g. in accounts of epistemic justice, where harm consists in discrediting the testimony of someone and keeping them from contributing to collective knowledge production as epistemically harmful (e.g. Fricker 2007). In cases where this is due to prejudice against marginalized groups, this presents an epistemic injustice (ibid.). There is, however an extent of overlap between Fricker's account and the definition I offer insofar as keeping (structurally marginalized) agents from contributing to collective bodies of knowledge most likely presents a harm to an epistemically important social system (Fleisher, Šešelja 2023: 8; see also Carey 2023), thereby indirectly obstructing the epistemic success of agents. For an application of a perspective on deepfakes that also engages with potential harms to knowns, see Kerner and Risse (2020).

violation of a negative right means it is limited to specific behaviours that obstruct epistemic success. Conceiving of epistemic harm as the violation of a positive right, on the other hand, would render any failure to contribute to such success harmful, making the range of suspect behavior much broader than I intend to cover in my analysis.[6] Second, limiting the locus of epistemic harm to areas of interest further allows me to more neatly limit the scope of epistemic harms in a specific area, in this case political discourse online. While posting an innocuous deepfake of a cat may otherwise still satisfy the above definition, I am not concerned with such cases here. I will return to this concept of epistemic harm in the coming chapters.

Deepfakes

Definitions of deepfakes in the philosophical literature largely converge on the following: *Deepfakes are realistic pieces of synthetic audiovisual media produced through machine learning showing someone doing or saying something that did not occur* (*e.g.* Fallis 2020; Rini 2020; de Ruiter 2021; Harris 2021). This definition is useful as it captures a variety of aspects that are relevant for the epistemic harms deepfakes may cause in digital information environments. It captures a broad range of applications of deepfake techniques instead of isolated use-cases, such as face-swapping, or a single medium, such as video. There is a wide variety of applications which make use of a combination of techniques (Millière 2022: 9; *see* Mack 2018) for a variety of purposes (*see* Kietzmann et al. 2020; Vasist, Krishnan 2022). Further, defining deepfakes as pieces of machine-learning enabled synthetic audiovisual media distinguishes them form products of generative large language models, such as ChatGPT, and conventional means of audiovisual media manipulation, such as Photoshop.[7]

However, the above definition also has shortcomings. First, deepfakes do not necessarily need to show a human subject to be politically significant and cause (epistemic) harm. One may imagine a deepfaked image of a hospital that has presumably been hit by an airstrike being spread on social media during an escalating conflict. Second, deepfakes do not need to be realistic to cause epistemic harm, though the degree to which a deepfake is realistic likely amplifies the epistemic harms it may cause (*see* Ch. 3). Therefore, I slightly broaden the above

[6] Should such a positive right exist my analysis would need to be extended, but not be void.

[7] Outside of philosophical discussions, deepfakes and LLMs are often discussed in concert (e.g. van Huijstee et al. 2021). Nonetheless, audiovisual media differs from text in its ability to be realistic (*see* 2.3), as well as its ability to eliciting associations (*see* 3.3). As both are relevant for the epistemic harms deepfakes may cause (*see* Ch. 3) LLMs are excluded from my discussion here.

definition while keeping in mind the most prominent uses of deepfakes indeed are showing someone doing or saying something. In this thesis, I understand deepfakes as *(mostly) realistic pieces of audiovisual media produced through machine learning showing something or someone in a counterfactual way*. This definition can be broken down into the following aspects:

1. Deepfake Technology (‘*synthetic pieces of audiovisual media produced through machine learning’*),
2. Subject Representation (‘*mostly realistic*’, ‘*showing something*’),
3. Narrativity (‘*showing something or someone in a counterfactual way’*).

In the next three sections, I will explore these aspects further in order to build a foundational understanding of deepfakes for the following chapters.

2.2 Technological Foundation

Deepfakes are a subclass of synthetic audiovisual media produced through relatively novel machine learning techniques. Most sophisticated deepfake systems are rooted in deep-learning (DL) and make use of neural network architectures (Millière 2022). In such systems, input data is abstracted into specific characteristics through a series of consecutive layers of data processing units, so called nodes, allowing to system to learn abstract but complex representations its training datasets (LeCun et al. 2015; Buckner 2018, 2019). DL has been leveraged for a variety of applications, including data classification, attempts to predict the behaviour of complex systems, and systems that are able to generate novel outputs based on their training data (LeCun et al. 2015; Crawford 2021; Millière 2022). Deepfake applications usually extend DL architectures for classification (Buckner 2018, 2019) toward a capacity to generate novel outputs (Millière 2022: 6).

This is achieved through utilizing large amounts of digital recordings in a given medium. Digital recordings of speech, images and video are quantized physical signals represented by "discrete numbers on a machine-readable data storage" (ibid.: 4). Any medium-specific data format – e.g. JPEG or MP3 – has parameters through which a given piece of media is described in a machine-readable way. These parameters can be abstracted into a large, but finite set of possible expressions. The boundaries of this set delineate a space which contains all instantiations possible in a given medium as points (ibid.: 6). However, most points in this space would be mere noise. DL systems are capable of discerning patterns in the characteristics they abstract from their input data (Buckner 2018, 2019; LeCun 2015; Millière 2022).

Depending on how the dataset is structured, DL systems to distinguish characteristics that correspond with specific object classes, e.g. cats or apples (Millière 2022: 21). In the higher-dimensional space, there is a much smaller set of points that are cat images rather than noise or non-cats. Object classes can be understood as localized regions in high-dimensional space, and the system learns to associate certain regions with them (Buckner 2019: 10). The patterns of spatial distribution representing object classes identified by a model are called 'latent space'. Whether a given image (of a cat) is identified as belonging to an object class ('cat') depends on whether the point in high-dimensional space it corresponds with is located within or in proximity to 'catness' (ibid.; Millière 2022: 6).

Generative models can create new representations of object classes from this latent space. If a user would want to produce the image of a cat, a generative model that has been trained on cat images generates a novel image that is represented by a point in its latent space for 'catness'

(Millière 2022: 6). Examples of an early generative model walking the line between noise and discernible objects can be seen below (Figure 1).

(a) Hen

(b) Ostrich

Figure 1. Images generated by an early generative model. Dai et al. 2016: 7.

The principles laid out above also apply to video and sound generation, as well as local manipulations of pre-existing material, such as replacing a face in a video by another. However, audio and especially video generation are more demanding as they require considerable temporal and/or spatial integrity to be convincing (Millière 2022: 12).

Deepfakes build in foundational principles of machine- and deep-learning which are principally agnostic when it comes to the medium and the purposes they are deployed for. Through the complex representation of the characteristics of their training data, deepfake models can depict and manipulate a vast variety of objects so long as they develop a sufficiently vast latent space. As they imitate the characteristics of their training data, generative models trained on realistic recordings, deepfake models strive to produce realistic outputs. While many models have not quite achieved indistinguishable realism yet, it is commonly perceived as likely that they will at some point (Millière 2022; Rini 2020).

2.3 Authentic Recordings and Realistic Deepfakes

Now that the technological foundation of deepfakes is established, I turn to the second aspect of their definition on which a large part of their epistemic significance hinges: the specific way they represent their subjects ('*(mostly) realistic*', '*showing something or someone'*). Deepfakes, like other kinds of media, convey information. In doing so, they represent their subject in a specific way. In the existing literature, the way deepfakes do so and why it is epistemically significant is usually explained under reference to the epistemology of photographs (*e.g.* Rini 2020; Fallis 2021; Pierini 2023). In this section, I summarize this argument.

To start, some terms need to be introduced: *recording*, *footage*, *authenticity* and *realism*. In contrast to common usage where they may be used interchangeably, these terms have specific meaning in the context of this thesis and are used to distinguish deepfakes and non-fake pieces of audiovisual media. The latter are referred to as '*recordings*'. They differ from manipulated audiovisual media, including deepfakes, because they have the property of being '*authentic*'. This is due to epistemically significant characteristics of their production process which are described below. '*Footage*' serves as an umbrella term that includes recordings, deepfakes, and other kinds of manipulated audiovisual media. This accounts for the ambiguity recipients may face when they encounter pieces of audiovisual media online. Lastly, the '*realism*' of footage describes the property of conforming to the perceptual expectations of a recipient regarding how authentic footage of something or someone would look like. This will also be elaborated on below. A piece of footage is realistic either if it is authentic, or if it sufficiently approximates the characteristics of authentic footage for a recipient not to be readily able to identify the difference.

Recordings are deeply embedded in epistemic practices and frequently used as evidence to validate or challenge other kinds of evidence, such as testimony (Walton 1984; Rini 2020; Schauer 2022: 139-144). Despite this, relatively little work has been done on the epistemology of recordings, under the exception of photographs, which may serve as an analogue (Rini 2020: 9f; Fallis 2021: 636).

Photographs occupy a privileged position in epistemic and evidentiary practices (Rini 2020; Schauer 2022). They provide what some epistemologists have called perceptual evidence (Hopkins 2012; Cavedon-Taylor 2013) which is particularly valuable because it justifies belief under less stringent conditions compared to other kinds of evidence, such as testimony

(Cavedon-Taylor 2013: 288f).[8] Fallis (2021) argues that the epistemic status of photographs is rooted in their production process. When using an analogue camera, "light reflected from a physical object is directed by lenses and mirrors onto a photographic surface" (ibid.: 623) resulting in a photograph. Decisions of where the camera is pointed, its angle, when the process of transposing light is initiated, and other choices reflect values of the photographer (Sontag 1973; Schauer 2022: 141). However, once these decisions are enacted, light reflected from objects will be transposed in a way that is not mediated by the mental state of the photographer, capturing also objects that escaped the photographer's attention (Fallis 2021: 636; *see also* Cohen, Mesik 2004; Cavedon-Taylor 2012; Walden 2005, 2012). Photographs are for the most part consistent with those past material states they depict (Hopkins 1998: 72f; Abell 2010: 83).[9] It is in this sense that they are *authentic*.

This process of transposition has analogues in video and audio format (Cohen, Meskin 2004; Rini 2020). Insofar as the epistemic status of photographs is rooted in the *authenticity* they acquire from a consistent transposition process it applies to other kinds recordings as well – so long as they result from a similarly consistent process. Importantly, it the process that needs to be consistent, not its outcome. This distinguishes photographs of poor quality from photorealistic paintings. Individual paintings may be indistinguishable from high quality photographs, and even surpass poor photographs in accurate detail, but the process of painting is not consistent. Instead, it depends on the skill of an artist and their recollection of a scene, among other things (Fallis 2021). The same goes for manually altered footage, such as photoshopped images or (most) movie footage.

To be consistent, a process of transposing past material states must be sufficiently sensitive to differences in material states and the resulting representation must sufficiently reflect this sensitivity. I take the criterion of sufficiency to be met insofar as any existing procedural inconsistencies do not generally lead to counterfactual interpretations regarding the represented material states. This may be the case because inconsistencies are either minor or well-known. For example, say a digital camera would always produce a handful of randomly distributed black pixels when taking a photograph. In a photograph with several million pixels this would

[8] For testimony to serve as a justification for knowledge, additional factors must be considered, such as the trustworthiness of the testifier (Levy 2022), or incentives for giving false testimony (Harris 2022).

[9] In fact, the desire of this capacity was arguably what largely drove the development of photograph technology in the first place (Kingslake 1989). The mechanical process by which this is achieved was therefore uniquely tailored to this objective.

hardly matter for its interpretation. Similarly, if the cameras lens had a scratch that would somewhat transform the resulting photograph, this would not jeopardize the authenticity of the resulting picture. The possibility of some minor inconsistencies or limitations is known for many forms of recording.

Limitations and inconsistencies may, however, lead to an absence of specific information. Recordings can only capture information on a certain scope and scale. A standard photograph cannot give an indication whether there were airborne disease vectors present when it was taken. Further, recording technologies are designed and calibrated to be selectively sensitive. Until the 1980s, design choices within the photographic process lead to the systematic misrepresentation of the skin tone of non-light skinned people to the extent of rendering their identity illegible (Roth 2009, 2019, *cited in* Habgood-Coote 2023: 13). A recording can only speak to some past material states, and not others. For those it does speak to, however, it does so in a way that is authentic and allows recipients to obtain knowledge. Even a photograph which results from such a – technically and morally – flawed process as to render the identity of non-light skinned people illegible allows one to gain knowledge about the presence of certain light skinned people in a scene, but not about the presence of a specific non-light skinned person. However, absence of information is not the same as counterfactual information. The photograph would still be authentic as the transposition process that created it, though unjust (*see* Liao, Huebner 2021), is consistent.

When recipients encounter footage, they are usually not aware of how it was created. However, there is a powerful cue for ascribing authenticity to a piece of footage. Authentic footage is generally *realistic*. Recordings provide evidence of past material states in a way that corresponds to how we would have perceived them: sounds are transposed into an auditory medium, reflections of light into a visual one. This allows recordings to conform with our expectations regarding how an authentic representation of an object would be like.

> *If*
>
> 1. the physical states of the world behave in a consistent manner,
> 2. the access of human perception to (some of) those states is sufficiently consistent,
> 3. we are able to form sufficiently consistent perceptual expectations on the basis of our perceptual experience,
> 4. and we are familiar with the limitations of a recording medium,
>
> *then* a recording will most likely comply with our perceptual expectations, with the prevalence of exceptions depending on how prone to error our expectations are.

Insofar as the above argument holds, the aesthetic properties of recorded representations – perspective, shade, sound patterns, organization of objects, spatio-temporal consistency, etc. – overlap with perceptual expectations for the recorded object. Recordings have, in some sense, a minimal difference to their objects (Watson 1984). It is in this sense that they seem *realistic* to us.

However, realism does not necessitate authenticity. Recordings that do not correspond with perceptual expectations may be rare, but occasional shortcomings in perceptual expectations are likely. Alternatively, footage may meet perceptual expectations (be *realistic*) without being authentic. As described above, deepfake models strive to generate novel footage that is indistinguishable from their training data by simulating its properties. If successful, a deepfake appropriates the appearance of its subjects (Poulsen 2021) with minimal differences (Walton 1984). The resulting realistic footage may indistinguishable form a recording for a recipient. As will become apparent in the next chapter, many epistemic issues scholars raise in relation to deepfakes are rooted in their ability to produce such realistic, but inauthentic footage (*see* 3.1).

2.4 Deepfake Authorship

Deepfakes represent their subjects in a way that strives for realism (*see* 2.2) making it harder or even impossible for a recipient to distinguish them from recordings (*see* 2.3). In this capacity, they are created by someone, a *producer*, for the purpose of conveying certain meaning about their subjects. In digital information environments, this purpose generally also entails sharing a deepfakes with (unaware) others to communicate that information. Deepfakes allow producers to tell a story about their subjects ('*showing something or someone in a counterfactual way'*).

In itself, this is not novel. Recordings may also be used to "lie, palter, fudge, hedge, slant, and embellish" (Schauer 2022: 143), e.g. because they are framed in a misleading way. Nonetheless, the information recordings can communicate about their subjects is somewhat constrained by the past material states they depict. Deepfakes, on the other hand, transcend these constraints. In addition to established ways of influencing the content of footage – giving instructions to recorded subjects, cutting, software editing etc. – deepfake applications commonly provide an array of means to directly shape the elements present in their outputs. Examples include text prompts (Galatolo et al. 2021; Patashnik et al. 2021; Ramesh et al. 2021), the selection of reference and target material (Perov et al. 2021; GitHub 2023), fine-grained parameters to shape the generation process (ibid.; Millière 2022: 15), or simply regenerating material until satisfied. Instead of being bound by past material states, DFAs give producers a great degree of creative freedom in how they want to represent a subject. Insofar as it accessible to manipulate the generated outputs, deepfakes require much less time and resources to produce high quality results than traditional means of audiovisual media synthesis such as Photoshop (Millière 2022: 14f).

Deepfakes allow their producers to tell a story about their subjects in a way that is not immediately recognizable as being the result of a creative instead of an authentic process (*see* 2.3). Due to their potential to be realistic, deepfakes pose the risk that recipients mistake the story for something that happened. Much of the literature on the epistemic harm of deepfakes comes down to this risk of deception. Nonetheless, the success of the story a deepfake producer seeks to tell about their subject depends on how recipients make sense of the footage they are presented with. This process is the subject of the next chapter.

2.5 Concluding Remarks

In this chapter, I defined digital information environments (*platforms allowing users to publicly post and react to audiovisual content*), epistemic harm (*unjustified obstruction of epistemic success in a relevant area*) and deepfakes (*mostly realistic pieces of audiovisual media produced through machine learning showing something or someone in a counterfactual way*)(2.1). I further established that deepfake applications (DFAs) strive to produce realistic novel outputs based on imitating the characteristics of their training data (2.2). The argument for their epistemic significance made in the epistemic literature on deepfakes hinges on this capacity to be realistic (comply with perceptual expectations). Through it, deepfakes may be indistinguishable from recordings, which have a privileged position in epistemic practices because they are authentic (resulting from a process that consistently represents past material states). Because they are realistic, deepfakes may be confused with recordings (2.3). At the same time, DFAs give producers a great degree of creative freedom in shaping their outputs. As a result, DFAs can be used by the producers of deepfakes to communicate certain meanings about their subjects to recipients in digital information environments who may be unaware that they are encountering a deepfake (2.4).

Aside from these key insights, a few aspects of the overall phenomenon of deepfakes bear pointing out for the following chapters. Deepfakes as a socio-technical phenomenon are constituted by a variety of elements: Deepfake technology (the underlying techniques that enable deepfake applications and the data they are trained on), producers (those who create deepfaked media content), subjects (what or who is portrayed by a deepfake), recipients (those who encounter and engage with deepfaked content) and the digital information environment deepfakes are posted in. The epistemic harms caused by deepfakes, which I discuss in the next chapter, emerge as an interplay of these elements.

3. The Epistemology of Deepfakes

This chapter present the theoretical component of this thesis. Here, I engage with the epistemic harms deepfakes cause in digital information environments to answer the first research question: *How do deepfakes interact with the beliefs of their recipients and which epistemic harms may arise from this*?

In this effort, I summarize how the epistemic harm from deepfakes is understood in the current epistemic literature on deepfakes. Though some arguments point to the possibility of further epistemic harms from deepfakes (*e.g.* Harris 2021), I argue that scholars have so far mainly been occupied with epistemic harms in the form of deception, the undermining of recordings as evidence and the erosion of trust, all of which depend on the ability of deepfakes to pass as recordings (3.1). However, the epistemic literature on deepfakes does not provide a mechanism for how recipients interpret footage and may incur epistemic harm as a result (*see* 2.1). I provide such a mechanism for how recipients make sense of deepfakes drawing on Peircean semiotics and the framework of epistemic vigilance (3.2). Building on this account, I argue that there are two additional ways in which deepfakes may cause epistemic harm: *cognitive resonance* and *polarized fellowship* (3.3). I conclude with summarizing the key insights of this chapter and answering the first research sub-question (3.4).

3.1 Deepfakes as Deception

There is a variety of accounts on why, how, and to what extent deepfakes are epistemically harmful in the philosophical literature.[10] However, when it comes to the kinds of epistemic harm deepfakes cause, these accounts largely converge on three forms: deception, undermining recorded evidence and erosion of trust.

Deception, according to the literature, occurs when deepfakes leading their recipients to acquire false beliefs. Due to being realistic, recipients may mistake a deepfake as a recording and think it is an accurate representation of past material states (see 2.3). Alternatively, recipients aware of the existence of deepfakes may fail to acquire true beliefs from a recording because they doubt its authenticity. Whereas the realism of footage was once a reliable indicator for its authenticity, and therefore truthfulness of the depicted past material states, deepfakes decrease the probability that realism and authenticity coincide (*compare* Fallis 2021). Insofar is the failure to acquire true beliefs from recordings is not an individualized incident, this dynamic may jeopardize the role of recordings as evidence (ibid.: 631f; see also Atencia-Linares, Artiga 2023; Matthews 2022, 2023; Pierini 2023; Rini 2020). Deepfakes therefore jeopardize the role of recordings as evidence in epistemic practices and may, according to Rini (2020), also give someone caught on a recording the ability to claim being the subject of a forgery, the so-called 'liar's dividend' (ibid.; Chesney, Citron 2019).

Aside from these potential impacts on the beliefs of individuals, the role of recordings and the reliability of epistemic practices, deepfakes may further lead to an erosion of trust. Recipients may end up being overly sceptical toward others due to the perception that they are either susceptible to deepfakes or actively using them to deceive (Diakopouos, Johnson 2021; Rini 2021). As a result, individuals, institutions, and epistemic practices are at risk of losing trust from recipients (ibid.; *see also* Rini 2020; Diakopoulos, Johnson 2021; Fallis 2021; Atencia-Artiga 2023; Matthews 2022, 2023; Pierini 2023).[11]

[10] Again, epistemic harm, here, is understood the unjustified obstruction of the epistemic success of an agent in the area of politically relevant information (*see* 2.2).

[11] Deepfakes undermining trust is partially backed by empirical evidence obtained from Vaccari and Chadwick (2020) for a representative sample of the UK population. Empirical evidence obtained by Altay and Acerbi (2023) in the UK and the US suggests that perception of the vulnerability of others to misinformation was a potent predictive factor for how respondents perceived the threat misinformation poses.

All three of these instances present distinct epistemic harms. Deception understood as the acquisition of false beliefs presents an obvious obstruction of a recipients' epistemic success. Insofar as the acquired belief pertains to a relevant area, which is by definition the case for the kinds of deepfakes I am concerned with, this presents an epistemic harm. How jeopardizing the role of recordings as evidence is harmful is slightly less clear. If recipients fail to acquire true beliefs from a recording and recordings become a less reliable source of information (*see* Fallis 2021), those who populate the digital information environment have obstructed the epistemic success of recipients in the sense that epistemic success either was not achieved – because they did not acquire a true belief – or is harder to achieve because the effort recipient needs to invest to achieve it has increased (*compare* Kerner, Risse, 2020: 98).[12]

A similar case can be made for the erosion of public trust. Trust plays an important role in epistemic practices, from those in science (Levy 2022) to those in democratic systems (Warren 2018). If an actor – or an institution – is trustworthy and competent we can unburden ourselves from double-checking (Levy 2022). Nonetheless, trust in institutions with important epistemic functions, such as science, courts and governments has suffered in recent times (Hanson et al. 2019). Deepfakes may exacerbate this dynamic, for example through rendering it plausible that others, including institutional actors, are susceptible to fall for a deceptive deepfake (Diakopoulos, Johnson 2021) or are actively using deepfakes to engage in deception themselves. Insofar as the trust deepfakes erode is conducive of epistemic success – e.g. because it results in an epistemically beneficial division of epistemic labour (Levy 2022) or is necessary to uphold epistemically beneficial institutions and social systems (Warren 2018) – this is epistemically harmful.

Jeopardizing evidence and the erosion of trust present distinctive instances of epistemic harm. However, both are dependent on the (perceived) ability of deepfakes to deceive. If there is no (perception of) deepfakes threatening to deceive someone, there is little reason to suspect that the epistemic status of recordings may suffer, that evidentiary practices cease to be reliable, or that others, including institutional actors, will make false decisions based on deepfakes or use them to deceive others. For both kinds of epistemic harm, it must be plausible that deepfakes can pass as recordings. Few would argue a cartoon is produced in an attempt to provide false

[12] In other words, a practice – transmitting information through recordings – that was conducive of epistemic success was harmed (*see* 2.2).

evidence for the event it portrays.[13] The epistemic harms of jeopardizing evidence and eroding trust hinge on deepfakes being (perceived as) able to deceive. In turn, this perceived risk of deception will depend, at least to a considerable extent, on the realism deepfakes can achieve. Additionally, the risk of manifest deception from deepfakes – the acquisition of a false belief as a result from encountering a deepfake – also plausibly depends on deepfakes achieving a high degree of realism. These epistemic harms foreseen in the philosophical literature hinge on the ability of deepfakes to deceive. In light of this, it is unsurprising that many of the definitions in the literature refer to deepfakes as realistic pieces of synthetic audiovisual media (*see* 2.2).[14]

However, the notion of deception that underpins the epistemic harms from deepfakes drawn up by the epistemic literature is incomplete. Even if one concedes that the loss of a true belief is virtually equivalent to acquiring a false belief, it only covers three of the four conceptions of deception as defined by Chisholm and Feehan (1977: 143ff): Acquiring a false belief, loss of a true belief, and prevention of the acquisition of a true belief. The fourth kind of deception, preventing the loss of a false belief (ibid.: 144), is absent.[15] This suggests that there is room for additional kinds of epistemic harms beyond the current understanding in the literature.

Indeed, some authors have already suggested the possibility of epistemic harms that do not hinge on the capacity of deepfakes to deceive. Öhman (2022) points to such a possibility in cases where a deepfake does not represent an existing person, but only a synthetic non-existent stand-in. He argues that such footage may nonetheless evoke and reinforce prejudice held by recipients toward the community the synthetic stand-in is perceived to belong to (ibid.: 5). Insofar as such prejudice are false reinforcing them present an epistemic harm insofar as it obstructs the epistemic success of losing them.[16] Harris (2021) similarly highlights that

[13] Of course, cartoons may still be used to suggest what they depict is truthful, but the cartoon will derive its credibility from somewhere else, e.g. being based on an eye-witness account.

[14] Again, false beliefs may also be acquired from an unrealistic piece of footage, but such footage would need to be able to derive credibility from a source different from a false extension of the epistemic status of recordings. This will become relevant in the coming sections.

[15] Chisholm and Feehan strictly speaking define eight kinds of deception, as they further distinguish between instances of active causal contribution and passive allowance of the four kinds of deception (1977: 143ff). However, the passive allowance of deception is not easily reconciled with the understanding of epistemic harm as the violation of a negative right (*see* 2.2). I therefore only focus on active kinds of deception here.

[16] Habgood-Coote (2023: 18f) similarly worries that it may not be mainstream political discourse that will suffer the most from the epistemic harms deepfakes could enable, but marginalized communities. He stresses the relevance of the social contexts deepfake technology is introduced to.

deepfakes may have harmful associative effects, whether they are believed or not. He further argues that the recipients of deepfakes take the context in which they encounter them into account, e.g. the source that has posted them (*see* Bode 2021). Accordingly, Harris (2021: 13381) suggests that recipients will likely not incur epistemic harms from being deceived by deepfakes, so long as they have appropriate patterns of trust toward the source of the footage they encounter.

Taken together, Öhman (2022) and Harris (2021) suggest that there is more to the reception of deepfakes than their immediately present content and subjects that feature in it. However, the process of how recipients form beliefs from footage, including deepfakes, and what these beliefs pertain to, is not examined thoroughly in the philosophical literature. In the following section, I provide an account for how recipients make sense of footage, including deepfakes, drawing on Peircean Semiotics as described by Chandler (2022), as well as the framework of Epistemic Vigilance (EV), established by Sperber, Mercier and colleagues (Mercier 2017, 2019; Sperber, 1997; Sperber et al., 2010).

3.2 Making Sense of Footage

As has been established in prior sections, deepfakes carry information about their subjects (‘*showing someone doing or saying something’*)(*see* 2.3, 2.4). However, how what is depicted in deepfakes relates to the meaning and beliefs recipients derive from them has so far not been examined thoroughly, either theoretically or empirically.[17] In this section, I seek to close this gap by providing a conceptual framework for how recipients *make sense* of footage they encounter online.

The process of ‘*making sense*’ of footage can be separated in two parts: *interpretation* (grasping the content of footage) and *evaluation* (appraising its credibility).[18] The process of interpretation is accounted for by Peircean semiotics, whereas EV gives insight into how beliefs are formed based on communication. Semiotics and EV come from quite different domains. Whereas the former is rooted in philosophy of language, epistemic vigilance emerged from evolutionary psychology.[19] Nonetheless, both frameworks are concerned with how recipients engage with information and allow to point to vulnerabilities in this process that may make recipients susceptible to epistemic harms from deepfakes, even if those deepfakes do not manage to deceive them. Combining both frameworks is useful, as Peircean Semiotics, following Chandler (2022), does not provide a detailed account for how recipients come to form believes based on footage but encompasses both audiovisual media and other forms of communication. The framework of EV on the other hand provides a detailed account for how recipients form believes but was developed looking at verbal communication. So, while semiotics bridges between verbal communication and (audio)visual footage, EV tackles the issues of how information is evaluated with respect to believability. Both frameworks therefore complement each other.

[17] Some exceptions are the empirical studies done by Ahmed (2021), as well as Vaccari and Chadwick (2020).

[18] I do not assert that these processes are separated *in vitro*. Nonetheless, conceptually separating both is useful to understand the relation between understanding the meaning of a deepfake and forming a belief based on it.

[19] While evolutionary psychology faces considerable critique from other disciplines (*see e.g.* Confer et al., 2010; Huneman, Machery, 2015), I need not affirm the field’s methodology or epistemic premises in order to give credence to the mechanisms of epistemic vigilance. Irrespective of the origin of these mechanisms, they are logically coherent and corroborated by empirical evidence (Mercier, 2017, 2019).

Peircean Semiotics

Semiotics is concerned with "how meanings are made and how reality is represented […] through signs, sign systems, and processes of signification" (Chandler 2022: 2). Signs and sign systems are understood very broadly. A sign is anything that stands for something else (ibid.) and any act of interpretation can be conceived of as a semiotic process ('*semiosis*'). Signs mediate much of, if not all, human experience, including verbal communication (ibid.: 37) and obtaining information from footage.

While it is still contested what the right model for the relation of signs to meaning is, I will draw on the triadic model originally proposed by Peirce as described by Chandler (ibid.). Peircean semiotics principally applies to all kinds of signs, not just language, and is therefore suitable to be applied to recordings and deepfakes. Following Peirce, any sign can be conceptualized as a triad composed of three elements: *subject*, *representation*, and *reference* (*see* Figure 2).[20]

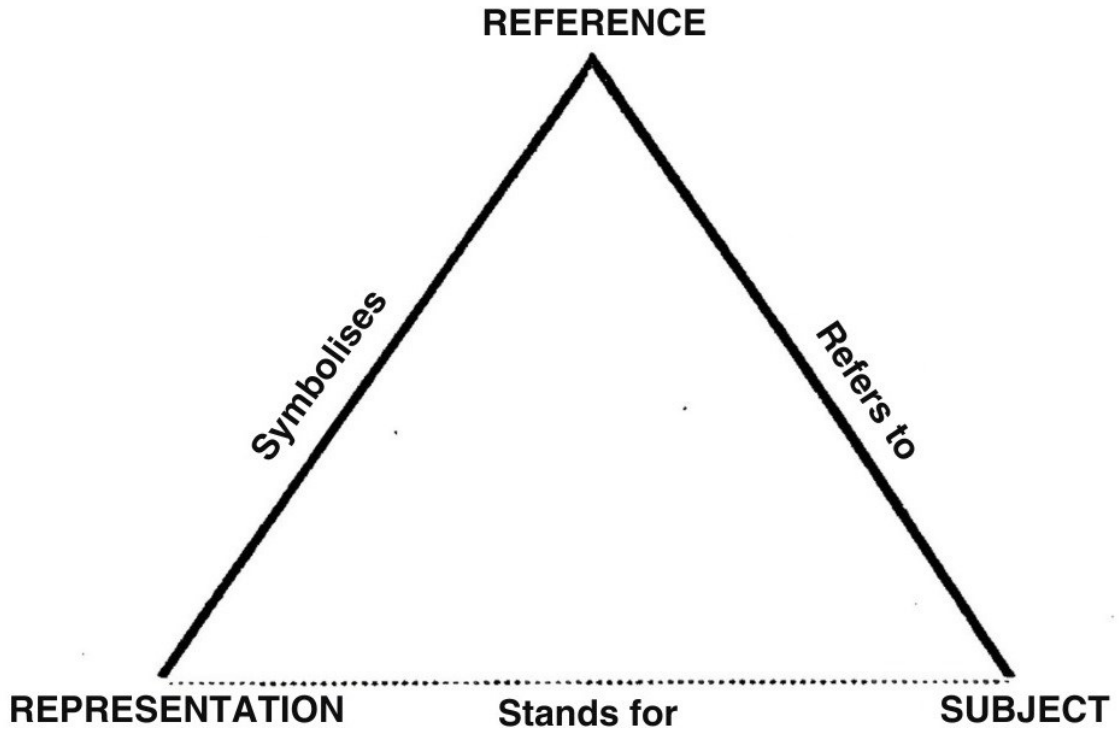


Figure 2. Adapted from Richards, Ogden (1923: 11).

'*Subjects*' are the thing(s) in the world as they are. They can be concrete, such as cat, or abstract, such love or a nation state (ibid.: 13). '*Representations*' describes the (usually physical) artifacts

[20] Peirce's language is slightly different. He differentiates between objects (here subject), representamen (here representation) and references (Chandler, 2022). To provide coherence with the discussion in the previous sections, the terminology was adapted.

that ‘stands for’ subjects, e.g. a flag, facial expression, word or pieces of footage. ‘*References*’ are the contingent notions of the subject that are evoked in the mind of a recipient when encountering a representation. References are socially acquired and grounded in experience (ibid.: 34f). When forming a sign, the three elements irreducibly coincide; each element mediates between the other two (ibid.: 32-39).[21]

To illustrate, consider the following example. Imagine you are holding a photograph of your late cat. Among the *subjects* of this sign is the actual animal individual, but also the species cat, pets generally, etc. The *representation* is the material photograph you are holding. The photograph is characterized by certain attributes: size, thickness, weight, colour, blur etc. Seeing this photograph evokes certain *references*: the cat’s names, what you like about cats, loss, etc. In others the same representation may likely evoke (some) different notions. A person who was bitten by a cat may recollect pain. An acquaintance may share some references, e.g. the cat’s name, but not others, e.g. the feeling of loss.

References like happiness or loss can themselves be understood as kind of sign that needs interpretation, evoking another set of related notions. References are interrelated with other references in a so-called *frame of reference* (FoR). Through it, the interpretation of a representation unfolds into a series of related references to form *meaning*. This interpretational framework is grounded in experience and socially acquired because “the meanings of signs arise in the context of use” (ibid.: 36f). Use of signs and FoR are subject to social convention and therefore socio-culturally and historically contingent.[22]

The success of intersubjective communication depends on an overlap in the FoR between the communicator and a recipient (ibid.: 9). However, as are socio-culturally contingent, frames of reference differ to some extent not only between cultures and groups, but also between individuals. The meaning of a sign is polysemious; the same sign evokes specific references in some interpreters and not others, depending on frame of reference and context. Recipients may also be able to choose to draw on some references and not others in their interpretation (*see*

[21] “*The* [representation] *as the conveyer of meaning mediates between the* [subject] *and the* [reference]; *the* [reference] *mediates between the* [representation] *and the object to interpret the meaning; the object mediates between the* [reference] *and the* [representation] *to ground the meaning.*” (Daniel 2008: 437, terminology adapted).

[22] However, the interpretation of signs and the use of sign systems is – at least in Peirce’s understanding – limited by the nature of their objects, though the true nature of a subject is not fully knowable (Chandler 2022: 36-39).

Barthes 1977: 274).[23] Producers need to carefully consider how they communicate the information they intend to convey through a deepfake (*see* 5) and how to tap into the FoR of their intended recipients. Seeing the same deepfake, different recipients are likely to arrive at different judgements regarding what they are seeing, including whether it is a truthful depicting of an event or not.

Regardless of whether a piece of footage, or any other sign, is understood to be a truthful representation of its subject, it may likely evoke additional references that are related to the subject (ibid.; Barthes 1977). Consider again the arguments made earlier (*see* 4) regarding the potential harm deepfakes may cause due to evoking prejudice and affect (Habgood-Coote 2023; Harris 2021; Öhman 2022). What is represented by a deepfake is not an objective fact but depends on the recipients' FoR and the context in which it is encountered. This insight is relevant for the next section. Before developing upon it, however, I need to establish how the believability of footage is evaluated.

Epistemic Vigilance

Epistemic vigilance (EV) describes a "suite of cognitive mechanisms" (Sperber et al., 2010: 559) which is routinely engaged when one encounters new information to judge whether adapting one's beliefs is warranted. EV makes use of a variety of cues based on informational content, its source, and the pre-existing beliefs of a recipient. Generally, EV makes people hesitant to adopt new beliefs (Mercier 2017, 2019).[24] This is due to a potential conflict of interest between interlocutors. Overall, people stand to benefit greatly from communication with each other as they most likely have an interest that can be served by communicating. A given recipient may obtain information they did not have before, and a given communicators may benefit from their audience having certain information and acting accordingly (Sperber et al. 2010: 359f).

However, the interest of the audience is only served when the communicator is competent (has accurate information) and honest. The communicators interests on the other hand may not

[23] Bathes (1977) makes this point specifically regarding images. However, this argument plausibly holds for other kinds of signs.

[24] Whereas recipients of disinformation are often (explicitly or implicitly) cast as too accepting of new information (Habgood-Coote 2019; Mercier 2017, 2019: 1-14), EV casts doubt on this notion, both from a theoretical and empirical perspective (Mercier 2017, 2019; Sperber 1997; Sperber et al. 2010).

be best served by being honest, but rather by producing certain effects in their addressees and getting them to act in certain ways. When the overarching interest of communicator and audience align, e.g. when they coordinate collective action, there is little incentive for dishonesty. However, in many situations the interest of a communicator and their recipients diverge, and the effects intended by the communicator come at the detriment of the recipients. Therefore, there are risks to uncritically adopting new information from others (ibid.). To mitigate the risk of falling for false information at the detriment of their own interests, people exercise epistemic vigilance (EV)(Sperber et al. 2010; Mercier 2017, 2019).

Principally, recipients of new information need not make the beliefs they may acquire from it the spontaneous basis of their behaviour. Instead, they may acquire a belief with an embedded qualification that can either be affirming or discrediting, e.g. '*this untrustworthy person posted footage showing X*'. Recipients may still contemplate X or even affirm it in conversation, but it would not be the basis of their actions without them considering the justification for the belief X (Mercier 2017: 105; Sperber 1997). In addition, there are several complimentary mechanisms of EV that appraise the source (EV-S) and the content (EV-C) of new information in order to determine whether it should be adopted (Sperber et al. 2010: 369).

Source-based EV

Mechanisms of EV-S judge the trustworthiness of a source based on perceived competence, benevolence and tracking of commitments by the source to its statements and audience (ibid.; Mercier, 2017: 106). A source is perceived to be competent if it is dispositioned to have formed accurate beliefs and as benevolent if it is perceived as contributing positively to one's interest (ibid., *citing* Barber, 1983). Competence and benevolence depend on context. A source may be competent about some things, not others, and benevolent to one audience, but not another. Accordingly, "[t]rust should be allocated to informants depending on the topic, the audience, and the circumstances" (Sperber et al., 2010: 369). Often, there is little evidence to judge how much a source should be trusted precisely. EV-S therefore relies on various cues; epistemic ones for competence and normative ones for benevolence. Table 1 gives examples for these cues.

Type of cue	Cue	Example question(s)
Competence	Access to information	How would the source know about this?
	Intelligence	Does the source even have the capacity to understand what it is trying to communicate?
		Would the source be able to pick up on inconsistencies when encountering this information?
	Diligence	Did the source do an appropriate amount of research to corroborate this?
		Would the source moderate its claims proportional to the evidence?
	Credentials and judgment of trustworthy others	Does the source have the skills necessary to evaluate its evidence?
		Do other competent individuals and institutions vouch for the competence of the source?
		Does my friend who also studied this subject agree with the source's conclusion?
		Do most trustworthy people in my circle agree with the source?
Benevolence	Moral evaluation	Is the source a good person?
	Possible intentions	Why could the source want me to believe this?
		How does this information serve the source's interests?
		How does adopting this information serve or jeopardize my interests?
	Allegiance	Does the source belong to a coalition I belong to?
		Does the source belong to a coalition that opposes me?

Table 1. Source-based ques to trustworthiness. Mercier, 2017, 2019; Sperber et al., 2010.

These cues are then corroborated by tracking the commitments of a source. If a source commits to a piece of information – e.g. through purporting to believe in it – recipients take notice and impose costs should the information turn out to be misleading. They may reduce how much they trust the source in the future, how much they are willing to cooperate, and may share information regarding its unreliability with others, inflicting reputational costs (Mercier, 2017: 106, 2019: 87-90). This in turn diminishes the communicator's ability to exert influence, thereby incentivizing them to be reliable (ibid.).

Communicators, including malevolent and incompetent ones, understand, in some capacity, that their audience is looking out for the cues described above to judge the believability of what they are saying (Mercier, 2017, 2019); they may even use this to their advantage. Communicators may leverage the cost associated with violating commitments to gain the trust of (some) recipients, e.g. by expressing confidence in a statement or making promises of

commitment. Because recipients and communicators are, to some extent, aware that this increases potential costs, this increases the incentive for the communicator to speak from competence and follow through on their commitments (Mercier, 2017: 106).

Exploiting the same dynamic, there is another way a communicator can signal benevolence and commitment to a specific group of recipients. Both cues are also used to assess whether someone is suitable for collaboration in the pursuit of shared interests. They give an indication whether someone is only looking to reap benefits or is also willing to share costs (Mercier 2019: 192). By deliberately '*burning the bridges*' to other groups of potential collaborators "[communicators] can credibly signal to the remaining groups that [they will] be loyal to them, since [they] don't have any other options" (ibid.: 193). Mercier suspects this may be the motive for some to purport to hold absurd beliefs:

> "*When a writer suggests that Kim Jong-il can teleport, he doesn't expect his audience (least of all Kim Jong-il) to literally believe that. The point, rather, is to make the groveling so abject that even other North Koreans find it over the top. By signaling to other North Koreans that he's willing to go beyond what's expected in terms of ridiculous praises, the writer is telling the audience that he would rather seek Kim Jong-il's approval than that of a broader base of more sensible people.*" (ibid.)

Abject flattery is not the only shape this may take, other examples include appearing incompetent, or making morally repelling statements. Any statements that make one seem less trustworthy as a source of information to some audiences but not others are principally suitable (ibid.: 193f). At its most effective, burning bridges offends "the intelligence or moral standing of those who disagree with the beliefs used to burn bridges" (ibid.: 195). As I will argue in the next section, deepfakes may be especially useful for this strategy. Nonetheless, the trustworthiness of a source is only part of evaluating the believability of new information through EV.

Content-based EV

Some pieces of new information we encounter, on the merit of their content alone, are more believable than others. This believability based on content is judged through the two main mechanisms of EV-C: reasoning and plausibility-checking (Sperber et al., 2010: 374; Mercier, 2017). Through reasoning, pre-existing beliefs that are held as certain knowledge and arguments that speak for and against adopting new information are weighed (Mercier, 2019:

53; Sperber et al., 2010: 374). Information that is internally logically consistent according to reasoning is tentatively deemed believable whereas information that is internally contradictory is tentatively deemed unbelievable. Such reasoning may take place individually or in the exchange with others (Mercier, 2017: 109).[25] Empirical evidence suggests that people are rather good at recognizing weak arguments in cases where their conclusions are relevant to their interests (Petty, Wegener, 1997).

Nonetheless, the available evidence for reasoning through new information is often lacking. EV-C cannot always rely on internal logic and relevant knowledge to judge veracity (Sperber et al. 2010: 374). Often, "[t]he believability of newly communicated information must be assessed relative to background beliefs which are themselves open to revision" (ibid.). Accordingly, EV-C often relies on a checking the coherence of new information with existing related background beliefs to assess its plausibility (ibid.: 374ff). In other words, recipients draw on their frame of reference to appraise the plausibility of new information (*see* above). Should inconsistencies arise, there are three possible outcomes depending on the perceived trustworthiness of the source of the new information (ibid.: 375).

1. If the source is regarded as untrustworthy, the new information is likely to be rejected outright.
2. If the source is regarded as trustworthy and conflicting background beliefs are not held with strong conviction, background beliefs will be revised.
3. If the source is regarded as trustworthy and conflicting background beliefs are held with conviction, then either the perceived trustworthiness of the source will be revised, or those background beliefs. This requires conscious effort.

The mechanisms of EV-C, reasoning and plausibility checking, assess the believability of new information on the basis of its internal logic and its coherence with pre-existing beliefs. Even when new information conflicts with existing beliefs, it may be adopted, and the existing beliefs may be revised – as much as is needed to establish coherence (ibid.: 375f) – depending on how entrenched those existing beliefs are, and how trusted the source is.

Whether new information is accepted by recipients is a function of the perceived trustworthiness of a source, as well as the internal logic of that information and its coherence with existing beliefs. Either one of these mechanisms may be decisive for the adoption of a new

[25] Though the latter is generally more effective (Liberman et al. 2012; Minson et al. 2011, *both cited in* Mercier 2019: 54).

beliefs. In all of this, the interpretation of what is portrayed in footage and the evaluation the information that is conveyed thereby through mechanisms of EV are contingent on the individual recipients. The recipients' FoR, including their pre-existing beliefs, are historical and socially acquired and contexts in which new information is encountered differ. Further, judgements of trustworthiness depend considerably on additional information, e.g. presumed access to information, track record of commitment, reputation. While the mechanisms of making sense of footage may be widely shared among the human population, this means they will still likely result in varying judgements about a given piece of information. Additionally, the process of making sense draws on a broad range of beliefs, experiences and attitudes towards the subject, representation and source in question, all of which may be epistemically significant, as I discuss in the next section.

3.3 Harmful Ways to Make Sense of Deepfakes

If EV is effective in preventing recipients from acquiring false beliefs from deepfakes that would be good news as they would then pose less of a threat with respect to epistemic harm. However, even if EV may largely be successful in shielding recipients from being deceived by deepfakes, I argue the way recipients make sense of footage (*see* 3.2) also introduces some underappreciated vulnerabilities for epistemic harm from deepfakes.

As the mechanism of EV suggest, the realism of footage alone may not be the decisive factor regarding whether the content of footage may be believed or not. Realism suggests that footage is authentic and authentic footage is principally generally in a promising position with respect of clearing the mechanism of EV. Authenticity means that the content of footage is necessarily internally consistent as it portrays factual past material states (*see* 2.3), therefore strongly suggesting a positive evaluation by the mechanisms of EV-C. Further, having access to authentic footage is a cue that a source communicates with competence (*see* Table 1). This in turn suggests the source of the recording is trustworthy (EV-S)(*see* 3.2). However, that does not necessarily take precedent over other mechanisms of EV. Should the content of a recording clash with strongly held pre-existing belief and therefore fail in the process of plausibility checking, or should a recipient deem the source of a recording untrustworthy, e.g. because the take then to be a bad person that belongs to an opposing position and have lied to them before, a recipient may still end up rejecting a recording as false.

Vice versa, the fact that footage is visibly unrealistic may make it generally less likely for it to receive a positive evaluation during the process of EV, reasoning, plausibility checking, as well as cues to benevolence and commitment still factor in its evaluation and may take precedent against what others may see as better judgement. As a result, recipients may still end up believing an unrealistic deepfake and incur the associated epistemic harms.

However, I argue there are at least two additional ways in which deepfakes may cause epistemic harm: *Cognitive resonance* and *polarized fellowship*.

Cognitive Resonance

As established in the previous section, making sense of footage renders a broad range of background beliefs and attitudes salient in their recipients (*see* 3.2). Whether or not footage is evaluated as believable, there is significance to this. During the process of interpretation, deepfakes elicit associations toward their subjects (Harris 2021) and may end up presenting a convenient justification for pre-existing desires (Mercier 2019). As both effects are based in a resonance between the content communicated by a deepfake and pre-existing attitudes in their recipients, I unify them under the label of *cognitive resonance*.

That representations in pieces of media can induce associations attached to their subjects is backed experimental evidence (Foroni, Mayr 2005; Wittenbrink et al., 2001; *see further* Huebner, 2016). These association may cause non-epistemic harm to subjects, e.g. because they indignify them (de Ruiter, 2021; Rini, Cohen, 2022), but they also have an epistemic dimension. For one, provoking a strong emotional response may increase the believability of false information, thereby increasing the risk of deception (Vlasceanu et al., 2020). However, visual communication also plays a significant role in politics. For example, recipients evaluate the suitability of a politician as a leader partially on the basis of non-verbal cues displayed in television broadcasts (Bucy 2011). Deepfakes have been shown to be able to influence both the explicit and implicit attitudes people hold toward their subject, even when footage is known to be fake (Hughes et al. 2021). However, there is also evidence suggesting that recipients react more negatively to deepfakes that portray politicians in a way that deviates significantly from the recipient's beliefs about that politician prior to encountering the deepfake (Hameleers et al. 2023).[26] It is unclear whether the formation of associations is subject to mechanisms similar to those of EV, but it is plausible that they do to some extent. If a recipient already doubts the suitability of a politician as a leader, encountering a humiliating deepfake of them plausibly reinforces those doubts to some extent.

Insofar as deepfaked footage biases a recipient towards its subject or reinforces pre-existing prejudices, it causes epistemic harm as it hinders that recipients in coming to an epistemically sound judgement of that subject.[27] Recipients of humiliating footage of a politician might no

[26] Hameleers and colleagues (2023) did, however, only ask recipients to assess the believability of footage, meaning their findings might not transfer to associations.

[27] Compared to other kinds of text-based disinformation, deepfakes may be especially potent to induce such associations (Harris 2021: 13388; Rini 2020: 11), as (audio)visual media is particularly suitable to evoke memories (Vaccari, Chadwick 2020: 2).

longer take them seriously (Harris, 2021: 13388). This may further cause epistemic harm on a collective level, for example if the introduction of such footage veers public discourse toward addressing it (Rini, Cohen, 2022: 148-153) at the expense of discussing other relevant information.

However, that deepfakes can evoke such associations does not mean that they linger. If these associations are only short-lived, the epistemic harm that stems from them might not be significant. Unfortunately, there is reason to believe that these associations may linger. Mercier (2019: 204-208) argues that belief does not categorically precede interest. People do not just want act in certain ways due to the beliefs that they hold, e.g. not vote for a politician because they think they are not a suitable leader, but rather some beliefs are held in order to justify actions that were already desired before adopting them. In this case, people would come to believe that a politician is not a suitable leader because they do not want to vote for them. The believe that the politician is not a suitable leader serves as a *faux-justification* for not voting for them, though this is what one was going to do regardless.

Mercier argues (ibid.) such faux-justifications are adopted, not because a recipient is convinced by their veracity, but because they have utility. This argument is corroborated by results obtained in a study by Kim and Kim (2018) on the rumour that Barack Obama is Muslim. The authors conclude that the rumour only resonated with recipients that already disliked Obama before they happened upon the rumour. Similarly, in a study by Nyhan and colleagues (2017), supporters of Donald Trump did not adjust their level of support, even when they were exposed to and accepted corrections of false statements made by the former US president. Agents who oppose Obama or support Trump may face criticism that is hard to reconcile in absence of the ability to produce a justification. Faux-justifications provide reasons to engage in behaviour that might otherwise be shunned (Mercier, 2019: 206). The need to be able to produce a justification that makes shunned behaviours more acceptable gives rise to a kind of "market for justification" (ibid.). Given the epistemic status of recordings and the ability of deepfakes to approximate the realism of recordings to an extent (*see* 2.2, 2.3) that may give an inclined recipient plausible deniability regarding its veracity, deepfakes may be a particularly potent way to make offerings on this market.

Faux-justifications cause epistemic harm in two ways. First, because they allow those who hold them to deflect justified criticism and second, because, insofar as they are believed, they preclude that agents reflecting on their desired actions in epistemically productive ways. However, as the second kind of harm tethers to the acquisition of false beliefs from deepfakes

– though for a unique reason – I will hold that *faux*-justifications are a part of the epistemic harm of cognitive resonance because they provide agents with the utility of deflecting criticism toward their desired actions, without the need of being deceived into actually believing them.

Polarized Fellowship

Deepfakes provide a means to their producers to communicate meanings about the subjects they portray (*see* 2.4) to their recipients in digital information environments. In these environments, groups of recipients "gather […] to discuss shared cultural objects and worldviews" (Bode 2021: 921). Once posted, deepfakes are not just passively received, but recipients are able to engage with each other and conversations form. The modalities of this engagement depend on the technological affordances a platform provides. Beyond this, "cultural, ethical, and aesthetic spoken or unspoken rules" (Bode 2021: 922) that emerge from among the audience. In turn, the conversation that gathers around footage online influences its reception (ibid., *citing* Walther et al. 2010: 25-26). Deepfakes become anchor points for impromptu epistemic communities, inviting discussion of, among other things, their content, subjects and source, the epistemic status of the footage, and the reception in the comments (Bode 2021). This aspect of the phenomenon is part of what deepfakes are when they are posted online.

These characteristics of media reception in digital information environments introduce the second underappreciated way in which deepfakes may cause epistemic harm. Producers can leverage deepfakes to shape discussions online in a way that serves their own interest but causes epistemic harm in the process. As mentioned above (*see* 2.4), deepfake applications give producers the means to craft footage representing subjects in a myriad of ways. This allows them to – consciously or unconsciously – play into references shared with the audience of their deepfakes, like any form of communication requires (*see* 3.2). Producers will have certain expectations about their audiences' reception and – while this may not play out exactly as planned – can elicit certain interpretations. This applies to the content of a deepfake, but also to the evolving discussion to the producer themselves (or their online presence) as the source of the deepfake. The source a footage is part of the discussion that emerges during its reception in digital information environments (*compare* Bode 2021).

For producers who have malignant intentions, deepfakes lend themselves particularly well to shaping conversations around them in ways that are likely to cause epistemic harm. First, deepfakes plausibly evoke cognitive resonances in inclined audiences (*see* above). Second, they

are also likely to be epistemically contested (*compare* Bode 2021; Hameleers et al. 2023). Third, creating a non-consensual deepfake of someone is morally contested (*e.g.* de Ruiter 2021; Rini, Cohen 2022). These factors make deepfakes highly useful for burning bridges in an effort to cue benevolence to a subgroup of recipients (*see* 3.2). Deepfakes that portray their subjects in a derogatory or otherwise harmful way are likely to both cause moral backlash and epistemic contestation. Here, political and pornographic deepfakes have significant overlap, conceptually as well as practically.[28] Recipients who engage positively to this tactic will also signal to opposing voices that they are willing to go along with the harm endured by the subjects of such a deepfake. A producer of such a deepfake and recipients who respond to this tactic positively will mutually signal themselves that their interest do not align with those in opposition but do align with each other while driving opposers away. As a result, deepfakes can effectively drive a wedge between recipient communities. In doing so, however, deepfakes do not need to deceive anybody. In fact, them being abjectly false may very well amplify this dynamic (Mercier 2019: 192ff). Through inviting the contestation of their epistemic status and moral outcries, deepfakes may serve as anchor points around which like-minded recipients form epistemic communities (ibid.: 208). It is this dynamic of drawing in like-minded recipients and antagonizing dissenters that I want to call *polarized fellowship.*

How does polarized fellowship cause epistemic harm? First, it might amplify the diffusion of faux-justification resulting from a deepfake (*see* above). Communities that emerge around morally and epistemically contested deepfakes may be a good marketplace of faux-justifications as inclined recipients are likely to produce additional reasons for why it may be justified to inflict harm on the deepfake's subject or antagonize those who oppose the whole undertaking in the comments. Some of these additional reasons may be compelling to other recipients, serving to further entrench existing beliefs.

Secondly, evidence from two meta-studies shows that in discussions within like-minded groups, arguments tend to accumulate on one side of an issue and viewpoints on this issue tend to become more extreme in line with the groups' predisposition (Isenberg 1986; Myers, Bach 1974; Vinokur 1971, *all cited in* Mercier 2019: 209). This poses an epistemic harm insofar as it is highly dubious whether what emerges from this dynamic qualifies an epistemic success (*see* 2.1). This may further exacerbate the wedge between recipients that concur with the purpose and/or the content of the initial deepfake and its opposition. Ultimately, this may lead

[28] Pornographic deepfakes have e.g. been leveraged to retaliate against investigative journalist Rana Ayyab for perceived transgressions against the Indian BJP party (Ayyab 2018).

recipients into an echo chamber of sorts in which they either are themselves no longer receptive to perspectives that differ from the views held in their new-found epistemic communities or in which others are no longer willing to share those perspectives because recipients who are part of *polarized fellowships* are, rightfully, seen as hostile. Either way, this precludes opportunities for encountering epistemically valuable information from such recipients who fell for but were not deceived by a deepfake.

3.4 Concluding Remarks

In this chapter, I analysed the current state of the literature on the epistemic harm of deepfakes, concluding that it is mainly occupied with the capacity of deepfakes to be mistaken for recordings. The epistemic harms from deepfakes – deception, jeopardizing recordings as evidence and eroding trust – are thus rooted in the ability of deepfakes to deceive. However, an account that specifies how recipients interpret the meaning of and form beliefs based on footage – how they make sense of it – is absent from the epistemic literature on deepfakes (*see* 3.1).

Building on the frameworks of Peircean semiotics and epistemic vigilance, I provided such an account. When interpreting footage, recipients draw on a wide range of socio-culturally and historically contingent references. The evaluation of the believability of footage is based on its content and its source. Content is evaluated according to internal logic and coherence with pre-existing beliefs, whereas the source is evaluated based on cues to competence, benevolence and adherence to previous commitments (3.2).

Resulting from the process of making sense, deepfakes may evoke associations regarding their subject. These associations may be epistemically harmful insofar as they lead a recipient to (continue to) perceive that subject in a biased manner, or if a deepfake serves its recipient as a faux-justification that allows them to deflect criticism. I call these dynamics *cognitive resonance*. Further, deepfakes enable their producers to signal benevolence to some part of their audience, while antagonizing others. In turn, recipients who react positively to this tactic will themselves antagonize observers. In this capacity, deepfakes form anchors for epistemic communities that may reinforce their respective beliefs while precluding opportunities to encounter other perspectives. This is what I understand as the *polarized fellowship* that deepfakes may be uniquely suitable in facilitating. Both of these kinds of harm do not depend on deepfakes being able to deceive.

Having made these arguments, I can answer the third research sub-question: *how do deepfakes interact with the beliefs of their recipients and which epistemic harms may arise from this*? Deepfakes evoke a range of pre-existing beliefs and attitudes in their recipients. These pertain to the subjects of deepfakes, the believability of their content and the trustworthiness of their source. Depending on how content and source are evaluated, deepfakes can result in either an adjustment of these pre-existing beliefs or the formation of new beliefs. Additionally, deepfakes may evoke associations about their subjects. However, none of these effects occur necessarily. The epistemic harms that arise from this process are either based in deception –

including jeopardizing the role of recordings as evidence and erosion of trust – or a preclusion of epistemic success through the dynamics of *cognitive resonance* and *polarized fellowship*.

Having concluded the theoretical component of this thesis, I now turn to the empirical analysis and my second research sub-question: *How is the epistemic harm of deepfakes understood and addressed in relevant EU policy documents*?

4. Method

In this chapter, I lay the groundwork for answering these questions. First, I clarify the specific case I am research: Deepfake policy in the EU. As I show below, deepfake policy in the EU is not a straight-forward matter, but rather wrapped up in the larger field of disinformation and artificial intelligence policy (4.1). After having delineated my case, I describe and justify the means through which I conduct my research. In section 4.2, I describe the selection of policy documents which I analyse. In section 4.3, I specify the method for analysis, quantitative content analysis following Kuckartz and Rädiker (2023). As the empirical analysis will build on the insights gathered in previous chapters, I will specify how it does so in section 4.4. Again, I conclude with a summary of the research design (4.5).

4.1 Case Description

In the EU, deepfake policy is scattered across a variety of complementary regulatory frameworks (van Huijstee et al. 2021: 37-47). In their report to the European Parliamentary Research Service on deepfakes, van Huijstee et al. (ibid.) list a total of twelve policy frameworks that to some extent apply to deepfakes: General Data Protection Regulation, copyright law, image rights, criminal law, Audio Visual Media Directive, e-Commerce Directive and Digital Services Act, Code of Practice on Disinformation, Action Plan on Disinformation, European Democracy Action Plan, proposed Artificial Intelligence Act, select Parliamentary Resolutions. Of these, only the proposed AIA and some Parliamentary Resolutions address deepfakes directly (van der Sloot, Wagensveld 2022: 7).

This has significant implications for the present research interest. As there are not just a few distinct policies that tackle deepfakes comprehensively, deepfake policy is entangled with other policy discourses in which the diverging frameworks that are relevant to it are enrolled.[29] Deepfakes are, as a heterogenous phenomenon comprised of a variety of elements (*see* 2.5), situated at the intersection of various policy fields, which is reflected by the list above. However, of these fields, the two most prominent are artificial intelligence and disinformation policy. As products of machine-learning systems, deepfakes are clearly implicated by artificial intelligence policy, and given the great degree of creative freedom they provide (*see* 2.4) and the epistemic harms they may bring forth (*see* Ch. 3), it is unsurprising that deepfakes raise concerns about use in disinformation campaigns.

In recent years, disinformation and artificial intelligence policy have seen considerable activity in the EU and have developed a considerable intersection as concerns regarding the potential use of AI systems for disinformation have been raised (*e.g.* AIDA 2021; Chesney, Citron 2019; Smith, Mansted 2020). In the following, I will present the mayor developments in both areas. I begin with disinformation policy, as it is overall the most relevant policy field when it comes to kinds of deepfakes I am concerned with (*see* Durach et al. 2020; Datzer, Lonardo 2022; Justo-Hanani 2022).

Disinformation policy has become a salient issue in EU policy since Russia invaded eastern Ukraine in 2014 (Datzer, Lonardo 2022: 757; van Huijstee et al. 2021: 42). Since, policy-

[29] It also means the study deepfake policy needs to be approached by drawing on a range of policies that are mostly concerned with other issues and are only partially relevant to the questions I seek to answer, which will be addressed in the following sections.

makers have been quite active in the field, often explicitly tying the phenomenon to digital information environments (*see e.g.* EC 2022a; Regulation (EU) 2022/2065). Disinformation, in terms of EU policy, is defined as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm" (Wardle, Derkahshan 2017: 20; *see also* EC 2018d). It is distinguished from misinformation by an intention to mislead or profit economically and from malinformation by not being based in reality (ibid.).

To counteract disinformation, EU institutions have implemented numerous measures. In 2015, the EEAS published an Action Plan on Strategic Communication (EEAS 2015) which set up the Eastern Strategic Communication Task Force. The task force specifically engages with countries in the eastern neighbourhood of Europe to mitigate the influences of disinformation on the region (Helding 2021: 7, *cited in* Datzer, Lonardo 2022: 757). Efforts to counteract disinformation were accelerated again after the success of the Donald J. Trump presidential campaign in 2016 (Datzer, Lonardo 2022: 758), resulting in heightened regulatory activity in the following years. In 2017, a High-Level Expert Group on Fake News and Online Disinformation (HLEG FNDI) was set up (EC 2018d: 3) and a voluntary Code of Practice on Disinformation (CoP) was proposed and signed by private actors including large platforms such as Google and Facebook (EC 2019a). Further, an Action Plan against Disinformation was put forth in 2018 (EC 2018a, 2018b). The EC also published two key communications setting out the strategic agenda and specific measures to counteract disinformation (Durach et al. 2020: 10; EC 2018c, 2018d). Next, a European Democracy Action plan was unveiled in 2020 (EC 2020a). Aside from these activities on the regulatory side, media literacy campaigns play a pronounced role in the EU's approach to disinformation (Datzer, Lonardo 2022; Sadaba, Salaverria 2023).

The two most relevant developments in European disinformation policy, however, are is the recent update of the CoP of 2018 (EC 2022a) and the adoption of the Digital Services Act (Regulation (EU) 2022/2065). The strengthened Code of Practice (sCoP) is a reaction to criticism regarding the efficacy of and compliance with the initial CoP (ERGA 2020), resulting in a more stringent co-regulatory framework. The newly adopted Digital Services Act (Regulation (EU) 2022/2065) is set up to be the landmark piece of legislation for the digital economy and online platforms. In this effort, it also tackles targets (illegal) disinformation (Datzer, Lonardo 2022: 757ff) and sets the foundation for embedding the sCoP as a Code of Conduct, increasing the EU's influence in the co-regulatory scheme (*see further* Appendix 5).

Similar to the heightened activity in disinformation policy, artificial intelligence policy has been a busy are for the EU in past years. AI has been labelled by the EC as "one of the most strategic technologies in the 21st century" (2018e: 1). Compared to disinformation, European artificial intelligence policy dates back further, back to 2010 when the EU began drawing up its digital agenda (EC 2010). Beginning with a parliamentary resolution in 2017 (EP 2017), the field has seen increased activity. but has recently seen increased regulatory activity, Following the resolution in 2018, the EC issued a communication on artificial intelligence (EC 2018e) and the High-Level Expert Group on Artificial Intelligence (HLEG AI) drafted ethical guidelines for AI (HLEG AI 2019a). In 2019, an EC communication on "building trust in human-centric artificial intelligence" (EC 2019b) was issued. The following year, the EC published a whitepaper for AI regulation (EC 2020a) which would lead into the proposal of a legislative framework for the regulation AI, the so-called Artificial Intelligence Act (AIA)(EC 2021a; Justo-Hanani 2022: 146-150). As "the first ever legal framework on AI" (EC 2021b), the proposed AIA takes a risk-based approach to the regulation of artificial intelligence systems and applications. AI systems are either characterized as low-risk, high-risk, or prohibited. Following this classification, different obligations and liabilities for the providers and users of such systems arise (EC 2021a). Again, I will return to these documents in the next section and now turn to deepfakes in particular.

From this brief introduction to deepfake, disinformation and artificial intelligence policy in the EU, the report by van Huijstee and colleagues (2021), and the understanding of deepfakes established in chapter two, a few (possible) approaches to deepfake regulation can be identified:

1. Targeting producers and technology providers through liabilities resulting from rights of subjects (e.g. privacy rights, image rights, criminal law)
2. Targeting platforms that deepfakes may be shared on
3. Targeting subjects, producers and recipients through educational measures (e.g. media literacy trainings)

Scholarly literature on deepfake policy in the EU is scarce, especially when it comes to the role of the latter two approaches. However, the role of privacy rights and the GDPR in deepfake policy has been analysed by van der Sloot and Wagensveld (2022). For the production and spreading of a deepfake to be in compliance with the GDPR, either the portrayed subject has to consent (Regulation (EU) 2016/679 Art. 9), or the purpose of the deepfake has to meet the legitimacy principle and the subject has to be informed about the production (ibid.: Art. 5, 6, 13, 15). Successfully claiming the legitimate purpose of a deepfake will most likely rely on the

producer being able to evoke their right to freedom of expression (van der Sloot, Wagensveld 2022: 10f). In absence of consent or a successful claim to legitimate purpose subjects of deepfakes can evoke their privacy rights or base appeals on the ground of criminal or tort law (ibid.). However, there are considerable issues of enforceability. Legally establishing that a given person is indeed the subject of a given deepfake (Öhman 2022) and causally relating a deepfake to a harm caused, e.g. violence against minority groups, are likely to proof difficult (van der Sloot, Wagensveld 2022: 10). Further, much of the enforcement of privacy and image rights, copyright and criminal law depends on identifying an actor to hold liable, which should not be taken for granted, especially in a disinformation context (ibid.).

4.2 Document Selection

There are two key challenges the empirical investigation in this thesis is faced with. First, as mentioned in the previous section, van Huijstee and colleagues (2021: 37-47) have already conducted an analysis on deepfake policy in the EU. My analysis will need to be distinct and provide additional insights. Second, the scattered nature of EU deepfake policy has implications for how researching the conception of epistemic harms from deepfakes and how they are met is best approached. Because I am interested in uncovering the understanding of epistemic harm (*see* 2.1) in EU deepfake policy, a qualitative approach is warranted. For such an approach, the details of which will be addressed in the next section, the fact that deepfake policy is distributed across a variety of regulatory frameworks that only apply to this specific issue is small parts carries the risk of bloating the sample beyond what I can reasonably cover here. Apart from specifying the analysed documents, the purpose of this section is to address how these challenges are met and therefore justify the selection of documents and overall analysis.

Generally, my selection follows a qualitative sampling plan oriented toward the relevance of the content of the selected documents to the issue of governing the epistemic harms of deepfakes in the context of digital information environments on the EU level (Döring, Bortz 2016: 303f). This means that the documents must be official documents by EU legislative institutions (commission, parliament, council, or commissioned advisors) that either tackle deepfakes directly or are otherwise relevant to the phenomenon because they tackle disinformation in digital information environments or artificial intelligence (*see* 4.1). However, as a sample only following these two criteria would still be too large for a qualitative analysis, additional adaptation of the sample is needed.

As mentioned above, the report to the European Parliamentary Research Service 2021 (van Huijstee et al. 2021) already provided a detailed analysis of EU policy on deepfakes at the time of its writing. Instead of repeating an analysis of the policy documents that were considered in it, the report itself will be analysed as it fits the criteria above and may serve as an approximation of the contents of policy documents it considered. This is justified beyond pragmatic reasons as there are additional reasons for not subjecting all policies covered by the report to an in-depth analysis.

Since the publication of the report, the proposed Digital Services Act (EC 2020b) has become law and a new Strengthened Code of Practice (EC 2022a) was set up. Both supersede the original Code of Practice on Disinformation (2018a), the Action Plan on Disinformation (2018b), the European Democracy Action Plan (2020a) and the e-Commerce Directive

(Directive 2000/31/EC) and will therefore be considered in their stead. The Audiovisual Media Services Directive (Directive 2010/13/EU) applies to the distribution of violent and pornographic imagery (van Huijstee et al. 2021: 42), which are not the focus of this thesis.[30] Further, the role of the GDPR (Regulation (EU) 2016/679) in deepfake policy has already been analysed by van der Sloot and Wagensveld (2022) in addition to van Huijstee et al. (2021: 38f). Both will therefore not be considered in detail. As criminal law, image rights and copyright are largely in the hands of member states (ibid.: 40), they will also not be considered here as they are out of scope of the research question.

Apart from these omissions and adding the DSA and sCoP, there are some other relevant documents which are absent from the analysis by van Huijstee and colleagues (2021) but will be considered for analysis. The recitals of the DSA (Regulation (EU) 2022/2065) and proposed AIA (2021a) provide motivations for the regulatory measures they embody and will hence be included in addition to their articles. While they yield some insights regarding the conceptual understanding of the epistemic harm caused by deepfakes, strictly regulatory documents such as the sCoP, DSA and proposed AIA only sparingly introduce or define (most of) the concepts that they are deploying.

Instead of solely relying on them, I will also turn to documents that have been produced precisely to engage with those concepts within the discourses on disinformation and artificial intelligence policy. These include communications by the European Commission (2018d, 2018e, 2018f, 2019b, 2021c, 2021d), relevant documents by the High-Level Expert Groups on Fake News and Disinformation (HLEG FNOD 2019) and Artificial Intelligence (HLEG AI 2019), as well as two studies on disinformation commissioned by the parliamentary committee on foreign interference and disinformation (Bayer et al. 2021; Wigell et al. 2021). Further, the recent report on the implementation of the sCoP (ERGA 2023) and a working paper by the parliamentary committee on AI (AIDA 2021). Aside from this, there is a considerable amount of parliamentary resolutions that do concern deepfakes.[31] However, they are not considered here. During the initial stages of the empirical analysis, parliamentary resolutions have been excluded because, as they cover a broad range of issues, they yield comparatively limited additional insights on deepfakes compared to other documents. The final list of documents which are analysed in the empirical component can be seen in Appendix 1.

[30] Though this kind of footage may overlap with the issue of epistemic harm from political deepfakes in some instances (*see* 3.4).

[31] These are: EP 2017, 2018a, 2018b, 2019a, 2019b, 2020a, 2020b, 2021a, 2021b, 2021c, 2022a, 2022b, 2023.

4.3 Qualitative Content Analysis

As stated earlier, the objective of the empirical component of this thesis is to arrive at an understanding of how epistemic harms from deepfakes are understood and addressed in the selected EU policy documents (*see* 1). Uncovering this requires interpreting these documents. To ensure that this interpretation is intersubjectively credible, it needs to be systematized. In this effort, I will conduct a qualitative content analysis (QCA) following Kuckartz and Rädiker (2023). While there are various competing approaches of QCA (*e.g.* Mayring 2014, 2021; Schreier 2012), Kuckartz and Rädiker are more focused on interpretation and the gradual development of a category system instead of quantifying results (*compare* Mayring 2014, 2021). In the following, their approach is described in detail. As they encourage researchers to tailor their methodological approach to a given research project (Kuckartz, Rädiker 2023; Stamann et al. 2016), this section will further detail and justify adaptations made in service of my research interest.

In a succinct definition, QCA can be described as "the systematic analysis of the meaning of material in need of interpretation by assigning it to the categories of a category system" (Stamann et al. 2016: para. 9). Systematized categories are a central analytical output of QCA (Kuckartz, Rädiker 2023: 34). How these categories are developed depending on methodological specifications: deductively, inductively or in a combined approach. Whereas deductive categories are based on previous knowledge and theories developed before engaging with empirical material (ibid.: 51), inductive categories are developed during close engagement with the material (ibid.: 21). Either kind of category is assigned to specific segments of the material in a process referred to as *coding*.

Coding describes how QCA engages with data. Researchers partition empirical material into segments – phrases, sentences, paragraphs, units of meaning, whole texts – and ascribe certain categories to those segments in an iterative process. Engagement with data is the basis of generating categories in inductive approaches. However, also in deductive approaches the application of categories can make modifications necessary or analytically useful (ibid.: 59). At the end of the coding process, QCA yields a category system, in which categories are usually organized either in a hierarchy that consists in layers of so-called child- and parent-categories, or a network of interrelating categories (ibid.: 40f). This category system then serves as the main point of reference for analysis, though coded segments are utilized to illustrate findings (ibid.: 211f).

In line with the pursuit of this thesis, a rich interpretative description, the specific form of QCA used here makes use of deductive and inductive codes, is oriented toward themes, and structures the coded material. The partition of data happens along paragraphs (ibid.: 46). This is sensible as it renders the amount of coded material more manageable than sentence-by-sentence coding would, while retaining sufficient specificity.

The previous sections of this thesis have built extensively on the existing scholarly literature on deepfakes. This literature already offers a broad overview of harms, their sources, as well as potentially remedies and alleys for governance. These are obviously relevant points of references when analysing policy documents that are relevant for deepfakes. This background is acknowledged as the basis for the construction of deductive categories (ibid.: 14; *see* 4.4). These categories are largely descriptive and can be described as themes (ibid.: 36). However, developing inductive categories further builds on and adds to the category system by generating insights from within the analysed material (ibid.: 36). This stands to enrichen the resulting description.

Lastly, the QCA done in this thesis draws on the process of structuring QCA (ibid.: 100-122). Apart from data selection (*see* 4.2), structuring QCA generally follows seven phases (Figure 4).

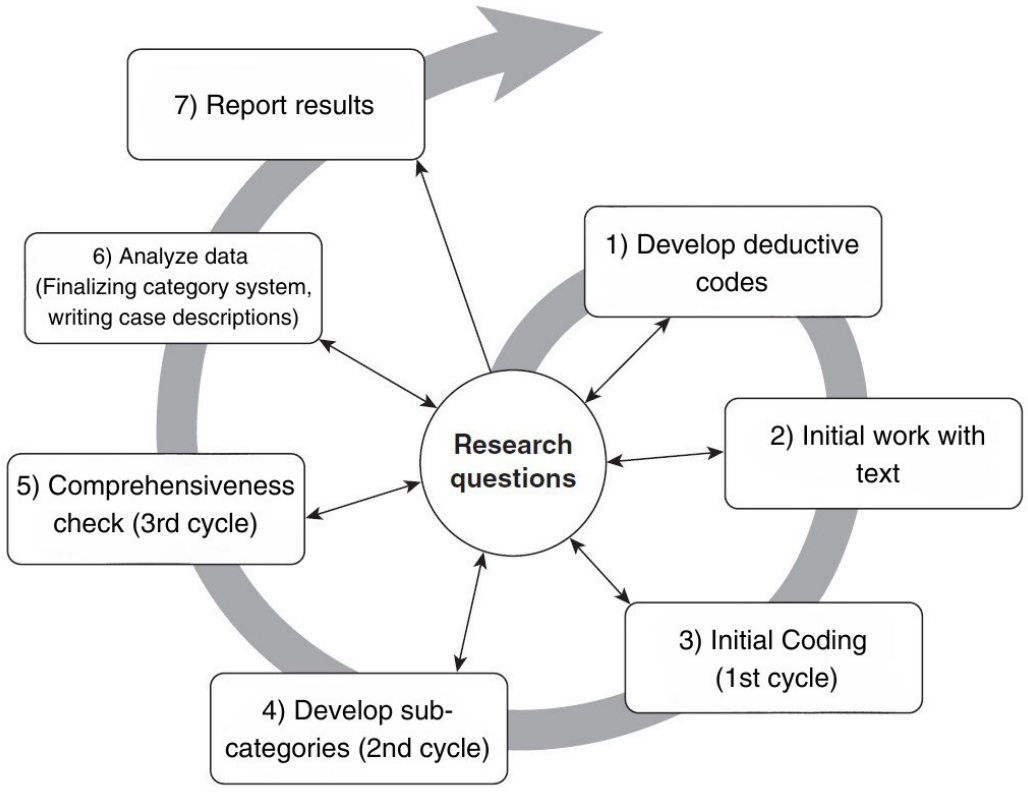


Figure 4. Phases of the structuring QCA conducted in this thesis. Adapted from Kuckartz, Rädiker 2023: 102.

Here, phase one consisted in the development of deductive codes (*see* 4.4) based in work in previous sections. Then, the selected policy documents (*see* 4.2) received an initial reading, beginning with the report to the EPRS (van Huijstee et al. 2021) and proceeding by date of publication.[32] This second phase focused on understanding the subjective meaning of the documents and the identification of relevant segments, central terms, and text progression. Where appropriate, research memos were made to serve as pointers for the following phases (Kuckartz, Rädiker 2023: 91-97). In the third phase, main thematic categories were applied using Atlas.ti. These codes were either based on the deductive categories or developed inductively (ibid.: 102f). In phase four, these main thematic codes were further differentiated in sub-categories which were largely developed inductively (ibid.: 106f). Phase five consisted in a third coding cycle that focused on comprehensive assignment of the developed codes across the entire material. In phase six categories the coded data was analysed. Categories were ordered into a hierarchical category system (ibid.: 37) and grouped categories compared and contrasted between documents through systematic case summaries (ibid.: 100, 112-114). In the final phase, results were reported in Chapter 5, including case summaries (Appendix 5). Throughout the process described above, measures were taken to ensure quality of research. These are detailed in Appendix 2.

[32] The report to the EPRS (van Huijstee et al. 2021) received the first reading as the document provides insight into EU deepfake policy prior to the more recent developments described above as well as being the most targeted document with respect to deepfake policy.

4.4 Developing Deductive Codes

The previous chapters provide a basis on which some categories for the initial phase of analysis. These deductive codes are adapted and supplemented by inductive codes during the analysis process. The full initial deductive category system can be found in Appendix 3. In this section, I want to offer a rational for this initial category system and how it develops out of the research questions and work in earlier chapters.

As has been established in Ch. 2, the overall phenomenon of deepfakes is heterogenous and consists of a variety of elements: producers, deepfake technology, deepfaked media content, subjects, platforms and recipients. Further, European deepfake policy is dispersed across a variety of policy frameworks and intertwined with policy discourses on disinformation and artificial intelligence. Consequently, the analysed documents, the issues they raise, and the measures they put forward address different aspects of the overall picture. As such, the first category of the initial deductive scheme will delineate which elements of the overall phenomenon are addressed (*Elements Addressed*). This category contains codes that different between different instances of these elements that might be covered in the documents (*e.g.* audio-, image-, and video-based deepfakes, or passive, sharing, and commenting recipients).

Based on the second research question – How is the epistemic impact of deepfakes conceived of and addressed in relevant EU policy documents? – two further categories are necessary: *Understanding of Epistemic Harm* and *Policy Measures*. Codes in the former category build on section Ch. 3 to enrol the different concepts of epistemic harms in the empirical analysis (*e.g.* jeopardizing evidence, cognitive resonance etc.). The category of policy measures builds on section 4.1 and the measures suggested by van Huijstee and colleagues (2021). It is divided in sub-categories for measures that hold actors potentially involved in the production of deepfakes accountable (*Accountability*), measures that introduce transparency e.g. into the process of reception of footage (*Transparency*), education on the respective issues addressed (*Education*), and *Content Moderation* of the content recipients encounter online. Codes in these sub-categories are based on measures floated in the broader discourse on deepfakes and disinformation and will be adapted in line with the analysed material. Lastly, one category will delineate the different kinds of documents that are part of the sample.

Overall, the system of deductive categories I devised has four layers: categories, sub-categories where applicable, corresponding codes, and sporadic sub-codes. Each category includes a residual category to aid in the development of inductive codes in the later stages of the research.

4.5 Concluding Remarks

In this chapter, it has been established that EU deepfake policy is scattered across a variety of frameworks and intertwined with other policy discourses, most significantly for the present research interests here are disinformation and artificial intelligence policy (*see* 4.1). On this basis, sixteen policy documents have been selected for analysis based on their relevance to the regulation of deepfakes, disinformation and artificial intelligence, as well as the conceptual understanding of epistemic harm in the EU (*see* 4.2). These documents will be analysed using Qualitative Content Analysis following Kuckartz and Rädiker (2023). This analysis will combine deductive codes (*see* Appendix 3) with an inductive approach. During coding, deductive codes will be adopted and complemented by inductive codes into the final coding system (*see* Ch. 5, Appendix 6).

5. Deepfakes in EU Policy Discourse

In this chapter, I present the results gathered from the empirical analysis (*see* 4.3, 4.4) the selected policy documents (*see* 4.2). I lay out how the epistemic harm of deepfakes is understood in the analysed policy documents (5.1) and through which policy measures they are addressed (5.2). I conclude this chapter with an answer to the second research sub-question – *How is the epistemic harm of deepfakes conceived of and addressed in relevant EU policy documents*? – in section 5.3. General results of the analysis, such as case descriptions and the final category system, are presented in Appendix 5 and 6. In this chapter, I focus on the most significant insights regarding the understanding of epistemic harm and the policy measures that address it.

5.1 Epistemic Harm in EU Deepfake Policy

As was to be expected from the fact that deepfake policy largely is dispersed across several frameworks from two primary policy areas, deepfakes in EU policy discourse are largely subsumed under the categories of disinformation and artificial intelligence more broadly. Deepfakes and epistemic harms that are specific to them, such as jeopardizing the role of recordings as evidence, are rarely regarded separately. Different kinds of deepfakes – audio, image, video – are also usually not differentiated. Even in cases where deepfakes are singled out how their impact is understood does not significantly differ from other kinds of disinformation. Deepfakes are, for the most part, treated as just another kind of disinformation, that is only different insofar as it is "more potent" (EC 2018d: 5) and "[makes] the fight against disinformation even harder" (Bayer et al. 2021: 99). They present an acceleration of the harms of disinformation, not as a unique phenomenon.[33]

> "[D]*eepfakes may produce a feeling of general distrust, contributing to an information environment where the veracity of information feels impossible to know. This lack of trust may have far-reaching consequences for democracies.*" (Bayer et al. 2021: 25)

> "*Images are effective communication means, because audiences can create and retrieve memories more easily when exposed to visuals. Therefore, audio-visual media have a strong appeal for their audience and may have a unique psychological power*." (van Huijstee et al. 2021: 22, *citing* Vaccari, Chadwick 2020)

As a result, the understanding of the epistemic harm from deepfakes in the analysed documents aligns with the understanding of epistemic harm from disinformation more broadly. This understanding, in turn, primarily takes the shape of harms to epistemic goods, deception, manipulation and the erosion of trust (*see* Appendix 7). The epistemic good the policy documents see under threat can be differentiated into two areas: democracy and fundamental rights.[34] Specifically, it is argued or – which is the more frequent case – implied that disinformation harms key democratic institutions such as elections, pluralistic public discourse, the media ecosystem, and next to a wide variety of fundamental rights.

---

[33] Nonetheless, as products of AI systems, there are some differences in how their impact is addressed (*see* 5.2).

[34] Fundamental rights are all rights included in the Charter of Fundamental Rights of the European Union (EU 2010).

"*From its inception, the EU approach to countering disinformation has been grounded in the protection of freedom of expression and other rights and freedoms guaranteed under the EU Charter of Fundamental Rights*" (EC 2021d: 1).

"[D]*isinformation harms democracies by hampering the ability of citizens to take informed decisions and participate in the democratic process. This is a major problem, given that technological developments have made possible the dissemination of disinformation at unprecedented scale and speed*." (Wigell et al. 2021: 8, *citing* EC 2018c)

The democratic institutions and fundamental rights the analysed documents refer to have epistemic dimension. Freedom to expression and information in public discourse and a pluralistic media system are widely seen as epistemically beneficial (*e.g.* Anderson 2006). However, instances in which documents specify *how* disinformation comes to negatively impact these epistemic goods are rare, though some examples exist (*see* below).[35] As a result, the understanding of how disinformation causes these epistemic harms that emerges from the documents is relatively shallow. Though policy-makers refer to disinformation as deceiving and manipulating recipients, these notions remain vague.

"*Disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public* […]." (EC 2018d: 3)

"[…] *providers should therefore pay particular attention on how their services are used to disseminate or amplify misleading or deceptive content, including disinformation*." (Regulation (EU) 2022/2065: para. 84)

"[…] *Relevant Signatories recognise the necessity* […] *to counter and limit impermissible manipulative behaviours and practices across their services, foreign information manipulation and interference, and hybrid threats to security*." (EC 2022a: 15)

How disinformation deceives, manipulates and erodes trust is usually not expanded upon in the documents produced by legislators. In fact, part of the policy agenda set out by the sCoP is that

[35] As can be seen in the examples, this critique is less appropriate for those documents which were produced by independent experts (Bayer et al. 2021; HLEF FNOD 2018; van Huijstee et al. 2021).

stakeholders come to a shared understanding of what constitutes "impermissible manipulative behaviours and practices" (ibid.).[36]

Policy documents on artificial intelligence contribute little to how the epistemic harm from deepfakes is understood, though they occasionally also stress risk from deception and manipulation from AI systems generally.

> "*In an AI context, freedom of the individual for instance requires mitigation of (in)direct illegitimate coercion, threats to mental autonomy and mental health, unjustified surveillance, deception and unfair manipulation*." (HLEG AI 2019: 10)

While dedicatedly epistemic harms only receive surprisingly sparse conceptual attention considering that disinformation is primarily an informational phenomenon, harms that are not primarily epistemic feature prominently.

> "*Along with political disinformation, the 'infodemic' has endangered people's health and livelihoods, including creating discrimination and hostility towards various minority groups, such as Asians, migrants, refugees, and elderly people*." (Bayer et al. 2021: 14)

> "*For example, deepfakes that enable a deceptive manipulation of reality, or are capable of inciting violence against people or causing violent social unrest*." (van Huijstee et al. 2021: 59).

Overall, policy-makers seem to have an instrumental understanding of disinformation, which may explain why they appear to be very alarmed with how disinformation may deceive and manipulate its recipients and the threat to democracy and fundamental rights without providing a strong conceptual account for how this is the case. That disinformation deceives, manipulates and undermines democracy is presumed to be a given, meaning it can be used as a tool for this purpose.

> "*Organisations and agencies of influence (be they undertakings, states, or non-governmental organisations with a stake in political and policy debates, including sources external to the EU) can use disinformation to manipulate policy and societal debates*." (EC 2018d)

This perspective is problematic for two reasons. First, it fails to be sensitive to the agency of recipients and their motives for why they engage with disinformation. Instead, this perspective

[36] The potential for deepfakes to jeopardize recordings as evidence is acknowledged only by van Huijstee and colleagues (2021: VIII), which is not entirely surprising given that deepfakes are rarely singled out.

centres on the producers of disinformation and the platforms on which it spreads.[37] Secondly, it frames the issue of disinformation in a manner that implies a relatively high degree of organization and collaboration among those producers. The data yield further evidence for the impression that policy-makers are primarily concerned with organized producers as the driving agents for disinformation. As mentioned above, European disinformation policy was, in large parts, a response to disinformation campaigns related to Russia's invasion of Ukraine in 2014 (Datzer, Lonardo 2022: 757; van Huijstee et al. 2021: 42). Correspondingly, at least part of the field characterized by the perception of disinformation as an organized, external phenomenon. This becomes apparent whenever disinformation is subsumed under larger concepts such as hybrid threats, foreign interference or influence operations.

> "*Furthermore, disinformation campaigns by third countries can be part of hybrid threats to internal security, including election processes, in particular in combination with cyberattacks*." (EC 2018d: 1f).

The role of domestic actors in such campaigns is acknowledged and considered to be growing (*e.g.* Bayer et al. 2021: 12). Nonetheless, disinformation is seen in close proximity to organized action. Policymakers also repeatedly stress the role of fake accounts in amplifying disinformation online (*e.g.* Regulation (EU) 2022/2065: para. 104). By contrast, the intersection and ambiguity between disinformation and misinformation is rarely addressed.

Overall, this suggests that, in the view of policy-makers, there is little nuance between recipients are either citizens who are deceived and have their fundamental rights violated by disinformation and actors who engage with disinformation in a manner that is self-interested, but not part of a larger organized scheme. There seems to be little room for nuance between being a citizen who needs to be protected and a Potemkin persona.[38] Recipients are either deceived or aware agents that intend to cause harm.

Nonetheless, in some instances, particularly in the analysed independent studies (Bayer et al. 2021; HLEG FNOD 2018; Wigell 2021), the motives of recipients receive more attention.

---

[37] Incidentally, the role of societal conditions in fostering the conditions for successful disinformation campaigns is addressed right before the provided quote. However, no conclusion regarding the role of recipients is drawn from it (EC 2018d).

[38] A Potemkin persona describes "inauthentic users who build a credible online presence across multiple platforms and mix their political messaging with banal posts about their supposed daily life" (Bayer et al. 2021: 12).

> "[…] *susceptibility to disinformation is predicted by many individual factors* […] *acquired and cultivated over a person's lifespan, starting in early childhood.* […] *Thus, all efforts to build resilience against disinformation should be planned and implemented in the long run across all societal areas (e.g. politics, economy, schools, the media system) – ideally, efforts should be preventive instead of curative*." (Bayer et al. 2021: 101)

> "*Initiatives to counter hybrid interference, therefore, need to include various means of supporting societal resilience, such as* […] *policies directed toward enhancing media literacy, social cohesion and welfare, particularly by integrating diasporas and minorities, who otherwise risk being used as proxies for hybrid interference efforts.*" (Wigell et al. 2021: 15f)

Correspondingly, these documents also raise sources of epistemic harm from disinformation that more closely resembles the non-deception-based kinds of harm discussed in Ch. 3.

> "*The more likely, bizarre, provocative, and entertaining a story is, the stronger emotional reactions (e.g. surprise, disgust) it generates in its recipients, and the more recipients share the story.* […] *Manipulated messages that support peoples' worldviews are more likely to be shared*." (Bayer et al. 2021: 99, 102; *citing* Calvillo et al. 2021; Faragó et al. 2020; Greifeneder et al. 2021, Vosoughi, et al. 2018)

However, these efforts are barely reflected in the discussion of epistemic harm in policy documents produced by legislators.[39] Overall, references to epistemic harms that are not based in deception, cognitive resonance and polarized fellowship, are exceedingly rare (*see* Appendix 7). This also reflects on the measures policy-makers address epistemic harms, which I will address in the next section.

Overall, the understanding of epistemic harm that emerges from the analysed policy document seem to be in line with a deception-centred understanding of disinformation and deepfakes reminiscent of how deepfakes have been discussed in the epistemic literature (*see* 3.1), though considerably less substantiated. At the same time, epistemic harm appears to be seen as – at least for the most part – as something that is done by organized and malicious producers to citizens who need to be protected.

---

[39] Polarization generally is discussed slightly more frequently (*e.g.* EC 2021d: 4; HLEG FNOD 2018). However, polarization and polarized fellowship were coded separately to differentiate between general mentions of polarization as opposed to instances where dynamics are described that resembled my concept of polarized fellowship more closely (*see* 3.3).

Before turning to the proposed policy measures, there is another aspect of the discourse on epistemic harms as embodied by the analysed documents that is conspicuous. Policymakers seem to constantly be torn between stressing the harms of disinformation and doing something against them and the fear to overly infringe on freedom of expression.

> "*On the one hand,* [political speech] *should enjoy enhanced privilege, in accordance with the principle of freedom of expression, and because it forms the necessary basis of democratic public discourse. But at the same time, if a politician takes advantage of false or polarizing content, that causes imminent harm to a country's cohesion, democratic process, and even individual human rights*." (Bayer et al. 2021: 120)

> "*Public reaction to censorship will backfire, as 'the establishment' or 'parties in power' could be (mis-)perceived as manipulating the news to their advantage*." (HLEG FNOD 2018: 30)

> "*The Signatories are mindful of the fundamental right to freedom of expression, freedom of information, and privacy, and of the delicate balance that must be struck between protecting fundamental rights and taking effective action to limit the spread and impact of otherwise lawful content.*" (EC 2022a: 1).

This aspect of the policy discourse has significant implications for the measures that are proposed by policymakers. To fully appreciate how the epistemic harms of deepfakes and disinformation are understood, it is therefore necessary to first specify how they are being addressed. I will therefore now turn to these measures before returning to the insights of this section in 5.3.

5.2 The EU Policy Response

The analysed documents cover a broad range of measures that either directly affect deepfakes and other kinds of disinformation online or implicate them. Before delving into the analysis, there are two characteristics of the data that bear pointing out. First, many of the policy measures in the data are presented in a way that is somewhat decontextualized. This makes it difficult to attribute which of the harms – both epistemic and non-epistemic – associated with deepfakes and disinformation are addressed by them. Second, due to the nature of the analysed documents – mostly consisting in communication by the European commission and independent studies – most measures are suggestions or intentions rather than enacted policy. As sCoP and DSA are effective pieces of (co-)regulation they present an exception. As the AIA, is in a progressed stage of the legislative process (AI Act 2023) it is likely that at least some of the measures in the analysed proposal are going to be effective policy. In this section, I therefore give a general overview of the proposed policy measures and then describe the measures presented in these three documents in more detail.

There is a substantial difference between measures proposed in those documents that mainly address disinformation and those that are concerned with AI. AI-focused documents are predominately concerned with issues of international competitiveness and fostering innovation. To this end, much space is occupied by soliciting investments, supporting AI research, training professionals and creating "innovation hubs" (EC 2018e: 7; EC 2018f; 2019b; 2021c). However, concerns about fundamental rights and, to a lesser extent, disinformation are addressed. Notably, the HLEG on AI has produced a set of guidelines for ethical AI that includes measures to enhance the transparency and explainability of AI systems and highlights the need to address fundamental rights concerns, including deception and manipulation, in risk assessments (*e.g.* HLEG AI 2019: 18).

> "*Like many technologies, AI systems can equally enable and hamper fundamental rights*. […] *AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning,* […]." (ibid.: 15f)

Transparency obligation and risk assessments may help to address the epistemic problems with deepfakes to some extent, though it is dubious that providers of DFAs have so far been are

unaware of the potential harmful uses and are willing to mitigate them.[40] Insofar as risk assessments remain voluntary or remain without consequences, it is implausible that this will have much impact. In terms of AI policy discourse, the proposed AIA goes considerably further than other documents in addressing the issues associated with deepfakes, as I discuss below. Importantly, documents focusing on AI focus on measures targeting the providers of AI systems whereas documents focused on disinformation tend to focus on platforms (EC 2018d: 2).

Documents concerned with disinformation propose a much broader range of measures compared to their AI-counterparts. A substantial part of the policies proposed in these documents does not focus on measures that mitigate epistemic harm per se, but rather are in support of this objective indirectly. This may be the case either because they provide policy-makers and other actors (researchers, fact-checkers, journalists) with *situational awareness* regarding disinformation events that have occurred or are occurring – e.g. through mandating platforms to gather data and make it available (EC 2018d: 10; EC 2022a; Regulation (EU) 2022/2065 Art. 40) – or to make sure that relevant entities in digital information environments, particularly platforms, comply with applicable measurements, e.g. through independent audits, producing compliance reports or risk and impact assessments (Regulation (EU) 2022/2065: Art. 42). Additionally, there are various proposals to facilitate stakeholder collaboration and to strengthen the capacity of relevant institutions (*e.g.* EC 2018d: 8f; EC 2019b: 9; 2021c: 5). However, here I will consider only those measures that are directly relevant to epistemic harms. To address these, there are two primary categories of measures: *content moderation* and *prevention*.

Content moderation

Beginning with content moderation, as mentioned in the previous section, EU policymakers are very concerned with preserving freedom of expression while tackling disinformation (*see* 5.1). Generally, fact-checking and labelling false information and engaging in the verification of sources are preferred compared to deleting content outright.

> "*A dense network of strong and independent fact-checkers is an essential requirement for a healthy digital ecosystem*." (EC 2018d: 9)

[40] Deepfakes emerged from within an online community with the explicit purpose of creating non-consensual deepfake pornography. Many DFAs are currently hosted in a way that purposefully avoids moderation (Winter, Salter 2020).

Nonetheless, policymakers also propose several measures in between merely labelling something as false but otherwise leaving it unaffected and deleting content and banning users. Platforms are also asked to limit the reach of and demonetize content (EC 2021d; HLEG FNOD 2018; van Huijstee et al. 2021).

> "*Policy-makers could consider obliging platforms to label detected deepfakes as such and/or to take down unlabelled deepfakes once the platform is notified by a victim or trusted flaggers following established procedures*" (van Huijstee et al. 2021: 62)

For all of these content moderation decisions, platforms are required to make decisions in an ideologically neutral way that accords to freedom of expression (Bayer et al. 2021: 14; HLEF FNOD 2018: 14) and to provide users that are affected by a content moderation decision with a way to appeal (*e.g.* Regulation (EU) 2022/2065: Art. 17, 20, 21; van Huijstee et al. 2021: 63).

> "[It] *is recommended that the DSA obliges online platforms* […] *to pay full respect to fundamental rights, in particular the right to freedom of expression, freedom of information, equality and non-discrimination, privacy, and dignity – especially in their content moderation decisions.*" (Bayer et al. 2021: 14).

Content moderation clearly addresses those epistemic harms of false information that occur due to recipients being deceived, either because they do not encounter false information in the first place, or because they are alerted to it being false.[41] Nonetheless, as content moderation is a reaction to manifest pieces of content, it is distinct from prevention.

Preventation

Preventative measures aim at increasing the resilience recipients of digital information environments towards disinformation before given narratives take hold.[42] By far the most prominent measures in this area are tailored to raising awareness the of recipients regarding issues of disinformation, and to increase their media literacy through dedicated educational

[41] In the case of deleting content, however, there may be some degree of epistemic harm caused by the content moderators as this keeps recipients from obtaining information regarding the source of that post. However, I will not explore this further here.

[42] During the analysis, measures that aim to disincentivize posting disinformation and deepfakes, e.g. by requiring users to identify themselves, were also coded. However, suggestions to this effect were marginal (*see* Appendix 8).

training (*see* Appendix 8)(*e.g.* EC 2018d; EC 2022a: 17). Civil society organizations play a pronounced role in this (HLEG FNOD 2018: 11, 17).

> "[R]*aising awareness and ensuring that people can differentiate between information and disinformation is of utmost importance*" (Wigell et al. 2021: 8)

Funding for these projects is provided by the EU and platforms (*e.g.* EC 2022a: 19). Aside from educating recipients, policymakers focus on supporting 'authoritative information' and journalism in a variety of ways (*e.g.* EC 2018d: 14; Bayer et al. 2021: 128) and increasing the access to accurate information, e.g. by asking platforms to algorithmically amplify its distribution (EC 2021d: 18f, 2022a: 20; HLEG FNOD 2018: 14).

> "*Funding and other support of European and national journalism and media pluralism by the European Commission and Member States remains crucial moving forward.*" (van Huijstee et al. 2021: 66)

> "*The most frequently deployed types of intervention tend to challenge disinformation by producing initiatives that help create resilience among citizens and empower the various actors impacted.* […] *Examples include initiatives to influence 'findability', privileging credible content in ranking algorithms,* […] *an enabling environment for news media and professional journalism, as well as investments in media and information literacy* […]." (HLEG FNOD 2018: 14)

Other measures

Deleting, demonetizing or restricting the visibility of content and accounts notwithstanding, policy documents devote little attention to measures that hold the producers of disinformation accountable. This is likely the case because relevant bodies of law, e.g. criminal law and copyright, are largely in the domain of member states (EC 2005). International sanctions are the exception in this regard.

> "[S]*tates may actively use deepfakes in disinformation campaigns.* […] *If diplomacy does not yield sufficient results, a policy option is to impose well-considered economic sanctions.*" (van Huijstee et al. 2021: 61f)

This is somewhat surprising, given that producers are clearly seen as causing considerable harm. More stringent accountability for producers would likely serve as a deterrent. However, it may also invite critique on the grounds of freedom of expression. If policy-makers are already

hesitant to delete false information, they will be even more hesitant to legally prosecute producers.

Again, the proposed measures above are not necessarily binding. They are, however, partially embedded in the AIA, DSA and sCoP, which also introduce considerable means to hold platforms and potentially providers of DFA accountable in the form of fines should they fail to comply (*see* Appendix 5). I describe these frameworks in detail below, beginning with the AIA.

Proposed Artificial Intelligence Act (AIA)

The AIA, as a legal framework for the regulation of AI systems, is not primarily concerned with disinformation. Further, though nonetheless highly relevant for reason which will become clear in the following, it also only applies to deepfakes to a limited extent. This is because only a few systems are outright prohibited under the AIA and more stringent obligation only apply to systems that are categorized as high-risk (EC 2021a). Despite the considerable role concerns regarding fundamental rights play in how the threat of disinformation – and by inclusion deepfakes – is perceived (*see* 5.1), the AIA does not currently classify DFAs as either prohibited or high-risk (EC 2021a: Art. 52).

Why this is the case is not entirely clear as both the criteria for prohibited as well as high-risk systems seem to apply to deepfakes according to how their epistemic harms are understood in the policy discourse (*see* 5.1).

> "*The following artificial intelligence practices shall be prohibited: (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm*." (EC 2021a: Art. 5(1a))

While harms in the form of *cognitive resonance* arguably present a kind of subliminal technique (*see* Hughes et al. 2021) and may arguably cause harm to the subjects of deepfakes, apparently this does not meet the necessary threshold to result in prohibiting DFAs. Not classifying DFAs as high risk similarly seems to fail due to meeting a threshold of harm.

AI systems are classified as high-risk either because they are deployed in a sensitive area[43] or because they negatively impact fundamental rights in a significant way (ibid.: Art. 6, 7).

> "[High-risk] *AI systems pose a risk of harm to the health and safety, or a risk of adverse impact on fundamental rights, that is* […] *equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III*." (ibid.: Art. 7(1b)).

Again, this suggests that policy-makers do not consider deepfakes to pose a sufficient risk to fundamental rights that would warrant the additional obligations of high-risk systems. As classified now, the AIA only introduces labelling obligations for deepfakes (ibid.: Art. 52(3)).

However, this classification might not prevail. The commission may, at any point, reclassify DFAs as high-risk AI systems (EC 2021a: 13). Numerous other documents already argue that their classification should be revised (*e.g.* EC 2021d: 12; van Huijstee et al. 2021: 59). As the AIA is still not adopted in its current form, DFAs may well be reclassified. That it will be is plausible under circumstances in which deepfakes become a significant aspect of the overall phenomenon of disinformation. It is therefore sensible to consider measures that would mitigate epistemic harm if DFAs were classified as high-risk systems in the present analysis.[44]

Providers of high-risk AI systems need to conduct risk and conformity assessments (EC 2021a: Art. 9, 19) either themselves or through an independent auditor (ibid.: Art. 33, 43) and provide (technical) records of their systems for the purpose of monitoring conformity (ibid.: Art. 11, 12). Additionally, they need to put risk management measures in place that address the risks that have been identified for a system through this process.

> "*In identifying the most appropriate risk management measures, the following shall be ensured: (a) elimination or reduction of risks as far as possible through adequate design and development; (b) where appropriate, implementation of adequate mitigation and control measures in relation to risks that cannot be eliminated*." (ibid.: Art. 9(4a, b))

> "*The risk management measures […] shall be such that any residual risk […] of the high-risk AI systems is judged acceptable, provided that the high-risk AI system is used in*

[43] These are systems in the area of biometric identification and categorization of natural persons, critical infrastructures, education, employment, access to essential services and benefits, law enforcement, migration and border control, administration of justice and democratic processes (EC 2021e: 5)

[44] Alternatively, providers of DFAs may choose to comply with the additional obligations voluntarily by entering into a code of conduct (ibid.: 15).

> *accordance with its intended purpose or under conditions of reasonably foreseeable misuse*." (ibid.)

In the analysed policy documents, banning DFAs or the machine-learning architectures that enable them is rarely considered (*e.g.* van Huijstee et al. 2021: 61). However, should DFAs be reclassified as high-risk systems, it might be the case that they can no longer be offered on the European market without heavily restricting the degree of creative freedom they afford to producers (*see* 2.4). Nonetheless, aside from being classified as high-risk, this also depends on how this re-classification is justified and, accordingly, what is considered an unacceptable risk.

Digital Services Act (DSA)

Whereas the AIA focuses on the providers of AI systems, the DSA primarily regulates "intermediary services" (Regulation (EU) 2022/2065: Art. 3(g)) in the European shared market. In doing so, the DSA covers a variety of issues and actors, includes providers of technological services, market spaces and search engines. However, here I only discuss measures that are relevant to content posted in digital information environments (*see* 2.1), which the DSA describes as platforms. While the DSA also addresses disinformation, throughout the legally binding body of the regulation measures apply for the most part only to '*illegal content'*.

> "[This regulation aims to ensure] *a safe, predictable and trusted online environment, addressing the dissemination of illegal content online and the societal risks that the dissemination of disinformation or other content may generate, and within which fundamental rights enshrined in the Charter are effectively protected and innovation is facilitated*." (Regulation (EU) 2022/2065: para. 9)

Illegal content is defined as follows:

> "[A]*ny information, which, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State, irrespective of the precise subject matter or nature of that law*." (ibid.: Art. 2(h))

Correspondingly, insofar as pieces of disinformation are not also considered illegal – which is not per se the case in most member states – large parts of the DSA do not apply (Strowel, de Meyere 2023: 74).

However, there is a plausible argument that deepfakes ought to be considered illegal – e.g. because they are found to infringe on privacy rights, are violent or pornographic. If deepfakes were generally considered illegal, the DSA requires platforms to remove them upon being notified by political authorities or users (Regulation (EU) 2022/2065: Art. 9, 16). However, it is unlikely that all kinds of deepfakes will be criminalized. Policy-makers have so far shown hesitancy to impair freedom of expression and introducing liabilities for the producers of disinformation. Criminalizing all deepfakes may not be feasible, e.g. on the grounds of protecting their use as a means of satire (van der Sloot, Wagensfeld 2022).

Nonetheless, the DSA also introduces two sets of measures that ultimately may prove impactful, even if the production of deepfakes remains legal. First, similar to the AIA, the DSA requires platforms to conduct risk assessments (Regulation (EU) 2022/2065: Art. 34) and implement risk mitigation measures (ibid.: Art. 35) through which they are particularly obliged to account for issues related to disinformation.

> "[Platforms] *can be used in a way that strongly influences safety online, the shaping of public opinion and* […] *should therefore assess the systemic risks stemming from* […] *potential misuses by the recipients of the service, and should take appropriate mitigating measures in observance of fundamental rights*." (ibid.: para. 79)

> "*A third category of risks concerns the actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes, as well as public security.* […] *A fourth category of risks stems from similar concerns relating to the protection of public health, minors and serious negative consequences to a person's physical and mental well-being, or on gender-based violence. Such risks may also stem from coordinated disinformation campaigns related to public health* […]." (ibid.: para. 82f)

Mitigation measures include design changes, content moderation by '*trusted flaggers*', content removal and awareness raising (ibid.: Art. 35). While deepfakes are not mentioned explicitly, platforms are asked to implement labelling practices that applies to them.

> "[Platforms must ensure] *that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful is distinguishable through prominent markings when presented on their online interfaces* […]." (ibid.: Art 35(1k)).

Lastly, the DSA set a legal basis for inviting platforms and other stakeholders into a co-regulatory code of conduct which set risk mitigation commitments for participants (ibid.: Art. 45). Failure to comply with such a code results in an invite by the Commission "to take the necessary action" (ibid.: Art. 45(4)). While this seems to have little force behind it, it may hint toward ambitions for further regulation should the DSA not yield its desired effect. The sCoP present an obvious candidate for such a code of conduct. The DSA might therefore strengthen its position as a co-regulatory framework.

Strengthened Code of Practice on Disinformation (sCoP)

Due to limitations of the AIA and DSA – though both frameworks are by no means irrelevant because of them – the sCoP currently presents the most relevant (co-)regulatory framework for disinformation, including deepfakes. The sCoP (EC 2022a) presents a voluntary co-regulatory framework for various actors in the digital information ecosystem – platforms, advertisers, technology providers, researchers, fact-checking and journalistic organizations, and civil society actors, as well as EU representatives – to engage around issues of disinformation. Signatories of the code commit themselves to several measures and agree on specific indicators through which their compliance with them is assessed. Additionally, the sCoP introduces stringent reporting and transparency obligations (ibid.). Signatories convene regularly in a permanent task force to report on their progress and exchange information as well as best practices (ibid.: 37f).[45]

Primarily, the sCoP focusses on addressing '*impermissible manipulative behaviours'* (ibid.: 15f), which includes tactics, techniques and procedures (TTPs) related to disinformation and misinformation, as well as other practices related to the artificial amplification of content and deepfakes (ibid.). However, the sCoP does not thoroughly specify what these practices consist in, beyond providing some examples (*compare* 5.1). Instead, it is part of the sCoP is facilitating that signatories come to a shared agreement on what these behaviours consist in.

> "*Relevant Signatories will convene via the Permanent Task-force to agree upon and publish a list and terminology of TTPs employed by malicious actors, which should be updated on*

[45] Though in the first assessment report on the sCoP, the compliance with these reporting commitments was criticized (ERGA 2023).

an annual basis, and consist in a shared understanding of manipulative behaviours and practices not permitted on their service to-date." (ibid.: 16f).

The sCoP suggests this common understanding should be built on existing frameworks such as AMITT (ibid.: 15f), which is open source and compiled by a non-profit organization.[46] Notably, the commission does not appear to want to specify what counts as impermissible and defer the responsibility to delineate which behaviours exactly are subject to content moderation to a multi-stakeholder process instead.[47]

As a response to content that has been found to fall under the to-be-determined definition of impermissible manipulative behaviour, signatories commit themselves to a variety of content moderation measures. Disinformation content is demonetized by withholding ad revenue from actors spreading it and signatories commit to prevent disinformation from being spread in the form of advertisements (ibid.: 5-8).

"*The Signatories recognise their collective and individual accountabilities to work together to defund Disinformation in advertising* […]." (ibid.: 4)

Further, signatories seek to limit the spread of disinformation through "prohibiting, downranking, or not recommending harmful false or misleading information" (ibid.: 20). Recipients are also supposed to be presented with a wide range of fact-checked and verified information regarding the content they encounter (ibid.: 21-24).

"*Relevant Signatories commit to* […] *equip users to identify Disinformation. In particular* […] *Relevant Signatories commit to facilitate* […] *user access to tools for assessing the factual accuracy of sources through fact-checks* […], *as well as warning labels from other authoritative sources*." (ibid.: 21f).

Through these measures, users of platforms that signed on to the sCoP are supposed to be presented with a safer information environment (ibid.: 20). The sCoP also devotes one commitment specifically to deepfakes.

---

[46] Acronym stand for Adversarial Misinformation and Influence Tactics and Techniques (GitHub n.d.). The framework is an open source project maintained by a foundation (ibid.; Disarm Foundation n.d.)

[47] Outsourcing the responsibility for content moderation decisions appears to also occur on the side of platforms themselves: "[Content moderation] *requires evaluation of content, which is always debatable. To alleviate the risks and accusations of bias and subjectivity, most platforms have outsourced fact-checking*." (Bayer et al. 2021: 82)

> "*Relevant Signatories that develop or operate AI systems and that disseminate AI-generated and manipulated content through their services (e.g. deepfakes) commit to take into consideration the transparency obligations and the list of manipulative practices prohibited under the proposal for Artificial Intelligence Act.*" (ibid.: 17)

The way deepfakes are being referenced here is curious, given that the AIA itself does not consider deepfakes as prohibited or high-risk systems (*see* above). While it has to be seen what to make of this reference in the sCoP, this might indicate that the EU is either moving toward a more stringent position on DFAs since the proposal of the AIA in 2021, or the commission is seeking to tackle problematic deepfaked media content without necessarily regulating the systems than enable them.

As mentioned above, the group of signatories spans various actors in the digital information environment, including fact-checking organizations, researchers and civil society actors (EC 2022b). The sCoP requires commercial actors to collaborate with them in various ways. Most prominently are obligations to share data gathered around impermissible manipulative behaviours and provide funding for researchers. Platforms must provide access to such data through APIs (EC 2022a: 27ff).[48]

Recipients are addressed in the sCoP through some dedicated measures. First, it seeks to give recipients more agency by enabling them to choose the content that they are being recommended (ibid.: 21), report content themselves (ibid.: 25) and providing information about their practices in a transparency centre (ibid.: 35f). Users are further provided with means to appeal content moderation decisions (ibid.: 25).

> "*Relevant Signatories commit to inform users whose content or accounts has* [sic] *been subject to enforcement actions (content/accounts labelled, demoted or otherwise enforced on)* […] *and provide them with the possibility to appeal against the enforcement action at issue and to handle complaints in a timely, diligent, transparent, and objective manner* […]. (ibid.: 25)

Notably, the first of the above quotes highlights again that, similar to content moderation decisions, appeals are to be settled in a way that is neutral. Lastly, the signatories of the sCoP further agree to engage in media literacy campaigns (ibid.: 19).

[48] Application Programming Interfaces. Allows two separate applications to communicate with each other (IBM n.d.), e.g. for automatically downloading updated data etc.

> "*Relevant Signatories will develop, promote and/or support or continue to run activities to improve media literacy and critical thinking such as campaigns to raise awareness about Disinformation* […] *among the general public across the European Union, also considering the involvement of vulnerable communities*." (ibid.: 19).

While vulnerable communities are mentioned, it is not specified who is addressed by this precisely. Having now provided a through picture of epistemic harms are discussed and addressed through policy measures in the analysed documents, I will turn to answering the second research sub-question.

5.3 Concluding Remarks

To summarize the results presented above, I now turn to the answer of the second research sub-question: *How is the epistemic harm of deepfakes understood and addressed in relevant EU policy documents*?

Generally, the understanding of epistemic harm in the analysed policy documents is conceived as a threat to epistemic goods in the form of fundamental rights, particularly freedom of expression and information, and democratic institutions such as public discourse. Harms to these aspects presents epistemic harms in the terms of this thesis insofar as they are conducive of epistemic success (*see* 2.1). While the analysed documents do not go into depth how these epistemic good are harmed through disinformation and deepfakes, the discussion of both phenomena suggests that this is the case because they deceive and erode trust, which is in itself – as I have argued in Chapter 3 – a function of the deceptive capacity of deepfakes.

From how disinformation and recipients are discussed in the analysed policy documents, it further emerges that epistemic harm is primarily understood as something that is done by producers who act intentionally and are generally perceived as contributing to an organized effort (*see* 5.1). Little attention is paid to why individual recipients, who are not enrolled in a larger campaign, would engage with or incur harm from disinformation.

Policy-makers seek to mitigate the problems caused by disinformation primarily through obliging platforms moderate content and through preventative measures. While content moderation practices include the removal of content or accounts under some circumstances, overall policy-makers tend to be hesitant to interfere with the kinds of content users can post and view online. Instead, measures that provide information that contextualizes supposedly false content are preferred, e.g. labelling and fact-checking. Preventative measures centre on providing better access to 'authoritative information' from credible sources, e.g. through supporting media organizations and algorithmically amplifying their content. Users are further given more agency giving them means to influence the information they are shown as well as appeal content moderation decisions they are subjected to (*see* 5.2).

6. Comparing the Empirical Results with the Theoretical Perspective

In this chapter, I will compare the theoretical perspective on epistemic harms from deepfakes (*see* Ch. 3) with the results obtained from the empirical analysis (*see* Ch. 5) to answer the final research sub-question.

To reiterate, five kinds of epistemic harm were identified in this thesis: deception, jeopardizing recordings as evidence, erosion of trust, cognitive resonance and polarized fellowship. I have shown that the first three of these harms are based in deception, whereas the latter two are not dependent on recipients being deceived (*see* Ch. 3). The results of the empirical analysis have shown that EU policy-makers perceive the epistemic harm of disinformation, and by inclusion deepfakes, to primarily consist in harms to fundamental rights and democratic institutions that can be seen as epistemic goods (*see* 2.1) though deceptive and manipulative practices. Non-deception-based harms are only considered at the margins (*see* 5.1). In the analysed documents, policy-makers propose (and adopt) a broad range of measures to mitigate the epistemic harms they see (*see* 5.2). *Do these policies also address the epistemic harms that were identified in the theoretical component of this thesis*?

On a superficial level, all five kinds of epistemic harm are, to some extent, addressed. However, there is a significant difference in the extent to which this the chase and how salient the different kinds of epistemic harm are in the overall policy discourse. How this is the case becomes apparent when considering how recipients are seen and which aspects that contribute to the epistemic harm from deepfakes are not meaningfully covered by the proposed measures, though they may also be affected to some extent.

Content moderation and preventative measures (*see* 5.2) affect which content recipients encounter content online, how this encounter is framed and how citizen can exercise their agency in it. Supporting actors that provide accurate information in digital information environments and amplifying the spread of such content change the overall makeup of the information encountered online. Likewise, deleting content or decreasing its spread do the same. As such, if effective, the likelihood of encountering accurate information when dwelling in digital information environment will increase whereas the chance to encounter false information will decrease. Labelling and fact-checking, as well as adjacent measures such as marking trustworthy sources, change the context in which information is encountered. If successful, this gives recipients a reliable cue for the source to be speaking with competence (*see* 3.2). I call this set of measures the '*minimal interference approach*'.

The success of the measures above, however, relies on a certain kind of recipient; one that engages with public discourse in digital information environments trying, though maybe sometimes failing, to obtain accurate information from good sources and who has a sufficient degree of trust in the institutions that implement these measures and provide the information recipients are asked to see as 'authoritative'. If such a recipient is navigating public discourse online and encounters footage that is labelled as false, or that is presented alongside a notification pointing to a fact-checking article, they will likely be more vigilant about believing the information conveyed by that content (Walter et al. 2020). Similarly, such a recipient is unlikely to specifically seek out information from non-authoritative sources. If implemented successfully, all of the measures mentioned above are likely to prevent a substantial degree of the epistemic harm from deepfakes for this recipient.

Unfortunately, there are two interrelated problems. First, recontextualization only prevents deception-based epistemic harms. Second, even if one concedes that the above accurately describes most recipients in digital information environments, it certainly does not describe all.

As established above (*see* 3.3), recipients may still experience epistemic harm even if they are not deceived regarding the veracity of the content they encounter (*see* 3.3). A deepfake may still resonate with the pre-existing beliefs, attitudes and interest of a recipient despite being labelled as a deepfake (*see* Hughes et al. 2021). In this case, epistemic harm in the form of deception may be mitigated, while cognitive resonance is not. Similarly, posting a deepfake to *burn bridges* and serve as an anchor point for a problematic epistemic community likely works just as well whether its content is fact-checked or not, as what matters here is the perception of benevolence and commitment, not factuality (*see* 3.3).

The possibility of the formation of polarized fellowships even when the measures of the minimal interference approach are successful leads into the second issue with the above measures. Epistemic vigilance suggests that the minimal interference approach relies too heavily on recipients trusting the actors responsible for content moderation and producing 'authoritative information'. This trust should not be presumed. There is reason to doubt that either state institutions, media organizations, many civil society actors or platforms enjoy much trust among recipients that most heavily engage with disinformation narratives (Imhoff, Lamberty 2018; Klebba, Winter 2021; Pierre 2020, *all cited in* Bayer et al. 2021: 100f). To the contrary, lack of trust in these institutions compared to source that oppose them might be what renders recipients susceptible in the first place (Buchanan, Benson 2019; Faragó et al. 2020, *both cited in* Bayer et al. 2021: 100). Insofar as this lack of trust is due to the perception that

the institutions moderating content and providing 'authoritative information' belong to an opposing coalition, have previously violated a commitment to the recipient or pursue interests that go against their own, the measures of the minimal approach are unlikely to overcome the mechanisms of epistemic vigilance of those vulnerable recipients. In other words, such recipients are unlikely to believe the contextualization and authoritative information that is made available to them.

Despite the minimal interference approach appearing as the preferred policy option, measures that go beyond it are also part of response to disinformation (*see* 5.2). Non-deception-based harms would be addressed more appropriately by deleting deepfaked content or limiting its range. If recipients do not encounter deepfaked content, they will not form associations on its basis or form problematic epistemic communities around it. This would, however, present a greater infringement on the freedom of expression of recipients, putting policy-makers into a difficult position.

There may, however, be a resolution. Instead of 'censoring' content that is epistemically harmful – which some recipients are likely not happy about (Oremus 2022) – or resting on the assumption that recipients ought to trust the sources of information that policy-makers designate as 'authoritative' – sound as that designation may be – policy-makers should actively engage in building rapport with vulnerable recipients. (Re-)building trust with recipients who are susceptible to disinformation narratives will need to be predicated on understanding and targeting the normative dimensions of what it means to *be regarded as* an authoritative source of information (*see* 3.2). Likely, this will also require policy-makers to nuance their understanding of what it means to be a producer and a recipient of disinformation and to be harmed by it (*see* 5.1). Only then can minimal interference work. In addition, media literacy campaigns may incorporate *normative* literacy that raises critical awareness regarding whether the purveyors of disinformation online truly have the best interest of their recipients at heart. Nonetheless, such efforts are faced with the same problems of being seen as a trustworthy source of information. One needs to be trusted to be believed when telling someone not to trust a third party.

Against this background, one more aspect of the analysed policy documents needs to be acknowledged. As argued above, the perceived benevolence of a source is part of how the credibility of a piece of information is assessed by a recipient (see 3.2). Building trust and trustworthiness are referenced as a policy goal several times across multiple documents (Bayer et al. 2021: 101; HLEG FNOD 2018: 38; Wigell et al. 2021: 8). Nonetheless, strikingly few of

the proposed or adopted measures tangibly contribute to recipients perceiving the institutions that enact the measures that are supposed to mitigate the problem of disinformation as benevolent. This is despite several of the independent studies pointing out that e.g. socio-economic conditions are a factor for the appeal of disinformation (Bayer et al. 2021: 99-101; Wigell et al. 2021: 8f).

If the interpretation of deepfakes is based on the pre-existing beliefs and references, as well as the interests and perceived allegiances of their recipients and deepfakes, as a result of this process, cause epistemic harms including, but also going beyond, deception (*see* Ch. 3) then it is dubious that the measures proposed or enacted in the current state of EU disinformation policy (*see* Ch. 5) will succeed in the comprehensive mitigation of epistemic harm from deepfakes.

This plausibly not only applies to deepfakes, but other kinds of disinformation as well, though deepfakes may be especially suitable to be exploited for these specific kinds of epistemic harm (*see* 3.3). It may take measures outside of the traditional domain of artificial intelligence or disinformation policy to tackle the circumstances that make pieces of disinformation appealing to some recipients because they appear to them as benevolent whereas established sources of information are seen as not trustworthy. However, if EU policy-makers pursues to comprehensively tackle the epistemic harms of deepfakes and disinformation, these circumstances need to be understood and addressed. In absence of this, EU policy on deepfakes risks to only tackle some of the symptoms the technology may aggravate, and to miss the underlying causes.

The answer to the third and final research sub-question – *are the identified kinds of epistemic harm caused by deepfakes addressed by EU policy*? – therefore must be: partially. Deception, jeopardizing evidence, and erosion of trust – insofar as it results from deception – are addressed, however cognitive resonance and polarized fellowship are certainly impacted to some extent, though it is unlikely that they can be effectively remedied by the current set of policy responses.

7. Conclusion

In this thesis I investigated the epistemic harms[49] that deepfakes may produce, as well as how these harms are understood and addressed in relevant EU policy documents. Deepfakes, as pieces of synthetic audiovisual media created through leveraging machine- or deep-learning techniques (*see* 2.2), imitate the realism of authentic recordings and in doing so, provide the producers of deepfakes with a great degree of creative freedom in how they want to present their subjects (*see* 2.4). Recordings are deeply embedded in everyday epistemic practices and recordings enjoy a privileged status as evidence as they reliably represent past material states. Deepfakes are expected to reach a degree of realism that makes them hard to distinguish or even indistinguishable from recordings. Recipients will then no longer be able to tell whether a piece of footage is a recording or a deepfake on the merit of its realism. Commonly, this seen as the root of epistemic harms caused by deepfakes (*see* 2.3).

Building on this understanding, the epistemic literature on deepfakes proposes three distinct kinds of epistemic harm as a result of deepfakes: deception, jeopardizing the role of recordings as evidence and erosion of trust. I argue all of these harms depend on the (perception of a) capacity of deepfake to deceive recipients into holding them as truthful (*see* 3.1). However, the frameworks of Peircean semiotics and epistemic vigilance suggest that the reception of footage is more complex than judging the believability of footage merely on the basis of realism. Instead, it depends on the socio-culturally and historically contingent experiences of a recipient, the context in which footage is encountered in, pre-existing beliefs, mechanisms of inference, the perceived competence and benevolence of the source that presents the footage and a recipient's history with that source (*see* 3.2). This process of interpreting and evaluating the believability of footage, on the one side, means that recipients may not believe realistic footage, on the other side, however, it also introduces vulnerabilities that result in the possibility of a deepfake causing epistemic harm despite not deceiving its recipient. These harms consist in evoking associations toward the subject of a deepfake that bias the recipient against it, as well as providing faux-justifications for pre-existing attitudes and desires – which I call *cognitive resonance* – and in allowing the producers of deepfakes to publish an inflammatory deepfake in an effort to signal allegiance toward an audience that is in opposition to the deepfake's subject and alienate an audience that is in opposition to those who engage in or condone such behaviour – which I call *polarized fellowship*. This leaves me with five kinds of epistemic harm caused

[49] I define epistemic harm as follows: *Epistemic harm is caused if an agent obstructs, without legitimizing reason, the epistemic success of another in the area of politically relevant information* (*see* 2.1).

by deepfakes: *deception, jeopardizing the role of evidence, erosion of trust, cognitive resonance* and *polarized fellowship*.

After identifying these kinds of harm, I conducted a qualitative content analysis of policy documents that were identified to be relevant to deepfakes within EU deepfake policy discourse (*see* Ch. 4) to uncover how EU policy-makers understand the epistemic harms caused by deepfakes and through which policies they are addressed. In this discourse, deepfakes are primarily seen as a part of the overall phenomenon of disinformation. The conducted analysis shows that their epistemic harm is primarily understood from a perspective of harms to epistemic goods in the form of freedom of expression, freedom of information, as well as key democratic institutions. Again, these harms seem to be rooted in the capacity of disinformation to deception its recipients, as well as manipulate their behaviour. Unfortunately, the uncovered notions of epistemic harm lack conceptual specificity. Epistemic harms that clearly do not depend on a capacity for deepfakes, or disinformation more broadly, only played a marginal role. It also emerges from the policy documents that the harms from disinformation are seen as something that is predominantly done by producers with an intention to harm and to recipients who, in turn, need to be protected. This leaves little room for recipients who may perpetuate some epistemic harms, e.g. through engaging in maladaptive epistemic communities around *polarized fellowship* or who proclaim false justifications for belief because they resonate with them, but also incur epistemic harms at the same time (*see* 5.1).

Policy-makers aim to mitigate the harms from disinformation primarily through *content moderation* and *preventative measures*. In doing so, they are hesitant to restrict freedom of expression in digital information environments. As a result, the most prominent and preferred policy option presented in the discourse are: labelling and fact-checking false content, increasing the access to accurate information, e.g. through algorithmically amplifying it or otherwise supporting trustworthy sources. These measures primarily contextualize information that is deemed to be false and provide access to 'authoritative information' that is deemed to be accurate. Nonetheless, more restrictive content moderation policies such as deleting and limiting the reach of disinformation content are also considered in some cases (*see* 5.2).

When comparing the perspective on the epistemic harms from deepfakes in Ch. 3 to the results of the empirical analysis in Ch. 5, policy-makers appear to be overlooking the crucial role that trust in the source of a piece of information plays in the process of how recipients make sense of information in digital information environments. For contextualizations and access to 'authoritative information' to mitigate epistemic harms, recipients need to believe them, which

will depend on how much they trust the instructions that present them. However, as this cannot be taken for granted. By contrast, no meaningful measures are introduced that are suitable to improve the rapport of established sources of information in digital information environments in a meaningful capacity. Further, though harms of *cognitive resonance* and *polarized fellowship* may be somewhat mitigated by policies that limit the reach or delete disinformation, contextualization and alternative sources of information are unlikely to be enough to remedy both problems. Therefore, the answer to the overarching research question of this thesis – *does EU policy address the epistemic harms caused by deepfakes?* – is: No, not in a comprehensive manner.

Policy-makers need to compliment measures for content moderation and improving the accessibility of accurate information in digital information environments with measures that (re-)build trust in the very political institutions that set out these measures. Media literacy training, though not a default response as it suffers from the same problem regarding the trustworthiness of the institution that provides it, may be a suitable starting point so long as policy-makers can find a way to go through actors that vulnerable recipients trust. However, ultimately measures that signal benevolence will go a long way. Instances in which people perceive political institutions to act against their interests will result in the information those institutions put out loosing credibility. Stress and perceived lack of control aggravate this issue (Bayer et al. 2021: 99-102). Certainly, policy-makers cannot act in a way that simultaneously aligns with everybody's interest and avert every crisis. However, measures that alleviate socio-economic grievances, personal stress and help individuals manage crises, though not in the traditional domain of disinformation or artificial intelligence policy, may address cases where existing measures fall short. The whole-of-society approach described by Wigell and colleagues (2021) stresses the need for a wholistic approach to resilience against hybrid threats (2021). Policy-makers may extend this wholistic perspective to policy response to disinformation as well.

Lastly, I want to address some limitations of my research. Beginning with the scope of the conceptual inquiry, I have focused in this thesis on political deepfakes. I have chosen a broad understanding of what counts as political but distanced myself from including not strictly epistemic literature on deepfakes for my considerations. I have thus excluded literature on gender-based harms from deepfakes (*e.g.* Maddocks 2020). There is a pitfall in making this distinction. As can be seen in the example of investigative journalist Rana Ayyab, deepfakes may be leveraged to retaliate against others for perceived transgressions (Ayyab 2018). Ultimately, this may have a chilling effect on the participation of populations who are

particularly vulnerable to being targeted with deepfakes in public discourse. Epistemic harms that result from silencing effects that selectively affect some potential participant in public discourse have not been examined here but – considering the literature on epistemic injustice (Fricker 2007) – this presents a promising direction to further develop the understanding of epistemic harms from deepfakes (*see also* Kerner, Risse 2020).

Further, more research – both conceptual and empirical – is needed to corroborate the two additional kinds of epistemic harm I proposed in this thesis (*see* 3.3). Questions of interest regarding *cognitive resonance* are in how far associations formed on the basis of (known) deepfakes persist and in how far deepfakes are able to entrench pre-existing beliefs. With respect to *cognitive resonance*, empirical research should look at the persistence of ad-hoc epistemic communities that engage with polarizing deepfakes and whether those recipients form an epistemic community beyond the individual posts they engage with.

Further, there are some limitations to the scope of the empirical analysis. First, my analysis only reflects EU policy discourse as it manifests in select documents and studies. Positions of administrative agencies, such as the European External Action Service, or individual parties are not represented. Similarly, the perspective of member states, civil society actors and the private sector are not represented in any further depth. Work that seeks to expand on my analysis of the EU policy discourse on deepfakes may expand in these directions.

Additionally, there are efforts to regulate deepfakes underway in other jurisdictions, e.g. the USA and China (Geng 2023). These efforts may present an interesting area for further (comparative) research, not only because these policies themselves stand to have a profound influence on deepfakes as a whole, but also because the EU has repeatedly voiced its conviction to be a frontrunner in human-centred AI policy (EC 2019b, 2021c). Looking at how EU deepfake policy compares to that of other jurisdictions may present an interesting case study for the EU's larger strategy in the area of emerging digital technologies. Lastly, I have not endeavoured to judge the measures stipulated by the analysed policy documents regarding their efficacy of achieving their specific goals (*see* 5.2). While they appear to address various kinds of epistemic harm, whether they actually do so is open for further inquiry.

I want to close with echoing the argument Harris (2021) makes for why deepfakes are not as big of a problem as some commenters suggest: "*through appropriate patterns of trust, whatever epistemic threat deepfakes pose can be substantially mitigated*" (ibid.: 13373). All that is needed to make this argument into a reassuring one is to make sure the patterns of trust of the recipients of deepfakes are appropriate.

References

Abell, C. (2010). The Epistemic Value of Photographs. In Philosophical Perspectives on Depiction. Oxford University Press.

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics and Informatics, 57, 101508.

Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2019). The State of Deepfakes. Deeptrace.

Anderson, E. (2006). The epistemology of democracy. Episteme, 3(1-2), 8-22.

Atencia-Linares, P., & Artiga, M. (2022). Deepfakes, shallow epistemic graves: On the epistemic robustness of photography and videos in the era of deepfakes. Synthèse, 200(6), 518.

Ayyab, R. (2018). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. Huffington Post, 21.11.2018. Available at: https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316. Accessed 06.12.2023.

Barthes, R. (1977). Rhetoric of the image. In Heath, S. (ed. and trans.), Image, Music, Text. London: Fontana Press, pp. 32–51.

Bateman, J. (2020). Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios. Carnegie Endowment for International Peace.

Bayer, J., Holznagel, B., Lubianiec, K., Pintea, A., Schmitt, J., Szakács, J, & Uszkiewicz, E. (2021). Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the EU and its Member States. Available at: https://www.europarl.europa.eu/thinktank/en/document/EXPO_STU(2021)653633 (Accessed 06.12.2023).

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. Convergence, 27(4), 919-934.

Bradford, A. (2012). THE BRUSSELS EFFECT. Northwestern University Law Review, 107(1).

Bradford, A. (2020). The Brussels effect: How the European Union rules the world. Oxford University Press, USA.

Brattberg, E., Csernatoni, R., & Rugova, V. (2020). Europe and AI: leading, lagging behind or carving its own way?. Carnegie Endowment for International Peace. Available at:

https://carnegieendowment.org/2020/07/09/europe-and-ai-leading-lagging-behind-or-carving-its-own-way-pub-82236 (Accessed 06.12.2023).

Brown, É. (2018). Propaganda, misinformation, and the epistemic value of democracy. Critical Review, 30(3-4), 194-218.

Buchanan, T., & Benson, V. (2019). Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"? Social Media + Society, 5(4).

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. Synthese, 195(12), 5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10), e12625.

Bucy, E. P. (2011). Nonverbal communication, emotion, and political evaluation. In: Döveling, K., von Scheve, C., & Konijn, E. (eds.) The Routledge Handbook of Emotions and Mass Media, 195-220.

Calvillo, D. P., Garcia, R. J., Bertrand, K., & Mayers, T. A. (2021). Personality factors and self-reported political news consumption predict susceptibility to political fake news. Personality and individual differences, 174.

Carey, B. (2023). Misinformation and Epistemic Harm. Social Philosophy Today.

Cavedon-Taylor, D. (2013). Photographically based knowledge. Episteme, 10(3), 283-297.

Chandler, D. (2022). Semiotics: the basics. Routledge.

Charlet, K., & Citron, D. (2019). Campaigns Must Prepare for Deepfakes: This Is What Their Plan Should Look Like. Carnegie Endowment For International Peace. Available at: https://carnegieendowment.org/2019/09/05/campaigns-must-prepare-for-deepfakes-this-is-what-their-plan-should-look-like-pub-79792 (Accessed 06.12.2023).

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev., 107, 1753.

Chisholm, R. M., & Feehan, T. D. (1977). The intent to deceive. The Journal of Philosophy, 74(3), 143-159.

Cohen, J., & Meskin, A. (2004). On the epistemic value of photographs. The Journal of Aesthetics and Art Criticism, 62(2), 197-210.

Congressional Research Service (2022). Deep Fakes and National Security. Available at: https://apps.dtic.mil/sti/pdfs/AD1171722.pdf. (Accessed 06.12.2023).

Confer, J. C., Easton, J. A., Fleischman, D. S., Goetz, C. D., Lewis, D. M., Perilloux, C., & Buss, D. M. (2010). Evolutionary psychology: Controversies, questions, prospects, and limitations. American Psychologist, 65(2), 110.

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Dabbous, A., Tarhini, A., & Harfouche, A. (2023). Circulation of Fake News: Threat Analysis Model to Assess the Impact on Society and Public Safety. 2023 IEEE International Symposium on Technology and Society, 1-9.

Dai, J., Lu, Y., & Wu, Y. N. (2016). Generative modeling of convolutional neural networks. Statistics and Its Interface, 9(4), 485-496.

Datzer, V., & Lonardo, L. (2023). Genesis and evolution of EU anti disinformation policy: entrepreneurship and political opportunism in the regulation of digital technology. Journal of European Integration, 45(5), 751-766.

De Ruiter, A. (2021). The distinct wrong of deepfakes. Philosophy & Technology, 34(4), 1311-1332.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. New Media & Society, 23(7), 2072-2098.

Directive 2000/31/EC. Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'). Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32000L0031 (Accessed 06.12.2023).

Directive 2010/13/EU. Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services

(Audiovisual Media Services Directive). Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0013 (Accessed 06.12.2023).

Disarm Foundation (n.d.). DISARM is an open framework for those cooperating in the fight against disinformation. Available at: https://www.disarm.foundation/ (Accessed 06.12.2023).

Dunn, S. (2020). Technology-facilitated gender-based violence: an overview. Centre for International Governance Innovation: Supporting a Safer Internet Paper.

Durach, F., Bârgăoanu, A., & Nastasiu, C. (2020). Tackling disinformation: EU regulation of the digital space. Romanian journal of European affairs, 20(1).

European Commission (n.d. a). European Democracy Action Plan. Available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en (Accessed 06.12.2023).

European Commission (n.d. b). The Digital Services Act: ensuring a safe and accountable online environment. Available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en (Accessed 06.12.2023).

European Commission (n.d. c). High-level expert group on artificial intelligence. Available at: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (Accessed 06.12.2023).

European Commission (n.d. d). Coordinated Plan on Artificial Intelligence. Available at: https://digital-strategy.ec.europa.eu/en/policies/plan-ai (Accessed 06.12.2023).

European Commission (2005). Judicial co-operation in criminal matters: mutual recognition of final decisions in criminal matters. Available at: https://eur-lex.europa.eu/legal-content/HU/ALL/?uri=LEGISSUM%3Al33131 (Accessed 06.12.2023).

European Commission (2010). A Digital Agenda for Europe. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52010DC0245R%2801%29 (Accessed 06.12.2023).

European Commission (2018a). Code of Practice on Disinformation. Available at: https://ec.europa.eu/newsroom/dae/redirection/document/87534 (Accessed 06.12.2023).

European Commission (2018b). Action Plan on Disinformation. Available at: https://commission.europa.eu/publications/action-plan-disinformation-commission-contribution-european-council-13-14-december-2018_en (Accessed 06.12.2023).

European Commission (2018c). A Europe that Protects: Countering Hybrid Threats. Available at: https://www.eeas.europa.eu/node/46393_en (Accessed 06.12.2023).

European Commission (2018d). Tackling online disinformation: a European Approach. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236 (Accessed 06.12.2023).

European Commission (2018e). Artificial Intelligence for Europe. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN (Accessed 06.12.2023).

European Commission (2018f). Coordinated Plan for Artificial Intelligence. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0795 (Accessed 06.12.2023).

European Commission (2019a). Annual self-assessment reports of signatories to the Code of Practice on Disinformation 2019. Available at: https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019 (Accessed 06.12.2023).

European Commission (2019b). Building Trust in Human Centric Artificial Intelligence. Available at: https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence (Accessed 06.12.2023).

European Commission (2020a). European Democracy Action Plan. Available at: https://commission.europa.eu/document/download/63918142-7e4c-41ac-b880-6386df1c4f6c_en (Accessed 06.12.2023).

European Commission (2020b). White paper on artificial intelligence - A European approach to excellence and trust, COM(2020)65 final. Available at: https://ec.europa.eu/info/sites/default/files/commission-white-paperartificial-intelligence-feb2020_en.pdf (Accessed 06.12.2023).

European Commission (2020c). Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Available at: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52020PC0825 (Accessed 06.12.2023).

European Commission (2021a). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence

Act) and Amending Certain Union Legislative Acts. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (Accessed 06.12.2023).

European Commission (2021b). Proposal for a Regulation laying down harmonised rules on artificial intelligence. Available at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (Accessed 06.12.2023).

European Commission (2021c). Fostering a European Approach to Artificial Intelligence. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM%3A2021%3A205%3AFIN (Accessed 06.12.2023).

European Commission (2021d). Guidance on Strengthening the Code of Practice on Disinformation. Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0262 (Accessed 06.12.2023).

European Commission (2021e). Annexes to the Proposal for a Regulation of the European Parliament and of the Council – Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF (Accessed 06.12.2023).

European Commission (2022a). Strengthened Code of Practice on Disinformation. Available at: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation (Accessed 06.12.2023).

European Commission (2022b). Signatories of the 2022 Strengthened Code of Practice on Disinformation. Available at: https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation (Accessed 06.12.2023).

European External Action Service (2015). Action Plan on Strategic Communication. Available at: https://www.eeas.europa.eu/sites/default/files/action_plan_on_strategic_communication.docx_eeas_web.pdf (Accessed 06.12..2023).

European External Action Service (2019). Report on the implementation of the Action Plan Against Disinformation. Available at: https://www.eeas.europa.eu/sites/default/files/joint_report_on_disinformation.pdf (Accessed 06.12.2023).

European Parliament (2017). Civil Law Rules on Robotics. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html (Accessed 06.12.2023).

European Parliament (2018a). Media pluralism and media freedom in the European Union. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2018-0204_EN.html (Accessed 06.12.2023).

European Parliament (2018b). Online platforms and the Digital Single Market. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0272_EN.html (Accessed 06.12.2023).

European Parliament (2019a). A comprehensive European industrial policy on artificial intelligence and robotics. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2019-0081_EN.html (Accessed 06.12.2023).

European Parliament (2019b). Follow up taken by the EEAS two years after the EP report on EU strategic communication to counteract propaganda against it by third parties. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2019-0187_EN.html (Accessed 06.12.2023).

European Parliament (2020a). Framework of ethical aspects of artificial intelligence, robotics and related. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.html (Accessed 06.12.2023).

European Parliament (2020b). Intellectual property rights for the development of artificial intelligence. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html (Accessed 06.12.2023).

European Parliament (2021a). Artificial intelligence in education, culture and the audiovisual sector. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0238_EN.html (Accessed 06.12.2023).

European Parliament (2021b). Artificial intelligence questions of interpretation and application of international law. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.html (Accessed 06.12.2023).

European Parliament (2021c). Combating gender-based violence: cyberviolence. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0489_EN.html (Accessed 06.12.2023).

European Parliament (2022a). Artificial intelligence in a digital age. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0140_EN.html (Accessed 06.12.2023).

European Parliament (2022b). Foreign interference in all democratic processes in the European Union. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0064_EN.html (Accessed 06.12.2023).

European Parliament (2023a). The EU priorities for the 67th session of the UN Commission on the Status of Women. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0048_EN.html (Accessed 06.12.2023).

European Parliament (2023b). Foreign interference in all democratic processes in the European Union, including Disinformation. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0219_EN.html (Accessed 06.12.2023).

European Regulators Group for Audiovisual Media Services (2020). ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice. Available at: https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf (Accessed 06.12.2023).

European Regulators Group for Audiovisual Media Services (2023). ERGA report on the first year of the Strengthened Code of Practice on Disinformation. Available at: https://erga-online.eu/wp-content/uploads/2023/07/ERGA-SG3-report-CoP_June-2023_as-adopted.pdf (Accessed 06.12.2023).

European Union (2010). Charter of Fundamental Rights of the European Union. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (Accessed 06.12.2023).

Fallis, D. (2021). The epistemic threat of deepfakes. Philosophy & Technology, 34(4), 623-643.

Faragó, L., Kende, A., & Krekó, P. (2020), We Only Believe in News That We Doctored Ourselves. The Connection Between Partisanship and Political Fake News. Social Psychology, 51, 77-90.

Fleisher, W., & Šešelja, D. (2023). Responsibility for collective epistemic harms. Philosophy of Science, 90(1), 1-20.

Flick, U. (2009). An Introduction to qualitative research. SAGE. Fourth Edition.

Floridi, L. (1996). Brave. Net. World: the Internet as a disinformation superhighway?. The Electronic Library, 14(6), 509-514.

Foroni, F., & Mayr, U. (2005). The power of a story: New, automatic associations from a single reading of a short scenario. Psychonomic Bulletin & Review, 12(1), 139–144.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Frost-Arnold, K. (2023). Who should we be online. A social epistemology for the internet.

Galatolo, F. A., Cimino, M. G., & Vaglini, G. (2021). Generating images from caption and vice versa via clip-guided generative latent space search. arXiv preprint arXiv:2102.01645.

Geng, Y. (2023). Comparing "Deepfake" Regulatory Regimes in the United States, the European Union, and China. Georgetown Law Technology Review 7(1), 157-178.

GitHub (n.d.). AMITT Disinformation Tactics, Techniques and Processes (TTP) Framework. Available at: https://github.com/cogsec-collaborative/AMITT/tree/main/archived_version_of_AMITT (Accessed 06.12.2023).

Greifeneder, R., Jaffé, M. E., Newman, E.J. & Schwarz, N. (2021, Eds.). The psychology of fake news. Accepting, sharing, and correcting of misinformation. Routledge: Oxon, New York.

Habgood-Coote, J. (2019). Stop talking about fake news!. Inquiry, 62(9-10), 1033-1065.

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. Synthese, 201(3), 103.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2023). They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes. European Journal of Communication.

Hamilton, L. (2022). Conspiracy vs. Science: A Survey of U.S. Public Beliefs. Available on: https://carsey.unh.edu/publication/conspiracy-vs-science-a-survey-of-us-public-beliefs. (Accessed 06.12.2023).

Harris, K. R. (2021). Video on demand: what deepfakes do and how they harm. Synthese, 199(5-6), 13373-13391.

Harris, K. R. (2022). Real Fakes: The Epistemology of Online Misinformation. Philosophy & Technology, 35(3), 83.

Hart, H. L. A. (1968). Punishment and responsibility: Essays in the philosophy of law. Oxford: Oxford University Press.

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed 06.12.2023).

High-Level Expert Group on Fake News and Online Disinformation (2019). A multi-dimensional approach to disinformation. Available at: https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation (Accessed 06.12.2022).

Hopkins, R. (1998). Picture, image and experience: A philosophical inquiry. Cambridge University Press.

Hopkins, R. (2012). Factive pictorial experience: What's special about photographs?. Noûs, 46(4), 709-731.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), Implicit Bias and philosophy—Volume 1: Metaphysics and epistemology, 47–79. Oxford University Press.

Hughes, S., Fried, O., Ferguson, M., Hughes, C., Hughes, R., Yao, X., & Hussey, I. (2021). Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions. Available at: https://www.osti.gov/biblio/1780812 (Accessed 06.12.2023).

Huneman, P., & Machery, E. (2015). Evolutionary psychology: Issues, results, debates. Handbook of Evolutionary Thinking in the Sciences, 647-657.

Hunter, T. (2023) AI porn is easy to make now. for women, that's a nightmare., The Washington Post, 13.02.2023. WP Company. Available at: https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/ (Accessed: March 26, 2023).

Hwang, T. (2020). Deepfakes – A grounded threat analysis. Center for Security and Emerging Technology.

IBM (n.d.). What is an API?. Available at: https://www.ibm.com/topics/api (Accessed 06.12.2023).

Imhoff, R., & Lamberty, P. K. (2018). How paranoid are conspiracy believers? Toward a more fine-grained understanding of the connect and disconnect between paranoia and belief in conspiracy theories. European Journal of Social Psychology, 48, 909-926.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. Journal of personality and social psychology, 50(6), 1141.

Justo-Hanani, R. (2022). The politics of Artificial Intelligence regulation and governance reform in the European Union. Policy Sciences, 55(1), 137-159.

Kappel, K. (2013). Epistemological dimensions of informational privacy. Episteme, 10(2), 179-192.

Kerner, C. & Risse, M. (2020). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. Moral Philosophy and Politics 8: 81-108.

Khalifa, K., & Millson, J. (2020). Perspectives, questions, and epistemic value. Knowledge from a human point of view, 87-106.

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

Kingslake, R. (1989). A history of the photographic lens. Academic press.

Klebba, L., & Winter, S. (2021). Selecting and sharing news in an “infodemic”: The influence of ideological, trust- and science-related beliefs on (fake) news usage in the COVID-19 crisis.

Kuckartz, U., & Rädiker, S. (2023). Qualitative Content Analysis: Methods, Practice and Software. SAGE.

Kugler, M. B., & Pace, C. (2021). Deepfake Privacy: Attitudes and Regulation. Northwestern University Law Review, 116(3), 611-680.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Levy, N. (2022). In trust we trust: epistemic vigilance and responsibility. Social epistemology, 36(3), 283-298.

Liao, S. Y., & Huebner, B. (2021). Oppressive things. Philosophy and Phenomenological Research, 103(1), 92-113.

Mack, D. (2018). This PSA About Fake News From Barack Obama Is Not What It Appears. BuzzFeed, 17.04.2018. Available at:

https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed (Accessed 06.12.2023).

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. Porn Studies, 7(4), 415-423.

Marmor, A. (2015). What is the Right to Privacy?. Philosophy & Public Affairs, 43(1), 3-26.

Matthews, T. (2022). Deepfakes, intellectual cynics, and the cultivation of digital sensibility. Royal Institute of Philosophy Supplements, 92, 67-85.

Matthews, T. (2023). Deepfakes, Fake Barns, and Knowledge from Videos. Synthese, 201(2), 41.

Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution.

Mayring, P. (2021). Qualitative content analysis: A step-by-step guide. SAGE.

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. Review of General Psychology, 21(2), 103-122.

Mercier, H. (2019). Not born yesterday. In Not Born Yesterday. Princeton University Press.

Michailidou, A., Eike, E., & Trenz, H. J. (2022). Journalism, truth and the restoration of trust in democracy: Tracing the EU 'fake news' strategy. In Europe in the Age of Post-Truth Politics: Populism, Disinformation and the Public Sphere (pp. 53-75). Cham: Springer International Publishing.

Millière, R. (2022). Deep learning and synthetic media. Synthese, 200(3), 231.

Oremus, W. (2022). How social media 'censorship' became a front line in the culture war. The Washington Post, 28.10.2022. Available at: https://www.washingtonpost.com/technology/2022/10/09/social-media-content-moderation/ (Accessed 06.12.2023)

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2085-2094).

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., & Zhang, W. (2021). DeepFaceLab: Integrated, exible and extensible face-swapping framework. Available at: https://arxiv.org/abs/2005.05535. (Accessed 06.12.2023).

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. Annual review of psychology, 48(1), 609-647.

Pierini, F. (2023). Deepfakes and depiction: from evidence to communication. Synthese, 201(3), 97.

Pierre, J.M. (2020). Mistrust and Misinformation: A Two-Component, Socio-Epistemic Model of Belief in Conspiracy Theories. Journal of Social and Political Psychology, 8, 617–641.

Plasilova, I.; Hill, J.; Carlberg, M.; Goubet, M., & Procee, R. (2020). STUDY FOR THE Assessment of the implementation of the Code of Practice on Disinformation. Available at: https://digital-strategy.ec.europa.eu/en/library/study-assessment-implementation-code-practice-disinformation (Accessed 06.12.2023).

Poulsen, S. V. (2021). Face off–a semiotic technology study of software for making deepfakes. Σημειωτκή-Sign Systems Studies, 49(3-4), 489-508.

Öhman, C. (2020). Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. Ethics and Information Technology, 22(2), 133-140.

Öhman, C. (2022). The identification game: deepfakes and the epistemic limits of identity. Synthese, 200(4), 319.

Ogden, C. K., & Richards, I. A. (1923). The meaning of meaning: A study of the influence of thought and of the science of symbolism.

Oltermann, P. (2022). European politicians duped into deepfake video calls with mayor of Kyiv. The Guardian, 25.06.2022. Available at: https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko (Accessed 06.12.2023).

Pritchard, D., Turri, J. & Carter, J A. (2018). The value of knowledge. The Stanford Encyclopedia of Philosophy. Zalta, E. N. (ed.). Available at: https://plato.stanford.edu/archives/spr2018/entries/knowledge-value (Accessed 06.12.2023).

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. & Sutskever, I. (2021, July). Zero-shot text-to-image generation. International Conference on Machine Learning, 8821-8831.

Regulation (EU) 2016/679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed 06.12.2023).

Regulation (EU) 2022/2065. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065 (Accessed 06.12.2023).

Rini, R. (2020). Deepfakes and the epistemic backstop. Philosophers' Imprint, 20(24), 1–16.

Rini, R. (2021). Weaponized Skepticism. Political epistemology, 31.

Rini, R., & Cohen, L. (2022). Deepfakes, Deep Harms. Journal of Ethics and Social Philosophy, 22(2).

Roberts, T. (2023). How to do things with deepfakes. Synthese, 201(2), 43.

Roth, L. (2009). Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. Canadian Journal of Communication, 34(1), 111-136.

Roth, L. (2019). Making skin visible through liberatory design. In: Benjamin, R. (Ed.), Captivating technologies, 275–307. Duke University Press.

Sádaba, C., & Salaverría, R. (2023). Tackling disinformation with media literacy: analysis of trends in the European Union. Revista Latina de Comunicación Social, (81), 17-32.

Schauer, F. (2022). The Proof: Uses of Evidence in Law, Politics, and Everything Else. Harvard University Press.

Schreier, M. (2012). Qualitative content analysis in practice. SAGE.

Simonite, T. (2022). A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. Wired, 17.03.2022. Available at: https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/ (Accessed 06.12.2023).

Smith, H., & Mansted, K. (2020). Weaponised deep fakes: national security and democracy. Australian Strategic Policy Institute.

Sontag, S. (1973). On Photography. New York: Dell.

Sperber, D. (1997). Intuitive and reflective beliefs. Mind & Language, 12(1), 67-83.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & language, 25(4), 359-393.

Stamann, C., Janssen, M., & Schreier, M. (2016). Qualitative Inhaltsanalyse – Versuch einer Begriffsbestimmung und Systematisierung. Forum: Qualitative Social Research 17(3).

Steup, M., & Neta, R. (2020). Epistemology. The Stanford Encyclopedia of Philosophy. Zalta, E. N. (ed.). Available at: https://plato.stanford.edu/archives/fall2020/entries/epistemology/ (Accessed 06.12.2023).

Strowel, A., & De Meyere, J. (2023). The Digital Services Act: Transparency as an Efficient Tool to Curb the Spread of Disinformation on Online Platforms?. J. Intell. Prop. Info. Tech. & Elec. Com. L., 14, 66.

Telegraph (2022). Deepfake video of Volodymyr Zelensky surrendering surfaces on social media. 17.03.2022. Available at: https://www.youtube.com/watch?v=X17yrEV5sl4 (Accessed 06.12.2023).

The AI Act (2023). Developments. Available at: https://artificialintelligenceact.eu/developments/ (Accessed 06.12.2023).

van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review, 46, 105716.

van Huijstee, M., van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M., Fatun, M., Kool, L., & Gerritsen, J. (2021). Tackling deepfakes in European policy. available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039 (Accessed 06.12.2023).

Vasist, P. N., & Krishnan, S. (2022). Deepfakes: an integrative review of the literature and an agenda for future research. Communications of the Association for Information Systems, 51(1), 14.

Vinokur, A. (1971). Review and theoretical analysis of the effects of group processes upon individual and group decisions involving risk. Psychological Bulletin, 76(4), 231.

Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. Synthese, 201(1), 1-20.

Vlasceanu, M., Goebel, J., & Coman, A. (2020). The Emotion-Induced Belief-Amplification Effect. CogSci.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Walden, S. (2005). Objectivity in photography. The British Journal of Aesthetics, 45(3), 258-272.

Walden, S. (2012). Photography and knowledge. The Journal of Aesthetics and Art Criticism, 70(1), 139-149.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. Political Communication, 37(3), 350-375.

Walther, J. B., Carr, C. T., & Choi, S. S. W. (2010). Interaction of interpersonal, peer, and media influence sources online: A research agenda for technology convergence. A networked self, 25-46.

Walton, K. L. (1984). Transparent pictures: On the nature of photographic realism. Critical inquiry, 11(2), 246-277.

Wardle, C., & Darekhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe. Available at: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html# (Accessed 06.12.2023).

Warren, M. (2018). Trust and democracy. In: Uslaner, E. M. (ed.): The Oxford handbook of social and political trust, 75-94. Oxford University Press.

Wenar, L. (2023). Rights. The Stanford Encyclopedia of Philosophy. Zalta, E. N. & Nodelman, U (eds.). Available at: https://plato.stanford.edu/archives/spr2023/entries/rights/ (Accessed 06.12.2023).

Wigell, M., Mikkola, H., & Juntunen, T. (2021). Best Practices in the Whole-of-Society Approach in Countering Hybrid Threats. Available at: https://www.europarl.europa.eu/thinktank/en/document/EXPO_STU(2021)653632 (Accessed 06.12.2023).

Winter, Rachel, and Anastasia Salter. "DeepFakes: uncovering hardcore open source on GitHub." Porn Studies 7, no. 4 (2020): 382-397.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. Journal of Personality and Social Psychology, 81(5), 815–827.

World Health Organization (2022). Infodemics and misinformation negatively affect people's health behaviours, new WHO review finds. Available at: https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds (Accessed 06.12.2023).

Appendix 1 – List of analyzed documents

Document Title	Document Type	Year of Publication	Authors	Corresponding Institution	Citation
Tackling Online Disinformation: A European Approach	Communication by the EC	2018	European Commission	European Commission	EC 2018d
Artificial Intelligence in Europe	Communication by the EC	2018	European Commission	European Commission	EC 2018e
Coordinated Plan on Artificial Intelligence	Communication by the EC	2018	European Commission	European Commission	EC 2018f
A Multi-Dimensional Approach to Disinformation	Report	2018	High-Levcl Expert Group on Fake News and Disinformation	European Commission	HLEG FNOD 2018
Building Trust in Human-Centric Artificial Intelligence	Communication by the EC	2019	European Commission	European Commission	EC 2019b
Fostering a European Approach to Artificial Intelligence	Communication by the EC	2021	European Commission	European Commission	EC 2021c
Guidance on Strengthening the Code of Practice on Disinformation	Communication by the EC	2021	European Commission	European Commission	EC 2021d
Ethics Guidelines for Trustworthy Artificial Intelligence	Other	2019	High-Levcl Expert Group on Artificial Intelligence	European Commission	HLEG AI 2019
Working Paper on Artificial Intelligence and the Future of Democracy	Other	2021	European Parliamentary Committee on Artificial Intelligence in a Digital Age	European Parliament	AIDA 2021
Tackling deepfakes in European policy	Report	2021	van Huijstee, M., van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M., Fatun, M., Kool, L., & Gerritsen, J.	European Parliament	van Huijstee et al. 2021
Best Practices in the Whole-of-Society Approach in Countering Hybrid Threats	Report	2021	Wigell, M., Mikkola, H., & Juntunen, T.	European Parliament	Wigell et al. 2021
Disinformation and Propaganda: Impact on the Functioning of the Rule of Law and Democratic Processes in the EU and its Member States	Report	2021	Bayer, J., Holznagel, B., Lubianiec, K., Pintea, A., Schmitt, J., Szakács, J, & Uszkiewicz, E.	European Parliament	Bayer et al. 2021
Artificial Intelligence Act	Regulatory document	2021	European Commission	European Commission	EC 2021a
Digital Services Act	Regulatory document	2022	x	x	Regulation (EU) 2022/2065
2022 Code of Practice on Disinformation	Regulatory document	2022	Signatories (*see* EC 2022)	European Commission	EC 2022a
Report on the Implementation of the 2022 Code of Practice on Disinformation	Report	2023	European Regulators Group for Audiovisual Media Services	European Commission	ERGA 2023

Appendix 2 – Ensuring Research Quality

Arguably, common quality standards in (quantitative) research – objectivity, validity, and reliability – do not apply in an identical manner in qualitative research (Kuckartz, Rädiker 2023: 14f, 194ff). Especially when it comes to the development of inductive categories, ensuring intersubjective reliability can be challenging, if not impossible (ibid.: 59). Therefore, alternative standards of quality need to be followed. However, which standards one should follow is controversial. Kuckartz and Rädiker side with the position that qualitative research warrants the development of specific quality standards (ibid.: 194). Their proposed standards, insofar they apply to the present research interest and design, will be described in the following.

Broadly, there are two sets of standards of quality for QCA, internal and external standards. The former focus on authenticity and credibility of the research. They include fit and justification of the method, as well as its implementation (ibid.: 196). While method fit and justification have been covered above, ensuring quality of research in the implementation requires further considerations regarding the coding process, category system, and how the results of research are being reported (ibid.).

Coding would ideally be done by multiple researchers and moderated in case of disagreements. This can yield more precise category definitions and make their assignment to relevant text segments more reliable (ibid.: 106). However, within the boundaries of this thesis, that is not possible. Instead, precise category definitions and reliability were pursued through reflecting on the coding process during the supervision period and making the coded material and memos available to assessors, as will be described further below. Written coding guidelines for the main categories developed in the process can be found in Appendix 4. Further, the coding process, category definitions and the category system are critically reflected in the conclusions and some limitations are offered (*see* 7). As mentioned above, the category system is at the heart of QCA. Apart from coding all relevant segments and applying the correct categories, the quality of analysis largely depends on the adequacy and added value of the categories and the category system. Relevant standards of quality here are: relation to research question, analytical depth, inner coherence and logical relations between categories, exhaustiveness, precise category definitions and distinctiveness, plausibility and understandability (ibid.: 46-49, 196).

With respect to internal quality standard for the reporting of results, the inclusion of typical and atypical examples in the reported results, as well as archiving research material and making it available for auditing (Kuckartz, Rädiker 2023: 196) are to be considered. The former is

reflected in section 5. The latter will be met through making the final project file available to supervisors.

External standards focus on generalizability or transferability of a study, though generalizability is not necessarily applicable across QCA designs (ibid.: 207ff). As this thesis primarily seeks to make a claim regarding EU policy, generalizing claims to other jurisdictions – e.g. member states or other nations – is not desired. However, through the analysis of the selected documents, claims regarding EU policy are pursued. There are a few caveats to this. First, only publicly available documents by legislative institutions have been considered, under the exception of the report to the EPRS by van Huijstee et al. (2021). Public consultations, parliamentary discussions, contributions from individual member states on the matter *et cetera* have not been included in the sample, arguably diminishing the generalizability of results.

Nonetheless, applications of QCA should strive for transferability (ibid.: 209; *see also* Flick 2009: 400-412). Due to the selection of policy documents being based on the various elements of deepfakes and their intersection with disinformation and artificial intelligence policy more broadly (*see* 4.1) and the orientation of the category system toward those elements and intersections, the approach to QCA taken in this thesis should principally transfer to the study of deepfake policy in jurisdictions other than the EU.

The extent to which insights gathered from this analysis transfer to other phenomena is more challenging. As has been argued before and is evident from the selected documents, deepfakes are a quite unique phenomenon that intersects with a broad array of issues. Some of the insights gathered through the analysis will be relevant and transferable to other contexts beyond the research questions of this thesis, e.g. the application of synthetic audiovisual media in the creative sector. Nonetheless, transferability should not be overstated considering the heterogeneity of these other phenomena. Though the specificity of the present research design may jeopardize generalizability and transferability, it allows for more precise claims regarding the research interest of this thesis.

Appendix 3 – Initial Deductive Codes

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Document Type	Categorization of the analyzed documents.	EC Communication	Document published as a communication by the European Commission.	x	x
		Regulatory Framework	Proposed or already enacted regulatory frameworks.	x	x
		Independent Council	Studies commission by EU institutions but carried out by independent researchers.	x	x
		Other	Documents that do not fit the other categories.	x	x
Year	Year the document was published.	2018	Published in 2018.	x	x
		2019	Published in 2019.	x	x
		2021	Published in 2021.	x	x
		2022	Published in 2022.	x	x
		2023	Published in 2023.	x	x
Elements Addressed	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment.	Actors	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, actors involved.	Passive Recipients	Recipients that do not engage with false information online any further.
				Sharing Recipients	Recipients that share false information online.
				Commenting Recipients	Rcipients that comment on false information online.
				Private producers	Producers of false information online that are not part of a wider organization that spreads false information online.
				Producers affiliated with a (criminal) organization	Producers of false information online that are part of a wider organization that spreads false information online.
				Producers affiliated with a state	Producers of false information online that are affiliated with a state asking them to spread false information online.
				Existing human subjects	Subjects of a deepfake that are existing persons.
				Non-existent human subjects	Subjects of a deepfake that are human but not existing persons.
				Non-human subjects	Subjects of deepfakes that are not human, e.g. animals or buildings.
				Platforms	Platforms are digital information environments in which users are able to publically share multimedia content with each other. Usually, users would also be able to comment on and share content from other users. There is some uncertainty on the margins what constitutes a platform, but the paradigmatic case are large social media platforms such as Facebook and TikTok or fora such as Reddit.

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Elements Addressed	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment.	Technology	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, aspects of the technology involved.	Deepfakes generally	Refers to deepfakes broadly, without a specified format.
				Image-based DFs	DFs in image format.
				Video-based DFs	DFs in video format, usually accompanied by deepfaked audio.
				Audio-based DFs	DFs that are purely auditory.
				DFAs generally	References to DFAs without further specification.
				Independent DFAs	DFAs that are available as stand-alone software.
				Embedded DFAs	DFAs that are embedded in other applications (e.g. Snapchat, TikTok).
				Other kinds of synthetic audiovisual media	Manipulated images, audio or video that have not been produced through machine- or deep-learning techniques.
				Generative Artificial Intelligence	Non-audiovisual generative AI, e.g. Large Language Models.
				Non-SAM, non-AI kinds of false information	False information that does not come in the form of audiovisual media, e.g. written misinformation.
				Deep-learning techniques	Know--how pertaining to deep-learning and neural networks.
				Machine-learning techniques	Know-how pertaining to machine-learning.
				Input data	Data necessary to train deepfake applications.
				Databases	Large sets of data that can be used for training a deepfake model accessible through a provider.
				Scraping	Gathering of input data from the open internet via scraping tools.
		False Information Online	Among the topics that are addressed in the respective documents are issues pertaining to false information online, including disinformation, misinformation and fake news. Specifically, different kinds of false content online.	Disinformation	False information produced and published with malicious intent. May include deepfakes.
				False information generally	False information online in general, irrespective of intent.
Understanding of epistemic harm	Notion of the specific kind of epistemic harm caused by deepfakes or (where not applicable) misinformation and disionformation more broadly.	Deception	Refers to then notion that false information online leads to recipients adopting false beliefs.	x	x
		Jeopardizing Evidence	Refers to the undermining effect that false information online may have to evidentiary practices, e.g. of recordings in criminal procedures, but also public discourse.	x	x
		Erosion of Trust	Refers to the loss of trust in epistemic institutions, practices and the competence of others as a result of false information online.	x	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Understanding of epistemic harm	Notion of the specific kind of epistemic harm caused by deepfakes or (where not applicable) misinformation and disionformation more broadly.	Cognitive Resonance	Refers to the acknowledgement that false information online may not only deceive recipients through causing a false belief, but may rather also be problematic in the notions and affects it evokes as well as being useful to recipients as faux-justifications.	x	x
		Polarized Fellowship	Refers to the notion of false information online signaling the trustworthiness and group-membership of a producer in order to form epistemically significant community with recipients.	x	x
		Other	x	x	x
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may onyl attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Transparency	Policies and other measures that are proposed or enacted in response to deepfakes. This includes policies that apply to false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly, as these policies will have implications for deepfakes as well. Specifically, this category is concerned with measures that aim at achieving greater transparency on various levels.	x	x
				x	x
				x	x
		Accountabiliy	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that hold actors involved in these phenomena for certain kinds of misconduct or that introduce new liabilities.	Provider accountability	Providers of input data or DFAs are made liable.
				Platform accountability	Platforms on which deepfakes are being presented to recipients are made liable.
				Producer accountability	Producers of deepfakes are made liable.
		Education	Various measures to educate the public on specific issues regarding deepfakes, AI or disinformation.	Raising awareness for producers	Potential producers of deepfakes are made aware of potential liabilities and harms that may result from deepfakes
				Raising awareness and media literacy training for recipients	Measures that are raising awareness or increasing the media literacy of potential recipients.
				Advising caution about sharing data	Potential subjects of deepfakes are cautioned against sharing footage of themselves or others.

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may onyl attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Content moderation	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with how content in digital information environments is moderated by various actors. Moderation can entail contextualization of information, e.g. through fact-checking and labelling of online content, or the deletion of content.	By professionals	Paid or otherwise institutionalized content moderators.
				By users	Users can flag footage as suspected to be fake.
				Automated detection	Automated systems try to filter footage for deepfakes.
				Labelling	Labelling false information online as (potentially) false upon discovery.
				Deletion of deepfakes and other false information	Deletion of false information online or disingenuous accounts upon discovery.
				Fact-checking	Fact-checking refers to calls for assessing the accuracy of information online and attempts to correct false information through providing accurate information on the matter. Fact-checking is usually done by journalists, academics, or civil society actors.
		Other	x	Research	(Support of) research into various aspects of the phenomena of false information online or (generative) artificial intelligence. This may include automated deepfake detection, or research into suitable responses to or the current state of false information online.
				Other	x

Appendix 4 – Coding Guidelines for Main- and Sub-Categories

Category	Category definition	Coding Guideline
Auxiliary	Codes that are in service to support the overall analysis alongside the codes in the other categories. Auxiliary codes help to build a fuller picture of the analyzed documents and aid in analyzing the documents in Atlas.ti.	These codes are applied either to documents as a whole or to segments in which issues are discussed that stray from the focus of the other categories.
Auxiliary - Document Type	Categorization of the analyzed documents.	Codes in this category are applied once per document to clarify which kind of document it is.
Auxiliary - Year	Year the document was published.	Codes in this category are applied once per document to clarify in which year a document was published.
Auxiliary - Main Issue	Main issue of interest in the respective documents.	Codes in this category are applied (usually) once per document to clarify the main issue of a document (deepfakes, AI, disinformation). If the document has more than one main focus, both will be labelled as main issues.
Auxiliary - Cautions	Segments in which various relevant aspects are reflected on critically. Codes included in this category occured more than five times.	Codes are applied to segments in which issues that are coded in other categories are reflected upon critically. This might be because there are constraints on the efficacy of measures, their enforcement or other issues that may arise from them.
Auxiliary - Relevant Context	Segments in which context is given to the issues of deepfakes, disinformation or artificial intelligence.	Codes are applied to segments in which context is provided for the issues of false information online, disinformation, deepfakes, or relevant actors that interact with these phenomena.
Elements Addressed	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment.	Codes are applied to segments in which elements that are directly connected to the phenomenon of deepfakes and false information online are referenced.
Elements Addressed - Actors	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, actors involved.	Codes are applied to segments in which actors that are directly connected to the phenomenon of deepfakes and false information online are referenced.
Elements Addressed - False information online	Among the topics that are addressed in the respective documents are issues pertaining to false information online, including disinformation, misinformation and fake news. Specifically, different kinds of false content online.	Codes are applied to segments in which different kinds of false information online are referenced.
Elements Addressed - Technologies	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, aspects of the technology involved.	Codes are applied to segments in which technological elements that are directly connected to the phenomenon of deepfakes and false information online are referenced.

Category	Category definition	Coding Guideline
Understanding of epistemic harm	Assessment which understanding of epistemic harms caused by false information online (disinformation, misinformation, fake news, deepfakes) is present in the analyzed policy document. Epistemic harms are defined as follows: Epistemic harm is caused if an agent obstructs, without legitimizing reason, the epistemic success of another in the area of politically relevant information.	Codes are applied to segments in which forms of epistemic harms are discussed in relation to the phenomena of deepfakes or other false information online.
Understanding of epistemic harm - Epistemic goods	Epistemic goods positively contribute to people forming accurate beliefs. This may be because recipients are exposed to accurate information and diverse viewpoints through a pluralistic information environment in which knowers and other discourse participants can freely express themselves and recipients are free to seek out information.	Codes are applied to segments in which forms of epistemic harms are discussed in relation epistemic goods that are implicated by the phenomena of deepfakes or other false information online. Codes in this category are also applied when harms to these epistemic goods are not mentioned explicitly but also where the relation is implied. The latter case was altogether more frequent.
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Codes are applied where documents propose, call for or establish measures that are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena.
Policy Measures - Accountability	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that hold actors involved in these phenomena for certain kinds of misconduct or that introduce new liabilities.	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) that seek to introduce ways through which actors involved can be held accountable for any kind of misconduct.

Category	Category definition	Coding Guideline
Policy Measures - Content moderation	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with how content in digital information environments is moderated by various actors. Moderation can entail contextualization of information, e.g. through fact-checking and labelling of online content, or the deletion of content.	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) that either support or introduce means of moderation the content that is posted in digital information environments with the goal of restricting or providing additional contextualization of false information online.
Policy Measures - Other	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly but do not fit the other categories.	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) are not covered by the other categories.
Policy Measures - Preventative measures	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that preempt the negative effects of false information online (disinformation, misinformation, fake news, deepfakes).	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) try to prevent (epistemic) harms from false information to materialize.
Policy Measures - Situational awareness	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that aim at producing a precise understanding about how these phenomena exist in the world and how regulation and guidelines that are relevant to them are applied and take effect.	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) that attempt to gather information and enhance the understanding of involved actors on how deepfakes, false information and artificial intelligence manifest and interact with digital information online.
Policy Measures - Transparency	Policies and other measures that are proposed or enacted in response to deepfakes. This includes policies that apply to false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly, as these policies will have implications for deepfakes as well. Specifically, this category is concerned with measures that aim at achieving greater transparency on various levels.	Codes are applied where documents propose, call for or establish measures that 1.) are directed at deepfakes, false information online, or other related elements that are covered in the respective category in relation to those phenomena and 2.) try to achieve greater transparency regarding the conduct of involved actors.

Appendix 5 – Case Descriptions

Ethics Guidelines for Trustworthy Artificial Intelligence by the High-Level Expert Group on Artificial Intelligence (HLEG AI 2019)

Similar to the communications by the European Commission on AI, the guidelines by the HLEG strike a markedly different tone compared to the analyzed documents that specifically pertain to disinformation. Whereas references to potential harms to fundamental rights in other documents had a markedly epistemic dimension – freedom of expression and information, democratic participation etc. – the references to fundamental rights in these guidelines highlight this dimension far less prominently. Instead, fundamental rights are to be understood in the context of human dignity, agency, privacy, discrimination and health (ibid.: 10f, 15ff). Nonetheless, potential impacts of AI systems on democratic processes are also considered, though only to a limited extent (ibid.: 19).

In line with this considerably broader scope of the ethics guidelines, the measures proposed by the ethics guidelines are more focused on facilitating transparent decisions by AI systems, which is less relevant in the context of deepfakes, and holding providers of AI systems accountable (ibid.: 9). Nonetheless, the application of the ethics guidelines by providers of AI systems can enable them to see potential risks their systems might entail, including the risk of deception and manipulation (ibid.: 12).

Working Paper on AI and the Future of Democracy (AIDA 2021)

The AIDA working paper documents the insights gathered by a joint meeting of the parliamentary committees on 'Artificial Intelligence in a Digital Age' and 'Foreign Interference in all Democratic Processes in the European Union including Disinformation' investigating the intersection of AI and disinformation across two panels on AI and democracy and regulatory approaches to technological developments (ibid.: 2). As such, this document presents a rare case in which both issues of AI and disinformation are centered at the same time.

The main insights and measures to mitigate the problems that emerge from the intersection of disinformation and AI are as follows. Participants see the main epistemic harm emerging from threats to democracy.

> "*Information critical to the survival of democracies cannot compete with the motivated disinformation propelled by increasingly influential platforms and AI-powered algorithms*." (AIDA 2021: 4)

Participants further see increases in the access to, as well as the commodification and presence of synthetic audiovisual media content online (ibid.: 5) and a more pronounced role of domestic actors (ibid.: 5) in the spread of disinformation.

Correspondingly, the panelists suggest that the standing of accurate information and media pluralism in digital information environments needs to be bolstered. This should be achieved through supporting independent journalists and media organizations (ibid.: 4) and (automated) content moderation (ibid.: 4).

At the end of the documents, representatives of the parties were asked to contribute a short statement on their party's stance toward the issues raised in the panels. (ibid.: 6f). Parties mainly stress the significance of the phenomenon for the integrity of democratic institutions and fundamental rights, while some – namely the 'European People's Party Group', the 'Progressive Alliance of Socialists and Democrats Group', and the 'Greens/European Free Alliance' – also highlight the relevance of freedom of expression.

Artificial Intelligence Act (EC 2021a)

The AIA (ibid.) was proposed by the European Commission as "the first ever legal framework on AI" (EC 2021b) in April 2021. Based on a whitepaper published in 2020 (EC 2020), the AIA is supposed to embody the long-term strategic vision of the EU with respect to AI (EC 2021b). More specifically, the proposed framework contains a legal definition for what is to be considered AI (EC 2021a, Article 3), and varying obligations depending on the kind of AI under question. Systems that not qualified as high-risk or prohibited only have to let users know they are interacting with or seeing the output of an AI system, including labelling for deepfakes (Art. 52). Additionally, the providers of low-risk systems may voluntarily enter into codes of conduct, rendering them susceptible to the increased obligations of high-risk systems (Art. 69).

AI systems are prohibited if they exploit their users' disabilities, are used for social scoring, engaged in real-time remote biometric identification (Art. 5(1b-d) or "[deploy] subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm" (Art. 5(1a)). Systems are classified as high-risk if they are listed in

Appendix III of the AIA (Art. 6). These are systems in the area of biometric identification and categorization of natural persons, critical infrastructures, education, employment, access to essential services and benefits, law enforcement, migration and border control, administration of justice and democratic processes (EC 2021e: 5). This list can be amended by the commission either because they are used in the areas covered by Appendix III (ibid.), or because "the AI systems pose a risk of harm to the health and safety, or a risk of adverse impact on fundamental rights [...] equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III" (Art. 7). As there is a case to be made that deepfakes pose risks that align with those of prohibited AI systems under Art. 5(1a), or high-risk AI systems under Art. 7(1b), it is unclear whether the classification of DFAs will remain in the category of low-risk in the long run (see also van Huijstee et al. 2021: 37f). Systems that are categorized as high-risk have additional obligations such as conducting risk and conformity assessments (Art. 9, 19), either themselves or through an independent auditor (Art. 33, 43), provide (technical) records of their systems for the purpose of monitoring conformity (Art. 11, 12), and human oversight (Art. 14). Further, high-risk systems are listed in a European database (Art. 60).

The AIA is implemented and monitored through national supervisory authorities (Art. 59) and coordinated by the European Artificial Intelligence Board (Art. 56-58). These institutions further monitor AI systems that were brought to market (Art. 61) and share information regarding incidents and malfunctions where obligations have been breached (Art. 62). In case of infringements upon the laid-out obligations, the providers of AI systems are to pay fines "up to 30.000.000 EUR or, if the offender is company, up to 6 % of its total worldwide annual turnover for the preceding financial year, whichever is higher" (Art. 71).

Digital Services Act (Regulation (EU) 2022/2065)

Adopted in October 2022, the DSA supersedes the e-Commerce Directive (Directive 2000/31/EC). It is supposed to, among other things, protect users (e.g. safeguarding of fundamental rights, reducing their exposure to illegal content) and increasing the obligations of platforms (transparency and accountability frameworks). The obligations of platforms progressively increase depending on whether they are classified as intermediary services (e.g. domain name registrars), hosting services (e.g. webhosting services), online platforms (e.g. marketplaces, social media platforms) or very large online platforms (online platforms whose EU customer base exceeds 10% of the EU population)(EC n.d. b).

The DSA is not primarily occupied with disinformation but rather with 'illegal content', referring to "any information that [...] is not in compliance with Union law or the law of any Member State [...] irrespective of the precise subject matter or nature of that law" (DSA Art. 3(h)). Should content qualify as illegal in this sense, the DSA presents 'intermediary services' (Art. 3(g, i)) with considerable obligations to act against it, e.g. through limiting access to their service or deleting content (Art. 23, 32). Focus of the DSA, and most relevant for the present research interest, are very large online platforms, defined as those platforms which have monthly average more than 45 million users or more than 10% of the EU population (Art. 33). Particular attention is further also devoted to offering users of intermediary services means to appeal any decision that the providers of such services (e.g. social media platforms) enact against them (Art. 20) and that intermediary services be impartial when enacting decisions and processing appeals. Intermediary services that violate the obligations set out by the DSA face up to "6 % of the annual worldwide turnover of the provider of intermediary services concerned in the preceding financial year" (Art. 52, 74).

To monitor enforcement and compliance, the DSA creates the new institutions of the Digital Services Coordinator in the member states and the EU (Art. 49-51), as well as a coordinating board (Art. 61-63), who are equipped with the power to make intermediary services provide a range of information regarding their services and conduct (see e.g. Art. 65, 85). Independent compliance audits are also part of the monitoring and enforcement schemes (Art. 37, 72). In addition, some of this information is made available for researchers (Art. 40 (8-13)). With respect to deepfakes, the DSA requires synthetic audiovisual media to be labelled (Art. 35(1k)). Intermediary services are further encouraged to enter into codes of conduct with each other to further specify the measures of how to enact compliance with the DSA (Art. 45). However, importantly, there is no obligation for intermediary services to actively engage in the seeking-out of illegal content in the DSA. Instead, intermediary services only need to act against it once it is brought to their attention.

Strengthened Code of Practice on Disinformation and *Implementation Report* (EC 2022a, ERGA 2023)

The sCoP (EC 2022a) presents a voluntary framework for actors in the digital information ecosystem – platforms, advertisers, technology providers, researchers, fact-checking and journalistic organizations, and civil society actors – to engage around issues of disinformation. The EC initially proposed a Code of Practice on Disinformation in 2018 (EC 2018a). However,

after considerable critique of the effectiveness of the initial code (ERGA 2020; Plasilova et al. 2020) it has been revised. The 2022 code has 34 signatories across industry, academia and civil society who agree on self-regulatory standards in an effort to combat disinformation (EC n.d. c). Signatories include Google, Meta and TikTok (EC 2022b).

Generally, there is a threefold focus to the measures in the sCoP. First, actors create increased situational awareness through gathering information on disinformation and sharing this information with other stakeholders, including researchers and fact-checkers (ibid.: 8f). Second, the sCoP specifies a variety of content moderation measures for disinformation. These include fact-checking, labelling (ibid.: 10f; 21ff) and verifying the sources of information (ibid.: 11f). These main measures are further flanked by demonetizing disinformation (ibid.: 5), deletion disingenuous accounts, increasing the reach of accurate information and providing media literacy trainings to recipients (ibid.: 19). Lastly, the sCoP specifies transparency obligations and monitoring mechanisms for nearly every commitment. The implementation of the code is further overlooked by the European Regulators Group for Audiovisual Media Services who publishes yearly reports on its implementation (ERGA 2023).

In their first evaluation of the sCoP, ERGA highlighted the efforts undertaken by signatories so far, but points out that across the board, there is room for improvement. Platforms still have ways to go in their efforts to e.g. set up transparency centers, data access, providing country-specific data and how they report their progress (ERGA 2023).

The sCoP might be elevated from a voluntary code of practice into a co-regulated code of conduct under the DSA (Regulation (EU) 2022/2065, Art. 45). As the DSA itself is occupied with illegal content (Art. 2, 3), not per se with disinformation, the sCoP presents arguably the central (self-)regulatory instrument in the disinformation space. Nonetheless, as the commission is highly invested in monitoring the sCoP and might be prepared to follow a more stringent regulatory strategy should it fail to provide the desired results, it should not be underestimated as a governance framework.

Report by the High-Level Expert Group on Fake News and Disinformation (HLEG FNOD 2018)

The HLEG on Fake News and Online Disinformation was initiated in 2018 to “contribute to the development of an EU-level strategy to tackle the spreading of fake news and disinformation” (EC n.d. c). In it, 40 representatives of social media platforms, the media,

academia, civil society and citizens came together to gather opinions on suitable measures (ibid.). At the end of 2018, the HLEG published a report to this end (HLEG FNOD 2018).

The report stresses the necessity for stakeholders to collaborate (ibid.: 5) and suggests a self-regulatory approach to disinformation, with stakeholders committing to a Code of Practices, which was realized to the initial Code of Practice (EC 2018a) and continues in the sCoP (EC 2022a). Primarily this report is concerned about the harms disinformation may pose to epistemic goods such as information and media pluralism and the integrity of democratic institutions (HLEG FNOD 2018: 5). As remedies, the high-level expert group suggests measures that enhance the situational awareness regarding disinformation in digital information environments, such as information sharing between stakeholders and enhanced transparency by platforms regarding displayed contents (ibid.: 22ff). Secondly, the HLEG recommends enhanced efforts in media literacy training (ibid.: 25f) and supporting independent journalists and media organizations (ibid.: 27-30). All the while, the HLEG is concerned with the impacts any given measure against disinformation may have on freedom of expression (ibid.: 18ff, 31f).

Tackling Deepfakes in European Policy (van Huijstee et al. 2021)

Published in 2021, this report was produced by a team of independent researchers for the Panel for the Future of Science and Technology (STOA)(ibid.). In its own right, the report presents an extensive analysis of the overall phenomenon of deepfakes, including technological background, societal implications, applicable policies and suggested measures on the basis of the scientific and professional literature on deepfakes, relevant policies and expert interviews (ibid.: I). The report is still the most comprehensive publicly available policy document specifically on deepfakes that has been produced by/for an EU institution to date. The report provides a summary of the technological underpinnings of deepfakes, covers a broad variety of impacts and gives numerous policy recommendations. Additionally, it critically reflects the efficacy of potential policies.

In terms of understanding of epistemic harm, the report echoes the overall perspective of the analyzed policy documents and focusses mainly on deception-based harms, next to fundamental rights. However, the report also calls attention to the potential of deepfakes to cognitively resonate with their recipients. There is no specific set of policy measures to mitigate the epistemic harm of deepfakes that stand out, instead the report gives a very broad overview of a variety of measures, including deepfakes as high-risk or even prohibited systems in the

proposed AIA (EC 2021a), international sanctions, automated detection (though the report overall has critical perspective on its efficacy (van Huijstee et al. 2021: 14)), labelling both of deepfakes and trustworthy sources, restricting the reach of deepfakes, empowering the subjects of deepfakes and media literacy training (ibid.: VII, VIII).

Disinformation and Propaganda: Impact on the Functioning of the Rule of Law and Democratic Processes in the EU and its Member States (Bayer et al. 2021)

Bayer and colleagues (2021) study the disinformation landscape in the European Union from 2019 to 2021. Through this, the authors want to current trends in disinformation practices and suggest a variety of measures to target the overall phenomenon. The authors find that current disinformation practices and influence operations rely more heavily on domestic actors and the impact of disinformation is especially significant when it is spread or perpetuated by figures of political authority (ibid.: 12f). Generally, the authors see considerable threats to fundamental rights from disinformation, including threats to live and livelihoods during the COVID19-pandemic (ibid.: 14), as well as to the integrity of democratic institutions (*e.g.* ibid.: 20, 113).

The authors also assess the psychological mechanisms of disinformation. Next to characteristics of a given narrative and repetition, they point toward the significance of the trustworthiness of a source, personal characteristics of a recipient, as well as a recipients' general level of trust (ibid.: 13). Building on this assessment, the policy measures suggested to mitigate the harms from disinformation center around media literacy, counter-narratives and fact-checking. However, the authors also point out that media literacy education, content moderation and measures designed to increase the exposure of recipients to accurate information – e.g. through supporting journalism – depend on how trusted the administration institutions are (ibid.: 63ff). As recipients who are susceptible to disinformation are likely to distrust such institutions, the authors recommend to further invest in (re-)building recipients' trust in the long term (ibid.: 102, 108). Further, media literacy trainings and counter-narratives should, according to the authors, be tailored toward the addressed audience (ibid.: 14; 107f).

Aside from measures targeting recipients, the authors focus on measures that increase the access to accurate information in digital information environments, e.g. by supporting existing trustworthy actors (civil society actors, educators, researchers, journalists, fact-checkers) and bringing them together into collaborative networks and putting out 'affirmative information' (fact-checked counter-narratives by trustworthy institutions with amplified reach).

Best Practices in the Whole-of-Society Approach in Countering Hybrid Threats (Wigell et al. 2021).

Disinformation can be deployed by foreign actors as part of hybrid threat strategies. In their study, Wigell and colleagues analyzes the current state of hybrid threats, as well as offers a variety of measures to facilitate societal resilience against disinformation in the form of 'hybrid interference' (Wigell et al. 2021: 14ff).

Special emphasis is given to the need for institutions of the member states and the European Union to cooperate, share information and increase their respective capacities to mitigate and increase resilience against hybrid threats. The authors recommend that policymakers put in place measures that help to identify hybrid threats and implement a so-called whole-of-society approach to hybrid threats. This includes proactive measures such as supporting civil society, independent media organizations and journalists, transparency obligations for social media platforms, awareness raising and media literacy training (ibid.: 41). Further, it includes putting foreign actors under sanction and deploying counter-narratives to foreign influence operations (ibid: 42).

With respect to counter-narratives and media literacy, the authors point to the need of targeting groups that are particularly vulnerable to disinformation narratives, e.g. through tailoring both measures to their needs and actively including members of those communities in their implementation. They also point to the need of actively (re-)building trust in and mitigation given socio-economical inequities in such communities (ibid.: 16, 22).

Appendix 6 – Final Category System

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Document Type	Categorization of the analyzed documents.	EC Communication	Document published as a communication by the European Commission.	x	x
		Regulatory Framework	Proposed or already enacted regulatory frameworks.	x	x
		Independent Council	Studies commission by EU institutions but carried out by independent researchers.	x	x
		Other	Documents that do not fit the other categories.	x	x
Year	Year the document was published.	2018	Published in 2018.	x	x
		2019	Published in 2019.	x	x
		2021	Published in 2021.	x	x
		2022	Published in 2022.	x	x
		2023	Published in 2023.	x	x
Main Issue	Main issue of interest in the respective documents.	MI Deepfakes	The main issue of the document are deepfakes.	x	x
		MI Disinformation	The main issue of the document is disinformation.	x	x
		MI Artificial Intelligence	The main issue of the document is AI.	x	x
Cautions	Segments in which various relevant aspects are reflected on critically. Codes included in this category occurred more than five times.	c. accountability	Critical reflections on holding actors accountable.	x	x
		c. automated detection	Critical reflections on automated detection.	x	x
		c. cooperation among stakeholders	Critical reflections on cooperation of various stakeholders.	x	x
		c. fact checking	Critical reflections on Fact-checking.	x	x
		c. information sharing	Critical reflections on information sharing.	x	x
		c. labelling	Critical reflections on labelling.	x	x
		c. monitoring enforcement / compliance	Critical reflections on monitoring the compliance of various actors to (proposed) measures.	x	x
		c. self-regulation	Critical reflections on the efficacy of self-regulation by platforms, advertisers, or providers.	x	x
		c. transparency	Critical reflections on measures that aim to enhance transparency.	x	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Relevant Context	Segments in which context is given to the issues of deepfakes, disinformation or artificial intelligence.	Changing media environment	Changing media environment' refers to mentions of the broader trend of changes in the media environment due to rapid technological change, the democratization of content production, and the challenges and opportunities brought about by digital platforms.	x	x
		Cyber security	Cyber security' refers to mentions of the broader issues of cyber security, e.g. the threat of hacking.	x	x
		Declining public trust	Declining public trust' describes the mention of the broader societal trend of a decline in trust in public institutions, democratic systems, experts, politicians and politics in general.	x	x
		Disingenuous accounts	Disingenuous accounts that amplify the reach of content online. Usually, this is achieved through fake online profiles or bots which engage with this content in this effort, e.g. through liking, sharing, commenting or reposting such content. Also includes trolls.	x	x
		Increase in false information online	Refers to mentions of the broader (perceived) phenomenon of an increase in the prevalence of false information online.	x	x
		Increasing access to deepfake technology	Refers to the (perceived) increased access to the means of deepfake production for interested parties.	x	x
		Increasing importance of visual communication	Refers to the (perceived) increase of importance of visual media to communicate information and media in general.	x	x
		Normalization of manipulated media	Mentions of the increasing prevalence and normalization of manipulated audiovisual media in digital information environments, e.g. through filters on social media apps.	x	x
		Political micro-targeting	Refers to the practice of tailoring political messages to specific segments of the population based on a variety of individual-level characteristics.	x	x
		Commercialization of deepfakes / disinformation	Commercial services that offer the production of deepfakes or access to their proprietary deepfake technology.	x	x
		Sensationalism	Refers to mentions of the broader (perceived) dynamic of increased sensationalism in media, media consumption and public discourse.	x	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Elements Addressed	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment.	Actors	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, actors involved.	Recipients	Recipients of content, potentially including false information, online.
				Subjects	Subjects are who false information online (disinformation, misinformation, fake news, deepfakes) or the products of artificial intelligence is about. They may be existent or non-existent persons or non-humans (e.g. animals, objects or buildings).
				Non-subject rights holders	(Legal) authors of input data that was (recognizably) part of the generation of a given deepfake but is not the source of the features by which the subject is identified (e.g. facial features).
				Foreign actors as producers	Producers of false information online that are (traceably) connected with a foreign state.
				Domestic actors as producers	Actors who produce false information online (disinformation, misinformation, fake news, deepfakes) and are affiliated with an EU member state.
				Producers generally	Refers to the producers of false information online (disinformation, misinformation, fake news, deepfakes etc.) without further specification.
				Platforms	Platforms are digital information environments in which users are able to publicly share multimedia content with each other. Usually, users would also be able to comment on and share content from other users. There is some uncertainty on the margins what constitutes a platform, but the paradigmatic case are large social media platforms such as Facebook and TikTok or fora such as Reddit.
				Advertisers	Actors who advertise on online platforms.
				Technology providers	Refers to the providers of the knowledge or technologies necessary to produce deepfakes or deepfake applications, the providers of either technologies involved in creating and spreading false information online.
				Civil society	Actors in civil society involved in various issues around AI, disinformation or deepfakes. This includes e.g. NGOs or activists.
				Media organizations and journalists	Established media broadcasters, such as TV or radio stations, newspapers etc. This may also refer to their online presences on other platforms.

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Elements Addressed	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment.	Technology	Elements relevant to digital information environments, deepfakes, artificial intelligence or false information online that are being addressed in a given segment. Specifically, aspects of the technology involved.	Artificial intelligence	Pertains to segments that mention artificial intelligence systems, including generative models.
				Artificial intelligence techniques	Knowledge about AI techniques and architectures that are suitable to achieve a given purpose, e.g. creating a deepfake.
				Deepfakes generally	Refers to deepfakes broadly, without a specified format.
				Audio-based deepfakes	Deepfakes that are purely in audio format.
				Image-based deepfakes	Refers to deepfakes in (single) image format.
				Video-based deepfakes	Refers to deepfakes in video format. Usually, these deepfakes are combined with deepfaked audio as well.
				Synthetic audiovisual media generally	Pieces of audiovisual media that are partially or entirely machine generated.
				Input data	Data necessary to train deepfake applications.
				Direct messaging	Direct messaging services for communication between individuals or non-public groups.
		False Information Online	Among the topics that are addressed in the respective documents are issues pertaining to false information online, including disinformation, misinformation and fake news. Specifically, different kinds of false content online.	Disinformation	False information produced and published with malicious intent. May include deepfakes.
				False information generally	False information online in general, irrespective of intent.
				Hybrid threats	Various measures that are being combined by a foreign adversary to exploit vulnerabilities in the target while keeping below the threshold of warfare. This is not limited to but also includes disinformation campaigns as part of influence operations.
				Influence operations and foreign influence operations	Attempts to influence political decision, e.g. an election, through covert means. Usually, these are undertaken by foreign adversaries but incorporate domestic actors. This includes, but is not necessarily limited to, spreading disinformation or otherwise amplifying ideas of domestic actors that are favorable to the adversary.
		Other	x	x	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Understanding of epistemic harm	Assessment which understanding of epistemic harms caused by false information online (disinformation, misinformation, fake news, deepfakes) is present in the analyzed policy document. Epistemic harms are defined as follows: Epistemic harm is caused if an agent obstructs, without legitimizing reason, the epistemic success of another in the area of politically relevant information	Deception	Refers to the notion that false information online leads to recipients adopting false beliefs.	x	x
		Jeopardizing Evidence	Refers to the undermining effect that false information online may have to evidentiary practices, e.g. of recordings in criminal procedures, but also public discourse.	x	x
		Erosion of Trust	Refers to the loss of trust in epistemic institutions, practices and the competence of others as a result of false information.	x	x
		Cognitive Resonance	Refers to the acknowledgement that false information online may not only deceive recipients through causing a false belief but may rather also be problematic in the notions and affects it evokes as well as being useful to recipients as faux-justifications.	x	x
		Polarized Fellowship	Refers to the notion of false information online signaling the trustworthiness and group-membership of a producer in order to form epistemically significant community with recipients.	x	x
		Polarization	Mentions of societal polarization without alluding to the specific dynamic of polarized fellowship.	x	x
		Manipulation	Mentions of the capacity of false information online being manipulative.	x	x
		Epistemic Goods	Epistemic goods positively contribute to people forming accurate beliefs. This may be because recipients are exposed to accurate information and diverse viewpoints through a pluralistic information environment in which knowers and other discourse participants can freely express themselves and recipients are free to seek out information. References to these epistemic goods, where they are not mentioned to be under threat directly, present an implicit understanding of epistemic harm from deepfakes and disinformation insofar as they describe what is put at risk through those phenomena.	Epistemic goods generally	General mentions of epistemic goods in the context of the potential impacts disinformation may have.
				Fundamental rights generally	General mentions of fundamental rights in the context of the potential impacts disinformation may have. These mentions need not always pertain to epistemically significant fundamental rights (e.g. freedom of expression) in particular.
				Democratic integrity	Democratic integrity' describes the functioning of key democratic institutions such as elections, legislative procedures, due process, and equality before the law. It also includes the recognition of the legitimacy of the democratic system overall in the eyes of the public. Insofar as democratic institutions are suitable to facilitate epistemically sound decisions, presenting a collective epistemic harm.
				Information and media pluralism	Reference to the role of information and media pluralism

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Understanding of epistemic harm	Assessment which understanding of epistemic harms caused by false information online (disinformation, misinformation, fake news, deepfakes) is present in the analyzed policy document. Epistemic harms are defined as follows: Epistemic harm is caused if an agent obstructs, without legitimizing reason, the epistemic success of another in the area of politically relevant information.	Epistemic Goods	Epistemic goods positively contribute to people forming accurate beliefs. This may be because recipients are exposed to accurate information and diverse viewpoints through a pluralistic information environment in which knowers and other discourse participants can freely express themselves and recipients are free to seek out information. References to these epistemic goods, where they are not mentioned to be under threat directly, present an implicit understanding of epistemic harm from deepfakes and disinformation insofar as they describe what is put at risk through those phenomena.	Freedom of expression	Freedom of expression' describes citizens' right to speak their mind in public discourse. It is recognized as a fundamental right in the European Union in Article 11 of the EU Charter of Fundamental Rights. However, as such it is constrained when it conflicts with other fundamental rights.
				Freedom of information	Freedom of information' describes a citizens' right to receive and impart information and ideas without the interference of a public authority inscribed in Article 11 of the EU Charter of Fundamental Rights.
		Other	x	x	x
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Accountability	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that hold actors involved in these phenomena for certain kinds of misconduct or that introduce new liabilities.	Accountability generally	General reference to the introduction of liabilities or other means of holding actors involved in issues of deepfakes, AI or disinformation accountable.
				Provider accountability	Measures that introduce liabilities for the providers of deepfake technology or means to hold them accountable for an established kind of misconduct.
				Platform accountability	Measures that introduce liabilities for platforms or means to hold them accountable for an established kind of misconduct.
				Producer accountability	Measures that introduce liabilities for producers of false information online or means to hold them accountable for an established kind of misconduct.
				International sanctions	The imposition of sanctions as a result of (a specific way of) spreading false information online.
				Co-regulation	Introducing a co-regulatory scheme for actors involved in digital information environments or AI e.g. through a code of practice or a code of conduct.
				Other	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Content moderation	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with how content in digital information environments is moderated by various actors. Moderation can entail contextualization of information, e.g. through fact-checking and labelling of online content, or the deletion of content.	Appeals	Means for those who find themselves at the receiving end of a content moderation decision, e.g. through having their account or a post deleted, to appeal their decision and have their account or content reinstated or receive an alternative remedy.
				Automated detection	Automated systems that are capable of detecting false information online or deepfakes in order to either remove them, flag them as false or to make them the subject of other kinds of content moderation.
				Content moderation generally	Content moderation in general, usually done by professionals. Instances where user reporting was explicitly mentioned are coded separately.
				Content moderation neutrality	Platforms (or contracted contact moderators) are mandated to moderate contract in a manner that is unbiased, non-discriminatory and ideologically neutral.
				Deletion of deepfakes, false information and disingenuous accounts	Deletion of false information online or disingenuous accounts upon discovery.
				Demonetizing	Cutting of monetary revenue from content or advertisements spreading false information online.
				Fact-checking	Fact-checking refers to calls for assessing the accuracy of information online and attempts to correct false information through providing accurate information on the matter. Fact-checking is usually done by journalists, academics, or civil society actors.
				Labelling	Labelling false information online as (potentially) false upon discovery.
				Source and authenticity verification	Measures that are designed to verify and indicate the trustworthiness or authenticity of a source. This may take the form of providing means to assess whether a poster online is who they claim to be or consist in certain technological markers that signal that footage is indeed authentic.
				Stunted visibility	Stunting the spread of false information online by de-emphasizing its algorithmic rollout to potential recipients on platforms.
				User reporting	Content moderation that is done by users, e.g. through flagging content that they perceive to be fake.
				Other	x

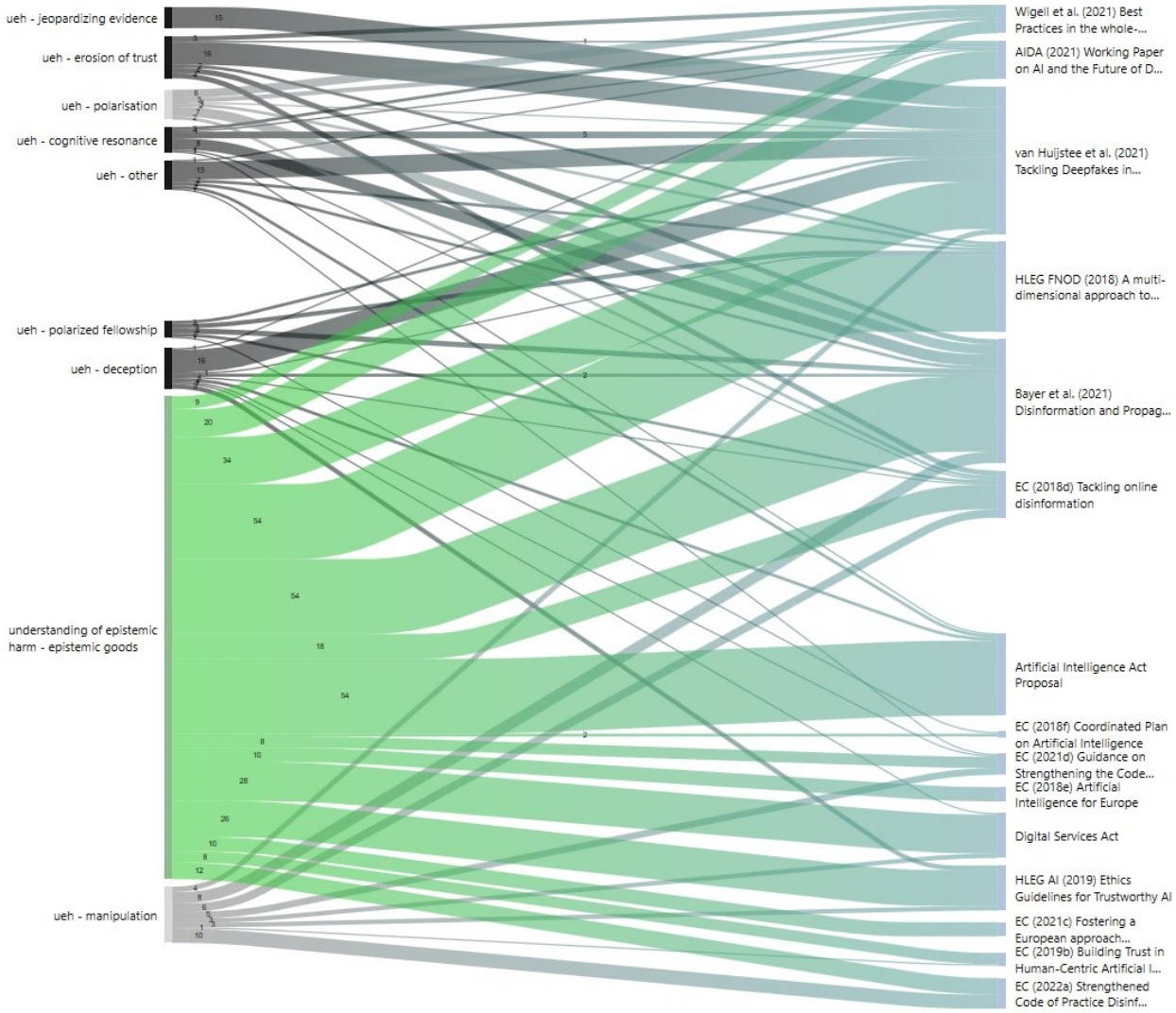
Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Other	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly but do not fit the other categories.	Cooperation among stakeholders	Cooperation between different stakeholders, e.g. platforms, state institutions, researchers, civil society, around issues of false information online and AI.
				Cooperation between institutions	Coordinated action between various institutions for varying purposes around false information online and AI. This may include institutions on the national or European level for purposes such as gaining better situational awareness, enforcement or sanctions.
				International agreements and cooperation	Cooperation with other states and instructions on issues of false information online and AI beyond EU member states.
				Strengthening institutional capacity	Various measures designed to increase the capabilities of institutions that are involved with false information online or artificial intelligence. This may include increasing their budget, providing trained personnel, training existing personnel, or increasing competencies.
				Institutional support for subjects	Institutionalizing support structures for the subjects of false information online.
				User choice	Measures to give users or increase their ability to influence which content they are being presented with in digital information environments.
				Counter-narratives	Publishing of persuasive messages designed to counter disinformation narratives online.
				Reforming or clarifying the application of existing regulation to deepfakes or other false information online	Extending the application of existing regulations or clarifying how it is to be applied to new phenomena such as deepfakes.
				Other	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Preventative measures	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that preempt the negative effects of false information online (disinformation, misinformation, fake news, deepfakes).	Raising awareness and media literacy training for recipients	Measures that are raising awareness or increasing the media literacy of potential recipients.
				Raising awareness and training for professionals	Measures that are raising awareness in professional circles, e.g. content moderators, law enforcement, journalists or public servants.
				Supporting journalism	Measures designed to support trustworthy journalistic work, e.g. through providing professionals more access to information, providing funding, securing revenue streams, increasing independence or protecting journalists from (threats of) violence.
				Improving access to accurate information	Measures that attempt to, in some capacity, boost exposure to information from trustworthy sources, e.g. journalists or institutions.
				Building trust	Mention of the need to build trust in public institutions and epistemic practices.
				Addressing recipient vulnerabilities	Measures that address the social, cultural or economic factors that might make some recipients susceptible to false information online.
				Identity authentication on platforms	Requiring users of platforms to authenticate themselves through providing a proof of identity, e.g. their ID.
				Banning (certain) DFAs or other technologies	Banning the use or provision of (certain) DFAs or other (AI) technologies in the context of deepfakes or false information online, rendering their use or provision in the EU illegal.

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Preventative measures	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that preempt the negative effects of false information online (disinformation, misinformation, fake news, deepfakes).	Adversarial attacks	Manipulation of potential input data to sabotage the production of a deepfake. The manipulation is (mostly) unnoticeable for a human observer but mislead an AI model regarding the subject of an image.
				Restricting knowledge on (certain) AI techniques that enable deepfakes	Restriction making the knowledge necessary for building DFAs available to the public, e.g. through (academic) publications or the provision of modular solutions on e.g. GitHub.
				Skepticism toward recordings	Suggestion to or education of recipients to the effect that they revise the epistemic status they attribute to recordings, e.g. by affording them less default credibility.
				Preventative measures generally	Measures that are supposed to act preventatively against the harms caused by false information online without specifying further.
				Other	x
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Situational awareness	Policies and other measures that apply to deepfakes, or false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly. Specifically, this category is concerned with measures that aim at producing a precise understanding about how these phenomena exist in the world and how regulation and guidelines that are relevant to them are applied and take effect.	Information sharing	Measures that make information regarding events in digital information environments accessible to actors other than those who are gathering this data. Also includes sharing of insights gathered through investigations, best practices, etc.
				Research	(Support of) research into various aspects of the phenomena of false information online or (generative) artificial intelligence. This may include automated deepfake detection, or research into suitable responses to or the current state of false information online.
				Situational awareness generally	Establishing situational awareness without further specification.
				Other	x

Category	Category definition	Code	Code-Definition	Subcode	Subcode-Definition
Policy Measures	Either proposed or effective policies that target misinformation broadly or deepfakes in particular. Measures may only attach toward one element in particular, e.g. advising caution regarding sharing data will target the potential subjects of deepfakes.	Transparency	Policies and other measures that are proposed or enacted in response to deepfakes. This includes policies that apply to false information online (disinformation, misinformation, fake news) and artificial intelligence (generative artificial intelligence, synthetic audiovisual media) broadly, as these policies will have implications for deepfakes as well. Specifically, this category is concerned with measures that aim at achieving greater transparency on various levels.	Transparency generally	Measures that are designed to increase transparency in general, without specifying further.
				Platform transparency	Measures that are designed to increase transparency on the conduct of platforms.
				Provider transparency	Measures that are designed to increase transparency on the conduct of the providers of technologies used for the production of deepfakes, disinformation or AI technology in general.
				Producer transparency	Measures that are designed to increase transparency on the conduct of the producers of false information online, e.g. requesting that they label their content.
				Compliance monitoring	Measures suitable to monitor how various actors enforce or comply with applicable regulations and guidelines.
				Compliance monitoring	Measures suitable to monitor how various actors enforce or comply with applicable regulations and guidelines.
				Risk and impact assessments	Platforms and providers creating risk and impact analysis on their products which includes whether they might eb conducive of or cause epistemic harm, e.g. through an impact on fundamental rights.
				Independent auditing	Monitoring whether and to what extent the conduct of platforms and providers is compliant with their respective obligations.
				Other	x
		Other	x	x	x

Appendix 7 – Sankey-Diagram Understandings of Epistemic Harm



This table only serves to illustrate this point as the insights that can be gained from quantifying the results of QCA are very limited (Kuckartz, Rädiker 2023).

Appendix 8 – Sankey-Diagram Preventative Measures

This table only serves to illustrate this point as the insights that can be gained from quantifying the results of QCA are very limited (Kuckartz, Rädiker 2023).