MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

the political deepfake:

a dance between conceptions, materialisations, and policy approaches

verfasst von / submitted by

Auriane van der Vaeren, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 906
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Science-Technology-Society
Betreut von / Supervisor:	Univ.-Prof. Sarah R. Davies

A C K N O W L E D G E M E N T S

Whereas I initially intended this thesis to be a sort of ultimate means to convince an audience of my intellectual capacity, it gradually transitioned to be a quest to self-expression. In many ways, the thesis tied in to an earlier quest of better finding my self. But such instance of converging interests is not happening just like that, out of the blue. The entire environment needs to allow for it to happen. However infinitesimally dispersed across space, time, matter, it are the elements of my environment that made this thesis become what it is. A thesis that would thus not have liberated the same flavours without the support that I have received.

I first and foremost wish to sincerely thank Professor Sarah R. Davies for her guidance all along this project and for her kindness and generosity upon allowing me to explore this project in all its mysterious ways. Truly, I do not believe that this project would have blossomed into a quest to self-expression as much as it did otherwise.

I further wish to thank the Department of Science and Technology Studies of the University of Vienna for critically enlightening some of the previously unexplored corners of my mind. As well as my interview partners for sharing their expertise and thereby further enlightening my mind—Professor Noah Giansiracusa, Professor Britt Paris, Global Disinformation Index. And Anke Obendiek for her early help on the policy proposal in this thesis.

I would like to dedicate special thanks to my peers, Celine, Fabian, and Luise, for their insightful feedback. And particularly to my partner, Sylwia, for her thorough reading and for having helped me to appreciate what it means to write clearly all the while remaining true to myself.

About the personal blossoming that happened throughout the thesis—whether it was with it, because of it, or stimulating it—it would by far not have been possible without the guidance of Pascale who opened me to new dimensions of my psyche and my being.

Of course, I would not be here without my parents. And I would not be the way that I am without the help of my brothers. All key figures in my early experiences of life. To all my friends that continue to bend my mind in ways that allow me to better appreciate life's exquisite texture and crunchiness. To the bullies that I encountered, despite (rehearsing) the traumas, I thank you for shaping me as I am. And it would be dishonest not to dedicate this thesis also to the victims of my unawareness who are important parts of the incremental maturation of my consciousness.

Lastly, who would I be as a Science-Technology-Society student without ending with a big shoutout to all the nonhumans that define me or shaped me in what I am today—the gut bacteria, the atmosphere, the chemical lab vials, the concrete roads, the mosquitoes, the taste buds, the brewery plants, the milky way.

To all the encounters that make of this incidental existence a moment of singularity.

Cheers.

T A B L E O F C O N T E N T S

I N T R O D U C T I O N

Wooooooow … wait, what?! Really? Is this true? Is this reality? Can I trust this? … Many claim that such constant state of doubt and distrust could become the new norm if political deepfakes were to freeride the online space. Political deepfakes—these seemingly authentic fake images, videos, audio files, or texts that would wreak havoc and undermine trust in the politics by spreading disinformation[1]. But despite the sensational sapidity of such prospects, still today, the effects of disinformation and political deepfakes are complex and understudied (Center for Information, Technology, and Public Life, 2022; Kwok & Koh, 2020). The question that thus initially spurred my curiosity for this topic was whether such deepfakes truly embody the potential for a generalised democratic bad trip. A curiosity that then little by little morphed into the present project which is about understanding how our conception of deepfakes influences our responses to their existence, and thus how our conception influences their regulation.

Given the writing style of this first paragraph, to prevent eventual further surprises, it proves useful to note that this thesis is not written according to the typical textbook writing rules. At times it will be more free-spirited. But it does not mean that it is less academically relevant as such. The thesis very much applies academic rigour and shall therefore remain intelligible and scientifically founded. What the sometimes more immersive writing style does, is that it applies the logic of freedom of expression to gently (t)ease the standards by which academic knowledge is produced—the standards by which knowledge is created, assembled, and delivered. It is not about shaking academic practices to their core. It is simply about aligning the thesis format to its substance[2]. This thesis explores how conceptions of reality feed our practices that go about that reality—how conceptions of the deepfake feed its policy approaches. It is thus similarly interesting to explore how conceptions of knowledge feed our practices of knowledge production. Knowledge is partial. Thence, so too are its production practices. By acknowledging that partiality, and by being part of the supply chain of knowledge through the very act of delivering this thesis, it becomes my responsibility too to tease our present-day practices of knowledge production. Because why would any proposed knowledge be rejected for the sole reason of its medium rather than its substance? Let's see if I manage to convince you of the validity and legitimacy of this—perhaps to some—queerer thesis genre.

Having settled the case for the thesis style, we can come back to the subject matter, the political deepfake. Let us therefore first ensure that we are terminologically aligned about the understanding of

[1] Disinformation is the *intentional* production of fake information either for malignant or non-malignant purposes. Misinformation is the *unintentional* production of fake information (Donovan, 2021).

[2] Scholarly work that has a similar endeavour abounds. For instance, do Mar Pereira, 2022, in ethnographic writing, or Svabo & Bønnelycke, 2020, in Science & Technology Studies.

the notion *political deepfake* as applied throughout this thesis. And it is perhaps good to start with a mark of caution about deepfakes in general. Because while they are intentionally produced false content through deep learning, so-called synthetic media, they do not have a harmful intent per se. Indeed, deepfakes can be used for various constructive uses too, such as for medical diagnoses and trainings (Shin et al., 2018; Williams, 2022). On the other hand, disinformation is the intentional production of false information for a deceitful purpose. Therefore, the sparkling duo of *deepfake disinformation* can be defined as online disinformation created through deep learning that purposefully seeks to deceive its audience. Adding to that the adjective political, *political deepfake disinformation* is deepfake disinformation that has a political intent[3]. By political intent is meant its destabilising, harmful, or deceitful character; one that seeks to undermine trust in a political individual, a political party, or a public institution. And to be complete, with politics is meant "any organized control over any level of human activity that is guided by human values" (Badiei & Fidler, 2021, p. 377). Unless specified otherwise, it is this understanding of political deepfake disinformation that will be used throughout the entire thesis, and which will hereafter simply be referred to as *deepfake*. The fact that the thesis is not interested in a particular deepfake case analysis but that it is interested in the study of the deepfake in its general being is also reflected in the main title of the thesis, *the political deepfake*.

Generally, because of our currently increasing reliance on online media for our consumption of news, the deepfake can have a broader and more targeted reach at faster rates than traditional disinformation. It therefore leads some to believe democracies to be at the mercy of deepfakes. However, the generous literature review that will open the curtains to this project will reveal a general inconclusiveness in the research findings concerning the ways that deepfakes affect society and the extent to which they affect trust in public institutions. Research indeed reveals that a fatalistic conception is inaccurate and reductive of reality. More accurate representations of the reality of the deepfake exist; representations that uncover the sociotechnical reality of the deepfake. This notion *sociotechnical* comes to highlight that the interest in understanding the deepfake should not merely reside in understanding how it influences social order but that it should equally reside in understanding how society influences the potential contained within the deepfake for democratic interference. Indeed, in a sociotechnical understanding, technology is not merely about its material dimension but also about how it is embedded in social practices, cultures, histories, systems, and infrastructures (Cover, 2022; Paris, 2021; Star, 1990). Understood as such, the deepfake both "shapes social structures" (Paris, 2021, p. 2) and functions as "a kind of social glue, a repository for memory, communication, inscription, actants, and thus has a special position in the net of actions constituting social order" (Star, 1990, p. 32). Such conception of

[3] There are other purposes for disseminating disinformation, such as commercial ones, but these are not of interest in present thesis. Similarly, besides political deepfake disinformation that serves to deceive a particular target audience and wishfully manipulate their political opinion, there also exist political deepfakes that have no disinformational purpose, such as clearly satirical and non-deceitful political deepfakes (Donovan, 2021; Taylor, 2021). Such *overtly* non-disinformational political deepfakes are also not part of the scope of this thesis.

the deepfake thus exists but it yet remains scarcely understood. And while it points at the need for theoretical research endeavours that would re-theorise the deepfake in order to re-align its conception with its empirical reality, such theoretical efforts remain even scarcer. It is in this research vacuity that I will root my intended research contribution. And it is also at this point that the subtitle of the thesis becomes insightful. *A dance between conceptions, materialisations, and policy approaches*, hints at the three milestones structuring this thesis in that contributing effort.

The first milestone is about the conceptual revisitation—the proposed re-theorisation—of the deepfake using Karen Barad's new materialist development (2007). This revisitation will stand at odds with the popular fatalist conception of the deepfake, and will therefore be argued to redress the shortcomings of this fatalist conception. Most notably, this Baradian revisitation will shake the widespread ideal of the clearcut fact-fake demarcation to its core—an ideal that also shows inconclusive according to present-day research findings but which nevertheless defines one of the European Union's (hereafter, EU) prime efforts to counter deepfakes.

To provide a means of grounding the Baradian conception of the deepfake, the second milestone will take us through an empirical exploration of the world of the deepfake in order to gather and assemble the ways through which it materialises in our society. The interest will thus not only be about its technological ways of becoming tangible, but it will also be about the social and societal ways of becoming real. The empirical dive will however not stop at an exploration of the deepfake in its sociotechnical materialisation (the so-called deepfake assemblage). It will also explore on which parts of this sociotechnical assemblage the EU grounds its policy approach to deepfakes (the so-called EU assemblage). In other words, it will also explore how the EU conceives of the deepfake. It is the nature of these policies—as being objects that have a power of enactment, a power of bringing ideas to their material realisation—that will be of further interest upon analysing the sociotechnical features of the deepfake that are of interest to the EU.

The third milestone is about combining the Baradian-inspired conceptual revisitation of the deepfake to the empirically crafted assemblages in order to come up with a policy proposal. This milestone is thus essentially about a comparative analysis between the deepfake assemblage and the EU assemblage, informed by the Baradian conception of the deepfake. In fact, it is the very act of applying the conceptual development to the analysed empirical assemblages that provides the reason for my experimentation with a policy proposal. A policy proposal that will be argued to be justified both ontologically and empirically. Which is not a minor claim given that the EU's policy approach will be argued to lack such justification. An ontological and empirical legitimacy that is however key in terms of safeguarding a trusting relationship between a governmental institution and civil society, since governing deepfakes is about governing speech and is therefore embedded in the fundamental principle of freedom of expression (de Vries, 2022). My curiosity is thus driven by that question of how to safeguard freedom of expression all the while governing speech. And which is also why experimenting with a policy proposal proves all the more insightful.

It is perhaps at this moment that one can appreciate the title of the thesis in its richest meaning. As not simply indicative of the milestone-structure of the thesis, but also in understanding that the reality of the deepfake is a dance between its conceptions, materialisations, and policy approaches. Appreciating how all three continuously (re)configure each other and themselves. So, although the thesis will eventually come to a physical end with the traditional conclusion, it hopes to suspend you, the reader, in a perhaps contemplative state. A state where one can wonder whether it is not advisable to be in touch with deepfakes to educate and challenge our minds. Whether doubt is not a healthy requirement to prevent a state of complacency. And whether us netizens ought not to be taken more seriously in our capacity for autonomous judgment.
Enjoy the ride.

INTERLUDE-TO-READ • examples of deepfakes

Before dashing straight into the literature review, I here provide some examples of how deepfakes can serve purposes of political deceit. Since the thesis is not about a case analysis of a particular deepfake, the examples merely serve as an illustration of *some* of the ways through which deepfakes can be instrumentalised for political objectives. The set of examples is thus not exhaustive.

As first such example, Image 1 shows the famous face-swapping technique. Besides static imagery, it can also serve to deceive participants during live videoconference calls (see for instance Roth, 2021, about how an EU member of parliament was deceived during an online video call). The deepfake palette therefore equally regards speech mimicry (including voice timbre, speed, intonation, …).

Image 1: Face-swap of former German Chancellor Merkel (left) with former USA President Trump (right) (Gensing, 2020).

Political disinformation existed throughout history. We can remind ourselves of the case of handcrafted pamphlets depicting Marie-Antoinette, wife of King Louis XVI, in the most caricatural and impudent ways (Image 2). One can thus easily imagine how pornographic deepfakes can fuel political chaos through discreditation. They have for instance already been instrumentalised for blackmailing purposes "against Malaysian Economic Minister Azmin Ali [where images of him purportedly] involved in gay sex were disseminated [to] government officials […] in a country where homosexuality is illegal" (Paris, 2021, p. 8).

Image 2: The political potential of deepfake pornography (Plume d'histoire, 2015).

In terms of foreign interference, geopolitical deception, and military strategy, the use of deepfakes is alluring for it can orient open wars, military coups, or diplomatic tensions. Respectively exemplified by the deepfake of Ukrainian President Zelensky (Image 3), by the case in Gabon where a deepfake was sought to prevent a coup (Ajder et al., 2019), or by the use of fake (satellite) images to justify a need for foreign intervention or attack (Image 4 and 5).

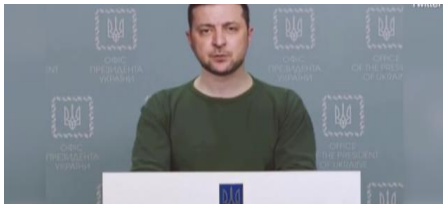


Image 3: Deepfake video of Ukrainian President Zelensky allegedly produced by Russians and streamed on the Ukrainian national television channel allegedly by Russian perpetrators (Newman, 2022).

Image 4: Deepfake to allege a country's possession of a particular weapon technology (The Economist, 2019).

Image 5: Deepfake posted by Chinese authorities of an Australian soldier killing an Afghan child (BBC, 2020a).

One might wonder why deepfakes would be worse than Photoshop or a handmade collage. Simply, deepfake technology makes it harder to detect the forgery. It thus takes more time to eventually debunk, while time becomes a scarce currency when events take a dark spin. Additionally, a deepfake image can be created out of the blue through loops of deep learning using entire image data sets. A deepfake image therefore does not have an original copy the same way that a Photoshoped image has, which makes the latter more easily detectable

(Giansiracusa, 2021, p. 19).

It is perhaps opportune to mind that the use of deepfake technology is certainly not secluded to the audiovisual realm however much that is its most examined form. Deepfake text generators are also very real and can be used for automated chatbots, for generating political speeches, for writing news articles and automatic email responses, etc. (see Image 6). Deepfake texts are already in use for email phishing scams and financial fraud. For instance, "[c]ybercriminals [impersonate] political organisations, mimicking their domains, slogans, and even getting people to donate to fake organisations" (Hannah, 2020; Bateman, 2020).

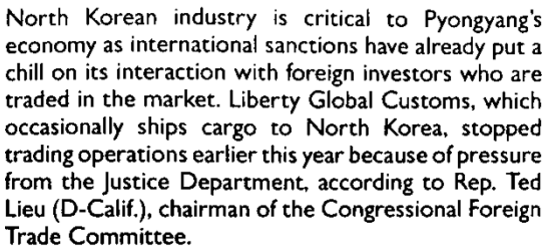

North Korean industry is critical to Pyongyang's economy as international sanctions have already put a chill on its interaction with foreign investors who are traded in the market. Liberty Global Customs, which occasionally ships cargo to North Korea, stopped trading operations earlier this year because of pressure from the Justice Department, according to Rep. Ted Lieu (D-Calif.), chairman of the Congressional Foreign Trade Committee.

Image 6: Deepfake text created with GPT-2 software from OpenAI (Giansiracusa, 2021, p. 25).

Deepfake technology is also already in use for political campaigning and is likely to gain in popularity. For instance, an Indian presidential candidate tested such application to voice-over his speech in different dialects to reach a broader public (Lyons, 2020). Another campaigning application has been probed by Andrew Yang upon campaigning in the metaverse for the New York mayoral elections in the USA (Hackl, 2021). A South Korean political candidate deepfaked its video appearances to make himself look 'cooler' in order to reach a younger audience (AFP, 2022). Besides directly applying deepfake technology to manipulate a political figure's image, it can also serve for the creation of a political message. For instance, the Belgian socialist party used a deepfake video of Donald Trump for climate change campaigning (von der Burchard, 2018). Not only will these uses expectedly blur the line that distinguishes deception from candour. But it provides yet another hint at why regulating deepfakes is not such unequivocal an affair given that public bodies themselves become users.

Lastly, while all the above illustrations frame the use of deepfakes as if these were singular one-off events, it is useful to mind ourselves that deepfakes can be part of broader disinformation schemes (e.g., the case of a 15-year Indian disinformation campaign; Kuchay, 2020). This also brings into perspective a popular conception of the danger of deepfakes being about single major incidents, while in fact it is more so about a continuous coordinated infusion (Radsch, 2022). And that is also the reason for which deepfakes do not have to be highly technological to have the sought-after deceptive effect (Giansiracusa, 2021)[4].

[4] Some therefore distinguish between *deepfakes* and *cheapfakes* (also shallowfakes). Deepfakes are created by machine learning while the rest is created by less tech-savvy techniques (Paris & Donovan, 2019).

THE DOMINANT CONCEPTION OF THE DEEPFAKE

Given that the Internet is the deepfake's natural habitat and given our increased reliance on the Internet as a medium for the consumption of information (Newman et al., n.d.), these combined features of contemporary society exemplify the need for understanding and problematising the deepfake properly in how it affects society. To do so, nothing better than a comprehensive literature review. It is titled *the dominant conception of the deepfake*, because there is a clear pattern emerging across research publications, expert articles, and news articles alike. A tendency to depict the deepfake as a parasite to democracy. A parasite that inevitably leads to the infestation of the welfare state if left unaddressed. All these publications unravel important questions, but their premise is generally unquestioned. As if it were evidence that deepfakes lead to tragedy if left dangling around freely in the digital space. But research shows inconclusive. Both about how deepfakes affect society and about the effectiveness of current means implemented to counter deepfakes. Which is why I call this conception that dominates most accounts of the deepfake, *deepfake fatalism*, to emphasise its tragic-deterministic a priori.

Literature exists that redresses most of the shortcomings of deepfake fatalism. A literature that delves in the sociotechnical reality of the deepfake, and which I call *deepfake realism*. This literature highlights the relation of co-dependency existing between society and technology, between society and deepfakes. It thus points to the idea that deepfakes influence social order but that deepfakes also have social precedence as they are embedded in larger social structures—deepfakes are cultural (Cover, 2022). Yet, the scarcity characterising this literature makes that the sociotechnical reality of the deepfake is less understood. And it also results in it still insufficiently redressing the shortcomings of deepfake fatalism. It is thus in that research vacuity that I will root my research contribution.

deepfake fatalism · a mothership of disinformation

> *Reality apathy is when it's so hard to make sense of what's happening [that] people just sort of give up. […] People won't know what they can trust.*
>
> — Aviv Ovadya (The Economist, 2019)

Quotes like Aviv Ovadya's and experts predicting infopocalypse scenarios abound. Upon delving in the literature review, it became obvious that deepfakes are repeatedly associated with doomsday prophecies where society would topple into a hyperreality where all would be constantly doubted and distrusted if deepfakes were left unaddressed (Cover, 2022; Neo, 2021; Taylor, 2021). To determine whether present-day scientific research findings confirm the fatalist narrative, we will explore in this section the diverse

ways in which deepfakes are known to be tangled with society. For matters of clarity, this section is subdivided in subsections, each providing a key feature of interest of the deepfake in its fatalist recount.

the deepfake as technological event

The fears with deepfakes revolve around their capacity to convert disinformation from a formerly normal and benign social feature into a generalised societal cancer (see Byman & Joshi, 2020; Chesney & Citron, 2019; Diakopoulos & Johnson, 2020; Fallis, 2020; McKay & Tenove, 2020; Ovadya, 2019; Vaccari & Chadwick, 2020; Van Bavel et al., 2020; Walsh, 2020; Yadlin-Segal & Oppenheim, 2020). Compared to other disinformation techniques, deepfake technology allows for a faster, broader, and targeted reach of the content on the online space (Dobber et al., 2021; Vosoughi et al., 2018). The use and circulation of deepfakes has boomed between 2018 and 2020, with 49 081 deepfake videos counted in June 2020 (Ajder, 2020; Ajder et al., 2019). This number is surely an underestimation, be it because it only accounts for audiovisual deepfakes. At the same time, this number covers deepfakes of all types of intent, not merely political ones. But this number is likely to keep on mushrooming. And while this number is a mere grain of sand relatively to all of the online content that generally approximates infinity if summed up, it does not make them less potent as such. Our digital network societies become increasingly reliant on online media for news consumption (Castells, 2004; Newman et al., n.d.; Rosenberg, 2022), and the technological features of the deepfake make it particularly agile in this information ecosystem[5]. The risk embodied by deepfakes is thus not so much about their existence as it is about their online circulation dynamics. But as we are about to uncover, these online circulation dynamics—importantly dictated by social and behavioural dynamics—are complex and not bereft of ambiguity.

believability and audiovisual media

Some findings show that "false information spreads faster, and to more people, than true information" (Fallis, 2020, p. 635). But this feature might simply be an expression of our characteristic fondness for emotionally enticing information (Gabriel, 2021). Spreading does not mean that disinformation achieves its deceptive purpose. Equally so, even though some findings conclude that us humans tend to believe audiovisual material because it would be enhancing content trustworthiness by a mechanism of "ease of

[5] Note that online news consumption does not mean that people all revert to social media platforms to consume their news. It simply means that people go online. For instance, according to a survey in the UK, only a minor proportion of the population uses social media for news consumption (Ross Arguedas et al., 2022). And Facebook itself is retracting from news publishing deals since Facebook users would in fact not be consuming news as much via its platform (Masnick, 2022c; Fischer, 2022).

processing" (Brashier & Marsh, 2020, p. 503; O'Neill & Smith, 2014; Vaccari & Chadwick, 2020), other research shows that "[video is] not necessarily more persuasive" (Hendrix, 2021). Audiovisual media—the main medium of concern for deepfakes—would thus not necessarily lead people to believe that the content is genuine. Similarly, some researchers find that there would be no significant difference in gullibility between different media; "deception detection is approximately the same whether the message is conveyed through text (e.g., a court transcript, an Internet chat log), an audio recording (e.g., a voicemail, a radio program), or a video (e.g., an interrogation video)" (Hancock et al., 2021). A recent study in The Netherlands also concluded that deepfakes on social media did not have an "added persuasive power […] compared to textual disinformation" (Hameleers et al., 2022, p. 7). Our relation to information uptake is thus generally unclear. And again, for one to see a deepfake does not mean that that deepfake achieves its deceptive purpose. Deepfakes may "reduce trust in news on social media" (Vaccari & Chadwick, 2020, p. 9), but they would be generating uncertainty rather than truly deceive and would be leading to scepticism rather than "cynicism and alienation" (p. 10). The question then is whether a certain level of scepticism within a society is not fruitful and moral a requirement (Sher, 2019; Taylor, 2021). This ambiguity leads some to argue that "the epistemic threat of deepfakes is often overstated" (Harris, 2022, p. 18). While others mind that "[e]ven if a deepfake is ultimately debunked, or never believed at all, it can still hurt the person [or symbolic entity] it falsely depicts by changing the discursive context around them" (Rini & Cohen, 2022, p. 148).

It is perhaps opportune to mind that humanity has not always been in touch with audiovisual media. And therefore, it is not because audiovisuals today would serve as a catalyst for "[the making of] significant decision[s]" (Fallis, 2020, p. 624), that it will necessarily remain such catalyst[6]. Additionally, where some fear that younger generations would become more prone to online deception because of their increasing reliance on online audiovisual media to consume news compared to other generations (Etlinger, 2019; Kelly, 2022), one might in fact wonder whether that heightened consumption could not in turn inoculate them with some level of immunity towards deepfake deception.

our predisposition to self-validation

Another consideration is that, in general, we all rely on others to make sense of reality and construct our opinions and beliefs (Goldberg, 2010). But where some research argues that we do so "even if it contradicts [our] own knowledge" (Chesney & Citron, 2019, p. 1765), other research argues that "people choose to hear from those who are politically like-minded on topics that have nothing to do with politics (like geometric shapes) in preference to those with greater expertise on the topic but have different

[6] Moreover, one could question the presumption that leads to the derivation of such conclusion. For instance, in a decolonial endeavour about gender, Oyèrónkẹ́ Oyěwùmí (1997) deconstructs the precept that sight is the prime vector of information uptake.

political views" (Marks et al., 2019, p. 83). More nuanced research finds that one's readiness to believe a deepfake to be a representation of reality importantly depends on one's prior opinions (Osmundsen et al., 2021; Yun Shin & Lee, 2022), and that a politically opinionated context can exert a certain impact on the viewer of an audiovisual content "even if it does not have a persuasive effect" (Hendrix, 2021). However, yet other research shows that believability can importantly be influenced by virality, since "[the widespread becoming of] manipulated images and videos [leads these to be] interpreted as evidence or markers of truth" (Paris, 2021, p. 8). Which in turn can mean that "even rather poorly written fake news can be extremely influential" (Giansiracusa, 2021, p. 28).

INTERLUDE-TO-READ • deepfake "immunity"

When some scholars argue that "individuals with lower cognitive levels do not consider the fallacious nature of the content and [are more] likely to share deep fakes [and are more prone to] interpret the information in ways consistent with their preexisting biases" (Ahmed, 2022, p. 103), I wonder who can claim to be absolved of bias. This is important, because anyone declaring itself immune to deepfakes and disinformation clearly underestimates the reality of things. Whether we like it or not, and whether we are willing to admit it or not, we avid netizens have already been tricked, be it about a trivial something. Image 7 illustrates how any mind can be lured and deceived—it is merely a matter of finding that mind's soft spot. Research also shows that people have a hard time distinguishing what is true from what is false when exposed to mis/disinformation, and this, independently from their cognitive ability (Breakstone et al., 2021). Even an "awareness of the problem of fakes may prevent one from forming true beliefs based on authentic information [because of an excessive scepticism]" (Harris, 2022, p. 11). Dismissing those who are captured in the nets of disinformation and deceit as stupid or uneducated rather than understanding their vulnerability (Cover, 2022) reinforces popular—though reductive—narratives and thereby fuels polarisation and bigotry.

Image 7: Let's pretend the lurer is a deepfake (Vauss, n.d.).

microtargeting, recommender systems, and more perplexity

This is where the microtargeting potential of deepfakes enters the show. Microtargeting and recommender systems—the online algorithmic mechanisms employed to nudge users to particular online content or behaviour—are not devoid of lugubrious fancies. However, "the extent to which the "algorithms" can be blamed for societal problems (from teen depression to radicalization to genocide) remains unclear" (Bengani et al., 2022). Similarly, where some argue that capitalist online surveillance practices lead to behaviour modification (Zuboff, 2019), others nuance this claim by finding that "Internet advertising is not all that effective at modifying consumer behavior" (Crain, 2022). In that line of thought, while microtargeted deepfakes could "amplify the effects of a deepfake", this amplification is not necessarily linked to a heightened ideological polarisation (Dobber et al., 2021, p. 69). This study on microtargeted deepfakes was conducted in the USA, which has a particular political landscape. Therefore, conducting the same research in another political and cultural context might lead to nuanced or different findings. Nonetheless, as we all know (or should be aware of), we humans willingly seek self-validating information. Thence, even if recommender systems were to disappear, we would still row our boats toward content that validates our beliefs.

This brings us to the topic of online echo chambers. Chambers of social cohesion jeopardy as some like to coin them. Chambers in which the processes of personalisation lead to polarisation (Relph, 2021; Walsh, 2020). Such perspectives of the alarmist kind are not to be refuted overall, but other findings suggest that online echo chamber effects are overstated (Dubois & Blank, 2018; Lewis et al., 2011), if not leading to an opposite effect in terms of influence on political opinion because the online space would favour a diversity in one's content encounters compared to offline echo chambers (Gladstone, 2021; Groshek & Koc-Michalska, 2017; Ross Arguedas et al., 2022). The echo chamber effect would thus also be dependent on the political context at hand; cf. "the position of a political party within the political system changes the way they operate online" (Gladstone, 2021; Krasodomski-Jones, 2016, p. 33). To conclude, as much as our interpretation of deepfakes is not straightforward, our online behavioural dynamics are not straightforward either.

social media platforms in the limelight

A tremendous number of publications looking into deepfakes and online disinformation is focused on social media (De Blasio & Selva, 2021) in their role of "spreading and generating false content" and in being at the origin of some of contemporary society's illnesses, but "[n]one of these assumptions hold up to scrutiny" (Marwick et al., 2021).

Firstly, research is currently only focused on a limited number of social media platforms. For instance, LinkedIn appears absent from any such research, while it is not free of fake content, be it about fake

profiles (Bond, 2022). This narrow focus also results in research methods looking for instance into Twitter—one of present-day's major competitors in the social media arena—being inadequate (Tufekci, 2014).

Secondly, while the majority of the attention is directed towards social media or the technology giants Google-Apple-Facebook-Amazon-Microsoft, as these either host digital populations that by far surpass nation-state populations[7] or account for substantial online traffic, social media are not the only actors part of the online information ecosystem. Search engines exert an equally important influence on the ranking of online dis/information (Masnick, 2022a). And so too, "[t]elecommunications and cable companies themselves are deeply embedded in ecosystems for transferring consumer and corporate data" (László et al., 2022). While being part of the online information ecosystem, these actors currently remain in the research blind spot.

Thirdly, while some argue that "[t]he purported "self-regulation" by social media and technology platforms has failed across a range of issues" (Radsch, 2022, p. 26-27)[8], others argue that "private regulation by online platforms transpires to be the most efficient regulatory measure available" (Kalpokas & Kalpokiene, 2022, p. 65). Further ambiguity arises when research on Telegram shows that an absence of content moderation on large platforms does not necessarily mean that all becomes rotten and dominated by misinformation, and would in fact be dependent on the trustworthiness of news (Herasimenka et al., 2022). And while "UNESCO report[s] that Telegram is rife with Holocaust denial, and that malicious actors have found ways to avoid content moderation" (EU DisinfoLab, email newsletter, July 26, 2022), it is however not because criminals follow the digital trend that the digital space therefore becomes criminal altogether. Both dynamics ought not to be confused, but it is however an alarmist narrative that some have observed to be a means to securitise the regulation of deepfakes under government authority (Neo, 2021; Taylor, 2021).

Fourthly, "there is no evidence small players are more ethical than large players[,] both small and large players need to be held accountable to certain minimum standards that monitor their business models for data extraction" (László et al., 2022).

Lastly, looking at social media alone in their role in the dissemination of deepfakes is insufficient given that "the Internet's prevailing economic structure has been heavily shaped by public policy" (Crain, 2022; Saurwein & Spencer-Smith, 2020). As some thus point out, "[i]f we continue to stay focused on content alone and not on systems, those spreading disinformation will continue using social media to take full advantage of the product features enabling their successful, profitable campaigns" (Institute for Strategic Dialogue, 2022).

[7] For instance, Facebook hosts more than 2.9bi active users and Linkedin has more than 800mi users (LinkedIn, 2022; Statista, 2022).

[8] Though, Radsch fails to provide an explanation or references for this claim.

the deepfake as financial asset

By way of engulfing any potential source of income, contemporary profitmaking practices similarly took hold of disinformation, and thus of deepfakes (Global Disinformation Index, personal communication, 26 April, 2022). Today, the online information ecosystem, like much of the rest of the online space, is governed by the online advertising model. Online content providers such as commercial platforms, social media platforms, news websites, and alike, generate income through the publication of advertisements on their webpages and applications. There exist various contract types between online content providers and advertisers. Advertising revenue can for instance be calculated based on the amount of traffic on a webpage (the number of views of an advert), or on the number of times that users effectively click on an advert. Which is how microtargeted adverts—adverts that track users based on user-generated data—become all the more economically appealing as they increase the likelihood of us viewing or clicking on adverts and thereby generate more revenue. In other words, if I were willing to generate revenue from the dissemination of deepfakes, I could set up a website and find advertisers who are willing to pay me for the placement of their adverts on my website. That is also how online platforms that (inadvertently) provide their users with disinformation benefit from it. And the more notorious the platform, the more consequential its benefit. At the same time, a bad content moderation notoriety can injure the brand's reputation.

the deepfake as content

What online content providers today do to prevent the dissemination of deepfakes on their platforms is engaging in content moderation practices. Content moderation is about ensuring that content posted on a platform abides by the terms of use of that platform, and is mostly practiced to safeguard a brand's reputation. It can take various forms, but when it comes to the dissemination of deepfakes, it is always fact-checking that strikes the top hit on the content moderation popularity scoreboard. Fact-checking is the verification of online content through (semi)automated systems. For deepfakes, there already exist multiple artificial intelligence-based and blockchain-based verification tools. The former operating through trained content detection models, the latter through the digital watermarking feature of the technology (Al-Saqaf, 2019; Fraga-Lamas & Fernández-Caramés, 2019; Hasan & Salah, 2019; Nguyen et al., 2019; Paris & Donovan, 2019). There also exist fact-checking tools based on metric systems, such as author reputation metrics (e.g., Hoaxy), metrics discerning between satire, opinion, sensationalism, … (e.g., FactFaker), and reference-credibility metrics (e.g., SciFact) (Giansiracusa, 2021).

The fact-checking process does not stop there. Once a content has been checked, it requires further moderation, which can happen in mainly two ways. Either verified false content is removed. But this does not prevent netizens to download the content prior to its removal (and luckily so for watchdogs of

political fraud for instance). Plus, the way that the removal is operated is itself subject to diverse techniques (for instance, shadow banning is when a platform bans a user's content but without that user knowing about it and so the content becomes invisible to all users except to the author; Nicholas, 2022). Alternatively, verified false content is left online but is corrected or labelled with a warning sign or explicative comment. And here again there is a diversity in how it is practiced. For instance, counterspeech is about implementing automated empathic "alternative, polite, and non-aggressive response[s]" to hate speech on social media platforms (Chung & Vidgen, 2022; Hangartner et al., 2021). Also technical measures can be implemented to ensure that the content is no longer promoted through algorithmic referential mechanisms (Gillespie, 2022).

Of course, the entire fact-checking process is not devoid of mishaps, and each approach has its personalised set of contestations (Giansiracusa, 2021; Gillespie, 2022; Kalpokas & Kalpokiene, 2022). Firstly, cases of mislabelling are not rare. Which causes online content moderation to turn into a generally sensitive matter in terms of transparency and public distrust. Mislabelling is then also abused by some to claim a content to be a deepfake while it is not (the so-called liar's dividend; Giansiracusa, 2021). Secondly, technologies exist that fool deepfake-detectors themselves (GAO, 2020; Juefei-Xu et al., 2022; Nguyen et al., 2019). A real Tom & Jerry show that is also not devoid of transparency and distrust hiccups given that deepfake-detection techniques and their success rates commonly remain black-boxed (Giansiracusa, 2021). Perhaps due to company secrecy concerns to remain a leader on the market or because untangling what is truly happening in deep learning is sometimes a bit of a mystery[9]. Thirdly, transparency itself is not easy a matter, be it because of anonymity issues or because open-sourcing deepfake code according to some findings would result in its broader usage for pornographic purposes (Winter & Salter, 2019). Additionally, for transparency efforts to truly result in informed publics, the publics that are to be informed ought to be literate on the matter in the first place.

While some deplore today's lack of content moderation standards (Berkman Klein Center for Internet & Society, 2022), some research findings indicate clear ambiguity in regard to the effectiveness of existing fact-checking techniques. In the case of content removal, findings show that it can lead to a backfire effect that fuels public discontent and leads to a worse of situation (Wong, 2021; Wood & Porter, 2019). In the case of content correction, while some findings show that "the average subject accedes to the [corrected content] and distances [it]self from the inaccurate claim" (Wood & Porter, 2019, p. 160), other research shows that people do not always acquiesce to the corrective comments, especially if insufficiently persuasive as "people [would] consistently counterargue attitude-discrepant information" (Dobber et al., 2021; Garrett et al., 2013, p. 631). A subsequent question is then what level of persuasion is necessary for a critical mass to adopt corrected content. A propagandist question that is ethically questionable and which would undoubtedly also open the gate to multiple perplexities.

[9] Although that argument might itself then be abused and become an excuse to secure secrecy and financial interests to the detriment of public interest.

Lastly, while fact-checking efforts today are largely implemented with the idea that fact-fake demarcations would eradicate deepfakes, or online disinformation more generally, other research shows that instead of providing an end, fact-checking efforts implemented by platforms in fact would induce a shift—for instance, a shift in the way that content is distributed to the audience relatively to the way that journalism fact-checking protocols influence how content is distributed (Cavaliere, 2021).

considerations about deepfake fatalism · the need to acknowledge other deepfake realities

The general ambiguity found in the above research findings, and the many subsisting uncertainties and unknowns, lead to a general inconclusiveness with regard to how deepfakes affect society and the effectiveness of the means implemented to contravene deepfakes. In other words, the present-day research findings lack validating the fatalist tragedy narrative. And the inconclusiveness also illustrates why the regulation of deepfakes is tricky an affair. Which is without having yet mentioned that deepfake technology is not devoid of beneficial applications, thereby adding yet another layer of regulatory complexity. For instance, deepfakes can be useful in education (e.g., healthcare training, media literacy, …), for medical uses (e.g., training, voice restoration, diagnosing mental disorder, …), for artistic exploration, as business applications (e.g., detection systems such as live object recognition for automated self-driving cars or to detect failures in solar panels, video game industry, communication industry, e-commerce, metaverse simulations, tourism industry, …), for research (e.g., a virtual reality deepfake has been used to reconnect a mother with her lost child), or in politics (e.g., as satires stimulating democratic politics) (Chesney & Citron, 2019; GAO, 2020; Kwok & Koh, 2020; de Ruiter, 2021; Shin et al., 2018; Smith, 2021; Taylor, 2021; Williams, 2022).

The general inconclusiveness allows us to make the following observations about the popular fatalist narrative.

(1) The deepfake is more than a given technological event

Research efforts tend to focus on the one-way relation of how deepfakes affect society. As such, these efforts tend to rehearse the conception of the deepfake as an external given technological phenomenon free of original sin—so to speak—that could find an end, be eradicated. But the inconclusiveness characterising these research findings—being caused by social and behavioural perplexities—brings us to consider that the deepfake has deeper connections with society. It brings us to consider that the deepfake is as much a social event as it is a technological one (Cover, 2022).

(2) The deepfake is more than a tragedy

The general lack of backup for the fatalist tragedy narrative is clearly indicative of an obliteration of other realities equally (if not more) characteristic of the deepfake. Deepfake fatalism clearly underappreciates the diversity of ways in which deepfakes ground themselves in our society.

Fatalism is therefore reductive and deterministic a take on deepfakes (Neo, 2021; Taylor, 2021).

> "[P]ublic response to deepfake's emergence has seized upon those qualities to depict it as an existential threat […] arising from the state's perpetual drive to control its own becoming" (Taylor, 2021, p. 13)[10].

This quote allows us to appreciate why this observation has important policy and regulatory implications. Regulating deepfakes is about regulating online content and is therefore about regulating freedom of expression. The governance of deepfakes thus ought to be done carefully if willing to promote society's emancipation (Reisach, 2021; Taylor, 2021). Because, while the fatalist narrative sustains the idea that deepfakes can be eradicated and that their implications therefore can find an end, their deeper societal rooting leads us again to appreciate why deepfakes are not simply a technological given but are inextricably entangled with society.

(3) The deepfake is more than content

While fact-checking and automated detection techniques remain in the spotlight as tools to counter deepfakes—thereby highlighting once again the technological dimension of deepfakes to the detriment of its social dimension—approaches to deepfakes should not focus on assessing content alone but equally so on assessing our relation to that content (Lecomte, 2021). Deepfakes are content. But they are also a medium of expression, bringing to expression some of society's characteristics more than it is at their inception (Silbey & Hartzog, 2019; Taylor, 2021).

(4) The target audience is more than a homogenous mass

In line with the observation that the deepfake is more than a tragedy, the inconclusiveness that characterises present-day research findings also brings us to the critique about how this research is dominated by Anglocentrism and therefore lacks representativeness of the local and marginal contexts of information networks (Gillespie, 2022; Nguyễn et al., 2022; de Seta, 2021). Not only does this lead us to consider that an Anglocentric perspective cannot be representative of the entire reality of the deepfake, thereby also confirming the partiality of the fatalist narrative. But decolonial insights on disinformation studies (e.g., Lenoir, 2022) also lead us to consider that the "[narratives of fear about disinformation] build on and reify pre-existing ideologies, frequently involving race and inequality" (Kuo & Marwick, 2021, p. 1). The unspoken unanimity characterising the fatalist narrative thus rehearses an unquestioned Western democratic hubris.

(5) Civil society is deprived of agency

The observation that the Western-centric fatalist narrative lacks accounting for the diversity of the netizenry—the online citizenry—also brings us to the related observation that the netizenry is generally underappreciated. The netizenry is systematically taken for granted as consisting of one

[10] Note that Taylor does not succumb to a dystopian narrative of authoritarianism. He very much acknowledges the need to address the case of deepfakes for safety and security matters. But he proposes to conceive of the possibility that "the liberal state has – so far at least – discouraged other possibilities for deepfake's development" (2021, p. 13).

homogenous entity. And the netizenry is systematically taken for granted as a gullible mass prone to calamitous fortunes. But such dramaturgical perspective is reductive of reality—were the netizenry truly that gullible and passive, the doomsday prophecy would have already been confirmed. And perhaps it provides a lead as to why research findings are so inconclusive, thereby hinting at a need to review our currently too simplistic understanding of the netizenry.

Given these observations, time has come to shake the fatalist narrative and to acknowledge other realities characteristic of the deepfake.

deepfake realism · sociotechnical entanglements

The previous section of the literature review provided an elaborate appreciation of why the fatalist tragedy narrative of the deepfake—however popular—lacks empirical validation and therefore is insufficient to describe the reality characteristic of the deepfake. To provide such fuller description, the present section offers a review of the yet scarce literature that digs deeper into the sociotechnical reality of the deepfake—in its reality inclusive of its social rooting rather than exclusively centred on its technological dimension. Present section is therefore called *deepfake realism*. It paves the way to a deeper epistemological questioning of the fatalist conception, for it paves the way to a questioning of the conceptual premises based on which the deepfake is believed to be known. Hence, this section is divided in subsections that each regard a key turning point with respect to the observations made for deepfake fatalism, before ending with considerations about deepfake realism.

the deepfake as more than a technological event

"Rather than viewing deepfake technology as external to culture [it is argued] that the deepfake can only be apprehended as an artefact and practice that emerges from within culture as a response to specific desires" (Cover, 2022, p. 610). And as put by others, "[the] proliferation [of disinformation] on social media has developed from a socio-technical mix of platform design, algorithms, human factors and political and commercial incentives" (Saurwein & Spencer-Smith, 2020, p. 823). This was also one of the observations derived from the fatalist stance. Namely, that the deepfake is as much a social event as it is a technological one. In that sense, the praised ideal of the technological fix to deal with deepfakes is implicitly bound to fail (Paris, 2020). Therefore, while remaining minoritarian in their endeavour, some scholars call for the need of a shift in attention where the focus should not be on "how to deal with deepfakes and online disinformation" alone, but also on "how to deal with the underlying societal issues and structural inequalities" (Institute for Strategic Dialogue, 2022; Jarvis, 2021; Kreiss, 2021; Paris & Donovan, 2019; Taylor, 2021). It is thereby argued that a shift to deepfake realism would allow a shift

towards a more effective and accountable approach. One example of such shift is provided by Silbey and Hartzog (2019), where they write,

> "The potential upside of deep fakes is that they might help muster the political will to address the larger, structural problems made worse by the inability to trust what we see and hear. [M]aybe an effective way to respond to the scourge of deep fakes isn't to target the creation and use of deep fakes themselves, but rather to focus on strengthening the social and political institutions they disrupt" (p. 961).

the deepfake as a medium of expression

Despite the mitigated and at times counter-productive results from fact-checking assessments, a generalised strong belief subsists that fact-checking technologies are an essential means to contravene deepfakes. Not only does it have to do with the trending vibrant techno-optimism that characterises our contemporary society (Paris & Donovan, 2019; Clyde, 2022). And not only does it have to do with the popular dualist perception of reality where the world could be accurately described through binary opposites—in our case, subdivided between fact and fake. But the generalised faith in fact-checking also has to do with a focus that remains centred on envisioning the deepfake as content, thereby lacking its appreciation as being a medium of expression, as being about speech. Not just any speech, but protected speech according to some (de Vries, 2022; Masnick, 2022b).

> "[T]here is no good reason why machine-generated speech, which could equally well contribute to a free exchange of ideas as human-generated speech, should be categorically excluded from the protective scope of freedom of expression" (de Vries, 2022, p. 2).

Focusing on the deepfake as speech not only underscores again its social nature. But it provides a means to move away from the vibrant techno-optimism; away from the "twentieth-century libertarian imaginaries [where] an identity-free and bodiless sociotechnical future shaped the Internet [as] a value-free neutral zone" (Paris, 2020, p. 8 & 9). A move away from the ideal of technology neutrality that would in turn allow to address the more pressing systemic questions (Institute for Strategic Dialogue, 2022). Focusing on the deepfake as speech would also provide a means to move away from the popular—though failing—dualist perception of reality by which the fatalist fact-checking ideal abides. Indeed, deepfake realism points out that deepfakes and fact-checking technologies alike are not just an external given—an immaculate technological *being*—but that they are a social *doing*—having social precedence. A conception that thus complicates the application of the dualist ideal since binary demarcations are no longer value-neutral (Kalpokas & Kalpokiene, 2022). Indeed, "[t]he distinction between illegal and harmful content [in practice] might be a line that is difficult to draw" (de Vries, 2022, p. 18). To paraphrase Kalpokas and Kalpokiene (2022), "the main point of concern [should thus not be the deepfake as content] as such but the capacity [it has] for bringing into existence different determinate configurations of the world" (p. 80). It should not be about its being, but about its doing.

considerations about deepfake realism · the need for a reconceptualisation of the deepfake

The above literature review for deepfake realism shows to be scarce. But it does not prevent us from observing that it makes up for several of the shortcomings observed for the fatalist narrative.

(1) The deepfake is a sociotechnical event

The realist approach has ample consideration for the sociotechnical nature of the deepfake. A consideration for both the social and the technological reality of the deepfake, rather than remaining focused on its technological dimension alone. Deepfake realism therefore allows us to appreciate why a technological fix to tackle deepfakes is bound to fail.

(2) The deepfake is more than a tragedy

This redress for the common depiction of the deepfake as a threat to national security is not always unanimously expressed across literature on deepfake realism (Cover, 2022; Kalpokas & Kalpokiene, 2022; Neo, 2021; Taylor, 2021). And where these scholars all call for more than a sociotechnical approach to deepfakes and call for conceptual approaches to better re-theorise the deepfake, to my knowledge, Cover and Taylor are the only scholars straightforwardly expressing the need to move "away from perceiving the [deepfake] technology as having a negative impact" (Cover, 2022, p. 609). Cover heads onward in a deconstruction of the tragedy narrative through cultural theory, while Taylor does so through securitisation theory. Although such theoretical endeavours are critically lacking, they are crucial given the general inconclusiveness in the empirical findings. Re-theorisations are necessary to better re-align the conception of the deepfake to its observed reality.

(3) The deepfake is a medium of expression

The social nature of the deepfake brings us to consider that the deepfake is no longer an immaculate given technological phenomenon—free of original sin. The deepfake is speech. It is value-laden, and therefore approaches to tackle deepfakes are intrinsically value-laden; cf. "[l]imitations to freedom of expression imposed by the legislator always have some normative, cultural or societal grounding" (de Vries, 2022, p. 13). Which is why deepfake realism calls for more attention to the systemic entanglements rather than remaining focused on the content itself. Also because while regulatory firewalls ought to be set up to ensure a secure and prosper welfare state, regulation should not morph into a more draconian approach prone to governmental abuses "curtailing free speech and other civil rights" (Vaccari & Chadwick, 2020, p. 9).

While making up for the above shortcomings in the fatalist narrative, deepfake realism still lacks more consideration for two important aspects in the reality of the deepfake.

(1) The netizenry is underappreciation

Despite acknowledging the sociocultural reality of the deepfake to an impressive extent, the realist

account however still lacks an actual consideration of the fact that the netizenry is neither a homogenous mass nor a mass prone to misfortune and deprived of any agency.

(2) The failure of dualism

The lack of empirical evidence for the efficacy of current fact-checking methods is insufficiently addressed in deepfake realism. It insufficiently questions the dualist premise that is foundational to the fact-checking ideal, all the while the lack of empirical evidence in fact very much highlights the need to re-examine this dualist premise—the need to re-examine whether dualism alone is sufficient to provide adequate means to tackle the challenge posed by deepfakes.

These persistent shortcomings open the way to my intended research contribution that I am introducing below.

my contribution

While research efforts on the sociotechnical reality of the deepfake are already scarce, with "a majority of the research articles on fake news [being] atheoretical" (Arqoub et al., 2020, p. 69), theoretical efforts are even scarcer. However, as we have just uncovered, the persistent shortcomings in deepfake realism—about how to conceptualise the deepfake in ways truthful to its observed reality—lead us to appreciate the necessity of theoretical contributions. Aligning with former scholars, I equally argue through this thesis that such theoretical approach is essential. Not only to better re-align the conception of the deepfake to its observed reality, but thereby also to better inform policy approaches to the deepfake. Approaches that better satisfy the expectations of a democratic society and that are in line with the EU's promotion of a democratic welfare state and a safeguarding of digital rights (see EC, 2022d). Indeed, policies shape society, and as Neo (2021) writes,

> "the long-term societal consequences of [disinformation] would be shaped not just by its actual empirical effects, but also by how hegemonic societal actors frame and discursively construe the issue" (p. 214).

A conceptual revisitation of the deepfake would also prevent the discussions to remain stuck in a glitchy ping-pong competition between pure politics-lashing criticism or pure tech-lashing criticism (Neo, 2021). Indeed, rather than seeking to dis/prove links of causality between deepfakes and society through the same prism that characterises the fatalist narrative and still seeps through the realist narrative, the proposed conceptual journey allows to refresh that prism altogether, and thereby moves the discussions beyond any ping-pong game and beyond any tragedy-driven narrative that is characteristic of current discussions. Repeating only one tone inhibits a blossoming of the melody; repeating only one narrative inhibits a recount of reality's complexity. The next chapter thus sets on this conceptual quest. I hope that you are as excited as I am.

ANOTHER CONCEPTION OF THE DEEPFAKE

While deepfake technology is new in its own way and creates certain sociotechnical mumbo jumbos that are specific to it, the literature review showed that in many ways it is a reshaping of existing societal phenomena that now also come to expression through deepfakes. In other words, deepfakes are an expression of societal features updated to contemporary society's technological advances. And while the literature review allowed to contextualise my thesis, the present chapter will allow to appreciate what it means to reconceptualise the deepfake. This conceptual revisitation of the deepfake is based on Karen Barad's new materialist development (2007). Which is why the Baradian-inspired conception of the deepfake is called *deepfake Baradianism*. The argument will go that this reconceptualisation allows to redress the previously considered shortcomings in both deepfake fatalism and deepfake realism, and that it therefore provides a conception of the deepfake that better aligns with its observed reality.

Before dashing straight into Barad's world, I will briefly sketch the philosophical movement of new materialism. And then, after having brought deepfake Baradianism to fruitful maturation, I will end this chapter with a summarised overview of what deepfake Baradianism means. At this point, simply remember that the notion *deepfake*—from which we depart in this revisitation—refers to synthetic political disinformation that seeks to undermine trust in public institutions.

a new materialism briefing

Let us very summarily sketch the materialism family tree to contextualise new materialism. In its very generic understanding, materialism is about the attribution of importance to matter. Our society is often coined as materialist for its consumerist behaviour, where one hoards evermore material objects as if symbolising one's social status. This understanding leads us to a first distinction to be made between *post*-materialism and *new* materialism. Post-materialism signifies a shift from an interest in material security to an interest in self-expression (Miller, 2013). A second key distinction is between *new* materialism and (what some call) *old* materialism. To keep things scanty, compared to old materialism, new materialism attributes more agency to matter[11]. In new materialism, matter is not just an amalgam of given, finite, hermetic entities or building blocks. And it is not passive either (Gamble, Hanan & Nail, 2019). Matter is a continuance of that which is; a constitutive essence in constant flux. Dynamic and changing. Matter *is* extension, and therefore dichotomies, that immanently seem real, transcendentally no longer hold (quite a shortcut at this point but this will become crystal clear soon).

From that premise, new materialism forks into yet another set of diverse conceptualisations, but it is

[11] And of course, this demarcation is itself subject to debate.

Barad's understanding as introduced in their seminal work *Meeting the Universe Halfway* (2007) that is elaborated on below. In a nutshell, Baradian new materialism is about attributing agency to matter all the while not falling back into the old materialist idea that matter has a pre-existence to which we have a "mediated access" (p. 152). In other words, matter is independent from human intervention but it is not existing out there independently of our interaction with it. Time is now ripe to discover how Barad actually conceives of reality. And therefore note that whenever I will distinguish the Baradian conception from a so-called *classical* conception of reality, it will signify its distinction from a non-Baradian (old) materialist conception of reality.

deepfake Baradianism · reconceptualising the deepfake

This section discusses the revisitation of the political deepfake through the Baradian new materialist lens. The interest here is thus to consider what type of meaning can emerge from the deepfake by using Barad. I will start by introducing Barad's concept of *diffraction*. Diffraction is essentially about considering reality as an assemblage of interferences, as one blurry ensemble, rather than as a combination of distinct entities constituting a whole. This reviewed take on reality will allow us to appreciate a revised understanding of the notion *sociotechnical*. And it will allow us to appreciate why the doomsday prophecy is a partial and therefore deterministic conception of the deepfake reality. This appreciation will smoothly pave the way toward Barad's next concept, the *phenomenon*. The phenomenon brings us to consider that all that constitutes reality is in permanent entangled reconfiguring. Since Barad understands the phenomenon as the ontological primitive relation making up reality, it will allow to derive more ontologically founded conceptions of the deepfake; those conceptions that diffraction would so far merely hint at. Next, we will dive into Barad's foundational query about the relation between matter and meaning, which will allow us to appreciate why the deepfake is better conceptualised as speech than it is commonly conceptualised as content. Lastly, we will then cover the remaining question of interest here. Namely, what is objectivity and is it attainable if all is diffraction and interference. The subsequent chapter will provide a recapitulation of how to understand the revisited conception of the deepfake.

Because of its theoretical nature, this section will be a little denser than the rest of the thesis. However, to shape this in an as pleasant conceptual stroll as possible, I will expand on Barad's work step by step, subsection by subsection. Each subsection will thus be split in two parts. In the first part, I will always develop their work as will be necessary to bring this revisitation to fruitful maturation. In the second part, I will then always apply Barad's thinking to the deepfake.

diffraction, exploring the world in its differences

Contrasting to a classical understanding of reality as being constituted of a variety of entities that combine into a whole, like a mosaic, diffraction is essentially about considering reality as an assemblage of interferences that constantly reshape that same reality. Not only will it lead us to appreciate the deepfake as being about a process, a doing. But it will also question the validity of one-sided stories that claim to provide the full picture of reality. It will thus already hint at the requirement to move beyond a dualist perception of reality, and will thereby allow to have a first appreciation for the reason why a demarcation of fact from fake information is not straightforward a process.

Barad on diffraction

To rethink the common way of being reflexive, Barad builds on the diffractive methodology as developed by Donna Haraway. In brief, where reflexivity is about reflecting on your self through the other, and leads to a mirror play that keeps us in a dualist understanding of self and other, and remains centred on the self, the notion of diffraction in Baradian terms is about diffracting that which is. Namely, it is not about understanding something as a sharp clearcut image, but as propagating waves, prolongations, interferences, and entanglements. Note that in Baradian terms, entanglement is not just about the complexity of the world, but also about its constant reshaping, its constant "topological reconfiguring" (2007, p. 160). In other words, diffraction is about "patterns of difference" (p. 71) that allow for a decentred exploration of the world. Diffraction is about relations rather than positions. Self and other are not opposite reflections but in a relation of difference, in attunement. So, while reflexivity remains in a representationalist understanding of the world that would consist of a pre-existing subject and object, in a diffractive understanding of the world, "subject and object […] emerge through intra-actions" (p. 89)[12].

Intra-action "signifies the mutual constitution of entangled agencies" (p. 33). This contrasts with *interaction*, which "assumes that there are separate individual agencies that precede their interaction" (p. 33). To Barad, there is no such thing as pre-existing entities, as much as there is no such thing as pre-existing agencies. "[A]gencies are only distinct in relation to their mutual entanglement; they don't exist as individual elements" (p. 33). In other words, agency is not understood as the traditional action in its resulting end (the causal *being*), but as the action in its (re)configuring (the causal *doing*). Since all is in constant diffractive entanglement, in constant intra-action, agency is the process, not the result.

No need to feel dizzy at this point, because diffraction is not a eulogy of some extreme transhumanist

[12] In my understanding, diffraction is a form of reflexivity that some already naturally practice. So, perhaps the understanding as described by Haraway and Barad simply complement each other and therefore more so allow to engage in *reflexive differences* or *diffractive reflections*? Aye! Why not complicate things.

megalomaniac extravaganza where us humans could undo ourselves from our envelopes. Barad is not on a journey in Trippy Queendom of Farce where bodies would melt into one another. Barad does not "denigrate separateness as mere illusion [but they do] not take separateness to be an inherent feature of how the world is" (p. 136). Things are in relation. "Difference cannot be taken for granted; it matters—indeed, it is what matters. The world is not populated with things that are more or less the same or different from one another" (p. 136). In decentring the human, diffraction is thus not about forgetting our human positionality, it is not about flying high in transcendental skies. Diffraction is very much about being grounded and attuned to other modes of doing; acquiescing other modes of agency besides the human.

From there, Barad then asks, "how to responsibly explore entanglements and the differences [these entanglements] make" (p. 74). Which is how diffraction shifts the research interest from a reflexive focus, that is interested in the "correspondence between descriptions and reality [between words and things] (e.g., do they mirror nature or culture?)", to a focus that is interested in "practices, doings, and actions" (p. 135).

the deepfake as diffraction

Now what does diffraction do to our deepfake?

Let us first appreciate how a diffracted understanding of reality very much aligns with the idea that the deepfake has a sociocultural dimension, and thus aligns with the idea that the deepfake is both a technological and a social matter. In diffraction, all is entangled and in constant intra-action. Therefore, the interpretation of the deepfake in its pristine sense as a given digital or technological entity (as a static) does not hold. The deepfake is understood in its processual sense as a constant sociotechnical becoming (as a dynamic). If I were to speak of the deepfake in its pristine sense, as the being, this would mean that the deepfake *is* rather than *becomes*. It would mean that the deepfake has a pre-existence relative to something else, that it has precedence, that it is an immaculate given rather than a configuring. Whereas in diffraction, the deepfake is understood in its sociotechnical nature as a doing. As an occurrence where social order and technology mutually interfere and where none has precedence over the other. Note thus the interpretative nuance compared with the common understanding of the sociotechnical dimension of the deepfake where either social order or technology have precedence over the other (and thus where either have pre-existence, unlike the Baradian interpretation).

Diffraction further also allows to question the validity of one-sided stories that claim to provide the full picture of reality. In diffraction, reality being about a constant intra-action that is made of ever-so different and entangled patterns, the perception of reality therefore is ever-so diffractively different to the agents part of that reality. In other words, the sociotechnical reality of the deepfake, its doing, is interpreted differently by different agents. If focusing on human agents, it means that different minds

experience the reality of the deepfake diffractively differently since every mind processes information diffractively differently given that every mind experiences reality however-so diffractively differently in space, in time, and in matter. This means that the popular doomsday prophecy is only one diffracted conception of the vast deepfake reality, of its vast pattern of difference. For instance, another conception of the deepfake reality is that deepfakes offer a way to enhance critical thinking, a critical engagement with reality due to the surprise and doubt that deepfakes engender in our minds, thereby making us question our thoughts and beliefs, be it to result in their subjective validation. Let us be mindful again that these diffracted conceptions are neither pre-existing nor impermeable finite pieces of the diffraction. These various conceptions can inhabit us all at once to varying degrees and in different nuances and shades. We are all an amalgam of various thoughts that at once can converge, conflict, reconfigure, …. And this is why we all develop different scales of evaluation of how to perceive reality and thus how to conceptualise it. It is the infinity characterising diffraction that leads us to appreciate the inaccuracy of the doomsday prophecy as being representative of the reality of the deepfake in its entirety. Not only does this conception that envisions the deepfake as a mothership of disinformation, as a given that ought to be excised from society, remain of the order of the conceptual, of the speculative. But the lack of validation of the doomsday prophecy in present-day scientific research, in fact, makes that the tragedy narrative is more than merely conceptual. It is of the order of the mythical, the imaginative, since it does not align with the observed reality of the deepfake.

Diffraction leads us to a third and last implication for the deepfake. Namely, thinking diffractively of the deepfake implies the need to move beyond an at times debilitating dualist conception of reality. Diffraction blurs any previously enacted clearcut demarcations as inherited from the classical conception of reality. Interfering entities cannot be one-another's mirror reflections, since by nature they are mixing, intra-acting. They cannot be opposites. Therefore, the popular fact-fake and information-disinformation dichotomies are an incomplete picture of reality. Fact and fake, and information and disinformation are complexly entangled, continuously reconfiguring composite stories that are therefore more than two. They are not "more than two" in a sense of adding up to more than two, but in a sense of not being distinguishable in more than one—since nothing pre-exists and all is confounded. At the same time, they are more than one since they are is diffraction.

phenomena rather than things

Building on diffraction, the phenomenon allows to appreciate more profoundly this diffracted conception of reality. To Barad, phenomena are the ontological primitive relations making up the world, without being finite since nothing pre-exists. In terms of implications for the deepfake, the phenomenon will lead us to appreciate the intelligible character of the deepfake. More than being a doing—as was derived through diffraction and which exacerbated the dynamic and participative character of the

deepfake—the deepfake now becomes endowed with a quality of agency. This will in turn bring us to understand the ontological foundation for the absence of externality. An absence which will imply that all knowledge is partial. An observation that not only again questions the validity of one-sided narratives, but which will lead to the more profound implication that the deepfake reality can never find an end since society and deepfake are ontologically inextricable.

Barad on the phenomenon

Since nothing pre-exists in a Baradian conception of reality, then, what is reality made of? All that is—all that we feel, touch, smell, see, hear—is not pure imagination or some sort of continuous hallucinatory state of being, as if the constructivist notion that thinking precedes reality (mental impressions) would hold alone. To conceive of what reality is made of, Barad departs from Niels Bohr's work and proposes to understand phenomena as the ontologically primitive relations making up the world. But! relations without pre-existing relata (ha!), since in a Baradian spirit all is intra-action. Nonetheless, hold your horses, phenomena can take the form of real physical bodies or perceptions of the human mind. It is just that these cannot be "fixed and separately delineated things" (p. 129). Phenomena are entangled material agencies; inseparable intra-actions that "don't exist as individual elements [but exist distinctly only] in relation to their mutual entanglement" (p. 33). Phenomena are distinguishable only in relation to their mutual entanglements[13].

But then, what is this—what I would call—inseparable distinctiveness? It does sound paradoxical, doesn't it? It is as if I could feel Descartes' ghost peeping from around a corner of my mind. As if my Cartesian upbringing could not leave me wander into other ether where things are not because of some Immaculate Material Conception but because of a coming into being that is not secluded to the thing itself but becomes through relations. At once it sounds understandable and of the order of wizardry. How Barad sees it, is that the inseparable intra-actions that characterise phenomena "are nonarbitrary non-deterministic causal enactments through which matter-in-the-process-of-becoming is iteratively enfolded into its ongoing differential materialization" (p. 234). Firstly, this *causal enactment* is not to be confused with an interaction between pre-existing secluded entities that would allow to distinguish between externality and internality. It is about an intra-action within phenomena. In other words, it can be interpreted as the proposition that there is no absolute externality, and that instead there are relative

[13] A certain discomfort kicks in at this moment (and never really finds absolution), because where Barad seeks to give matter its agential due independent of human intervention, I am at odds with this claim that phenomena are the ontological primitive relations making up the world, given that this claim is then founded in a human way of experiencing the world. Who knows how a stone or a galaxy or a penguin or a doorknob or a Praying Mantis or a solar flare or a yet unknown something or an outer-space creature experience reality? They have differential organicity. One that the human will never be able to grasp. For instance, we humans do not see infrared light. We make it visible to us through devices, but infrared could very much mean something else to something else. Our sensory topology only allows us to experience a certain set of reality.

ones within phenomena. Which is also why "scientific results are reproducible[. Otherwise] it would be impossible (or at least very difficult) to account for the reproducibility of experiments" (p. 131). Secondly, the *becoming* does not refer to change as commonly understood as "a continuous mutation of what was or the unravelling of what will be, or any kind of continuous transformation in or through time" (p. 179). It is about "the iterative differentiatings of spacetimemattering" (p. 179). Indeed, to Barad, space, time, and matter all get mutually intra-actively configured and differentiated within phenomena. With phenomena, matter thus becomes a dimension in and of itself, just like time and space. And it is by considering that dimensional role of matter that a distinction between subject and object can be made within phenomena; it is because matter is a dimension that objectivity is possible. Upon the enactment of a particular practice of observation within a phenomenon—a practice of knowledge-making—a distinction is then being differentially materialised.

However, again, let us not be mistaken that such enactment does not lead to a representation of "inherent properties of subjects or objects" (p. 208). Otherwise, we would be back into the swirls of old materialist thinking where matter is a given. Which further would lead us back to the—in a Baradian thinking—delusional belief that there is such thing as an omniscient Objectivity, an absolute Truth. The material differentiation that has been enacted by a particular knowledge-making practice is only "determinate for [that] given practice" (p. 155). And that is how inseparable distinctiveness is to be understood. Namely, as the idea that knowledge-making practices "do not uncover preexisting facts about independently existing things as they exist frozen in time like little statues positioned in the world[. Knowledge-making practices uncover] phenomena" (p. 90-91). "Reality is composed not of things-in-themselves or things-behind-phenomena but of things-in-phenomena" (p. 140). "[H]umans (like other parts of nature) are *of* the world, not *in* the world, and surely not outside of it looking in" (p. 206, original emphasis).

Importantly, Barad therefore "does not subscribe to a notion of truth based on correct correspondence [of the existence of descriptions and reality]" (p. 56), since words and things do not pre-exist. Thus, while in the common Western perception of reality, separability (determinacy) is a condition to objectivity, in Bohrian-Baradian terms, it is the *absence* of inherent separability that is the condition to objectivity! Namely, we cannot know as some mysterious external entity. It is about "knowing as part of being" (p. 341). Because how else are we supposed to know something if we are not part of it?

the deepfake as phenomenon

Understanding the reality of the deepfake as a phenomenon firstly implies an *ontological* inextricability of deepfake and society. With that understanding, the deepfake becomes more than a dynamic participative doing as derived from diffraction. It becomes endowed with agency. It becomes intelligible, capable of enacting a practice of observation, capable of differentially materialising its reality. The deepfake has an organic existence so to speak—it is alive. A conception of the deepfake

that will sometimes be emphasised by referring to the deepfake as *deepfake reality*.

Secondly, since any knowledge-making practice or act of differentiation materialises the deepfake in relation to that very practice alone, the deepfake will never materialise in a given entirety because of the impossibility to access an external stance. In other words, "since there is no outside to the universe, there is no way to describe the entire system" (p. 351). "[N]o observer inside the universe can see all of what is in the universe" (Smolin cited p. 351). Any knowledge produced about the deepfake, any conception of it, is thus always only partial.

This absence of externality brings us to a third implication. Namely, that any proposed interventionist act for the excision of the deepfake reality in its entirety is utopian. In a Baradian sense, seeking its excision would in fact equate to a suicidal act, an act of annihilation of humanity, since deleting the deepfake phenomenon would require a complete expunction of society since deepfake and society are ontologically inextricable. In absolute terms, there is no end to the deepfake reality. Even if deepfake technology would someday be supplanted by another technology, it will simply cause a differential reconfiguration of that reality. It will not end it. In a Baradian new materialist conception of reality, where reality is in the doing rather than in the being, in constant reconfiguring rather than reaching finitude, any interventionist act of its excision will only excise the deepfake in relation to the very practice of its excision alone. It will not excise it all-together.

how matter and meaning relate

All-in-all, what Barad does in their work is trying to establish an ontological foundation to make a stronger case for why matter matters. That is what brings them to then focus on the relation between matter and meaning. Because if all of reality is intra-action and nothing pre-exists, then how can we come to know something if that something is then already reconfiguring? In other words, how is knowledge possible in an ever-interfering world? But to be able to respond to this question—a question that regards the possibility for objectivity—we have to make a brief pit stop to fuel on Barad's notion of *meaning-mattering*. That is what we are about to do here. Essentially, it will allow us to appreciate that the deepfake is more than content; that it is speech. This is an important shift in conception that can have insightful regulatory implications for policy approaches that usually tend to focus on a regulation of online content, where content is understood in its classical sense as an excisable pristine given—devoid of context. In a Baradian conception of reality however, this understanding does not hold.

Barad on meaning-mattering

In the previous subsection, we discussed how knowledge-making practices are the processes through which a differential materialisation is possible within a phenomenon. A differentiation that is bound to

the particular practice that was enacted, and which therefore does not give access to "inherent properties of subjects or objects" (p. 208). In Baradian terms, "[m]eaning is not a property of individual words or groups of words but an ongoing performance of the world in its differential dance of intelligibility and unintelligibility" (p. 149). We can easily agree with this when thinking about how words or their definitions are never static but change with their practice, with their use.

The idea that *meaning* is not a property of intra-active *matter* has four implications: (1) everyone and everything has access to knowledge, (2) knowledge is peculiar as it depends on the particular intra-action at hand, (3) knowledge once acquired is at that same instant already obsolete, (4) knowledge is at once formed, performed, and reformed.

The first implication means that "the forces at work in the materialization of the bodies are not only social, and the bodies produced are not all human" (p. 33-34). For otherwise one would fall back into social constructivist precepts. This was also already expressed upon acknowledging the deepfake as intelligible. The second point implies as already discussed that objectivity is relative and that there is no omniscient Objectivity. The third implication hints at an issue of range. Not merely temporarily speaking but very much spacetimematter-wise. Because although reality is in constant spacetimematter reconfiguration, there is nonetheless the reality of our human composure that we cannot transcend. And this human composure (whether physical or spiritual) frames the range of our possibilities and, therefore, the range of accessible perspectives as much as the composure of blowflies frames their experience of reality in a way that will never be accessible to us. If understanding everything as diffractively intra-acting, time has no inherent meaning but we do grow old in spacetimematter for instance. Barad never truly addresses this question of range, but, how I see it—while knowledge once acquired is at that same instant already obsolete—there are infinitesimal spacetimematter brackets during which knowledge acquired remains valid even though it is already intra-acting, because of the very distinct relative ranges that characterise for instance reality's vastitude compared to our human composure, our human reality[14]. Not only is this question important for the crafting of science, of knowledge that remains valid across spacetimematter. But it is equally significant in terms of policymaking; in terms of matters that matter in the short term. We cannot remain in some sort of transcendental state where we could remove ourselves from any matter of range, however nice and ecstatic it feels. In fact, we could think that it is the very mutable nature of reality that keeps us grounded. For it is this constant dynamism and our relation to it that would keep recalling us of our human composure. Now, back to the relation of matter and meaning, the fourth implication is that which Barad dedicates most of their attention to. Namely, to the doing of knowledge, coined as *meaning-mattering*. Meaning-mattering is part of Barad's key theoretical development—which they called *agential realism*—to which all of that which is being

[14] Of course, we could also think of the idea that various intra-actions give rise to similar experiences of reality and thus allow for the event of shared common knowledge. Thus, having an experience of reality that is repeated rather than falling into obsolescence. But I would argue that remaining aware of our human composure is necessary, for otherwise we would risk slipping again into the murky waters of transhumanist extravaganza.

described in this chapter leads to, and more. But however much theoretically contemplative the depths of that "and more" are, I will not delve in those meanders which are of less relevance for the present intent being about reconceptualising the deepfake. In a nutshell though, agential realism is about acknowledging matter's agential due in the process of doing knowledge, in the process of meaning-mattering. It is about acquiescing that matter has an organic existence, that it is alive, in constant intra-action, be it with or without human interference. Because remember that knowledge-making practices (intelligibility) are not a feature of the Anthropos alone[15].

So, about meaning-mattering, how is this intra-action between matter and meaning to be interpreted? Barad addresses this question through the material and the discursive practices of knowledge-making as previously elaborated by Michel Foucault and Judith Butler upon working out whether (*if*) matter and meaning are entangled. Beyond the if-question, Barad now proposes a development about *how* matter and meaning are entangled. Not being familiar with theoretical movements about knowledge, I was finding it interesting at first that knowledge-making practices are only thought of in terms of material and discursive practices. Why? Are other meaning-mattering proceedings deemed irrelevant? Why though? Because if thinking of how I make sense of reality, there is an important part of sub/conscious sensorimotor participation. An in-between the material and the discursive. Material-discursivity thus feels somewhat of a reduction. But we will shortly see that Barad reconceptualises matter and meaning in a way that makes room for such broader, hybrid understanding.

Let us start with the *material practices* constitutive of the knowledge-making practices. Matter is classically understood as a finite, inert, pre-existing entity. To Barad, matter is not passive but dynamic and in continuous reconfiguring. Matter is intra-action. It is "[a phenomenon in its] ongoing materialization" (p. 151). The material practices are the becoming concrete. Remember that becoming does not refer to *change* in its classical sense as "a continuous mutation of what was or the unravelling of what will be, or any kind of continuous transformation in or through time, but [that it refers to] the iterative differentiatings of spacetimemattering" (p. 179). *Matter* is the agential intra-action that brings a phenomenon to its materialisation.

Similarly, the *discursive practices* are not understood as anthropomorphic linguistic projections, but as "ongoing agential intra-actions of the world" (p. 148-149). Discursive practices "enact causal structures through which some components (the "effects") of the phenomena are marked by others (the "causes") in their differential articulation" (p. 149). *Meaning* is the agential intra-action that enacts a knowing within the materialised phenomenon.

In other words, where the material practices are about "the iterative production of different differences"

[15] On a side note, the struggle becomes quite striking of this inevitable anthropomorphism upon seeking to acknowledge matter as a meaning-mattering phenomenon that does not require human intervention to do so. An anthropomorphism that Barad either deems unnecessary to address because it would be clear to them that using human language to utter a conception is in itself implying anthropomorphism, or because they are themselves at odds with it in the sense that they seek to provide an ontological foundation for why matter matters outside of our existence.

(p. 137) that bring phenomena to their materialisation, the discursive practices "[enact specific determinacies] within the phenomena produced" (p. 149). The material and the discursive cannot exist without each other—they are "mutually articulated" (p. 152), mutually hybridised. Matter and meaning co-become. In a way, one can understand this as a moment where Barad distinguishes between realities (where matter and meaning co-exist) and myths (where matter and meaning fall short of each other).

the deepfake as an instance of meaning-mattering

Meaning-mattering is the process of creating knowledge (meaning) out of something (matter); an intra-active process of enacting different differences (meaning) within phenomena (matter). What does this notion do to our conception of the deepfake?

Firstly, the deepfake reality cannot be about pre-existing material practices only. It is inherently infused with discursive practices. Which joins the previous reconceptualisation of the notion sociotechnical where neither deepfake nor society have precedence. Which in turn requires another shift in conception. Namely, by requiring us to move away from conceiving of the deepfake as a pristine entity and to acknowledge its processual nature, it requires us to move away from its conception as content and to consider it as speech, as a medium of expression. Commonly, the deepfake is conceptualised as if it were a given piece of digital fabric; a content lacking depth and which could therefore be excised from society through some magical surgical move. However, meaning-mattering brings us to appreciate the depth of the deepfake; its more than material being. It brings us to appreciate the meaning attached to the deepfake. A meaning that is attached to it because of its nature as medium of expression, as speech. The deepfake is a medium, it is a relation. The deepfake is indeed not defined by an intrinsic something; the deepfake is defined by our relation to it. It is therefore that the Baradian conception causes the deepfake to better align with the notion of speech than it aligns with the notion of content.

Secondly, a material-discursive understanding of the deepfake implies the inadequacy of fact-checking efforts. Indeed, by seeking to attribute a label of either fact or fake to a deepfake, fact-checking devices seek to break the relation between matter and meaning, between content and speech, between the technology and its sociocultural dimension. Whereas in a Baradian understanding, matter and meaning are ontologically inextricable.

the possibility for objectivity when content is ever reconfiguring

Having travelled all this way and having had a last conceptual pit stop at meaning-mattering, we can now come to fully appreciate the final drive down the Avenue of Objectivity before reaching our destination, the Baradian revisitation of the deepfake. This final drive looks into the question of how

knowledge is possible in an ever-interfering world. Key here will be an appreciation of why the popular idealised conception of fact-checking is erroneous and why its application is therefore bound to fail.

Barad on objectivity

Despite all previous elaborations, the ultimate question remains open. Namely, if all of reality is interference and in constant reconfiguring, then what is objectivity? To respond to this question, Barad departs from the quantum physics experiments on particle-wave duality as was intensely debated between Bohr and Heisenberg. I will not dwell in the details (and suggest reading Barad for that), but in short, in the early 20th century, experiments exposed that light at times not only behaves as a wave but also as a particle. And that similarly, a particle (be it an elementary particle, an atom, or a molecule) at times not only behaves as a particle but also as a wave[16].

Ultimately, why not. It is not because things are not imagined that they do not exist. But, let's be honest, for the curious animals that we are, not understanding is oftentimes bugging. And while this is not peculiar to the scientific kindred, scientists were set on a mission to conceive experimental setups in order to reconcile their theories with the observed reality. What this mission to adjust the experimental setups to new findings firstly means, is that "the nature of the observed phenomenon changes with corresponding changes in the [device for observation]" (p. 106). Which is not a shocking revelation[17], but it proves that no clear distinction can be made between internal and external stances of knowledge-making. It proves that matter and meaning are inextricable. In other words, it proves that that which is observed cannot have a determinate pre-existence but that it acquires a determinate existence through the specific intra-active configuring of the observation[18]. Note that to remain aligned with a Baradian thinking, "observations do not refer to properties of observation-independent objects" (p. 114). As we have seen before, knowledge-making practices do not give access to "inherent properties of subjects or objects" (p. 208). "[O]bservation is only possible on the condition that the effect of the measurement is indeterminable" (p. 113). This is easily acceptable in the sense that if it were determinable, it would no longer be an act of observation but an act of orientation.

Now, what an agential intra-active Baradian constant reconfiguring and becoming of the world also implies is that humans cannot "determine the outcome or play an "interventionist role"" (p. 131). Because however much practices of observation, of knowledge-making, are indexical material-

[16] Heisenberg deduced *uncertainty* from this phenomenon (i.e., that the more one knows about particle position, the less one knows about particle momentum; thus entailing that both parameters can be known simultaneously but at varying degrees of exactitude). Bohr deduced *complementarity* (i.e., position and momentum cannot be measured simultaneously, and which-one is observed is dependent on the device for observation). Both deductions are based on distinct conceptions of internality-externality.

[17] But which does not mean that it is always easy to sit with in all instances in life.

[18] For Bohr, this meant that ""wave" and "particle" are classical descriptive concepts that refer to different mutually exclusive phenomena, not to independent physical objects" (p. 198).

discursive constructs, they themselves remain phenomena. Observations do not follow human commands only, since humans are not the sole participants in this becoming. As written nicely by Barad, "[observations] are not merely about us" (p. 142). "[K]nowing is a matter of part of the world making itself intelligible to another part" (p. 185). But what us humans *can* do, is observe the world in its becoming, because, as we discussed earlier, meaning-mattering practices enact causal structures within phenomena. They enact an "agential separability—the agentially enacted material condition of exteriority-within-phenomena—[that] provides the condition for the possibility of objectivity" (p. 175).

the deepfake and relative objectivity

Transposing Barad's postulate to the case of fact-checking devices means that these devices enable to distinguish between an information input (the observed) and the reader (the observer, be it a human or a machine). There is an important logical derivation to remember here, for, in contrast to what most believe, fact-checking devices thus do not allow to distinguish between fact and fake! Fact-checking devices do not allow to distinguish between two pieces of content. What fact-checking devices do allow to do is to assess the objectivity of content relative to the observing entity—relative to the device itself. Which is why the construction and setup of any fact-checking device is key, all the while remembering that one can always only know as part of what it observes. Therefore, no fact can be determined by one without that one being implicated in it. There is no possibility for externality, no possibility for absolute factuality, and thus no possibility for a fact-checking that is devoid of context, devoid of sociocultural interference. That is why there is no absolute demarcation possible between the good and the bad, the right and the wrong, …, the fact and the fake. All is ever partial. And all is ever becoming. To which Niels Bohr added that "we can't know something definite about something for which there is nothing definite to know" (p. 118). To give an example, we all know of political content—whether intentionally fake or not—that was once considered factual in our Western history but which has been debunked in the meantime (although remaining perspectively true in some minds). For instance, racist, classist, sexist, … content. Which thus exposes how the issue is not the deepfake, but the instances of social interference. A conclusion that was also drawn in literature on deepfake realism, where scholars call on the need for attention to the systemic character of deepfakes.

Perhaps we could thus explore the deepfake in its diffractions rather than letting us be seduced by the doomsday prophecy, however alluring the latter might be sensationally speaking. An exploration that would allow to better "theoriz[e] the relationship between [information] and [disinformation] together, without defining one against the other or holding either [information] and [disinformation]" (p. 30). In their work, Barad originally talks about "theorizing the relationship between "the natural" and "the social" together without defining one against the other or holding either nature or culture as the fixed referent for understanding the other" (p. 30). And they do so in reference to Judith Butler's postulate

about gender, that "cultural assumptions regarding the relative status of men and women and the binary relation of gender itself frame and focus the research into sex-determination" (p. 61). Thus, transposing this to the case of deepfakes, cultural assumptions regarding the relative status of information and disinformation and the binary relation of objectivity of information itself frame and focus the research into fact-fake demarcation. An endeavour thereby questioned by a Baradian conception of the deepfake. And a questioning that in turn aligns with scholarly calls to not focus on assessing content alone but equally so on assessing our relation to that content (e.g., Lecomte, 2021).

deepfake Baradianism summarised

The whole point of taking you on this conceptual stroll was to develop an ontologically founded reconceptualisation of the deepfake. To develop the concept of deepfake Baradianism as means to redress the shortcomings observed in deepfake fatalism and to redress the subsisting shortcomings in deepfake realism. A necessary re-theorisation of the deepfake to better align the conception of the deepfake to its empirical observations—to re-align the deepfake to its reality, the matter to its meaning. A re-theorisation of the deepfake that I therefore argue to be an ontologically justified basis to serve as guiding premise for the crafting of policy recommendations.

This reconceptualisation happened gradually throughout the various explored concepts of Barad. Each concept allowed to derive refreshed understandings of the deepfake reality, refreshed understandings that are recapitulated here.

(1) Beyond dualism and beyond tragedy

In a diffracted reality, nothing is "fixed and separately delineated" (Barad, 2007, p. 129). A Baradian conception of reality thus problematises the ideal of clearcut demarcations; it problematises a dualist conception of reality. Reality is not made of binary opposites. And it is therefore inaccurate to describe it through dichotomies like fact-fake, information-disinformation, human-technological. Not simply because those dichotomies are not pristine entities—free of human interference—but also because in diffraction there is more to reality than just two sides since all constantly interferes and thereby creates an infinity of patterns of difference.

Because the deepfake reality is more than black-or-white or good-or-bad, narratives describing that reality as either the one or the other are advisably questionable. And it results in the fact that the popular fatalist tragedy narrative is not merely empirically inaccurate (given the lack of evidence in current research findings), but it is also conceptually inaccurate (at least, according to a Baradian conception of reality).

(2) The deepfake as an intelligible doing

Understanding the deepfake as a doing rather than a given requires endowing the nonhuman with a capacity for intelligibility, a capacity for producing knowledge. Insufflating the nonhuman with

agency thus implies the acknowledgement of the discursive dimension of the deepfake besides its material dimension. In other words, it implies the acknowledgement of the human dimension of the deepfake besides its technological dimension. The deepfake is no longer a sudden Immaculate Conception free of original sin. The deepfake is now an agent that has a material-discursive reality that is capable of differentially enacting its reality—the deepfake is speech.

(3) The deepfake and society as inextricable, and the agential character of civil society

The intelligible character of the deepfake requires regulatory attention to focus not only on its material agencies (its technological reality), but equally so on its discursive agencies (its social reality). Deepfake and society are inextricable, for the deepfake cannot be addressed without also addressing its social reality. The human and the technology (society and the deepfake) "[converge] into a composite entity 'that has unique properties not available to either'" (Kalpokas & Kalpokiene, 2022, p. 80). Deepfake Baradianism thus subscribes neither to technological determinism nor to human centrism.

Additionally, acknowledging the social inextricability of the deepfake requires a consideration for the agential character of civil society—an actor in the play that has been so far continuously neglected in both deepfake fatalism and deepfake realism. The popular doomsday prophecy bereaves civil society of its agential integrity. It incapacitates it. Thereby paving the way to policy approaches that found themselves in a paternalistic thinking where the netizen ought to be protected from something it would have no influence over. A paternalistic thinking that I thus argue to be unjustified given the lack of empirical evidence for the doomsday prophecy.

INTERLUDE-TO-READ • a reviewed understanding of the notion sociotechnical

Despite acknowledging the social character of the deepfake upon acknowledging its more than technological being, deepfake realism still abides by a classical understanding of the sociotechnical reality of the deepfake. A reality where deepfake and society are in interaction; where social order has technological precedence and where the deepfake simultaneously has sociocultural precedence. Not only does this classical conception of mutual precedence prevent an appreciation of the true ontological inextricability of deepfake and society. This classical conception also traps it in an interpretative paradox. Namely, if society and technology have mutual precedence, it means that both simultaneously pre-exist—that both at once precede—which thus implies that none can have precedence. By highlighting an interference between society and technology, a diffraction where none has precedence, the Baradian conception thus allows to move out of this paradox.

(4) The absence of externality

Understanding the deepfake as a Baradian phenomenon brings us to consider that when we seek to know the deepfake, we are part of that phenomenon. One cannot know from a remove. One cannot know without being in touch with that which it seeks to know. This absence of an externality to the observed phenomenon implies that knowledge about the deepfake is only ever partial and is bound to the knowledge-making practice of the observer. Policy-wise, this implication questions the efficacy of standardised top-down approaches that lack an appreciation of the idiosyncratic character of the deepfake.

The absence of externality also translates in an absence of Objectivity, an absence of impartial knowledge. Policy-wise, it points at the need to trust the netizen in its capacity for autonomous judgement, since in absolute terms there is no such thing as privileged knowledge, and thus no-one can claim to know better. Which joins the previous problematisation of the efficacy of standardised top-down approaches.

It also questions the ideal of a technological fix that rehearses a conception of the deepfake as a pristine entity that emerged externally to our social reality—free of social interference—and which rehearses a conception of deepfake detection devices as being equally free of a sociocultural dimension. An exclusively technological focus to address the deepfake is inherently ineffective.

(5) The deepfake will never find an end

The verb *doing* underscores that the deepfake has agency. That the deepfake is not simply an entity prone to be subjected to whatever external act is practiced on it, also because there never is a true externality in a reality where all is intra-action. The deepfake is not an excrescence that can be cut out of society. Any interventionist act for its excision is bound to fail if seeking to put an end to it, for any such act will merely reshape the deepfake reality differentially otherwise. This conception further paves the way toward what Kuhlman and Rip (2018) called *tentative governance*; i.e., "[a governance that is] provisional, flexible, revisable, dynamic [with] open approaches that include experimentation, learning, reflexivity, and reversibility" (p. 450). It is about crafting policies that are open-ended rather than aiming at finding closure.

(6) The deepfake as speech

While many if not all discussions about the deepfake mention its intricate relation to freedom of expression, its descriptions and proposed policy approaches almost systematically centre on the deepfake as content, devoid of social context, and which can therefore be separately delineated and excised. But we have seen from the material-discursive intelligible doing of the deepfake that this idea does not hold in a Baradian conception of reality. And that understanding the deepfake as a medium of expression, as speech, allows to redress this inaccuracy.

(7) Beyond fact-checking

Several of the above refreshed conceptual understandings of the deepfake converge toward the argument that fact-checking—in a Baradian logic—is bound to fail. Envisioning reality as

diffraction—and thereby invalidating clearcut demarcations and dichotomies, and acquiescing the absence of externality—means that fact-checking cannot provide the answers it is currently sought to provide. Fact-checking cannot demarcate fact from fake. Fact-checking can only demarcate content relative to the fact-checking device.

Any act of demarcation within reality is bound to the observing entity. An observing entity that has a value-laden nature because of its discursive and therefore sociocultural nature. It is therefore impossible in absolute terms to enact clearcut demarcations to distinguish between factual content and fake content, since any such enactment is only ever true to the observer enacting that demarcation. Therefore, the popular consideration that fact-checking allows to impartially—factually—demarcate fact from fake is ontologically inaccurate. Which in turn requires a more profound acknowledgement of the diffracted reality that is characteristic of the receptiveness of society to fact-checked content. Even if enacted with utmost integrity, any fact-fake demarcation will thus intrinsically by received with great dissimilarity by its audience. Therefore, again, fact-checking will not provide the commonly sought-after end to the deepfake.

Despite that the Baradian conception problematises fact-checking, and despite that research uncovers its limited effectiveness, approaches to countering deepfakes and to content moderation are today still predominantly focused on fact-determination. This is not to say that fact-checking efforts are unnecessary overall, for they do allow to enact instances of meaning-mattering that are informative. However, fact-checking should not remain the sole regulatory focus since those fact-checking enactments are only ever-so diffractively informative.

So, this is it for deepfake Baradianism. From now on, the notion *deepfake* will still refer to synthetic political disinformation that seeks to undermine trust in public institutions, but! in its Baradian understanding as an entanglement of material-discursive agencies in constant doing. The deepfake is thus not a pristine entity. The deepfake is an ontologically inextricable interference between society and technology, where none has precedence. The deepfake thus has an organic existence so to speak—it is alive—as it is neither merely technological (inert), nor merely human (dependent on human intervention). The deepfake will therefore at times be referred to as *deepfake reality* to emphasise this conception.

Following the development of deepfake Baradianism, the question now is how this reconceptualisation allows to rethink policy approaches to the deepfake reality. At this point, I will have to keep you on the edge of your seat for a bit, as this can only be addressed later in the thesis. I will therefore first introduce you to my research questions after which we will explore the deepfake in its processes of materialisation, in its processes of becoming part of reality rather than remaining suspended in the imaginary. It is this empirical fieldwork that will subsequently allow to re-unite the discursive dimension of the deepfake with its material dimension in the policy proposal. It will allow to re-unite the theoretical conception with the empirical observations, to re-unite meaning with matter.

RESEARCH QUESTIONS

Nothing is simple in reality. Nothing is simple in the sense that reality cannot be truthfully described by singular unidirectional relations of input and output, cause and effect. Reality is not a single-layered spacetime stratum. It can be modelled that way, but it is not that way. In a diffracted understanding of reality, any conceptual model is ever-so diffractively true. Therefore, none has primacy, all are approximations. However, some conceptions prove to be more aligned with the observed reality than others. An alignment that matters not only in terms of accuracy of our understanding of our reality, but also in terms of how that conception is then in turn used to organise society and to administer social order.

Hence, a question that poked my curiosity was to understand how different conceptions of the deepfake lead to distinct interpretations of its materialisation, and thus to distinct forms of intervention, to distinct policy approaches. And since online disinformation—under which falls the deepfake—is today essentially legislated at the EU level (Hinds, 2019), my thesis asks,

MRQ How does a Baradian conception of the deepfake re-inform the EU's policy approach to deepfakes?

This research interest requires the knowledge of two matters. Firstly, it requires an understanding of the deepfake in its instances of materialisation, in its instances of becoming real. Secondly, it requires an understanding of the current policy approach of the EU. But since the original interest resides in how distinct conceptions of the deepfake affect its policy approaches, it further necessitates the tentative development of a policy approach. It is this tentative policy proposal, as inspired by the Baradian conception of the deepfake, that will ultimately allow to inform the main research question.

The main research question is thus subdivided in three subquestions. The first asks how to assemble the deepfake in its instances of sociotechnical materialisation, in the instances where society and deepfake technology interfere (later called the deepfake assemblage). The second asks how the EU conceptualises the deepfake and therefore asks what sociotechnical characters of the deepfake are to be found in the EU's policy approach (later called the EU assemblage). The third and last subquestion builds on a comparative analysis between the deepfake assemblage and EU assemblage, and asks how the Baradian revisitation of the deepfake can inform a policy approach.

SQ1 What sociotechnical entanglements materialise the deepfake?

SQ2 What conception of the deepfake does the EU materialise in its policy approach?

SQ3 What policy approach results from a Baradian conception of the deepfake?

Now, let's see where this exploration will lead us.

EMPIRICAL DIVE · ASSEMBLING THE DEEPFAKE THROUGH ITS SOCIOTECHNICAL MATERIALISATIONS

The literature review exposed an existing ambiguity in scientific research findings about how deepfakes affect society and about the effectiveness of the current means implemented to counter deepfakes. It exposed the lack of empirical evidence to back the popular fatalist tragedy narrative. Literature that engages in a realist account of the deepfake—one that engages with both its technological and human dimensions—further brought to light the existing call for a conceptual approach. A call to re-theorise the deepfake in order to better align its conception with the empirical observations.

Karen Barad was then introduced and applied for this conceptual revisitation of the deepfake. The notion *deepfake* now still refers to synthetic political disinformation that seeks to undermine trust in public institutions, but in its Baradian understanding as an entanglement of material-discursive agencies in constant doing. The deepfake is thus no longer conceived of as a pristine entity. The deepfake is an ontologically inextricable interference between society and technology, where none has precedence. The deepfake thus has an organic existence so to speak—it is alive—since it is neither merely technological (inert), nor merely human (dependent on human intervention). The deepfake will therefore at times be referred to as *deepfake reality* to emphasise this conception. In essence, deepfake Baradianism implied that (i) the tragedy narrative is reductive, (ii) the deepfake has both a material and a discursive reality, both a technological and a social reality, (iii) the netizenry is not a uniform gullible mass prone to calamity but has agency, (iv) any act of knowledge-making or intervention is only ever partial, (v) the deepfake reality can thus never find an end, (vi) the deepfake is speech more than it is content, and (vii) fact-checking is intrinsically inefficient despite its popular praise.

Having then introduced my research interest being about understanding how a Baradian conception of the deepfake can re-inform the EU's policy approach to deepfakes, I cannot meaningfully engage in this quest without investigating how the deepfake becomes real, without investigating how it materialises or takes root in society. And that is the purpose of this chapter. Firstly, it serves to assemble the deepfake in its sociotechnical relations, in its instances of interference between deepfake technology and society. Secondly, it serves to assemble the sociotechnical relations of the deepfake reality that the EU retains in its policy approach.

To provide an informative description of this empirical investigation, this chapter is subdivided in three sections. The first one describes my methodology. The second section describes the application of the methodology upon engaging in my quest to assemble the way that the deepfake materialises (the so-called deepfake assemblage). The third and last section describes the application of the methodology upon examining what parts of the materialisation of the deepfake are of interest to the EU upon regulating the deepfake (the so-called EU assemblage).

methodological design

The methodology that guided the empirical proceedings was inspired by three frameworks: Michel Callon's chain of translation (1984), Joseph Dumit's implosion exercise (2014), and John Law's pinboard approach (2006). Some of these frameworks are more conceptual than others. And it was their combination that provided a smooth means to concretely apply the more abstract among them. Combined, they constitute either the deepfake assemblage (when I will be assembling the ways through which the deepfake sociotechnically materialises), or the EU assemblage (when I will be assembling the sociotechnical sites of interest from the EU's policy approach). It is after developing on key ethical considerations that I will introduce the visual approach before then proceeding to a separate description of each of the three inspirational frameworks.

key ethical considerations

The overall endeavour of combining—of assembling—all three frameworks can be qualified as a bricolage, a tinkering, a methodological experimentation[19]. I believe such experimental approach to benefit research integrity in the sense that an experimental doing exhibits more overtly a mind's situatedness. Any mind operates along interests, flaws, neutralities, and indifferences that inevitably transpire in the results of any research endeavour. It is also because some of the frameworks that I here use are little systematised that my mind ought to rely more on itself. Which means that my personal prejudices will read more transparently upon describing my empirical proceedings compared to the case where I would follow a predefined more detailed stepwise proceeding. Note however that it is not because my approach is a bricolage, that it is therefore a less rigorous research approach. It simply means that my approach will more transparently expose my thinking during my doing rather than having my thinking hiding behind some form of externality. I am and will remain biased in those matters that I will remain unaware of, and I will therefore be obscuring certain findings and observations more than others—a play of light of which the patterns will reflect my research integrity because of the originally greater liberty of this methodological approach. Doing so also makes all the more sense given the topic of this thesis—as Barad would say, I cannot write about a phenomenon without being part of that phenomenon. Moreover, claiming that I would be delivering an impartial thesis would equate to creating a deceitful disinformational thesis[20].

It is also stemming from a place of research integrity that I adopted a more personal writing style. One

[19] Bricolage comes from Claude Lévi-Strauss' *La pensée sauvage* (1962), translated *The savage mind* (1966).
[20] Coined *deepfake methodology* by Wæver and Buzan (2020; although they developed it as a critique on scholarly work that was in their opinion fallaciously arguing that securitisation theory is founded on racist premises).

that proved at times to be more graphic or immersive. A writing style that felt closer to my being, one where my affective dimension upon producing knowledge could transpire in my recount of this process. By affective dimension, I mean "the deeply affective conversations we have inside our heads with various embodied and disembodied voices" (do Mar Pereira, 2022). For instance, transcribing how my thinking was affected during the process of producing knowledge, and by these proceedings. Especially that doing so provided a more accurate description of my empirical enterprise and therefore provided more insights into how I wove together both the deepfake assemblage and the EU assemblage. Also because remaining faithful to my self allowed me to remain more transparent to you as a reader, since I will have less re-edited or made up my proceedings. Adopting this writing style thereby allowed me to reach a deeper level of research integrity. I do not pretend that this allowed me to incarnate purity—I have my fair share of lumps, bumps, and flatness—but I hope that this more overt exhibition of my status as bricoleur provided a home to my textures to reveal themselves more comfortably (be it advertently or not). As Tommy Genesis would say, "I'm taking off my skin, I'm speaking from within" (2021). Also because I cannot really pretend to be writing for anyone but for me, since whatever I write might already be evident to others. At best, this thesis shatters my own modes of being, knowing, and doing.

The use of a visual approach was a way to further engage in that site of tension situated at the limen between research integrity and research performativity. It served to both better attune to my visual mind and to better allow a diffractive kind of reflexivity. The thesis still develops alongside an old-school way of unilaterally proposing to the reader a representational logic (as opposed to making it a more inter/intra-active medium for instance; Svabo & Bønnelycke, 2020). But what I nonetheless hope for is that instead of singling out a something, it can propose a web of connections and depths (Strathern, 2002)—a web of tendencies and backgrounds (Ahmed, 2006)—to better represent the fluidity of a reality that is in constant flux instead of operating alongside the modernist mantra of the single epistème (Mol & Law, 2002, p. 8). Not as means of proposing a better way of doing (for there is no best way since anything can be broken down into tradeoffs, and I will not be more knowledgeable than experts on the matter), but as means to give a more accountable "account of the diversity and complexity that are embodied and entangled in these types of political action" (Pérez-Bustos, Sánchez-Aldana, & Chocontá-Piraquive, 2019, p. 371).

These reflections on researcher performativity also bring us back to Barad's development on meaning-mattering. Transposing their development to the meaning-mattering nature of this thesis, allow to consider the doing of this thesis as an enactment of a relation of causality. A relation of Baradian causality, which is not about enacting an interaction between pre-existing given entities, but about enacting a relation of intra-action within the deepfake reality. The thesis enacts such relation (i) between the deepfake and my methodology (between the deepfake's doing and *my doing*), and (ii) between the EU's policy approach to deepfakes and my methodology (between the EU's doing and *my doing*). This implies that I, as researcher-performer, cannot transcend *my doing*—my practices of observation. I

cannot remove myself from my methodology and provide a conclusion that would be the absolute Truth. I can only ever be partial, be me. And, given the topic of this thesis, claiming otherwise would again be cynical, as it would equate to creating a deceitful thesis. What this thesis allows to do through that dual endeavour of looking both at the deepfake and at the EU's approach, is that it allows to engage in a reflection of how my understanding of the deepfake compares to that of the EU; a comparative reflection that enables me to put my meaning-mattering practices into perspective.

Now, besides the fact that the purpose of proposing to understand the deepfake reality as diffracted is to eventually better re-theorise the deepfake to redress the observed shortcomings of its present-day popular conception, and to uncover how this could re-inform policy approaches to address deepfakes in a legitimate way[21], the very act of proposing such diffracted understanding or ontological multiplicity entails a risk. A risk because deepfakes are a controversial material, and especially because I am looking at deepfakes that have a political intent. Academic theses always have a political dimension since they always argue for something in certain ways. These argumentations and their way of being assembled always exhibit ideological premises. And therefore, the very act of proposing an ontological multiplicity is a political act. And that is the reason for which this thesis could cause harm even though I am never directly engaging with individuals that would have a vulnerable stake in the deepfake saga. Hence, while not causing direct moral harm and remaining ethical during my research practices, the risk entailed by this thesis comes down to its potential for misinterpretation. A misinterpretation that could be voluntary or not, but which could serve an ideological misuse to deceitfully criticise policy approaches to deepfakes and undermine a government. An ideological deceitful criticism that would hamper the welfare state while the very purpose of this thesis is about securing that welfare state. I cannot control what someone else does with my thesis, just like no writer can. But what I can control at my level to prevent such misinterpretation of my thesis, is to provide an as clear, articulate, and transparent recount as possible. To deal with this risk, I sought external contribution in addition to the reviews by my supervisor and peers. Two individuals unaccustomed to the discipline of Science & Technology Studies allowed to indicate instances lacking clarity.

a visual approach

My goal through this thesis is to give voice to the ontological multiplicity present in the deepfake reality. It is to destabilise the dominant fatalist conception of the deepfake given its lack of empirical backup. However! In doing so, the purpose is not one of perpetuating an old-school modus operandi of constructing competing narratives reminiscent of a patriarchal sentiment of superiority-inferiority. It is

[21] Having observed the inaccuracy of the present-day conception, knowing its inadequacy, in a way it would be unethical for me to not provide a means to redress it.

about exhibiting a diffracted view of the deepfake—a panel of nuanced perspectives. Any point of interference between distinct ontologies in the deepfake reality—whether converging, clashing, convoluting, …—is a point inviting us to see otherwise, a point inviting us to tune in to other modes of being and doing. Points of interference are not spaces of othering. They are bridges. It was in that sense that using a visual approach was part of this diffractive exercise that seeks to move beyond the "dominant scriptocentric epistemologies" (Svabo & Bønnelycke, 2020, p. 1; see also Harding, 2016, on the pluralism and multiplicity of scientific knowledge). Such approach is of course charming a method given my mind's soft spot for visual phantasia. But more importantly it is a means for me to remain more diffractively alert and (hopefully) less prone to fall into the trap of othering and binary thinking that would undermine an acknowledgment of reality's multiplicity, of reality's diffracted doing. The visual approach is to allow to bring the elements and relations of reality's "shapeless matrix" more into conversation (Greenhouse as cited in Strathern, 2002, p. 91). To better allow the patterns of difference constitutive of the world to exist without seeking to add "a correlative sense of unity or wholeness to the individual parts" (Strathern, 1991, p. 109). And to remain attentive to the differences of the deepfake reality that interfere diffractively—somewhere along the spectrum of constructive and destructive interference. An exercise of alertness and awareness that would be less convenient for me if going by writing only, since, as Strathern wrote, "[reality does not add up and does not] form social wholes the way the textual analogy leads us to think" (p. 109).

the chain of materialisation

The notion *chain of materialisation* is proposed as a combination of Karen Barad's previously elaborated new materialist conceptual development (2007) and Michel Callon's chain of translation (1984). Where Barad provides a way of ontologically founding a refreshed mode of understanding matter—an understanding that stands at odds with the classical conception of matter as being constituted of fundamental finite particles—Callon provides a framework to structure my empirical fieldwork. Essentially, where Barad conceives of reality as an "entangled state of *agencies*" (2007, p. 23, emphasis added), Callon in comparison could be said to conceive of reality as an entangled state of *entities*. And it is in that sense that Callon provides a way to systematically investigate the deepfake in its sociotechnical entanglements. The chain of materialisation is thus a framework that allows an empirical examination of the deepfake to assemble the ways in which it materialises; the sociotechnical instances through which the deepfake is part of reality, through which it is grounded in our society.

In Callon's words, the deepfake reality would be described as a complex network or web of interrelations where science and technology play a significant role. Callon's interest in such networks is focused on the power relations that cause some actors to govern a system and others to comply with it (e.g., in his work on the scallops of Saint-Brieuc, "the defenceless larvae are constantly threatened by predators";

1984, p. 209). His framework—coined sociology of translation—is about understanding how "the identity of actors, the possibility of interaction and the margins of manoeuvre are negotiated and delimited" (p. 203). It is structured in four moments of translation: problematisation, interessement, enrolment, and mobilisation. These moments of translation constitute what some scholars have dubbed the *chain of translation*, to refer to the nature of this descriptive framework being about a concatenation—a sequencing of the network in actors and relations.

Problematisation is the moment where actors seek "to establish themselves as an obligatory passage point in the network of relationships" (p. 204). To have a successful problematisation, actors ought to craft adequate sequences of relations and adequate modes of ordering the other actors as means to anchor the latter to the interests of the former. This is the moment of so-called *interessement*. Interessement is the stratagem employed to build the relations necessary to secure the sought-after problematisation. And when interessement is successful, the elements are *enrolled*; i.e., the web of interrelations becomes coordinated. The final phase of translation, the *mobilisation*, is the instance during which other actors are being spoken for, represented. It is the moment where represented actors actively support the representatives and which results in those represented actors becoming mobile—they are displaceable to other contexts (e.g., in Callon's work, the scallops of the Saint-Brieuc Bay are figuratively being displaced to conference rooms). Callon concludes by stating that "[c]losure occurs when the spokesmen are deemed to be beyond question" (p. 220). This whole process is the chain of translation.

I thus propose to rebrand this as *chain of materialisation* to graft Barad's thinking onto it. The point is to concatenate the multiple ways in which the deepfake sociotechnically materialises, the ways in which the deepfake and society interfere. Thus, where Callon's focus resides in the power relations, my focus will reside on the sociotechnical relations. Where Callon determines a set of constitutive actors that acquired a certain authoritative poise in the network, I will determine a set of sociotechnical elements that constitute the materialisation of the deepfake regardless of their poise in the assemblage. Where Callon describes a network in terms of problematisation, interessement, enrolment, and mobilisation, I will be describing the assemblage of the deepfake in a Baradian spirit of doings and intra-actions. But, like Callon, the ultimate network or assemblage will be nothing but dynamic and reiterative. Indeed, "thinking in terms of assemblages [...] involves [treating the deepfake reality] not as [a product] but as [a process] involving many simultaneously interacting components that must themselves be understood as assemblages" (Kalpokas & Kalpokiene, 2022, p. 76).

I will engage in a detailed description of the application of this framework when I will be describing my empirical proceedings upon assembling the chains of materialisation for the deepfake and for the EU alike.

the implosion

Although initially proposed by Donna Haraway[22], the concept of the implosion used here is the one as further elaborated by Joseph Dumit (2014). An implosion is essentially about an unpacking of an object (and in fact can also be applied to concepts) through its world histories (p. 344). It is about following the object in its connections whether these be historical, material, symbolic, …. Dumit proposes various dimensions to structure the proceedings of the implosion along those histories or connections[23].

Departing from there, and to fit the current interest being about how the deepfake and society relate, the implosion was structured along a slightly reviewed version of Dumit's *technological dimension*. It is rebaptised as *sociotechnical dimension* and it is about following "[the kinds of technologies and machines that enable the deepfake to be produced and maintained; the technologies or devices that are joined with it; who has access to these machines and technologies; the sorts of information technologies that are involved; the political, economic, social, and societal dimensions of these technologies and how they help constitute it]" (p. 352).

For the implosion of the deepfake, note that in a Baradian thinking the implosion is not about breaking up the deepfake down to some imagined elementary particles that would define its assemblage. There are no definite pre-existing entities since reality is diffraction and thereby implies indefiniteness. Which can be confirmed by the fact that the breaking up could go on endlessly and is limited by our knowledge only. It is limited by our partiality only rather than because of reaching an often-fantasised position of remove. Therefore, the implosion of the deepfake is about following the sociotechnical relations that materialise the deepfake; the relations that tie the deepfake to society, to its behaviour, its modes of organisation, and vice versa. As Ahmed wrote (2006), "objects become alive not by being endowed with qualities they do not have but through a contact with them as things that have been arranged in specific ways" (p. 164). As was expressed by Barad making the same point, knowledge-making practices do not give access to "inherent properties of subjects or objects" (2007, p. 208)—the implosion does not insufflate life by endowing the deepfake with qualities it does not have but by way of assembling its relations.

For the subsequent implosion of the EU policies, the implosion is about identifying the sociotechnical elements that will be mentioned in those policies.

I will provide a detailed description of the application of this framework upon describing my empirical proceedings when imploding the deepfake and the EU policies. However, it is clear at this point that the

[22] See Donna Haraway's Modest_Witness@Second_Millennium .FemaleMan©_Meets_OncoMouse™: feminism and technoscience (1996).

[23] Note that while those dimensions are structuring pillars, these are no less in relation to one another and are extensions of one another. So, of course, these dimensions are somewhat artificially demarcated, and this is not omitted, but they nonetheless serve well as means of structuring the analysis. These dimensions operate as vectors but not as determinants.

implosion is little systematised. Which is what I believe to make it such a transparent mode of doing research, since any of the choices that I would make in the empirical endeavour will be guided by my mind, not by a pre-given methodological framework. The implosion will expose my thinking all the more visibly since any justification provided for any methodological choice upon weaving the assemblage will have to come from me.

the pinboard

Given that the ultimate interest of the empirical exploration of the deepfake is to craft its chains of materialisation—by concatenating the deepfake through the sociotechnical dimension—I used John Law's pinboard approach (2006) as a tool to organise my data collected through the implosion. The pinboard approach is a "juxtapositionary practice" that allows for an organisation of the data that is flexible; one that "can be easily revised" and thus "can be done differently by different people" (p. 4). A pinboard does not seek the "crafting [of] very coherent fractiverses" (p. 9-10), and therefore it leaves space to oddities and matters that seem unfit. In Baradian terms it would translate as being about elucidating some of the diffracted reality of the deepfake. If doing otherwise, I would be crawling back to a classical thinking of providing you with some separately delineable part of reality.

Assembling a pinboard allows for a physical interaction with the data. It allows to do more than mental work as it allows for an attunement with the material. Physically speaking, the pinboard is thus a very dynamic process where the elements are moved around, displaced, eventually re-labelled—looking for different modes of ordering and constituting the assemblage. "There is no authorised ordering of the data files, and these may be chosen and assembled in different ways by different users to generate different spaces and the times" (p. 9), and different materialisations. The pinboards thus have the experimental function of being "a set of learning surfaces" (p. 10). Learning surfaces that will expose the manyfold sociotechnical entanglements through which the deepfake materialises. The pinboard is extra interesting, as "[it] forces us to modesty because it is very particular and because it doesn't cohere well" (p. 21). It will thus force me to be more clement later upon analysing how the EU approaches the deepfake. Which aligns well with the purpose of the thesis in two ways; (i) it is not about mapping out the complexity of the deepfake reality to create an artificial singularity or a generalisation, and (ii) it is not about providing a superior recount of the deepfake reality in a sort of residual spirit of competition. It is about proposing to understand the deepfake in its diffraction, in its multiplicity. A proposition to understand the deepfake otherwise. This bodily interaction with the data will soon become clear once that I will detail the way that I applied this framework in practice.

It is also during the pinboard that three expert interviews served to assess my knowledge. The interview questions were therefore drafted such as (i) to pinpoint whether a relevant element might have slipped through the cracks of my mind during my investigation, (ii) to pinpoint whether a particular aspect

would require more in-depth investigation, and (iii) to challenge, verify, and rectify assumptions that were shaping in the process of my investigation of the deepfake. The videoconference interviews were therefore conducted in the form of an informal discussion. Guiding interview questions were prepared but the ultimate discussion went through the questions in a different order and in different formulations, and additional questions were asked when insightful input emerged during the 1-hour interviews. There were no questions about information that would qualify as confidential. A consent form was provided to the interviewees prior to the interviews to inform them precisely on the format and the content of the interviews. The interviews were not transcribed as there was no coding necessary given their use. The three experts were two scholars and a policy expert whom were all originally contacted during the literature review phase: Professor Noah Giansiracusa (the author of *How algorithms create and prevent fake news*, 2021), Professor Britt Paris (who has done a lot of research on deepfakes), and an expert from Global Disinformation Index (an online-media-rating institute that calculates a website's risk of providing disinformation and that also provides advice to policymakers). Although the interviews are not publicly available, and are therefore not verifiable to the reader of this thesis, it is however the public character of the research of these three experts (through their personal or institute's publications) that provides a public means of assessing their stance on the deepfake reality.

the deepfake assemblage

In this section, I will describe how I proceeded upon applying my methodology, my assemblage of the three frameworks. The description of the deepfake assemblage will develop in the chronological order of my stepwise proceedings. Starting with the implosion of the deepfake, followed by the pinboard of the deepfake, and ending with the resulting deepfake chains of materialisation. Overall, the point was to follow the deepfake in its sociotechnical relations to thereby pinpoint the elements constitutive of the ways in which the deepfake materialises—the ways in which the deepfake and society are inextricably anchored to one another. Remember indeed that in a Baradian understanding of such relations of materialisation, it is not about understanding these relations as interactions between pre-existing and delineable entities that would allow to demarcate the deepfake from society—to demarcate an external from an internal. In a Baradian understanding, the process of materialisation is about an intra-action within the deepfake reality and thus within society since our reality is ontologically inextricable from the deepfake reality.

At this point in the thesis, I was alien to the EU regulatory landscape and to its approach to address the deepfake, and I remained alien to it for the entire duration of this empirical phase. I was not unaware of some conducting policy lines of the EU, since (i) some of the research publications previously read for the literature review concerned policy recommendations on how to address deepfakes or online disinformation, and that (ii) policies are discussed by numerous experts and institutes of which I was

following the work. However, I did not engage at this point with any EU documentation and thus remained unfamiliar with any of their described interests, narratives, terminology, sociotechnical specificities, nuances, etc.. Doing so was to prevent picking up an EU logic before engaging myself in assembling the deepfake. A doing to engage in an independent exploration of the deepfake multiplicity first, rather than to risk binding myself to already elicited materialisations of the deepfake that I might have interpreted otherwise or that might have obstructed my view on other conceptions of the deepfake reality.

the deepfake implosion in practice

The implosion was entirely based on publicly available information; on information that anyone can access provided that one has access to the Internet. The data that I was collecting consisted of keywords, quotes, or short sentences that fitted the previously defined sociotechnical dimension. The data was thus entirely consisting of text; there were no images, sounds, videos, or any type of digital medium other than text. The data poured in from several media. Mostly from literature, whether from academic research papers, news articles, expert articles, expert opinion pieces (e.g., Techdirt), academic books, academic think tanks (e.g., The Shorenstein Center on Media, Politics and Public Policy), non-academic think tanks (e.g., Upturn.org), newsletters from stakeholders in the digital content distribution industry (e.g., Digital Content Next), online magazines (e.g., Aeon). But also from online talks, webinars, and conferences organised by the think tanks that I was following. And from a thinking outside-my-box by exploring the digital information space, either departing from the previously enumerated sources of information, or aided by my personal activity logs on LinkedIn and Facebook that thereby provided me at times with insightful input (among which I consider memes to be part of; cf. Image 7). One could interject here that using my thinking outside-my-box is not publicly accessible material. Indeed, it is not. But just as it is for any researcher, it is also the purpose for engaging in a transparent recount of my empirical fieldwork. Hence, where publicly available documentation already publicly exhibits its political stance (although surely there are various level of transparency), it was by being transparent in my proceedings that I in turn also sought to publicly advertise my political stance. I cannot undo myself from my opinions, and it is also my opinion that importantly influences my way of reading and interpreting the publicly available information. Which again connects to my pursuit for integrity upon adopting a writing style that feels closer to my being—that is less made up.

The little restrictiveness characterising the implosion as methodology gave way to a sort of mild libertinism in the data collection, which in turn allowed for more serendipitous moments to happen, contrasting to the case where I would have engaged in a more constraining modus operandi. It also gave more chances to go queer rather than remaining aligned to a framework that might have risked keeping

me cognitively numb (cf. Ahmed, 2006). For sure I could remain conservative in my implosion[24]. But whatever I was, I was so more transparently, for the approach will have been my genuine approach from *within* as Barad would say (2007). Moreover, this liberty of the mind is what brought most pleasure in doing this thesis. A pleasure that had an important influence on its content as it reinvigorated my creativity and sharpened my analytical spirit—two oh-so important ingredients in research.

While barely starting with the fieldwork, it becomes evident however that the abundance of data can easily lead to a monstrous gallimaufry. It was time for some more organisation. And given the visual nature of the methodology, the first instinctual insight to structure the collected data was to centralise it in a digital mindmapping application, EdrawMind from Wondersoft 2022 (Image 8).

Image 8: Snapshot of a part of my EdrawMind mindmap.

At this point in the fieldwork, I was still discovering some new relations and elements. No feeling of saturation just yet. I therefore kept collecting relevant sociotechnical elements and kept mindmapping them. But digitally mindmapping the implosion was quickly hitting its limits as it became too vast to remain manageable for the size of a laptop screen. It became messy. A nice mess, but difficult to keep orderly. And that is where the pinboard jumped in.

the pinboard in practice

Contrasting to the implosion, which was mainly a data collection and centralisation phase, the pinboard allowed for a true engagement with the material. On a late afternoon, with some red wine, wasabi nuts, and deep techno in the background, uniting my senses, and caving in my rental studio, I started writing the data collected through the implosion on paper snippets. The movement of my hands combined to that particularly tendering setting is what induced some sort of a meditative state. A state that allowed me to zoom in, to be captive of the deepfake, and to forget about all except of this project. A state

[24] However much I might like to believe that I am thinking alternatively and creatively, in fact, I merely repeat that which I have been wired to repeat. But then, still. A potential for catharsis subsists, and forever will, since my very interaction with the world constantly reshapes my synaptic circuitry, for I myself am a phenomenon.

allowing me to be more consciously intra-acting with the material. I thus kept on writing down sociotechnical elements on paper snippets without looking at my EdrawMind mindmap where the data was previously centralised. Letting my mind wander and the thoughts circulate as they come and as they go gave birth to new ideas. Clearly, although technically speaking I was in the early stage of crafting the pinboard, I was thus still also elaborating the implosion.

At this point, I did not care yet whether I would retain these elements in the final chains of materialisation. However, at this point, I did start to feel an itch of saturation, because whenever coming across a new source of information, it did not provide a renewed understanding of the ways in which the deepfake sociotechnically materialises in our world.

The next morning, I woke up, had some coffee, and brushed my teeth. I felt fresh. I started playing some ambient synth music, and off I went with the paper snippets in one hand and tape in the other. I started pinning the elements on the biggest available empty wall in the rental studio. Moving the elements around. Exploring what could emerge from them (or should I say, from me). The physical engagement connected my mind to my body and my mind-body to the data. An attunement. Insh'Allah! let the bricolage fulfil me. During the pinning I could sometimes think of another element, which I then added to the pinboard. The pinboard was thus a means of orienting myself. I am unsure as to what extent I was going queer or not, for clearly I can but surpass my upbringing and education. But at this point, I simply tried to create an assemblage, a web of tendencies (Ahmed, 2006). The movement of my body in that space allowed me to become more aware of my practice of hierarchisation, of my practice of pushing some of the elements to the background, of my "mode of ordering" (Law, 1994). An awareness that struck me more than when I was proceeding purely digitally. And which made me realise more overtly that I was clearly seeking to bring forth the so far less visibly explored matters of the deepfake reality in literature (which is where the thinking outside-my-box proved particularly insightful).

At this moment, I was thus already engaging in a basic level of analysis where I tried to find patterns, to rewire the assemblage in other ways, and to pull together certain chains of materialisation. This physical explorative engagement coincides well with reality. Reality is not provided in an ordered set of juxtaposed 2D puzzle pieces as if consisting of a single flat stratum. Reality can be thought of as a bowl of cooked spaghetti or rice noodles. With sauce. Liquid. A lot of fluidity. Messy. (Partially) connected. Multi-layered. Having depth. Intertwined. Convoluted. Clashing. Overlapping. In tension. In harmony. Tasty. Rebutting. Or both. And so too is the practice of knowledge-making. Law's pinboard approach thus provided a nice means to join Barad and Strathern in their same quest for ontological multiplicity, for acknowledging the diffracted being of reality. After pinning the elements on the wall (see Image 9), I proceeded to a double check with my EdrawMind mindmap and noted down any elements that would have slipped through the cracks of my above proceedings. It is also during this timeframe that the previously described three expert interviews jumped in to pinpoint eventually missing relevant elements, to pinpoint whether something would have required further investigation, or to challenge my assumptions-in-the-making.

Image 9: Pinboard.

The current proceedings expose the limits of high-technology and the benefits of what would relatively be qualified as low-technology. Personally, it helped immensely to have things laid out physically in order to better interact and play with the material—to better experiment and explore how it shuffled and connected otherwise or not. Where the screen size of my laptop hindered a manageable organisation of my data, using a wall allowed to exceed computer screen sizes, and allowed to feel unhindered in the way of organising the data. The physical organisation of the pinboard did not have to abide by the rules inscribed in any computer software. No need of hassling with digital mapping features. No obligation to remain linear, aligned, connected. Although … not so fast.

Despite the wall-space, things did become frustrating at some point again because the wall still inhibited to depict full-circle relations. Any element relates in multiple ways to others; relations are not simply about going from one point to another (Barad would equally strongly oppose such linear conception). I had wished for a giant terrestrial globe here. Instead, I bought inflatable balloons. But it still did not allow to do what I had in mind. Alas. There is only so much one can do depending on the availability of physical resources. Hence, although it allowed for a more attuned and playful interaction with the data, thereby involving an already more analytical doing, the pinboards only partially reflect my thinking. Pinboards allow a more interactive engagement, a more generative doing (the body-mind engagement allows for a different kind of involvement and channelling of creativity than would a purely mental work for instance). But the physical world, or more accurately, the lack of access to certain physical resources, inhibits a full blossoming of this very generative doing. I felt frustrated by that because it felt regressive. I then let that frustration be for what it was.

Additionally, throughout the pinning, there was always a moment where I felt like giving birth to a cloud of gibberish, because of the continuous bubbling up of yet other ways of assembling the deepfake reality. A monstrous frivolity that kept on having yet other dimensions. Despite the feeling of saturation that had previously emerged, I had now lost that feeling because I felt that the pinboard exercise could go on endlessly. It made me feel nostalgic of moments of insouciance. And it is maybe exactly that feeling that made my frustration paradigm shift. Suddenly, I had made peace with that frustration. The bricolage

still led to a continuous bubbling up of nuanced understandings of the deepfake, because,

> "No matter how many perspectives are assembled, they all create perspective. The formal product is infinity" (Strathern, 1991, p. 108).

But that was also what I was looking for. As Law explained, the pinboard does not seek to craft "coherent fractiverses" (2006, p. 10) and it is therefore "easily revised" (p. 4). The "learning surface" (p. 10) thus seemed to do exactly what it was supposed to: teach me. Besides teaching me about the deepfake, it taught me to accept the limits of my knowledge. My mind cannot experience reality otherwise than through itself. Another realisation that eased the acceptance of this teaching was that while my mind cannot go disruptively opposite any of my inculcated and incorporated prejudices, even if I would have been able to go totally alien and propose outer-galactic understandings of the deepfake, it would be of little use for the reality along which I operate. It was this impossibility that kept me grounded to not subsequently propose surreal or unrealisable things (at least for my doing). And however partial my diffracted view on the deepfake reality is, it makes it no less relevant or valid. Surely, a too-little-informed view would jeopardise my future undertaking to craft a policy proposal. But I trust my current undertakings. So, although in absolute terms an infinity of pinboards could be crafted for the deepfake, my mind eventually could not move beyond the ultimately crafted pinboard. Saturation.

As such, during the finalisation of the pinboard, the main query as already addressed was about the required level of depth necessary upon breaking up any element. The more one breaks up, the more sites of potential anchoring, of *interessement* one discovers (Callon, 1984). Just like fragmenting a piece of rock in an infinitesimal number of particles makes it more reactive to acid rain, so it is true for the sociotechnical reality of the deepfake. Fragmenting the elements of the pinboard make them more prone to erosion, more docile to an act of interference. But! there is always a threshold. Infinitely diluting matter makes it unreactive in return. Which brings me to policymaking and to appreciate why it is such tricky a practice. Policymakers ought to find the right dosage of intervention, a dosage particular to a specific context. And that was equally the point I was seeking to reach here and now. A point I reached because (i) whenever breaking up an element, it no longer provided matters of interest for the aimed policy proposal, and (ii) I could feel a clear saturation in the sense that whatever information I was encountering was simply pointing back to something that was already there. In short, either irrelevant or repetitive.

Note that while Callon in his sociology of translation was focused on power relations and therefore was cutting down to what he deemed to be the key figures of the network (the so-called spokesmen or representatives), I here explicitly leave room to more elements to be part of the pinboard. Demarcating between who "speaks for" and who is "spoken for" is but a straightforward exercise. Especially for the case at hand being that of identifying potential sites of regulatory interest, and that assessing what nonhumans can or not do without a deeper network assessment is more of a whimsical doing than one of critical assessment of underlying tenets upon doing so. It would also go against Barad's conception

of intra-action that "signifies the mutual constitution of entangled agencies [and therefore does not assume] that there are separate individual agencies that precede their interaction" (2007, p. 33).

So, before even hitting the limits of the wall's available space, a final feeling of saturation emerged. I thus took a picture of the finished pinboard and indicated clusters via computer (it would have been fun to do this on the wall but … you know) (see Image 10). These clusters served to provide a first referential basis to subdivide the pinboard into the various chains of materialisation. They served as structuring means. Examples of such clusters were: corporate, deepfake taxonomy, modes of intervention, ….

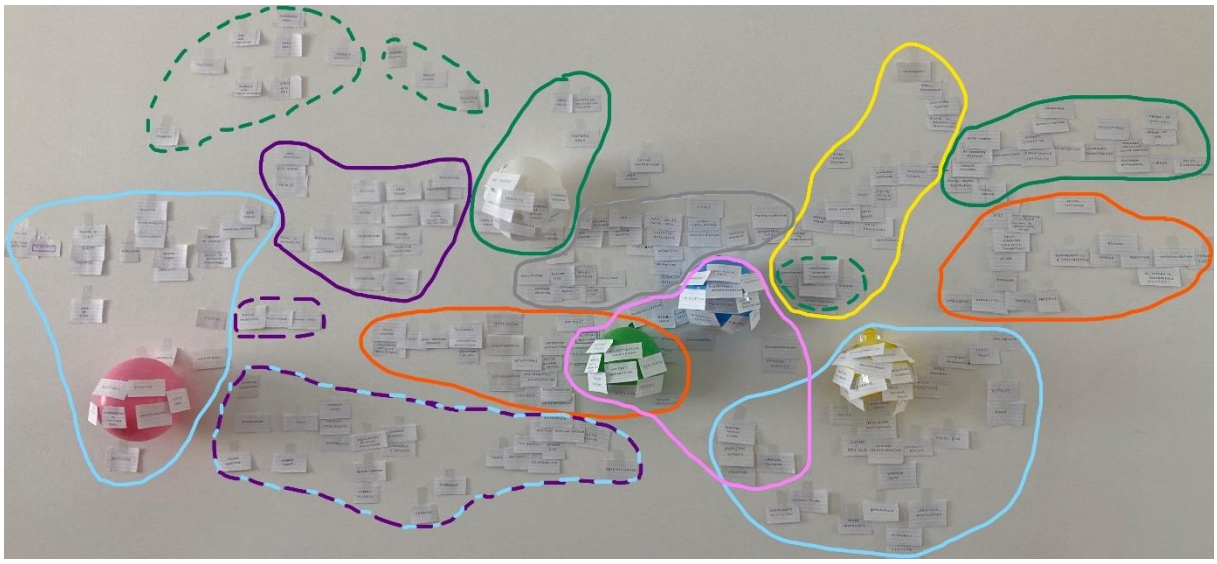


Image 10: Example of clustering.

the deepfake chains of materialisation

I was now standing in front of this big pinboard wall and manually sketched in a notebook some chains of materialisation. Alongside the sketches that I was drawing, I gradually unpinned the elements from the wall. Despite the previously uttered feeling of saturation, upon transcribing the handcrafted sketch over to the software Sketchboard.io (2022), at times, I happened to reshuffle some elements and regroup or relink them yet otherwise. True saturation was thus in fact consisting of a saturation trinity: during the implosion, during the pinboard, and during the chains. I suppose that this reality characterising my empirical work just proves Barad's point that nothing is a given but everything is a doing.

The resulting chains are depicted in Image 11. These chains reflect the process of sociotechnical materialisation of the deepfake, the ways in which it takes root in our society. Meaning, both the ways in which the deepfake influences social order but also the ways in which society influences the deepfake in the ways that it materialises. These chains are an epistemic construct so to speak. They reflect my way of assembling the knowledge gathered to construct them. And they will later serve as a tool to identify potential sites of interest for regulatory intervention upon crafting the policy proposal. I speak

of “potential”, because this stage in the research is still a screening phase excavating the buried sociotechnical entanglements, without yet looking at any EU policy document. Before moving on to the key takeaways from these chains, I first want to elaborate on how to interpret these chains, so that we are both aligned in terms of how to read them.

Firstly, the relations in the assemblage are not understood as causal relations in the Callonial sense of *inter*-acting entities where either one speaks for another (1984). They are here understood in the Baradian sense as instances of attunement, of agential *intra*-action (2007). Which brings us to a second consideration for the interpretation of the chains. Namely, that although Image 11 provides a particular photograph of the deepfake, the assemblage is no less dynamic. Not only because my mind would reshuffle some parts already otherwise sometime later (which is what effectively happened during the fieldwork), but because the deepfake is an ever-reconfiguring diffracted reality (Barad, 2007). Which is how some scholars speak of *distorted* and *distortable portraits* of reality “[because of the] incompleteness of or imprecisions in data” (Kalpokas & Kalpokiene, 2022, p. 83). Nevertheless, the spacetimematter scale of our human nature makes that certain things change more slowly than others. Which makes that the chains depicted in Image 11 are not simply invalidated because of the constant doing of the deepfake reality. In fact, it is the very tentacularity of the deepfake, the multiplicity of its assembled chains, representative of its deep societal implantation, that causes its stabilisation (Star, 1990). The deepfake is “'multifarious and finds itself realized in multiple forms across all fields of activity, and all scales of constitution of reality'” (Braidotti & Fuller cited in Kalpokas & Kalpokiene, 2022, p. 77). To take another example, the Internet is deeply rooted in our society because of the multiplicity of its infrastructural relations that are being reinforced not only on a material level (e.g., by the installation of cross-continent fibre optic cables), but also on a societal level (e.g., by the existence of companies providing broadband services, or because of scientific research endeavours that seek to optimise the ways the Internet can be utilised such as advancements in quantum Internet). The Internet is not therefore set in stone, it is not immutable. Be it because of existing efforts for a revisitation of the Internet infrastructure to cope with systemic inequalities (e.g., Monnet, 2020), or because of emerging digital technologies such as the metaverse that require a reapplication or revisitation of the Internet as we know it (e.g., Hackl, 2021; Smith, 2021). The growing multiplicity of its entanglements in our contemporary society nonetheless provides solid grounds for its stabilisation (e.g., the growing influence of digitalisation on society’s mode of organisation). These examples bring us to a third and last interpretative point. Namely, that not only the human has a capacity for translation, for meaning-mattering, but equally so the nonhuman (Barad, 2007; Callon, 1984). Hence, because the nonhuman too is intelligible, because it has the capacity to generate meaning and impose an identity, human inaction does not mean that the materialisation of the deepfake comes to a halt as such.

A total of seven chains of materialisation for the deepfake were assembled and are enumerated here: `chain of behavioural affects` (containing elements about how deepfakes and individual and social behaviour influence one another), `chain of Internet infrastructure` (containing elements that

constitute the Internet infrastructure), chain of deepfake technology (containing elements about how deepfake content is produced and the reasons for which they are used), chain of moderation (containing elements in relation to the various existing methods to moderate online content), chain of normalisation (containing elements reflecting on the normativity underlying content regulation), chain of e-feudalism (containing elements reflecting the monopolistic character of today's digital space), chain of information ecosystem (containing elements describing today's way of distributing, targeting, and consuming digital content). These seven chains are to be understood as a set of communicating vessels. They are not isolated from one another but are very much intra-acting, interfering.

Having described how to read these chains of materialisation, let us now discuss their key takeaways.

- There is no hierarchisation between the chains. They are all equally constitutive of the deepfake reality—equally constitutive of the way that the deepfake sociotechnically materialises. The different chains simply reflect distinct broader topics. And the labels of these chains provide a clear instance of transparency in terms of my personal a priori (e.g., e-feudalism) and interests (e.g., normalisation).
- So too, the elements in the chains are not subject to hierarchisation. All elements part of the assemblage are equally constitutive of the deepfake reality. It does not matter here whether some are recurring more often than others or have more relations to others. The goal of the deepfake assemblage is to provide an overview of what constitutes the deepfake reality, and to thereby provide a referential basis to subsequently assess in the next section the sites of specific regulatory interest to the EU.
- Given the interest in the sociotechnical aspects of the deepfake—having therefore followed the sociotechnical dimension of the deepfake[25]—these chains contain both elements of the material order (e.g., data servers, digital display screens, written legal acts, …) and of the immaterial order (e.g., confirmation bias, lobbies, information ecosystem, ...). The deepfake is an immaterial content (it consists of strings of digital code), created from both material and immaterial input (respectively, for instance, computer hardware and personal opinions), and that comes to both material and immaterial realisation (respectively, for instance, physical damage caused by bigotry and psychological influence).
- The material and immaterial doing of the deepfake is what brings us to appreciate it as more than an inert technological ensemble and to also appreciate its human nature, and hence its social and societal character. Not only does it lead us to appreciate why certain elements in the assemblage are more conceptual than others, and why it would be reductive and erroneous to dismiss these. But

[25] The *sociotechnical dimension* is about following the deepfake in "[the kinds of technologies and machines that enable the deepfake to be produced and maintained; the technologies or devices that are joined with it; who has access to these machines and technologies; the sorts of information technologies that are involved; the political, economic, social, and societal dimensions of these technologies and how they help constitute it]" (Dumit, 2014, p. 352).

it also aligns with the Baradian conception that matter (the material) and meaning (the discursive) become determinate together. It is not because an element is conceptual that it is therefore less precise or less meaningful in the assemblage (Dumit, 2014). Meaningfulness is only acquired through the relations in the network; even a materially tangible element remains obscure in the assemblage if not contextualised, if not put in relation.

- The chains of materialisation allow to be attentive to the fact that the way of concatenating the deepfake is very much bound to the limits of the knowledge amassed. One could keep on breaking up the chains into yet more elements. And while I discussed ceasing to do so upon feeling that doing so led to either irrelevance or repetition, this nonetheless brings us to the important consideration that no single element can be incriminated alone for the issues that deepfakes bring to expression. For instance, `social media platforms` cannot be the only ones blamed for ongoing societal issues (e.g., Institute for Strategic Dialogue, 2022). This will be elaborated on later in the analysis chapter.
- The diversity of the elements part of the set of assembled chains exhibits the plethora of potential sites of interest for regulatory intervention. For instance, where the literature review exposed interventionist interests to be particularly focused on `fact-checking` methods that centre on the detection and the hampering of the circulation of deepfakes, the sociotechnical implosion of the deepfake allowed to appreciate more crucially the fact that deepfakes rely on massive `data sets` for their creation and perfection. The chains thus allow to expand and diversify the regulatory options that could be not only of interest but also effective. For instance, with this example, it could mean that `data set providers` could find a fitting role in ways to address deepfakes.

Having finished the part concerned with the deepfake assemblage, from now on, I leave this assemblage entirely untouched. There is something scary about it since it leaves the door open to potential self-trolling. But simultaneously at sort of adrenaline thrill kicks in. I am curious of what the thesis further has to give.

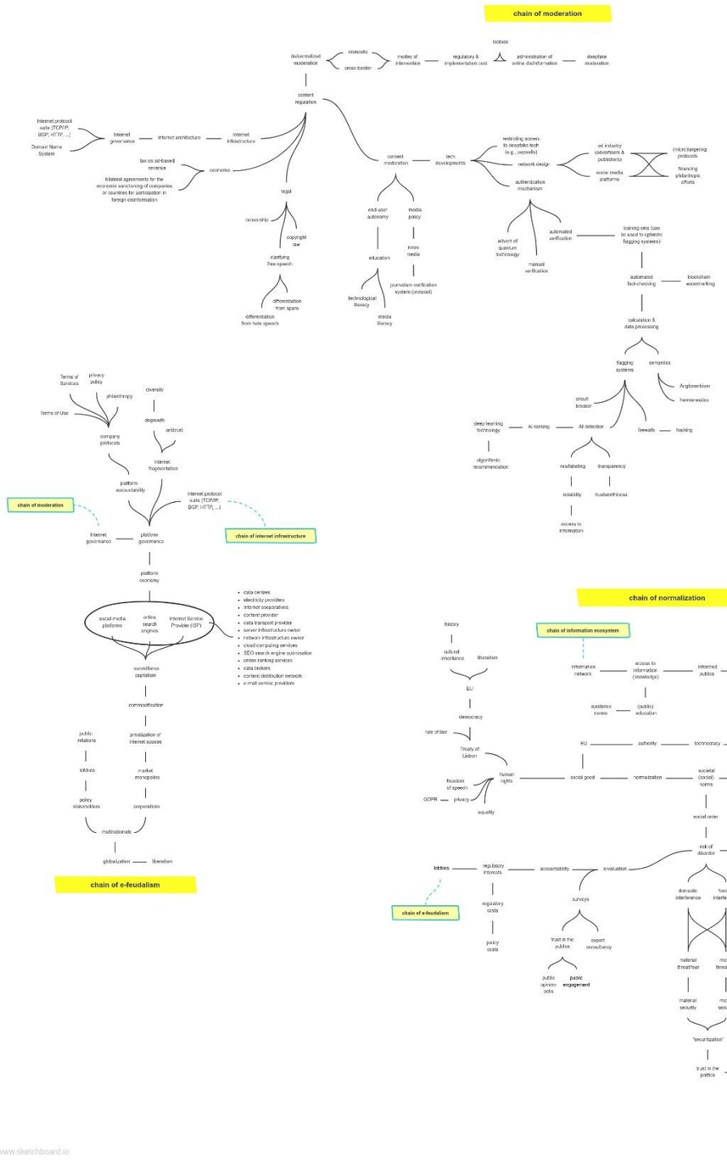


Image 11: Overview of the assembled deepfake chains of materialisation.

the EU assemblage

Congratulations. We have arrived at the second phase of the empirical chapter. This is where we will dive in the EU's approach to the deepfake. As a quick refresher, in the previous section on the deepfake assemblage, a set of chains of materialisation was assembled. These chains represent the sociotechnical ways in which the deepfake materialises; the ways in which the deepfake grounds itself in our reality and in which we ground ourselves in the deepfake reality. The mission in this section is to pinpoint which sociotechnical elements the EU mentions in its policy approach to address the deepfake. To this end, we will first identify the relevant EU policy documents. These documents will then be imploded through the same sociotechnical dimension as was used earlier to implode the deepfake[26]. The data collected through the policy implosion will then be grafted onto the previously assembled deepfake chains of materialisation, superposed. The resulting chains will be called the *policy chains of materialisation*. These policy chains will serve for the subsequent chapter to derive how the EU conceptualises the deepfake. In other words, it will serve to analyse on what grounds of the deepfake reality the EU roots—legitimates—its policy approach.

INTERLUDE-TO-READ • some regulatory terminology defined

I start from the etymological origin of *governance* to frame the way that the notions *policy* and *regulation* are used here. Governance comes from "[t]he Latin root word […] gubernare, which means to guide or pilot a ship[; it] is typically demonstrated by a collective group of people tasked with guiding an organization in alignment with its mission, vision, and values" (Nonprofit Quarterly, 2022). Governance is the sociopolitical undertaking to administer society's *vivre-ensemble*; its harmonious and prosperous living-together. In the EU, it is a top-down guidance that emanates from a democratically elected set of representatives.

Departing from that understanding, the term *policy* refers to the "bureaucratic and administrative rule making [but not the] legislative or judicial rule making [and thus also excludes] business-to-business regulation as well as civil regulation" (Levi-Faur, 2011, p. 6). A policy is thus not about criminal liability or prosecutorial charges, but about the guidelines provided at Eurocratic level that seek to safeguard a vivre-ensemble.

In turn, *regulation*, as part of a policy, is exclusively understood as the set of written and binding

[26] The *sociotechnical dimension* is about following the deepfake in "[the kinds of technologies and machines that enable the deepfake to be produced and maintained; the technologies or devices that are joined with it; who has access to these machines and technologies; the sorts of information technologies that are involved; the political, economic, social, and societal dimensions of these technologies and how they help constitute it]" (Dumit, 2014, p. 352).

rules that can be enacted by both public and private entities; e.g., EU institutions, EU member state governments, private companies. To distinguish however between public and private rules, whenever is spoken of *content regulation*, it is about the governmental administration of online content (be it at the level of the EU or at the level of its member states). And whenever is spoken of *content moderation*, it is about the non-governmental administration of online content (e.g., the administration of online content by private platform companies).

the relevant policies

Upon engaging in this exercise of identification of the relevant policies, I was no complete novice. I knew some of the documents of interest, be it because they were advised to me by Global Disinformation Index or because independent organisations such as the EU DisinfoLab, Techdirt, or Tech Policy Press mention these extensively when discussing online disinformation or deepfakes. But, again, at this point, I did not open any such document yet. The tricky part here was that EU inputs relevant to deepfakes were scattered across various documents. Not simply because the EU does not have an actual policy specific to deepfakes, but also because of the deepfake's tentacular reality. It involves deep learning technology, freedom of speech, illegal content, digital service providers, data training sets, …. Inevitably, the puzzle that I tried to assemble was thus going to miss some informative pieces that were to be found in policy documents that I did not implode. However, we will see upon discussing the policy chains later that these missing pieces would not have been key to understand how the EU addresses the deepfake, since the ultimately identified policy documents did cover the EU's approach to the deepfake well (we will discuss this in a second). And should these documents have proven insufficient to do so, I would have expanded my policy scope to cover for those missing parts.

To identify the relevant documents, I started by looking up which documents the European Commission (EC), the European Parliament (EP), and the Council of Europe (CoE) had written or which research units or consultative units were set up to advise the EU on how to address the deepfake. Since the query of a deepfake administration extends beyond purely deepfake technology-related matters, the policies estimated relevant to the deepfake do not always target deepfakes directly. Which was how the deepfake assemblage proved helpful to better pinpoint which regulatory domains are to be looked up (e.g., artificial intelligence).

Six policy documents were identified for their implosion through the sociotechnical dimension. And it proves already insightful to note that these policies all find their origins in the EC's priorities for 2019-24 for "a Europe fit for the digital age" (EC, n.d.). "The EU's ambition is to be digitally sovereign in an open and interconnected world, and to pursue digital policies that empower people and businesses to seize a human centred, sustainable and more prosperous digital future" (EC, 2021b).

(1) This year, in 2022, the EP and CoE reached a provisional agreement on the EC's proposed *Digital*

Markets Act (DMA). The DMA is about ensuring fair competition among the digital market players to guarantee equal digital market opportunities (EP, 2022a).

(2) Similarly, the EP and CoE reached a provisional agreement on the EC's proposed *Digital Services Act* (DSA). The DSA in turn is about "ensuring a safe and accountable online environment" (EC, 2022a) by the enactment of obligations for online service providers.

(3) An act that is surprisingly absent in discussions about the regulation of deepfakes and online disinformation is the *Artificial Intelligence Act* (AIA). The AIA acts as a legislative framework for AI and proposes a risk-based categorisation of various AI-based technologies (EC, 2021c). It also proposes some regulations for AI training models and data training sets.

(4) Similarly surprisingly absent is the *Data Act* (DA). The DA is a new proposal to put up with inadequacies of the preceding Data Governance Act. It seeks to "ensur[e] fairness in the allocation of value from data among actors in the data economy and to foster access to and use of data" (EC, 2022c, p. 2).

(5) The *Strengthened Code of Practice on Disinformation* (SCoPoD) complements the DSA[27] and is an updated version of its predecessor, the Code of Practice on Disinformation, as well as of previous European Democracy Action Plan that was concerned with "transparency in political advertising and communication", including transparency in the funding of political parties (EC, 2020a, p. 4). The SCoPoD focuses on defunding disinformation in general (EC, 2022b).

(6) The EP has set up the Panel for the Future of Science and Technology which requested a study on deepfakes specifically, *Tackling deepfakes in European policy* (EP, 2021). Although not being a true policy document, the study was included in the implosion as it was the only relevant EU document directly addressing deepfakes and that it included policy recommendations.

Image 12 illustrates how these six documents hang together. Going from a broader application, affecting more than just the deepfake, to a narrower application, directly applying to the deepfake.

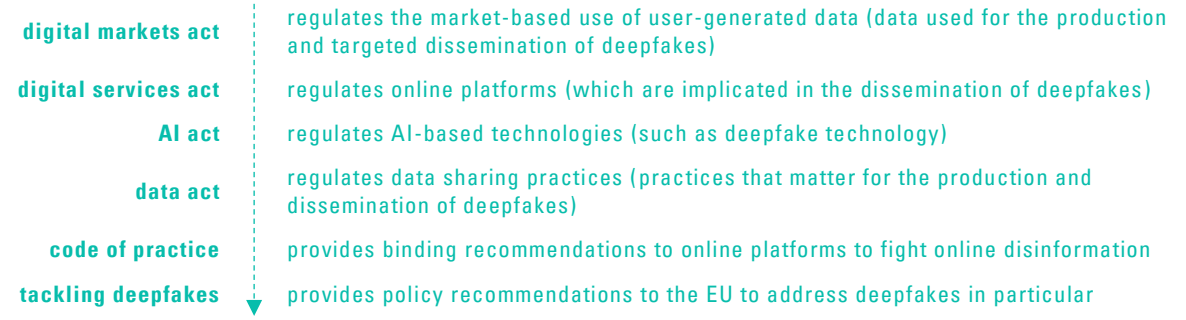


Image 12: The EU's policy approach to deepfakes.

Analysing these six documents was a *sine qua non* requirement if willing to acquire a comprehensive understanding of the EU's policy approach to the deepfake. Only sticking to one document would have been cheating the EU as it would by far not have been representative of its policy efforts.

[27] Specifically, it complements Article 35 of the DSA regarding Very Large Online Platforms (EC, 2022b, p. 2).

Other documents that discuss matters that some would perhaps have argued to relate to deepfakes were dismissed. Either (i) because they were not discussing policies or providing recommendations. Such as the European Digital Media Observatory hub that itself does not publish regulatory proposals, or the Special Committee on Foreign Interference in all Democratic Processes in the European Union, including Disinformation (abbreviated as INGE and INGE2) that does not provide recommendations even though it calls for more research in deepfake detection technology and for a labelling of deepfakes (EP, 2022c). Or (ii) because certain documents that would in fact be relevant to deepfakes are discussed in the documents part of the implosion in a more relevant and elaborate way. Such as the principles of transparency about data processing practices in the General Data Protection Regulation. And (iii) there also exist EU units that discuss deepfakes, but their angle of interest does not meet the interest of present thesis. For instance, the European Cybercrime Centre EC3 categorises deepfakes as "malicious uses of AI" (Sancho, Eira & Klayn, 2021), but it looks at deepfakes from the criminal perspective, such as identity theft, fraud, and pornography. Disinformation is not part of EC3's current priorities (even though they mention its potential hazard). And the European Audiovisual Observatory is focused on copyright infringement issues related to deepfakes (European Audiovisual Observatory, 2020), but it is not concerned with more societally nefarious aspects of the audiovisual properties of deepfakes.

the policy implosion in practice

The identified six documents of interest were now imploded using the same sociotechnical dimension that was used during the previous implosion of the deepfake[28]. Similarly to the deepfake assemblage where I had assembled my epistemic construct—weaving together the knowledge that I had gathered—the point here was to similarly derive the EU's epistemic construct from the data that I would collect through the implosions. However, the epistemic construct that was derived here was entirely policy-based. This thesis cannot have the pretence to have derived the entire actual epistemic construct of the EU, as that would have necessitated an extensively more laborious research enterprise. However, extracting the sociotechnical elements from the EU policies that address the deepfake accurately informed on what grounds of the deepfake reality the EU roots its policy approach. And it thereby provided the necessary means to understand how the EU conceptualises the deepfake.

The policies might not reflect all the underlying discussions, conflicts, ententes, and cross-member state clashes that the EU must deal with upon engaging in the enactment of any policy or regulation. And the policies might thus be argued to be a reduction of the EU's general knowledge and conception of the

[28] The *sociotechnical dimension* is about following the deepfake in "[the kinds of technologies and machines that enable the deepfake to be produced and maintained; the technologies or devices that are joined with it; who has access to these machines and technologies; the sorts of information technologies that are involved; the political, economic, social, and societal dimensions of these technologies and how they help constitute it]" (Dumit, 2014, p. 352).

deepfake. But these policies are those that the EU has communally agreed upon across member states and representatives[29]. And it is exactly that communal depiction that is of interest here, since it is only the communal agreement and enactment of policies and regulations that has a binding force. By enumerating the features of the deepfake reality that the EU communally agreed to enact as being the features that constitute the deepfake reality—the features pitched to require intervention—the EU as such provides its conception of the deepfake. Its version of the deepfake reality. And because any regulatory mode of intervention reshapes reality by enacting new tendencies and backgrounds (Ahmed, 2006), it is this communally agreed EU depiction of the deepfake that has a binding influence over the reality of the deepfake.

The implosion of the EU documents was less arduous than the previous implosion of the deepfake. Why? Because the policy implosion was bound to the content of the policy documents; the implosion could not drift out into the wild or into the unwritten parts of the policies contrasting to the deepfake implosion that was quite savage a bricolage. So, where the deepfake implosion was a self-reliant experimental exercise (and had to be experimental both to remain unprejudiced by EU's approach and to remain more transparent about my own opinionated situatedness), the policy implosion was an EU-reliant and policy-dependent structured exercise. In fact, the policy implosion essentially consisted of a coding exercise for which I used MAXQDA 2022 software. The codes consisted of keywords or short sentences fitting the criteria of the sociotechnical dimension. During the coding, I was also already analysing—noting down memos and reflections for each policy document. After finishing all the reading and collecting all the codes, I had then proceeded to a cross-check between these codes and my previously crafted deepfake chains of materialisation. It was at this point that the EU policy chains of materialisation were being worked out.

the policy chains of materialisation

Contrary to my initial intent, I did not assemble a separate set of chains of materialisation from scratch for the EU policies. If I would have, it would inevitably have closely resembled the chains crafted for the deepfake, given that I was the mastermind in both instances and that my concatenating logics would not suddenly have been disruptively altered. But also because upon reading the EU documents it became obvious that the similarities between the deepfake assemblage and the EU assemblage were more abundant than I had initially imagined. That was how I had decided to go by a superposition of the collected codes through the implosion of the EU documents on top of the previously crafted deepfake chains of materialisation.

Image 13 represents the outcome of this process. The blue stars mark recurring elements that are

[29] Except for the EP's study *Tackling deepfakes in European policy*.

mentioned oftentimes in the policies. The green stars mark elements that are only mentioned once or a handful of times. No star means that the element is absent from the EU policies. The pink texts indicate new material that was not part of the initial deepfake chains. At times, it served more so to highlight a nuanced understanding. Because it happened a handful of times that a matter expressed in an EU document was the same term as was expressed in the deepfake chains, but it had a different meaning because it was expressed in a different context (e.g., `open access` to data for economic fairness). As much as it happened a handful of times that a matter expressed in an EU document shared the same meaning with an element in the deepfake chains but was however expressed differently in the latter (e.g., `traceability` versus `flagging systems`). Both instances occurred only a handful of times, thereby backing the previous claim that the deepfake assemblage and the EU assemblage shared a mutual sociotechnical understanding of the deepfake in terms of the relations constitutive of the deepfake reality—in terms of the sociotechnical context or dimension of the deepfake.

Since the implosions operated alongside a sociotechnical dimension, the chains of materialisation obviously lack a representativeness of other types of contexts equally constitutive of the deepfake reality. However, the open nature of the dimensions structuring any implosion makes that the sociotechnical dimension was not impermeable to other types of histories, interests, or motives[30] (Dumit, 2014). The chains therefore did also bring to expression other types of contexts, even though less strikingly so. The most prominent and obvious example of such side-context—so to speak—was the EU's general underlying interest to create a strong digital Union market economy. Whereas my motive was to reflect on what elements are present or absent in the EU's conception of the deepfake. Both motives tie back to a care for the safeguarding of an idealised Western democratic welfare state. But they do so through different angles; a Eurocratic angle and a Science-Technology-Society angle.

While the EU assemblage might thus share a similar understanding of the sociotechnical context of the deepfake as is visible from the deepfake assemblage, a close scrutiny of the elements making up the policy chains, besides the relations, allows to make a few observations.

(1) Elements part of the EU policies and absent from the deepfake chains

Pink elements indicate matters that the EU mentioned and which I deemed relevant to the sociotechnical dimension upon imploding the policies, but which were absent from the deepfake chains. A closer investigation of the pink elements reveals the following considerations.

- The EU has an explicit focus on the economic market, which is unsurprising especially in terms of the Digital Markets Act and the Digital Services Act. This leads the EU to describe additional elements part of the deepfake reality, such as `data economy` (whereas related inputs in the deepfake chains were `platform economy` and `online advertisement models`), `consumer trust`, `brand safety` (whereas a related input in the deepfake chains was `consumerism`), and a

[30] Which reflected for instance in the sociotechnical dimension also being interested in the political and economic dimensions of the deepfake.

list of more detailed types of digital service providers considered for the legal acts, such as e-commerce, video sharing platform services, online intermediation services, ...

- The EU has a more technical language due to its administrative role, and therefore describes elements such as competition law, global standards, auditing bodies, illegal content. For the element illegal content, the deepfake chains did contain elements representative thereof—such as terrorism, hate speech, fraud, scams, criminal activity—but my brain not being wired in terms of what the EU considers legal or not caused this terminological absence.
- The EU pitches its enacted Data Act—that essentially regulates data sharing practices and data interoperability—also in the frame of the Internet of Things consumer products. I thus added this previously missing element, since deepfakes might at some point become part of the world of the Internet of Things (for instance, through speech mimicry in smart devices that provide voice servicing; e.g., Amazon Alexa). And their regulation could thus affect those deepfakes that have a politically deceitful purpose.

(2) Elements part of both the EU policies and the deepfake chains

For the elements that were mentioned by the EU and which were already part of the previously assembled deepfake chains of materialisation, the differences are chiefly about degrees of emphasis of these sociotechnical elements. This emphasis was demarcated using blue and green stars, respectively indicating an abundant recurrence of an element or indicating only a handful of such occurrences.

Where the elements part of the deepfake chains were not hierarchised in terms of order of importance (as it is not of interest), the distinct levels of emphasis found throughout the EU policies point at the EU's distinctive appreciation of certain elements in the deepfake reality. Indeed, by the repeated enactments of certain elements throughout its communally agreed policies, the EU gives away a conception of the deepfake where certain sociotechnical characteristics are deemed more important than others.

- The most notable blue stars (frequent elements)
  - The EU recognises the broader plethora of stakeholders part of the play as opposed to the common practice of focusing on social media alone. The EU thereby includes this diversity of actors in the deepfake reality rather than discharging them of their responsibility in that reality; e.g., AI systems providers.
  - The EU significantly emphasises fact-checking as a solution to online disinformation and deepfakes, and it exhibits a clear favouritism for a technology-based approach to fact-checking; e.g., automated fact-checking verification, AI detection, algorithmic recommender systems.
  - While having an abundant focus on technology (and unsurprisingly so since the imploded

documents are about a regulation of technology and digital markets), the EU significantly emphasises the need for transparency in digital practices to ensure fair business practices and to ensure an informed end-user with regard to the data it shares with digital service providers. As such, data protection is also a key focus of the EU.

- Being one of the fundamental EU rights (EU, 2012a), freedom of speech is key across the EU documents. It is also in that sense that one of the recommendations to address online disinformation and deepfakes is to clarify how hate speech distinguishes itself from free speech.
- In terms of internal security, to counter the dissemination of online disinformation and deepfakes, the EU highlights the need for cross-border data flows to help a faster intervention. Cross-border cooperation should also help the creation of Union standards in terms of content administration. It is interesting to note however that this fluidity in cross-border data flows are embedded in the EU's concern for fair digital market practices to ensure a strong digital Union market. A concern that distinctly primes over its concern to safeguard freedom of expression.

▪ The most notable green stars (rarer elements)

- The emphasis throughout the EP's policy recommendation on how to tackle deepfakes, is about the issue of generating distrust in the politics and in the media. But it does not identify financial fraud, pornography, or deepfake phishing as being equally part of potential deepfake scamming systems that have the intent of politically deceiving a target audience. The EU therefore shows to have a more limited appreciation of the deepfake reality than its actual currently known scope.
- This limited appreciation of the deepfake reality also shows in the fact that, when directly addressing deepfakes, the EU centres on the audiovisual dimension of the deepfake rather than also including its written occurrence for instance.
- Despite the Data Act discussing data sharing practices and interoperability across businesses and government agencies, and despite the Artificial Intelligence (AI) Act regulating AI systems, the EU lacks extended consideration for the regulation of open data and bulk data training sets, while these are the very food feeding AI systems.
- Besides the EU's call for the need of a heightened media literacy among the EU population, there is little about the use of media policies and journalism verification systems to counter deepfakes and online disinformation.

(3) Elements absent from the EU policies but part of the deepfake chains

The EU policies make no mention of the following elements that are however considered to be part of the deepfake reality (see the deepfake chains).

▪ The entire Internet infrastructure that allows for the Internet to function as it does is absent

from the EU policies upon regulating digital services and providers of these services (e.g., broadband, information technology coding, World Wide Web, or regulatory considerations in terms of environmental sustainability that come with an ever-growing digitalisation[31]). The EU therefore also does not consider Internet protocols as relevant sites for public administration. It is noted however that this absence might in fact be related to my process of identifying the relevant EU policy documents than it has to truly do with the EU's lack of its regulation.

- In general, there is little consideration for behavioural and psychological factors that influence a netizen in its online behaviour. The EU repeatedly mentions media literacy, critical thinking, and awareness, but lacks including matters such as opinion diversity, polarisation, confirmation bias, marginalisation, vulnerability, and so on. Furthermore, while mentioning media literacy, the EU lacks consideration for informed publics, trust in the publics, access to information, public engagement, the epistemic burdening of the netizen, and so on.
- Despite the EU being itself culturally diverse as it is composed of a diverse panel of languages and cultures, by being focused on securing a Union approach, it does not consider the idiosyncratic being of the Internet, and the Anglocentric and Western-centric nature of the data on which it founds its approach. It similarly lacks addressing upon what normative premises it founds its administration. Which could be considered unsurprising since the EU by nature is a normative institution. But at the same time, engaging in the normativity underlying any administration of online content and speech would be important in terms of regulatory transparency.
- Despite the EU's clear interest in regulating digital advertising practices and data sharing practices, and discussing abundantly matters relating to programmatic advertising and online microtargeting, it nonetheless makes no mention of data brokering practices, and it also has no explicit awareness about new market trends that shift toward contextual targeting (given that microtargeting nowadays operates with HTTP cookies, but that the use of these cookies is slowly disappearing).
- While every EU act ends with a section on the financial sanctions in case of non-compliance, the EU does not consider the possibility for a broader economic sanctioning of companies or countries that participate in foreign disinformation campaigns or their facilitation through bilateral agreements.

[31] Environmental considerations in relation to deepfakes might sound surprising, but recent research suggests that climate change and the dissemination of disinformation are no less relational than are the right and left brain hemispheres (Stechemesser, Levermann & Wenz, 2022). And as pointed out by Kalpokas and Kalpokiene (2022), "[w]ith the ever-growing natural resource requirements [...] and carbon footprint [...], issues pertaining to the environmental ethics of even benign uses of [deepfake] technology must be brought to the forefront" (p. 77).

- The EU mentions the need for public media literacy, but it does not provide insights on how to implement these schemes. For instance, it does not refer to a (possible) tax on digital ad-based revenue while multiple independent researchers and research institutes are exploring this possibility to use those revenues for public awareness campaigns on deepfakes (Giansiracusa, personal communication, 28 March, 2022; Global Disinformaion Index, 2022).
- Despite the EU emphasising the need for the technological industry's participation in the moderation of online content (i.e., platform accountability), and the need for the creation of global content moderation standards, the EU lacks considering yet other participative modes for the moderation of content and the debunking of deepfakes (e.g., open source intelligence initiatives).

The superposition of the EU elements onto the deepfake chains of materialisation allowed to observe on what grounds of the deepfake reality the EU roots its policy approach. Having now finished with the empirical dive, and having discussed primary observations, the next chapter will use the deepfake assemblage and the EU assemblage for an in-depth comparative analysis of both. A comparative analysis also based on the previous Baradian revisitation of the deepfake. The above primary observations, alongside the comparative analysis, and using the Baradian insights, will thus provide the means to derive how the EU communally agreed to interpret the deepfake in its sociotechnical materialisation. It will provide the means to derive the EU's enacted conception of the deepfake, before—in turn—leading us to the Baradian-inspired policy proposal.

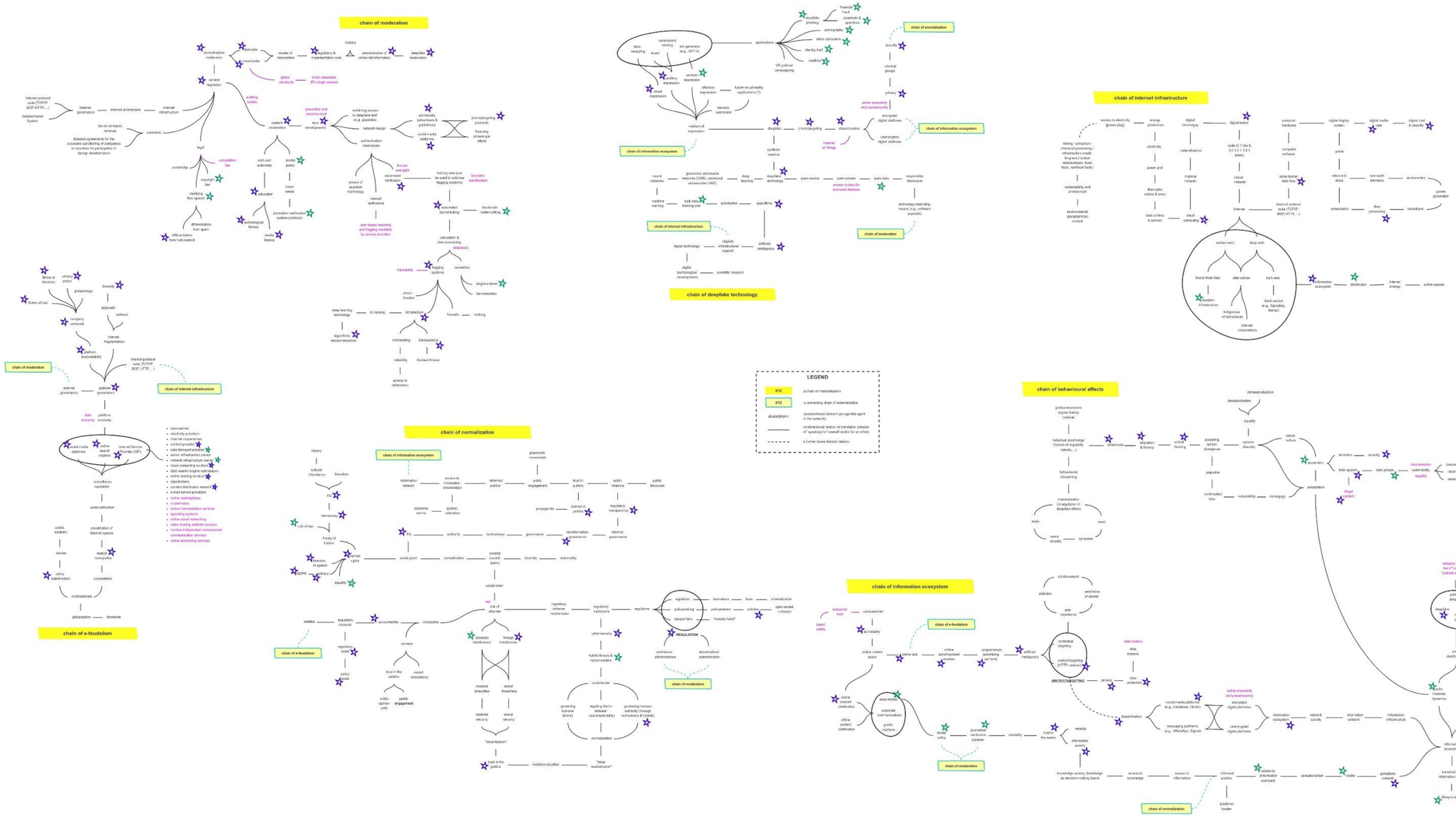


Image 13: Overview of the policy chains of materialisation.

ANALYTICAL DIVE · HOW CONCEPTIONS OF THE DEEPFAKE INFORM POLICY APPROACHES

Here we are! At last! Before moving on to the actual analysis, let us briefly recapitulate what has been done so far.

During the literature review, the notion *deepfake* was defined as synthetic political disinformation that seeks to undermine trust in public institutions. The literature review exposed an existing general inconclusiveness about how deepfakes affect society and about how effective the current means are to counter deepfakes. This inconclusiveness exposed how the popular conception of a deepfake tragedy lacked empirical evidence. This conception, dubbed *deepfake fatalism*, was thus argued to be a reduction of reality. And it led us to take a further dive into literature that looks at a more realist account of the deepfake; one that describes the deepfake in ways that better align to the more-than-tragic reality of the deepfake. The literature on this conception of the deepfake that was dubbed as *deepfake realism* pointed at the need for a theoretical research approach. But such theoretical efforts remain scarce, and that was how I introduced the purpose of present thesis to be about a contribution to efforts that experiment with another conception of the deepfake. A re-theorisation to better re-align the conception of the deepfake with its empirical observations. The purpose ultimately being about exploring how a conception of the deepfake that better matches its reality can therefore better inform its policy approach.

It was at this point that Karen Barad was introduced and applied for the conceptual revisitation of the deepfake. The notion *deepfake* now still referred to synthetic political disinformation that seeks to undermine trust in public institutions, but in a Baradian understanding that allowed for a redress of the shortcomings observed in both deepfake fatalism and deepfake realism. The deepfake was no longer conceived of as a pristine separately delineable entity. It was now conceptualised as an entanglement of material-discursive agencies in constant doing. The deepfake became an ontologically inextricable interference between society and technology, where neither has precedence. The deepfake thus has an organic existence so to speak—it is alive—since it is neither merely technological (inert), nor merely human (dependent on human intervention). A conception sometimes emphasised by referring to the deepfake as *deepfake reality*. This conception, dubbed *deepfake Baradianism*, implied that (i) the tragedy narrative is reductive, (ii) the deepfake has both a material and a discursive reality, both a technological and a social reality, (iii) the netizenry is not a uniform gullible mass prone to calamity but has agency, (iv) any act of knowledge-making or intervention is only ever partial, (v) the deepfake reality can thus never find an end, (vi) the deepfake is speech more than it is content, and (vii) fact-checking is intrinsically inefficient despite its popular praise. Most importantly, these implications offered a set of guiding premises to craft the upcoming policy proposal to address the deepfake. And I

say “guiding premises” since at that point it was still on the order of the conceptual, and thus yet to be adapted to the empirical findings. The choice of crafting a policy proposal emerged from the fact that the current theoretical contributions to the deepfake never provide actual detailed policy recommendations besides offering such guiding premises. Engaging in the crafting of an actual policy proposal—in addition to proposing guiding tenets—was to engage in a bridging exercise between theory and application that otherwise remain isolated.

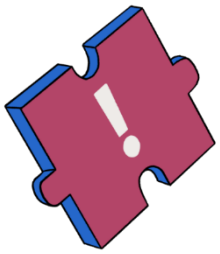
To thus provide the means of transforming the conceptual into the grounded—the guiding tenets into recommendations—an empirical dive into the world of deepfakes followed the Baradian revisitation. In effect, it was a study on how the deepfake sociotechnically materialises—on how the deepfake grounds itself in our reality and on how we ground ourselves in the deepfake reality. This empirical exploration of the deepfake reality gave birth to the *deepfake chains of materialisation*. Once so far, to be truly able of transforming the conceptual into the grounded and better inform *existing* policies, I thus needed to know how the EU addresses the deepfake. For that matter, the relevant EU policies were investigated to extract the sociotechnical elements of the deepfake that the EU enacts in its policy approach. These collected policy elements were superposed onto the former deepfake chains of materialisation. Which resulted in the so-called *policy chains of materialisation*. This superposition allowed for a comparative analysis between my way of assembling the deepfake reality and the EU’s way of weaving together the deepfake reality. Indeed, this comparison enabled to reveal which elements of the deepfake reality are taken into consideration by the EU; it revealed on what grounds of the deepfake reality the EU roots its policy approach.

It is at this moment that present analysis chapter jumps in. A chapter that essentially consists of an analytically elaborated policy proposal. Which was therefore chosen to be provided in the formatting style of an actual policy proposal (also to break the monotony of the thesis and give it a fresh breeze). The proposal will start with an executive summary of the situation to contextualise the policy approach to deepfakes. It will then proceed with a breakdown of the EU’s current policy approach. This analysis of the EU will then lead to the derivation of the EU’s conception of the deepfake. You will see that this conceptual derivation will be given in a so-called concepts box—a formatting decision made to clearly emphasise the importance of this part of the proposal although it is not a usual part in common policy proposals. After that, the policy proposal will move on with the actual proposal; i.e., the proposed policy approach followed by concrete recommendations for its application. The proposal will then be summarised in its concluding abstract.

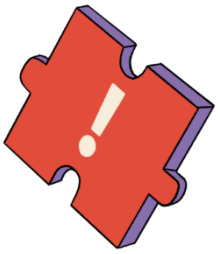
This proposal could not have seen daylight without the prior Baradian conceptual revisitation of the deepfake (to provide the guiding premises that redress the shortcomings of other conceptions). And neither could it without the comparative analysis of the previously assembled deepfake assemblage and EU assemblage (to provide a means to empirically ground the conceptual model of the deepfake). For, indeed, the overall endeavour is to elucidate how a Baradian conception of the deepfake re-informs the EU’s policy approach to the deepfake. Because as we are about to discover, the current EU approach is

based on a conception of the deepfake that tends toward that of deepfake fatalism, a conception that however showed to be empirically inaccurate. And which is how the upcoming Baradian-inspired policy proposal will be argued to provide the basis for both an ontological and an empirical legitimacy that the current EU approach lacks.

On a methodological note, while a preliminary Baradian-inspired policy proposal was drafted right after finishing the deepfake assemblage—to push myself through a Baradian thinking without yet being perturbed by an EU thinking—this draft was subsequently revised after finishing the EU assemblage. Note as well that for this policy analysis, since the relevant EU regulations and recommendations were scattered across six documents, I had first compiled the regulatory elements of interest to deepfakes into a single document, a *model EU regulation* (see Annex at the end of this thesis). This model EU regulation proved to be invaluably meaningful to assess the scope of the EU approach, its benefits, and its limitations. It is this model regulation that proved to be my true analytical basis to derive the EU's conception of the deepfake, because even though it is fictitious, this model EU regulation only contains content as provided by the EU.



ADDRESSING POLITICAL DEEPFAKES



EXECUTIVE SUMMARY

A **political deepfake** (hereafter referred to as deepfake) can be an image, video, text, speech, or sound produced with deep learning technology that seeks to induce political deception. By **political deception** is meant the undermining of public trust in public institutions—the government, the media, the science.

Society's present-day massive reliance on online media makes that deepfakes have the capacity for a rapid, large-scale dissemination of targeted political disinformation [1]. Their feared potential in terms of war propaganda, foreign interference, terrorism, the sapping of elections, disrupting democracy, bigotry, clickbait money-making, trolling, hampering progress in public health or environmental matters, and so on, have focused today's efforts on the prevention and mitigation of such feared incidents [2]. Yet, central to these fears lies the in fact still unanswered question of how exactly deepfakes affect a democratic society [3] [4]. Do they truly embody the potential for mass deceit, or does their capacity as medium of expression simply serve to express and exacerbate prior discontents or defamatory intents? Fake stories passing credibility checks have always existed, and they always will. And it is the very characteristic of deepfakes being a medium of expression that makes their regulation such sensitive a matter. Regulating deepfakes is regulating speech.

While leaving deepfakes operating in the wild is not an option, for it could translate into potentially consequential social and economic costs, the persistent inconclusiveness about how deepfakes affect democracies does not justify more constraining forms of intervention. A restrictive approach could do more harm by hurting public trust in public institutions, let alone be abused by demagogues. Present proposal thus suggests a two-speed approach, where deepfakes affecting national security would involve rapid action, and deepfakes regarding public opinion would involve the building of a long-term relation of trust. By seeking to complement the shortcomings of the EU's approach, the proposal centres on the neglected idea that the netizen ought to be entrusted more confidently in its capacity for autonomous cognition. After reviewing the EU's present-day approach to deepfakes, the recommendations will be developed through three structuring pillars.

Firstly, as disinformation shaped into a moneymaking practice through the influence of the online advertising business model [5], the first pillar focuses on dismantling that relation to lucre to ensure that the dissemination of deepfakes remains bound to political intents.

Secondly, having confined the dissemination of deepfakes to political objectives by removing financial incentives, the second pillar focuses on increasing efforts to increase the diversity of one's online content encounters to promote debates that lead to vindication rather than vilification. Because the trouble with deepfakes is not so much about the content as it is about our relation to that content [6].

Thirdly, to build a relation of trust with civil society and trusting the netizen's capacity for autonomous judgement, the third pillar therefore calls on the implementation of literacy schemes to better accompany online content uptake. The point is to implement content moderation practices that ensure and promote social cohesion, and prevent a restrictive turn of events.

01

defunding disinformation.
crack the business model and prevent deepfakes from turning into financial assets.

02

diverse online content encounters.
accustom society to the diversity opinions by increasing the diversity of online content encounters.

03

favouring literacy.
trust the netizen in its ability for autonomous cognition and raise public awareness about deepfakes.

THE CURRENT EU POLICY APPROACH

The EU has both direct and indirect approaches to deepfakes. Its direct approach is currently still at the level of recommendations [7]. Its indirect approach is a code of practice on online disinformation, and a set of legal acts that regulate the digital market, AI systems, and online data sharing practices [8].

When tackling deepfakes directly, the EU largely promotes an interventionist style focused on increasing fact-checking efforts. The EU thereby abides by a dualist conception of reality where the good and the bad, information and disinformation, are clearly separable. It also abides by the idea of society being a homogenous ensemble removed from any complexity. And it concentrates on the aspect of content rather than speech, envisioning deepfakes as pristine incidents rather than as echoes from civil society.

When tackling deepfakes indirectly through the code of practice on disinformation, the EU again largely emphasises the need for stronger fact-checking efforts. Here, it expresses a more acute concern about demonetising disinformation. But it does so from the perspective of ensuring brand safety more than it is concerned with the way user-generated data is used in those practices. And, here too, the EU remains rather elusive in its definitions and recommendations.

Finally, when tackling deepfakes indirectly through the regulation of the digital market, of AI systems, or of online data sharing practices, the acts focus on securing a strong digital Union market by favouring fair business practices, thus requiring transparency from businesses to ensure an open market. The EU relies on a two-scale approach, distinguishing between two categories of stakeholders. Only the first category really ought to abide by certain strict rules. The rest is encouraged to follow suit on a voluntary basis.

(i) **very large online platforms** and **gatekeepers**, respectively, platforms with a minimum monthly online reach of 45 million active recipients in the Union [9], and core platform services that "have a significant impact on the internal market" [10]; and

(ii) the rest.

Importantly, according to the AI Act, deepfake content ought to be labelled, unless it is used by law enforcement or for military purposes.

Very large online platforms are encouraged to have crisis protocols in case of "extraordinary circumstances […] that can lead to a serious threat to public security or public health in the Union or significant parts thereof. Such crises could result from armed conflicts or acts of terrorism, including emerging conflicts or acts of terrorism, natural disasters such as earthquakes and hurricanes, as well as from pandemics and other serious cross-border threats to public health" (see Annex, Article 20(5) [11]). The EU speaks of a voluntary engagement. However, in those extraordinary circumstances, the Commission "should be able to require […] service providers to initiate a crisis response as a matter of urgency" [11]. The EU lacks clarity on this nonetheless important point, since the EU would become the content administrator of the European Internet. Additionally, the EU lacks clarifying who can declare such state of emergency. Supposedly it is a decision that is to be enacted by each EU member state individually, but this then breaks the EU's present effort to consolidate an EU approach.

CONCEPTS BOX • the EU conception of the deepfake

The EU's policy approach to deepfakes allowed to make the following conceptual observations about how the EU conceptualises the deepfake.

- *The deepfake is excisable content*

  The deepfake has the potential to cause distrust in the politics and in the media. The EU's understanding of causation abides by a classical understanding of causality. One where the deepfake is a pristine incident that is separately delineable—external—and therefore also excisable from society through a regulatory surgical move. A move centred on fact-checking. Contrasting to those conceptions, in a Baradian logic, matter is inextricable from its context, and therefore matter cannot simply be excised.

- *An indisputable demarcation between fact and fake is possible*

  The EU strongly highlights fact-checking as *the* solution to deepfakes. The fact-checking ideal abides by a dualist conception of reality. One where information and disinformation are two separately delineable binary opposites rather than interfering parts of a diffracted ensemble. This ideal rehearses the thought that the issue with the deepfake is its content alone—an excisable matter. But a deepfake is as much about content as it is about speech and therefore social context. Even though the EU expresses concern over preserving freedom of speech, its recommendations lack crucial consideration for the social depths of the deepfake—for its discursive dimension. A shortcoming that comes to exposure for instance when the EU writes that "[certified trusted fact-checking organisations shall demonstrate] that they work in a diligent, accurate and objective manner" [12].

- *Deepfakes and the market logic*

  Market and economy are in the EU's limelight when it comes down to the regulation of technology, and thus of deepfakes. Evidently, one of the EU's primary objectives is that of securing a strong economic space. And it is embedded in that neoliberal logic of growth that the EU favours brand safety and data harvesting practices under the cupola of facilitated data sharing practices and digital services interoperability between business-to-business and business-to-government entities. The EU is thus centred on the market, and on businesses and public institutions, more than it is concerned with civil society and the safeguarding of freedom of speech. A market logic should however not overshadow other equally important logics.

  The neoliberal market logic of growth and harvesting further rehearses the (old) materialist conception of the deepfake as an entity apart, prone to both hoarding and excision. And it sustains the dualist conception of reality—the narrative of success or failure where one ought to either operate alongside the growth mantra or shall lag behind, be deprived of any agency, and have failed.

- *Trust can be built through a top-down approach*

  While the EU seeks to enhance public trust in the politics and in the media, it has little consideration for the netizenry—the online citizenry. The EU conceives of relationships of trust as a unidirectional matter; one that could be built through a top-down implementation. A conception that exposes the EU's paternalistic facet which importantly dismisses that trust is a two-way relationship. Not acknowledging this reality will inherently bind transparency and fact-checking efforts to failure.

- *Deepfakes can be tackled with technology*

  Techno-solutionism makes sense in an era where policies seek to lead society's digital turn. However, the type of techno-solutionism advanced by the EU abides by the idea of tech-neutrality. An idealised neutrality that has already been debunked at large [13], but that nonetheless persists at EU level; e.g., "[t]raining, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system" [14].

  Additionally, while focusing on content makes sense for the case of harmful and illegal content, a reliance on fact-checking technologies, organisations, and infrastructure alone maintains a dismissal of the netizenry in its capacity for autonomy and self-sufficiency.

- *Literacy is a given*

  While the EU centres on transparency requirements for advertising practices and algorithmic recommender systems, the EU gives very little attention to user literacy. However, how is a netizen supposed to navigate the information it is provided with, without the necessary background knowledge? Implementing transparency requirements without the implementation of a general user literacy scheme risks resulting in what some have coined as the epistemic burdening of civil society [15].

These observations lead us to conclude that the EU conceptualises the deepfake as a pristine piece of content, devoid of context, that can be separately delineated and excised from society through a top-down regulatory surgical move and thereby find an end. The EU thus operates alongside the observations made for the popular fatalist conception of the deepfake. And, although the fatalist tragedy narrative is not explicitly rehearsed in the legal acts that regulate the digital market space, data sharing practices, AI systems, and digital service providers, this tragedy narrative does appear in the EP's study on tackling deepfakes [16]. The EP's study does provide a fuller picture of the deepfake by expressing the various uses of it and providing some timid hints at its more-than-technological character—its social character. However, it still abides by the doomsday prophecy, and thereby legitimates a generalised dismissal of civil society's agential nature. Deepfake fatalism was however shown to lack empirical evidence. The EU's policy approach to deepfakes is thus based on a conception of the deepfake that does not align with observations.

ANOTHER APPROACH

The complementary approach—as inspired by the Baradian reconceptualisation of the deepfake—seeks to compensate for the two main criticisms vis-à-vis the EU's approach. The first is the complete disregard of the EU netizen as autonomous cognisant individual. Whereas, if the EU wishes to establish a relation of trust with its citizens, then it ought to trust them in their capacity for autonomous judgement and decision-making. The second criticism regards the EU's almost exclusive push for fact-checking, while research shows equivocal on the efficacy of this method [17], and exposes the systemic depths that deepfakes are implicated in, thereby informing on the need for other modes of intervention [18]. More precisely, the complementary approach seeks to put up with the following shortcomings.

(i) Fact-checking is a value-laden process. Whether implemented with or without content removal procedures, it does not necessarily provide the sought-after corrective effects. It can even prove counterproductive [19].

(ii) Relying on technology only is insufficient. Whether it is about algorithmic techniques to debunk deepfakes, automated content moderation, optimised training models, or the use of blockchain's watermarking feature, any of these technologies is guilty of mislabelling [20]. Not only is technology not devoid of sociocultural precedence (a reality that the EU currently fails to acknowledge more overtly [21]), but these technologies are themselves targeted by technologies that seek to counteract them [22]. Even if these are to be perfected with more target words, more context-sensitive training, …, they always only rely on training sets that are partial. Additionally, a tech-focus centres the attention on deepfakes as exterminable content rather than as evidence of systemic issues [23].

(iii) While data is at the basis of targeted advertising, a lot remains imprecise in terms of user-generated data brokering and data cross-combination through such brokering activities. So far, the EU only requires transparency in advertising practices, obliges gatekeepers and very large online platforms to provide its users with options to opt out of recommender systems without impeding on the quality of the service provided, and prohibits the nonconsensual transfer of user-generated data to third parties. However, the level of

detail to be provided to a user about its data upon transfer is imprecise. And so is the case of business buyouts; "what happens to corporate-collected data once a company is acquired or merged" [24]. Thirdly, providers of data training sets are categorised as non-high-risk AI systems that only need to comply with "minimum transparency obligations" [25].

(iv) The EU does not address the case of machine-generated data, such as artificial deepfake-generated data, that is currently only regulated by copyright protection.

(v) While deepfake systems are not prohibited, AI-based disinformation is prohibited [26]. However, arguably, content generated through deep learning falls under the scope of protected speech [27]. It does not mean that any deepfake is to be protected as such, but it does complexify the situation. As de Vries (2022) writes,

> "[to the EU, disinformation is] "verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public"[, but] this definition will only cover the very extreme cases of misleading information. Much of [machine learning]-generated news will fall in a normative grey area[,] where the boundaries of constitutionally protected and unprotected speech are not always easy to draw" (p. 2 & 17).

(vi) Even though the EP notes that "the most imminent danger [for disinformation operation] may be posed by deepfake text" [28], in its recommendations to tackle deepfakes, the EU is explicitly focused on audiovisual deepfakes.

(vii) The EU recommends the implementation of media literacy programmes to enable users to assess the factuality of online content. Firstly, this pushes the user into a dualist (polarised) perception of reality. Secondly, this is insufficient in regard of EU's transparency requirements for content moderation practices. The EU lacks a joint implementation with literacy schemes on labelling and ad-targeting practices; users need to be able to assess the technical information they are served with, however basic that information is.

(viii) While the EU enacts rules for online advertising practices, it is seemingly more permissive in terms of political ads than commercial ads [29]. Moreover, by focusing on the advertising practices of only very large online platforms and gatekeepers, it dismisses a whole panoply of stakeholders (the relatively smaller ones) that should equally abide by those same advertising practices, especially if those practices fund disinformation. "[S]mall platforms in the data economy are, in fact, equally of concern in many ways, not less" [24].

To compensate for these criticisms and better boost public trust in public institutions, the general proposition is to make room for debate and exchanges of opinions. This will favour a receptiveness towards other modes of thinking and will boost one to engage more critically with topics of interest. For that matter, the proposal recommends a two-level approach combined to a two-speed approach. Two-level in terms of content, like the EU, distinguishing between (i) harmful and illegal content and (ii) other content. And two-speed in terms of action-taking, unlike the EU, distinguishing between rapid intake and a long-term intake; distinguishing between deepfakes involving national security (hereafter, NS) and deepfakes regarding public opinion (hereafter, PO) [30].

It is because of the distinct level or risk involved by the welfare state for both categories that they involve a different temporal scheme. The NS category includes deepfakes used in conjunction with other means to cause a direct critical hit on a welfare state; e.g., war propaganda, deceitful international (public) relations campaigns, foreign interference, and large-scale cyberattacks aimed at disrupting state harmony. The PO category includes deepfakes used in conjunction with other means to deceitfully orient public opinion, be it on individual or community level. NS deepfakes involve a short-term on-the-spot approach that entails restrictive content moderation means. PO deepfakes involve a preventive long-term scheme consisting of means to raise public awareness on online encountered content and facilitating exchanges of opinions. The PO scheme includes means to intervene rapidly for cases of public political defamation, although these interventions would always only happen after a detected dissemination (as is the case today), and would be a publicly transparent practice to avoid fuelling public distrust.

For the same idea of avoiding to raise public distrust, it is discouraged to go by recommendations of having a centralised public administration of online content [31], or of regulating online chat spaces by "[l]imiting the number of users in (chat) groups" [32], or of creating third parties to intervene as "autonomous nonpartisan entities" to demarcate between what is right or wrong (the so-called *middleware approach* [33]). Firstly, because "platform authoritarianism" [34] should not be replaced by government authoritarianism or a top-down administration [35]. And secondly because there is no such thing as non-partisanship, especially when it comes to content moderation, and because this would add yet another substantive layer of middle(wo)men to an already dense landscape (overregulation), while the netizen is itself insufficiently acknowledged in its capacity for growing an aware online engagement.

It is on top of that temporal foundation that the three pillars of the proposal build.

01 Defunding disinformation

Disinformation is inexorable. It will always be part of humankind. But policies can dissuade money-enthusiasts from joining the drive by dissociating financial interests from disseminating deepfakes.

02 Diverse online content encounters

Except for matters of national security and cases requiring rapid intake, debates are be favoured instead of closing content down peremptorily, as the latter risks fuelling distrust. Favouring debates prevents worsening sentiments of resentment among individuals who already feel bereft of their freedom of expression. It also prevents false negatives resulting in the removal of content created by activists, satirists, informants, and whistle-blowers. And the avoidance of premature content removal also enables to better target systemic issues and "cultural misbehaviour" [36]. Moreover, while at times we willingly seek self-validating content, encountering other content might open our horizons and heighten our sense of perspective. Favouring diversity will also give more space and visibility to marginal voices of all kinds. It thus prevents online content to take a potential domineering and polarising turn that leads to bigotry rather than helps society acknowledge a diversity of opinions and accompanies society in that diversity. Lastly, a diverse online space will also result in a wider range of algorithmic cocktail models rather than "having so many citizens rely on the same algorithms to get their real-time news updates" [37].

Receptiveness to diverse content and other modes of thinking implies the need for AI diversification systems besides the AI recommender systems, and for a fragmented and decentralised Internet; cf. "no one player [should exercise] outsized power in making fine-grained decisions over content" [38]. A **diversification system** is an AI system that seeks to promote content diversification contrasting to the nudging purpose of recommender systems. AI diversification systems do not prevent algorithms to direct a netizen to disinformation, and the content one would encounter would still be a mix of opinionated voices, dissenting voices, harmed voices, and harmful voices. But it will make the Internet more diverse and inclusive. Which is why this pillar ought to be supported by the third pillar of online content literacy.

A shift towards a more participative and collaborative online space, that contrasts with today's monopolistic "privatized environment" [39], is already ongoing if thinking of Web3 or peer-to-peer Internet models. Some of those initiatives are funded by the EU [40]. Directing a netizen towards diverse content also does not prevent user-generated data to remain the main online currency. Therefore, it does not inhibit a participative interest from any stakeholder. At the same time, the AI diversification systems would operate based on content and not on user. Therefore, it does not prevent a data-degrowth mindset either if countries would someday be willing to move away from surveillance capitalism [41].

03 Favouring literacy

Too little consideration is given to the fact that the **netizenry**—the online citizenry—is not a uniform homogeneous mass deprived of critical thinking. And too little regulatory experimentation is done in that respect. We all process information differently, and while it is impossible to have a solution tailored to all, the aftereffects of opening the debates are better than closing content

based on unfit classifications. More attention ought to be provided to literacy to guide the netizenry in reality's diversity.

RECOMMENDATIONS

× 01 DEFUNDING DISINFORMATION ×

(1) Cracking the disinformation model

Efforts exist to disrupt the disinformation business model and to untie financial incentives from disseminating deepfakes. These efforts should be supported more fervently.

Providers of core platform services should be prohibited not to have ad-focused content moderation practices that delink ad-based revenue from illegal content, and should be encouraged to explore different approaches.

Core platform services include: "(a) online intermediation services; (b) online search engines; (c) online social networking services; (d) video sharing platform services[, music and podcasting platform services], (e) number-independent interpersonal electronic communication services; (f) operating systems; (g) web browsers; (h) virtual assistants [and virtual reality services]; (i) cloud computing services; (j) online advertising services, including any advertising networks, advertising exchanges and any other advertising intermediation services, provided by an undertaking that provides any of the core platform services listed in points (a) to (i)" [42]; (k) data brokering services; and (l) (video)gaming platform services.

Illegal content "refer[s] to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech [43] or terrorist content and unlawful discriminatory content, or that the applicable rules make illegal in view of the fact that it relates to activities that are illegal. Illustrative examples include the sharing of images depicting child sexual abuse, unlawful non-consensual sharing of private images, online stalking, the sale of noncompliant or counterfeit products, the sale of products or the provision of services in infringement of consumer protection law, the non-authorised use of copyright protected material" [44].

The exploration of moderating practices shall be supervised by an **Exploration Board**. One board per member state, each constituted of 5 representatives, stemming from different sectors of economic activity (no Eurocrats), and elected by the citizens of that member state. It shall publish yearly a baseline moderation protocol that shall have to be aligned to the findings of independent research. Board members shall be mandated for 4-year rotational shifts and shall be free of conflict of interest in regard of online advertising services and (intermediation) services.

Obliging content moderation is unlikely to drive up the price of online ad spaces, since providers of core platform services by nature apply content moderation processes. However, this time, they will have to orient their efforts towards delinking ads from illegal content. Letting them explore different means will favour plurality and diversity of content moderation and platform practices. This exploration will be well-serving if jointly calibrated with civil society awareness campaigns where citizens will be demanding such good conduct from the platforms. Deepfake literacy programmes, online content awareness campaigns, and research would be funded through a tax on digital ad-based revenue.

Providers of core platform services will want to follow a good conduct as they might otherwise face reduced online visibility, thus, reduced revenues. This reduced visibility will be ensured by encouraging through government incentives public, private, or mixed effort initiatives that tackle deepfakes and online disinformation, like the Global Disinformation Index initiative that downranks websites of disinformation offenders. Search engine optimisation agencies shall be obliged to collaborate with providers of core platform services, in the form of trainings and

consultations on the latest highlights in digital advertising systems and practices.

(2) Political advertising requirements

- **Data brokers**—companies whose primary service is the trade of machine-generated data and/or user-generated data, other than data regarding national security (such as data from a nuclear power plant), and which involves data collection, centralisation, transformation, or (re)organisation—shall not be limited or inhibited in their merger and acquisition endeavours. However, they shall be prohibited to provide their services for political purposes; i.e., the EU's enacted business-to-government data sharing regulation shall include a prohibition to share or sell such data for purposes of political campaigning. And governments are to be prohibited from buying or using data for campaigning. Besides the existing requirement from the EU that governments have to publish their requests made to private entities to share data, the Exploration Board shall have the ability to oppose any such governmental request for data if deemed not compliant with the above prohibition.
- Political entities shall publish all their purchases of online ad space [45]. Additionally, any content sponsored by a political party shall be labelled visibly (the same way that such markers already exist for industry-sponsored adverts).
- Whenever data broker services or core platform services shall process an amount of data that exceeds a certain threshold (to be established by the Exploration Board after expert consultation), whether it concerns sensitive data or not, that service shall automatically have to abide by the rules enacted for high-risk AI systems.
- Companies selling services to artificially boost content through fake accounts (e.g., boosting the number of followers or likes) shall be prohibited. And both private and public entities shall be prohibited from buying those services.

(3) Funding research

- While the business model in question is exclusively operating on the surface web because of the ad-based stimuli, the surface web only represents a fraction of the Internet. Too little is known about disinformation dynamics in other fragments of the Internet. Research should be funded to look at those dynamics. This would allow to monitor and assess timely the relevancy of further research in the deep and dark web.
- Given a rising trend in the publishing industry to push online registered users to become paid subscribers, an eye should be kept on changes in the online information ecosystem dynamics; whether some sort of reverse dynamic might emerge if paywalls become the new norm for quality content, and that deepfakes would strive on the free web.
- Discussions exist about the similarities between social media platforms and traditional media companies, with some arguing that both should be regulated similarly [46]. However, news consumption via social media platforms is not that straightforward [47]. More research ought to be done in both respects.
- "Direct bank transfers are reportedly the most frequently used online funding mechanisms [by some investigated hate groups]" [48]. Similarly, "crypto-funding is becoming an alternative way for malicious actors to get funds and keep producing false content" [49]. Therefore, as one can expect foreign interference campaigns to be a form of organised crime involving consequential money transactions, to dart bigotry-funding activities, research should also focus on developing detection protocols in banks, crypto-trading platforms, and crowd-funding platforms for the case of national security deepfakes.

× 02 DIVERSE ONLINE CONTENT ENCOUNTERS ×

(1) Introducing AI diversification systems

To heighten our sense of perspective and our to heighten our openness to diverse opinions, diversification systems ought to be

implemented (i.e., AI systems that seek to promote content diversification in contrast to current recommender systems). Techniques to help this already exist.

- *Propagation detection algorithms* serve to flag propagation dynamics regardless of content [52]. They could serve to detect nudging trends.
- *Filter bubble algorithms* detect content towards which a user is nudged [50]. Diversification systems could then be coded to add more variety to the content otherwise served to that user [51].
- *Circuit breaker algorithms* serve to halt the circulation of content once it reaches a certain virality threshold [51]. These should only be implemented for cases involving national security. It would not itself remove the content from the platform (which is also not the point of the algorithm), but it would inhibit debilitating cross-referral loopholes.

Diversification schemes should be monitored using algorithmic impact assessments and recalibrated or rethought accordingly by the joint association of providers of core platform services and the Exploration Board.

(2) Inhibiting cross-border disinformation

The EU currently prohibits the sharing of data between an EU established company and a third country unless international agreements exist. For reasons of national security alone, trade agreements should complement the existing EU regulation to administer "cross-border data flows" specifically to inhibit foreign disinformation interferences [53]; "private firms [should be banned] from producing and exporting disinformation as a service" [54].

(3) Decentralisation through participation

Society always operated along dynamics of leaders and followers. Because of this inevitability, where some are looking into means to decentralise the influencer economy, present proposal focuses instead on reminding individuals that they have a right to participate in the lead. In that line of thought, when a core platform service has its core business relating to content distribution—irrespective of the form of content—and that this platform service exceeds an average online visitation rate (online traffic) to be determined by the Exploration Board after expert consultation, the provider of this platform service shall have a democratically elected oversight board constituted of users of that platform and mandated for 1 year (what Ovadya coined *platform democracy*; [55]).

(4) Promoting workforce diversity

Content diversification cannot happen without a diversity in the workforce operative in whatever part of the deepfake supply chain (going from legislators to researchers to employees and employers in providers of core platform services etc.). Workforce diversity is not only about cultural diversity but also about academic background diversity; it is about both a multicultural and multidisciplinary workforce. Providers of core platform services shall be obliged to implement such diversity and employment experts shall be consulted for this implementation.

This shall be complemented with mandatory expert exchanges across the member state Exploration Boards during their mandate to encourage cross-Europe exchanges of best practices. The Exploration Boards shall also be obliged to have such consultative exchanges with experts from both private and public initiatives to counter deepfakes to ensure a better transfer of knowledge from research institutes to policymakers (e.g., Bellingcat, Global Disinformation Index, Institute for Strategic Dialogue, Office of the United Nations High Commissioner for Human Rights).

(5) Cyber defence cooperation

Cooperation shall be reinforced with cybersecurity institutes to elaborate content moderation approaches for national security.

(6) Funding non-Anglocentric research

- While research on deepfakes in general should focus more on the social aspect of associated information network dynamics and behavioural factors, these efforts should be particularly dedicated to non-Anglocentric content and dynamics. Especially that the EU is a multicultural and multilingual region. Optimised language- and culture-bound insights would allow for more idiosyncratically adapted content moderation practices (currently a pain point [56]), and a better implementation of diversification systems.
- In line with previous recommendations on Internet diversity, more research efforts ought to be dedicated to reviewed designs

of the Internet architecture and protocols [57]. This would allow for a more structural diversification of the online space.

× 03 FAVOURING LITERACY ×

(1) Informed transparency

Informed transparency here means transparency towards the end-user about the processes at work within core platform services while preventing an unnecessary epistemic burdening of the end-user, which is why this transparency scheme is further complemented with a literacy scheme.

- The EU currently requires explicative material on recommender systems and data sharing practices only from signatories of the SCoPoD, from very large online platforms, from gatekeepers, and from providers of high-risk AI systems. The proposal recommends that *any* provider of core platform services (as defined in the proposal) should be obliged to inform users of the functioning of its services, including the used profiling and targeting mechanisms, and it should open-source the recommender code.

  It is recommended that the disclosed information includes the following details.

  (i) Firstly, it should be published both in the form of manuals that include technical details and in the form of summarising videos or image slideshows providing the basics.

  (ii) Secondly, the information shall include: (i) "the interests (e.g. skiing, black-and-white movies, or bird-watching) and attributes (e.g. race, age, or sexual orientation) that the platform may have inferred about [the user], including how the system made these inferences"), (ii) "system-level documentation [like] ranking decisions applicable to the entire recommender system" ("documentation could include the types of content the platform prioritizes and downranks, and whether it is operating under heightened "break-glass" conditions due to external factors like elections or other events likely to increase civic unrest"), (iii) "[c]ustomers should also be informed if a platform is prioritizing its own brands or partners" [58].

  (iii) Lastly, unlike the current EU requirements, such transparency also ought to apply to the application programming interfaces that the users dispose of to look into the data.

- After consultation with experts, the Exploration Board shall publish a transparency act to make sure that this transparency scheme is well-framed to prevent perverse misuses. Therefore, it is also advised to implement a regulatory sandbox project before full-blown rollout. This transparency scheme should be complemented by legal and technical efforts to secure Internet protocols that "provide sufficient privacy protection" [59].
- Data brokers and providers of core platform services shall similarly publish their modes of cross-combining data.
- For deepfakes provided as a public service or provided as public content by a public institution or agency, the deepfake shall not only be labelled, but both the deepfake code and its data training sets—whether authentic or synthetic—shall be open source. The open-sourcing shall comply with the General Data Protection Regulation.
- Besides the ad-repositories required by the EU, the providers of core platform services should be obliged to publish (statistical) reports on removed content and make these reports easily accessible on their hosting services. This joins the previous obligation for providers of core platform services to have ad-based content moderation practices. These providers thus ought to have content moderation practices in general, besides the ad-focused moderation. To ensure a certain level of quality, providers are prohibited not to have content moderation practices that align with good practices in the field. These good practices will be published yearly by the Exploration Board.

(2) Implementing literacy schemes

To heighten public awareness about online content and online information ecosystems, literacy schemes shall be implemented in addition to the EU labelling requirement for deepfake content.

Few public awareness campaigns have already been introduced, such as the famous deepfake videos of Barak Obama and Queen Elizabeth II where both were ostensibly talking gibberish [60]. Such risk awareness is key to having a society that is critical in its engagement with digital media [61]. However, more frequent and true literacy programmes are necessary within schools and universities [62]. And equally so such regular trainings for information experts such as librarians to address deepfakes [63].

- High-school curricula shall include a few hours of online media literacy each year.
- Expert literacy programmes for librarians and government bodies.
- Online media literacy schemes shall be proposed yearly for free to reach beyond the classroom walls and to reach all generations. These shall be developed by providers of core platform service and initiatives to counter deepfakes, and shall be approved by the Exploration Board.
- Public deepfake awareness campaigns shall be developed yearly in both physical and virtual public spaces (in public squares, public transport, on smartphone applications, websites, …). These campaigns shall be funded through digital ad-based revenue taxes and shall be deployed by the Exploration Board.
- Open-source intelligence initiatives that are focused on deepfakes and political disinformation should be encouraged through financial incentives. This will inspire other civilians and grassroots movements to engage in debunking deepfakes (like the Bellingcat initiative). Similarly, community-based initiatives should be encouraged to be organised within private organisations (such as Twitter's Birdwatch). This will add to the plurality and diversity of types of content moderation practices.

ABSRACT

Encountering a deepfake can bring to expression dormant tensions and worries within a netizen. On the other hand, their removal can reinforce public distrust in public institutions. Well-crafted recommendations are therefore crucial to prevent a weakening of public trust.

Using a Baradian-inspired conception of the deepfake allows to compensate for two main criticisms made vis-à-vis the EU.

The first criticism is about the EU's complete dismissal of the netizenry in terms of its agential capacity. A dismissal that translates in a policy that tends toward paternalism (the netizen needs protection from the deepfake). And a dismissal that operates by the belief that a purely technology-based solution—fact-checking—is the answer to what would be a purely technological problem. But the deepfake is more than a technological event, and society can therefore not be dismissed.

The second criticism is about the EU's almost exclusive push for fact-checking efforts to address deepfakes, while it has been shown to be clearly insufficient and at times even counterproductive. Indeed, the fact-checking ideal abides by the classical ideal of objectivity, whereas anything that we humans do is inherently value-laden.

Using a Baradian-inspired policy approach allows to appreciate the netizen in its capacity for autonomous cognition and judgment. It offers approaches to the deepfake that give room to human agency, all the while securing means to address cases of bigotry and national security. Such approach is argued to only benefit a burgeoning trust in public institutions. And, therefore, entrusting civil society in its capacity to navigate the deepfake reality is as important as entrusting public administrations and private entities to do so.

ENDNOTES

[1] Dobber et al., 2021; Newman et al., n.d.; Vosoughi et al., 2018.
[2] Ayad et al., 2022; Chesney & Citron, 2019; Diakopoulos & Johnson, 2020; Donovan, 2021; Fallis, 2020; Ovadya, 2019; Vaccari & Chadwick, 2020.
[3] Bengani et al., 2022; CITAP, 2022; Cover, 2022; Hameleers et al., 2022; Kwok & Koh, 2020; Taylor, 2021; Yadlin-Segal & Oppenheim, 2020.
[4] Multiple studies depart from certain presumptions about deepfakes and hint at a need for a nuancing and thorough contextualising of those presumptions. For instance, encountering a deepfake does not necessarily affect one's political opinion (Hendrix, 2021), as one's uptake of such content is importantly determined by one's political a priori (Osmundsen et al., 2021; Yun Shin & Lee, 2022).
[5] Global Disinformation Index, personal communication, 26 April, 2022; Radsch, 2022.
[6] Dobber et al., 2021; Garrett et al., 2013; Lecomte, 2021; Wong, 2021; Wood & Porter, 2019.
[7] EP, 2021.
[8] EP, 2022a; EP, 2022b; EC, 2021c; EC, 2022b; EC, 2022c.
[9] EP, 2022b, p. 200.
[10] EP, 2022a, p. 92.
[11] EP, 2022b, p. 85.
[12] EP, 2022b, p. 54.
[13] E.g., Callon, 1984; Paris, 2021; de Vries, 2022.
[14] EC, 2021c, p. 29.
[15] Paris, personal communication, 8 April, 2022.
[16] EP, 2021.
[17] Dobber et al., 2021; Garrett et al., 2013; Wong, 2021; Wood & Porter, 2019.
[18] Gladstone, 2021; Institute for Strategic Dialogue, 2022; László et al., 2022.
[19] Wood & Porter, 2019.
[20] Juefei-Xu et al., 2022; Paris & Donovan, 2019.
[21] The EU's failure to acknowledge the cracks in the tech-neutrality ideal shows for instance upon writing that "Signatories providing trustworthiness indicators will ensure that information sources are being reviewed in a transparent, apolitical, unbiased, and independent manner" (EC, 2022b, p. 23). Or that "[t]raining, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system" (EC, 2021c, p. 29).
[22] GAO, 2020; Juefei-Xu et al., 2022; Nguyen et al., 2019.
[23] Lecomte, 2021.
[24] László et al., 2022.
[25] EC, 2021c, p. 3.
[26] EC, 2021c, p. 43.
[27] de Vries, 2022.
[28] Bayer et al., 2021, p. 25.
[29] See EC, 2022b, p. 10 & 11.
[30] The recommendation from the EU for very large online platforms to implement crisis protocols is also a type of temporal approach. However, where the EU proposes this as a minor addon, present proposal explicitly defines the entire policy approach around that temporal distinction.
[31] E.g., douek, 2022. Although a standardised and centralised content administration is also something I would argue to be furtively fancied by the EU (cf. EP, 2021).
[32] EP, 2021, p. 63.
[33] Fukuyama et al., 2020.
[34] Ajunwa, 2018.
[35] Goh & Soon, 2019.
[36] Logic, 2021.
[37] Fukuyama et al., 2020, p. 18.
[38] Fukuyama et al., 2020, p. 35.
[39] Bietti, 2022; Bowers & Zittrain, 2021; Goh & Soon, 2019.
[40] E.g., Life Itself Labs, n.d.; P2P Models, n.d..
[41] Zuboff, 2019.
[42] EP, 2022a, p. 86.
[43] Illegal hate speech is "the public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin" (Jourová, 2016, p. 1), and political orientation.
[44] EP, 2022b, p. 19.
[45] See Ó Fathaigh et al., 2021.
[46] Bollmann, 2022; Zilles, 2020.
[47] Masnick, 2022c; Fischer, 2022.
[48] Institute for Strategic Dialogue & Global Disinformation Index, 2021.
[49] Romero-Vicente, 2022.
[50] Stray, 2012.
[51] techDetector, 2021.
[52] Giansiracusa, 2021.
[53] Aaronson, 2021, p. 3.
[54] Aaronson, 2021, p. 17.
[55] Ovadya, 2021.
[56] E.g., Holroyd & Khatsenkova, 2022.
[57] Monnet, 2020; Paris, 2020.
[58] Bengani et al., 2022.
[59] Monnet, 2020.
[60] BBC, 2020b; Pfefferkorn, 2020.
[61] Azerikatoa Ayounga et al., 2022; Henley, 2020; Taylor, 2021.
[62] Henley, 2020; Ugland, 2019.
[63] Azerikatoa Ayounga et al., 2022.

Image credits: Ming Ong, C. (2021). *Fact-checking Modi's India* [Illustration]. Rest of the World.

considerations about the policy proposal

The proposal that you just read was only made possible because of the specific combination of the earlier chapters in this thesis. The literature review allowed to expose the lack of supporting evidence for the fatalist conception of the deepfake. The conceptual stroll guided by Barad brought to light new ways of conceptualising the reality of the deepfake thereby allowing for a particular kind of policy orientation. The deepfake assemblage allowed to visualise the extent and complexity of the sociotechnical reality of the deepfake; the multiplicity of its sociotechnical materialisations. The policy assemblage allowed to expose the elements of that complexity that are of interest to the EU; visualising on what grounds of the deepfake reality the EU roots its policy approach. And lastly, the model EU regulation provided in the Annex allowed to put together an analytically valuable document from which the policy proposal was crafted.

Engaging in this exercise of actually crafting a policy proposal—of bridging theory and application—made it all the more obvious that there is no single solution to the deepfake. It is not that there is "no" solution—that there is nothing to do about it—but it is that the solution is not singular to a reality that is tentacular. This also becomes visible if appreciating how the EU's approach was surely not dismissed altogether. The EU's regulations and recommendations are necessary in many ways that extend beyond the scope of present thesis. But, while there is no single solution to the deepfake, there are however solutions that are based on guiding premises that better align with the observed reality.

The analysis of the model EU regulation allowed to derive the EU's conception of the deepfake. A conception that was shown to abide by the fatalist conception of the deepfake. A conception that was however previously shown to be empirically invalid. The Baradian revisitation provided one such means to better re-align the conception of the deepfake with its observed reality. And which resulted in a policy proposal that has a foundationally distinct logic from that of the EU. Succinctly, where the EU thus anchors its policy approach on the fatalist conception of the deepfake that envisions it as a pristine and separately delineable content[32], I anchored my policy approach on deepfake Baradianism that envisions it as a social doing as well. A social doing that therefore makes it a tentacular phenomenon (a feature that came to empirical expression through the deepfake assemblage). This ever-reconfiguring tentacularity further implies that the deepfake reality can never find an end since the deepfake is intrinsically bound to society—it is itself social.

Departing from those conceptions, it was argued that an exclusive focus on fact-checking and technological efforts is bound to fail. Not only do present-day research findings already reveal their limited efficacy (empirical evidence). But the Baradian thinking now also shows that both efforts are

[32] To be complete, the EU conceptualises the deepfake as a pristine piece of content, devoid of context, that can be separately delineated and excised from society through a top-down regulatory surgical move to hopefully find an end.

intrinsically impossible (ontological evidence). Firstly, fact-checking cannot demarcate fact from fake. It can only demarcate content relative to the fact-checking device; a device that is itself not free of original sin, not free of human interference. Secondly, since the deepfake is more than just a technological phenomenon, it cannot be tackled through a technological approach only. It is for these reasons that I argue that the EU's current policy approach lacks both empirical foundation and ontological foundation; that it lacks both a material and a discursive anchoring in the deepfake reality. And I therefore in turn also argue that the Baradian-inspired proposal provides a means to redress this absence of legitimation.

What this means is that it is not only the policy approach that requires a shift, and neither is it only the narrative that requires a shift, but they both do, since both influence one another—both interfere. And such rethinking proves all the more crucial given that the mode of regulating deepfakes reshapes the way that political expression can be practiced by any of the actors or stakeholders in the assemblage. The same way that online platforms shape society through their content moderation practices (Cavaliere, 2021; Gillespie, 2022), so too does the EU shape society through its content regulation practices.

> "Whether it is selecting out and selecting for, through policy or through design, with whatever justification—all of it "moderates" not only what any one user is likely to see but also what society is likely to attend to, take seriously, struggle with, and value" (Gillespie, 2022, p. 8).

The fatalist tragedy narrative serves well as means to securitise the deepfake and encapsulate it under EU administration (Neo, 2021; Taylor, 2021). However, the doomsday prophecy remains on the order of the mythical more than it is part of reality. Hence, a general dismissal of the netizen's agential potential in the deepfake reality to in turn favour such paternalistic encapsulation of the deepfake under EU administration—where the EU thus positions itself as the curator of online content—is not justified. Neither empirically nor ontologically. Some argue that such positioning "is grounded in a liberal-humanist discourse that […] assumes that a 'clean' internet is one built on the veracity of recorded footage rather than the circulation of the deepfake[;] rather than the motivations, intentions and practices of users or the cultural antecedents that enabled deepfake technology" (Cover, 2022, p. 617). Adding to that, others also argue that focusing on technology to deal with deepfakes results in a strengthening of technocracy or expertocracy (Kim, 2020, p. 5), thereby again bereaving civil society of its agential integrity and participative capacity. It is as such that I thus further stretch my argument upon stating that such repeated neglect of the agential nature of the netizenry in the EU's approach intrinsically inhibits the accomplishment of public trust in public institutions. An accomplishment that is however one of the very original aims of the EU upon tackling deepfakes (EP, 2021).

While I do not subscribe to the idea of a greedy or Machiavellian government, it does not mean that a government cannot have certain ideological patterns or tendencies, however nuanced these may be. A questioning of the tragedy narrative is not about minimising the potential harm that deepfakes embody either. Such minimisation would underestimate the societal reality characteristic of the deepfake. And

while I do not subscribe to the idea of greedy or Machiavellian corporations either, such minimisation is a narrative that "technology firms [arguably embrace] in order to reject increased state regulation and oversight" (Neo, 2021, p. 221). I am convinced that governments, corporations, and civil society alike are all equally subject to behavioural quirks of all kinds. To thus frame things differently,

> "To analyse the deepfake [using a Baradian conception] is not to disavow the serious, problematic uses and misuses of a technology that has at times been weaponized as a form of disinformation and image abuse (including particularly women and minorities and those in the public sphere through pornographication and false attribution). Nor is it to proscribe a decidable position 'for or against' deepfake technology. Rather, it is to critique its conceptualization to strengthen the possible ethical responses by understanding culture's constitutive relationship with emergent technologies" (Cover, 2022, p. 610-611).

While the proposal here surely has its own shortcomings—be it because all is in intra-action, in constant interference—it thereby provides again evidence for the necessity of theoretical research approaches to the deepfake. Deepfake Baradianism will already become obsolete once the deepfake reality will have reconfigured. An obsolescence that is therefore already ongoing, be it because the thesis is but an instance of observation emanating from me as observing device. I as researcher-performer cannot transcend my practices of observation. And I as Baradian phenomenon cannot eclipse my nature as agency in constant reconfiguring.

CONCLUSION · AN ENCOUNTER BETWEEN DIFFERENT CONCEPTIONS OF THE DEEPFAKE

Time has come to greet you with a final round of applause, for this is where we will conclude the entire thesis. The thesis of which the main research question was to uncover how a Baradian conception of the deepfake re-informs the EU's policy approach to the deepfake. The thesis was therefore not interested in analysing a particular case of the deepfake. It was interested in the deepfake in its diverse occurrences and in its general influence on our society.

We started the peregrination by defining the deepfake as synthetic political disinformation that seeks to undermine trust in public institutions. It was with that understanding in mind that we explored existing literature on how deepfakes affect society and on current means to counteract deepfakes. And while the deepfake is popularly associated with doomsday prophecies where democracies would fall if deepfakes were left dangling around freely in the digital space[33], this literature review exposed the general inconclusiveness characterising current findings and it therefore exposed a lack of empirical evidence for this fatalist tragedy narrative. This absent validation of what was dubbed as *deepfake fatalism* led us

[33] The doomsday prophecy at times felt so obsessive that it pointed at an almost collectively held fetish and I suspect the click-bait economy to be at least partly responsible for that.

to observe five shortcomings about the conception of the deepfake in the fatalist narrative: (i) it is more than a technological event, (ii) it is more than tragedy, (iii) it is more than content, (iv) it is more than an Anglocentric event, and (v) it interferes with a netizenry—an online citizenry—that is more complexly and heterogeneously constituted than is currently acknowledged.

From those observations onward, we then explored literature that seeks to compensate for these shortcomings. A literature that provides a therefore more realist account of the deepfake—one that acknowledges its both technological and social character upon describing it. A conception of the deepfake that was therefore dubbed as *deepfake realism*. But literature on deepfake realism is scarce and it still comes short of (i) acknowledging the complexity of the netizenry, and (ii) questioning the dualist precept that conceives of reality as accurately describable through binary opposites, a precept that feeds the popular fact-checking ideal. It was in this two-fold shortage in understanding, as well as in the existing scholarly call for more theoretical research endeavours that today crucially lack, that I rooted my proposed theoretical contribution. A contribution that I thus argued to be necessary in order to provide a way of better re-aligning the conception with the deepfake to its empirical reality—of better re-aligning theory with observations.

That was how we headed on with a Baradian exploration of the reality of the deepfake. Karen Barad's new materialist development provided the means to ontologically found a reconceptualisation of the deepfake. Dubbed as *deepfake Baradianism*, this conception understands the deepfake still as synthetic political disinformation, but one that is in constant reconfiguring. The Baradian development might at times have read farfetched, but the resulting conceptual implications of this revisitation provided proof of the necessity for this theoretical maturation. The deepfake is no longer a given, pristine, finite, separately delineable entity. Synthetic political disinformation is now an entanglement of material-discursive agencies in constant doing. A diffracted pattern. An ontologically inextricable interference between society and technology where none has precedence. The deepfake thus has an organic existence so to speak—it is alive—since it is neither merely technological (inert) nor merely human (dependent on human intervention). A conception that was at this point sometimes emphasised by referring to it as *deepfake reality*. In essence, deepfake Baradianism implied that (i) the tragedy narrative is reductive thereby advisably questioning Manichean narratives, (ii) the deepfake has both a material and a discursive reality, both a technological and a social reality, (iii) the netizenry is not a uniform gullible mass prone to calamity but has agency, (iv) any act of knowledge-making or intervention is only ever partial, thereby making top-down approaches inherently limited in efficacy, (v) the deepfake reality can never find an end, it is not an excrescence that can be cut out, (vi) the deepfake is capable of differentially enacting its reality and is therefore speech more than it is content, and (vii) despite its popular praise, fact-checking is intrinsically inefficient as it cannot demarcate fact from fake but can only demarcate content relative to the fact-checking device that has itself a value-laden nature[34]. By providing a

[34] The popular consideration that fact-checking allows to impartially—factually—demarcate fact from fake is

theorisation of the deepfake that better aligns with its observed reality, since deepfake Baradianism redresses the shortcomings observed in both deepfake fatalism and deepfake realism, I thus argued at this point that a Baradian conception of the deepfake provides more accurate premises to inform policy recommendations—premises that are a more truthful representation of reality.

Before being actually able to maintain such argument, I could not do so without engaging myself with an empirical quest. It was here that I explored the deepfake in its present-day known ways of materialising in our world. I did so by following the sociotechnical instances through which both the deepfake grounds itself in our reality and through which we ground ourselves in the deepfake reality. The notion sociotechnical was thus not understood in its classical sense as being about an *inter*-action between society and deepfake where social order has technological precedence and where the deepfake has sociocultural precedence. It was understood in its Baradian understanding as an *intra*-action between deepfake and society where none has precedence. This empirical exploration first gave birth to the *deepfake assemblage*; a visual representation of my way of assembling the way that the deepfake sociotechnically materialises. This deepfake assemblage exposed the tentacularity of the deepfake reality. A tentacularity that indeed does not end at the material-technological dimension of the deepfake, but one that seeps into the confines of our society. A tentacularity that solidifies the deepfake and makes it therefore immune to interventionist acts seeking its swift excision. On top of which the intelligible nature of the deepfake endows it with a capacity to generate meaning and impose an identity. An absence of human intervention therefore does not mean that its tentacularity would be immobilised. But what this tentacularity entails is that while the deepfake reality will never observe external commandments to its entire annihilation, there are nonetheless multiple sites of potential regulatory interest.

This was the point where the empirical exploration gave birth to a second assemblage. The *EU assemblage*. Similarly to the deepfake assemblage, the EU assemblage was about exploring what sociotechnical elements of the deepfake's materialisation could be found in the EU's policy approach to the deepfake. A policy approach that was scattered across six documents: the Digital Markets Act, the Digital Services Act, the Artificial Intelligence Act, the Data Act, the Strengthened Code of Practice on Disinformation, and the study Tackling deepfakes in European policy. The EU assemblage was basically a superposition of the findings from those policies on top of the previously crafted deepfake assemblage. The result provided a means to visualise on what grounds of the deepfake reality the EU roots its approach.

In addition to the comparative analysis between both assemblages, a model EU regulation was put together to provide a valuable referential basis for the EU policy analysis. This policy analysis allowed the derivation of the EU's conception of the deepfake. A conception that was shown to match the fatalist

---

ontologically inaccurate. Which in turn requires a more profound acknowledgement of the diffracted reality that is characteristic of the receptiveness of society to fact-checked content. Even if enacted with utmost integrity, any fact-fake demarcation will thus intrinsically by received with great dissimilarity by its audience. Therefore, again, fact-checking will not provide the sought-after end to the deepfake.

conception of the deepfake, as it equally conceived of it as a pristine piece of content, devoid of context, that can be separately delineated and excised from society through a top-down regulatory move and thereby find an end. The EU lacked consideration for the deepfake as being more than a tragedy, more than a technological event, more than pristine content, more than a singularity, and as being constituted of a more complex netizenry than is currently acknowledged. Which were all observations made for deepfake fatalism that had been shown to lack empirical evidence. The EU did at times imply a conception of the deepfake as more than a technological event, but it remained embedded in the fatalist scenario of a postapocalyptic information anarchy where the netizenry is a gullible mass predisposed to calamitous fortunes. Since the EU's policy approach to deepfakes is based on a conception of the deepfake that does not align with observations, it was argued that the EU's approach is not only empirically unjustified but equally so ontologically unjustified. I however argue that such ontological and empirical legitimacy is key in terms of safeguarding a trusting relationship between a governmental institution and civil society, since governing deepfakes is about governing speech and is therefore embedded in the fundamental principle of freedom of expression (de Vries, 2022).

Using a Baradian lens, two key criticisms were formulated vis-à-vis the EU's policy approach. Firstly, the EU's almost exclusive attention to fact-checking technologies, and organisations using such technologies, as means to tackle deepfakes was argued inadequate given that the deepfake is neither simply a technological event nor is it simply a separately delineable occurrence prone to excision. This led to the argument that present-day efforts solely focused on fact-checking are bound to fail. Secondly, by paving the way to technocracy or expertocracy (Kim, 2020), the EU bereaves civil society of a first layer of agential integrity (of its participative capacity). And by considering not simply the deepfake as "passive and inert, requiring external (human) agency to do anything" (Kalpokas & Kalpokiene, 2022, p. 79), but very much also considering the human as passive and inert, the EU thereby bereaves civil society of a second layer of agential integrity. The EU completely disregards the netizenry's capacity for autonomous judgement. A dismissal that is further reinforced by the popular doomsday prophecy as it incapacitates civil society by being categorised as needing protection from something it would have no influence over. More than securitisation (Taylor, 2021), the EU thereby illegitimately intervenes by a form of exaggerated mothering. Illegitimately, because the EU's approach lacks both empirical and ontological foundation. I wish to repeat here that I do not subscribe to the idea of a greedy or malevolent government, for I believe it to be an as complex assemblage as the deepfake and to be composed of as many convoluting, clashing, converging, overlapping interferences as they are constitutive of any single mind. However, again, it does not mean that one cannot question some patterns in this interference, even if they are only partial representations of that assemblage. Especially that for the case of the governmental administration of deepfakes, it results in a reshaping of the way that political expression can be practiced by any of the actors or stakeholders in the assemblage.

At this moment, I thus further stretched the previous argument of the empirical and ontological absence of foundation to argue that a repeated neglect of the agential nature of the netizenry intrinsically inhibits

the accomplishment of public trust in public institutions. An accomplishment that is however one of the very original purposes of the EU upon tackling deepfakes (EP, 2021). Because why would anyone trust someone who has no faith in it in the first place? How are recommendations based on an unjustified form of paternalism where online content is to be curated supposed to secure a trusting society? It was from this questioning onward that the Baradian revisitation of the deepfake came into full force as it embodied the guiding premises of the policy proposal. Those premises that redress the shortcomings observed in both the fatalist and realist conceptions of the deepfake. The most important dimension of this policy proposal was thus to consider the netizen as an agent capable of orienting its digital destiny. A recommendation that I thus argued to be both anchored in the material and the discursive reality of the deepfake. And as it was about a policy proposal, it was also the reason for which the analysis chapter was provided in the format of such a policy proposal.

The Baradian-inspired policy proposal surely did not dismiss the EU's approach altogether, be it because the EU's approach is necessary in many ways that extend beyond the scope of present thesis. And it was also considered at the end of the Baradian-inspired policy proposal that this Baradian proposal itself inevitably falls short in multiple ways. Therefore, the same way that previous scholars called for more theoretical research on deepfakes, I thereby supported their call. Further research could focus on other means of reconceptualising the deepfake reality; other means that would provide ways of redressing the shortcomings of deepfake fatalism, deepfake realism, and deepfake Baradianism. Because in order to propose progressive ways of administering deepfakes, it is key to do so in ways that are accountable of the phenomenon's diffractive and intra-active reality. In ways that are thus also inclusive of non-Anglocentric contexts, which is a still rather missing consideration in present thesis. Notably, doing so would allow for an ontological questioning of the premise on which the conception of a 'clean' internet is based (Cover, 2022), and a questioning of the premise of an Internet singularity. Also because doing so could inform more than a deepfake administration and provide insights for the administration of online disinformation more generally.

The title of the thesis provides a nice way of more wholly appreciating this thought that the deepfake reality is ever reconfiguring, ever re-entangling in space, in time, and in matter. The conceptions of the deepfake, the materialisations of its reality, and the policy approaches to address it are in a perpetual dance of diffraction. A dance in which the EU—in its sovereign nature—has an important role to play. Because whatever conception of the deepfake it is that the EU enacts in its policies, it is this conception that crystallises, that becomes the premise on which the EU further builds, and therefore it is this conception that becomes part of our communal European history. The matter is thus not merely about what the EU knows. It is also about what the EU creates, what the EU performs, and therefore what *it* becomes. The italicised "it" has a triple meaning. Firstly, upon enacting a policy, the EU shapes the reality of the deepfake in its becoming by way of echoing a certain part of the reality of the deepfake in the policy—an echo that was here argued to reflect the fatalist version of the deepfake (the deepfake is an evidential doom). Secondly, the EU shapes its own sovereign becoming by way of echoing a certain

reality of itself in the policy—an echo that was here argued to reflect a certain paternalism (the deepfake is an evidential tragedy from which society ought to be protected). Thirdly, the EU shapes the European citizenry in its becoming by way of echoing a certain reality of the netizenry in the policy—an echo that was here argued to reflect a neglect of the complex assemblage characteristic of the netizenry (the netizenry has no agency and is evidentially prone to calamitous fortunes). It is by way of this particular three-fold enactment, this particular three-fold production of meaning, that the policy thus functions as a tool crafting the type of relation of trust that can exist between the EU and civil society. It is as such that the EU reshapes our collective European relation to trust. And hence it is as such that the deepfake reshapes our collective European relation to trust. A three-fold enactment that was here argued to inhibit a ripening of a relation of trust between the EU and civil society, while it is the very relation of trust that the EU seeks to cultivate. Therefore, contrasting to the popular conception, it is not the deepfake that intrinsically imposes a relation of distrust. It is our relation to the deepfake—the EU's relation to the deepfake for that matter—that generates a particular reality of dis/trust.

And where the introduction to this thesis opened with the common claim that a constant state of doubt and distrust could become the new norm if deepfakes were to freeride the online space, as written by Amoore,

> "[since] 'there is no unified authorial source of truth, but rather a distributed and oblique account of the impossibility of resolving truthfulness before the public'[,] then doubt becomes the default epistemological condition" (Amoore cited in Kalpokas & Kalpokiene, 2022, p. 81).

If doubt is the default, if it is inherent to life, then why not accept that reality rather than remaining suspended in the myth of a doubtless world? This thought brings me to a final consideration, for which I rewind back to the very original question that spurred my curiosity for this project. Namely, whether the deepfake truly embodies the potential for a generalised democratic bad trip. Because having uncovered with a Baradian conception of the deepfake that it is not the deepfake that generates distrust but our relation to the deepfake, it is however still too common to hear pejorative statements about individuals that fall for deepfakes or online disinformation. A belittling activity that dismisses the deeper intra-active social foundations of the deepfake. A moralism that leads to more polarisation and thereby in fact counteracts its original moralist intent to curtail the effects of deepfakes. And an activity that rehearses the need for an interventionist approach of external content curation. We all make mistakes. And that is how we learn. Denying that is denying our potential as individuals in development. In Baradian terms, it is denying our own nature as phenomena.

A N N E X · M O D E L E U R E G U L A T I O N

The model EU regulation, created through the compilation of the six policy documents that were imploded, starts on the next page. The six documents in question were: the Digital Markets Act (DMA), the Digital Services Act (DSA), the Artificial Intelligence Act (AIA), the Data Act (DA), the Strengthened Code of Practice on Disinformation (SCoPoD), and the study Tackling deepfakes in European policy.

This model EU regulation, although it is itself fictitious, only contains content as provided by the EU. Therefore, if anything in the model EU regulation remains unspecific or if anything internally clashes, it has solely to do with the EU. And, should clashing occur, that will probably have to do with the last document, *Tackling deepfakes in European policy*, as this is still about recommendations (unlike regulations). For clarity, any input emanating from this document will be preceded and succeeded by the following symbols respectively: "►", "◄". Note as well that the SCoPoD only applies to its signatories; it does not inherently apply to all providers of online platform services for instance. As for the acts, apart from the DA that targets any relevant service provider, the acts generally only apply to businesses that do not qualify as small-scale providers and users of the services under consideration. Thereto, all acts refer to the Commission Recommendation 2003/361/EC that defines micro, small and medium-sized enterprises on the basis of headcount, turnover, and balance sheet total.

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

on a European approach to deepfakes

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,
Having regard to the proposal from the European Commission,
After transmission of the draft legislative act to the national parliaments,
Acting in accordance with the ordinary legislative procedure,
Whereas,

(1) ►It is acknowledged that deepfake technology offers a lot of possibilities. At the same time, deepfakes can also cause "societal harm [in the form of news media manipulation, damage to economic stability, damage to the justice system, damage to the scientific system, erosion of trust in the politics, damage to democracy, manipulation of elections, damage to international relations, and damage to national security]" (EP, 2021, p. IV).◄

(2) "[P]olitical and issue advertising [are important] in shaping political campaigns and public debates […], particularly in forming public opinion, political and electoral debate, referenda, legislative processes and the voting behaviour of citizens" (EC, 2022b, p. 9). Because the online information ecosystem is primarily ruled by the digital advertising model it is crucial "[to ensure] political ads are run neutrally regardless of the political orientation or the issue addressed" (p. 9).

(3) "The conditions under which gatekeepers provide online advertising services to business users including both advertisers and publishers are often non-transparent and opaque. This opacity is partly linked to the practices of a few platforms, but is also due to the sheer complexity of modern day programmatic advertising. That sector is considered to have become less transparent after the introduction of new privacy legislation. This often leads to a lack of information and knowledge for advertisers and publishers about the conditions of the online advertising services they purchase[.]. [T]he costs of online advertising services are [therefore] likely to be higher [and] are likely to be reflected in the prices that end users pay" (EP, 2022a, p. 37). To secure a thriving digital market in the EU and a trusted e-commerce and online advertising market space, the Union ought to "fight fraudulent and deceptive commercial practices" (EP, 2022a, p. 28).

(4) The digital age is characterised by an increased reliance on AI systems. "[To] ensur[e] a safe, predictable and trustworthy online environment" (EP, 2022b, p. 15), and to "facilitate the development of a single market for lawful, safe and trustworthy AI applications" (EC, 2021c, p. 3), the Union ought to "[address] the opacity, complexity, bias, a certain degree of unpredictability and partially autonomous behaviour of certain AI systems, ensure their compatibility with fundamental rights and to facilitate the enforcement of legal rules" (p. 2).

(5) Ad-targeting and AI systems rely on data. "Data is a core component of the digital economy[.] The volume of data generated by humans and machines has been increasing exponentially in recent years. Most data are unused however, or its value is concentrated in the hands of relatively few large companies. Low trust, conflicting economic

incentives and technological obstacles impede the full realisation of the potential of data driven innovation" (EC, 2022c, p. 1). "In sectors characterised by the presence of micro, small and medium-sized enterprises, there is often a lack of digital capacities and skills to collect, analyse and use data" (p. 17). A regulation of data is necessary "to encourage and enable greater and fairer flow of data in all sectors, from business-to-business, business-to-government, government-to-business and government-to-government" (p. 2).

(6) "[N]ew technologies have emerged that improve the availability, efficiency, speed, reliability, capacity and security of systems for the transmission, findability and storage of data online, leading to an increasingly complex online ecosystem" (EP, 2022b, p. 26). In regard of this complexity, a regulation of the providers of services relying on those new technologies is necessary "[to ensure that] those activities are carried out in good faith and in a diligent manner[, the] condition of acting in good faith and in a diligent manner should include acting in an objective, non-discriminatory and proportionate manner, with due regard to the rights and legitimate interests of all parties involved, and providing the necessary safeguards against unjustified removal of legal content, in accordance with the objective and requirements of this Regulation. To that aim, the providers concerned should, for example, take reasonable measures to ensure that, where automated tools are used to conduct such activities, the technology is sufficiently reliable to limit to the maximum extent possible the rate of errors" (p. 24). Given the high and increasing reliance of society on very large online platforms for news consumption, these platforms need to abide by stricter obligations. Not only because these platforms ought to be liable for the presence of illegal content on their platforms, but equally so because their "considerable economic power in the digital economy" (EC, 2022c, p. 26) impedes an open market.

(7) "Almost 24% of total online trade in Europe is cross-border" (EC, 2020b, p. 5). "[This] cross-border nature of the use of data" (EC, 2022c, p. 7), combined with the "superior bargaining power [of core platform services that can lead to] to unfair practices and conditions for business users, as well as for end users of core platform services provided by gatekeepers, to the detriment of prices, quality, fair competition, choice and innovation in the market" (EP, 2022a, p. 5 & 27), makes a Union approach necessary to prevent market fragmentation and to ensure fair competition and business practices.

Have adopted this regulation,

TITLE I SCOPE AND DEFINITIONS

Article 1. Scope

(1) Deepfake systems and providers of "pre-trained models and data" (EC, 2021c, p. 32) are categorised as non-high-risk AI systems; i.e., AI systems that do not "pose significant risks to the health and safety or fundamental rights of persons" (p. 3). "[O]nly minimum transparency obligations" apply (p. 3). Providers of deepfake systems and pre-trained models and data, as well as "relevant third parties […] involved in the sale and the supply of software, software tools and components" (p. 32) are encouraged "to follow [on a voluntary basis] a code of conduct [that is] mandatory for high-risk AI systems" (p. 9). This code of conduct spans "requirements [on] data, documentation and traceability, provision of information and transparency, human oversight and robustness and accuracy" (p. 9). To favour best practices for content moderation and content dissemination, deepfake systems providers and providers of intermediary services also need to implement strategies to prevent their potential liability through "own-initiative investigations" (EP, 2022b, p. 21).

(2) The only exception to Article 1(1), is the use of AI systems to debunk deepfakes in the context of law enforcement (e.g., to verify the veracity of court

material). These AI systems are considered as high risk, since "accuracy, reliability and transparency [are] particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress" (EC, 2021c, p. 27).

(3) The regulation "[does] not apply to AI systems developed or used exclusively for military purposes" (EC, 2021c, p. 39). And stipulations on the regulation of data "shall not affect Union and national legal acts providing for the sharing, access and use of data for the purpose of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties" (EC, 2022c, p. 38).

(4) ►Audiovisual media are currently the dominant medium of the online information ecosystem. Audiovisual deepfakes are therefore of particular interest.◄

(5) Machine-generated data (such as synthetic deepfake data) is excluded from present regulation. Machine-generated data is only regulated by the Database Directive, which is the copyright protection for database authors but has no further obligations.

(6) Core platforms may request an exemption of (part of) their relevant obligations based on "grounds of public health and public security" (EP, 2022a, p. 121).

(7) Very large online platforms have to appoint at least one compliance officer to ensure their compliance to their relevant regulations.

Article 2. Definitions

(1) ►"**[D]eepfakes** are […] manipulated or synthetic audio or visual media that seem authentic, and which feature people that appear to say or do something they have never said or done, produced using artificial intelligence techniques, including machine learning and deep learning" (EP, 2021, p. I).◄

(2) A **deepfake system** is an AI system; a "software that is developed with [machine-learning] techniques […] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (EC, 2021c, p. 39).

(3) A **core platform service** includes: "(a) online intermediation services; (b) online search engines; (c) online social networking services; (d) video sharing platform services; (e) number-independent interpersonal electronic communication services; (f) operating systems; (g) web browsers; (h) virtual assistants; (i) cloud computing services; (j) online advertising services, including any advertising networks, advertising exchanges and any other advertising intermediation services, provided by an undertaking that provides any of the core platform services listed in points (a) to (i)" (EP, 2022a, p. 86).

(4) A provider of core platform services is a **gatekeeper** if "(a) it has a significant impact on the internal market, (b) it provides a core platform service which is an important gateway for business users to reach end users; and (iii) it enjoys an entrenched and durable position, in its operations, or it is foreseeable that it will enjoy such a position in the near future" (EP, 2022a, p. 92).

(5) "**Online platforms**, such as social networks or online marketplaces, [are] providers of hosting services that not only store information provided by the recipients of the service at their request, but that also disseminate that information to the public, again at their request. However, […] providers of hosting services […] where the dissemination to the public is merely a minor and purely ancillary feature that is intrinsically linked to another service, or a minor functionality of the principal service [are not considered as online platforms]" (EP, 2022b, p. 16).

(6) **Very large online platforms** are "online platforms which reach a number of average monthly active recipients of the service in the Union equal to or higher than 45 million" (EP, 2022b, p. 200). Included in this notion are very large online search engines.

(7) "[**Data**] means any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording" (EC, 2022c, p. 38). "[It] include[s] data in the form and format in which they are generated by the product, but [it excludes] data resulting from any software process that calculates derivative data from such data" (EC, 2022c, p. 20).

(8) "[**User**] means a natural or legal person that owns, rents or leases a product or receives [data processing] services" (EC, 2022c, p. 39).

(9) "[**Data recipient**] means a legal or natural person,

acting for purposes which are related to that person's trade, business, craft or profession, other than the user of a product or related service, to whom the data holder makes data available, including a third party following a request by the user to the data holder or in accordance with a legal obligation" (EC, 2022c, p 39).

(10) "[**Related service**] means a digital service, including software, which is incorporated in or inter-connected with a product in such a way that its absence would prevent the product from performing one of its functions" (EC, 2022c, p. 39).

(11) "[**Data processing service**] means a digital service other than an online content service" (EC, 2022c, p. 39); other than "an audiovisual media service […], or a service the main feature of which is the provision of access to, and the use of, works, other protected subject-matter or transmissions of broadcasting organisations, whether in a linear or an on-demand manner" (EU, 2017, p. 8).

(12) "[**Interoperability**] means the ability of two or more data spaces or communication networks, systems, products, applications or components to exchange and use data in order to perform their functions" (EC, 2022c, p. 40).

(13) **Illegal content** "refer[s] to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech or terrorist content and unlawful discriminatory content, or that the applicable rules make illegal in view of the fact that it relates to activities that are illegal. Illustrative examples include the sharing of images depicting child sexual abuse, unlawful non-consensual sharing of private images, online stalking, the sale of noncompliant or counterfeit products, the sale of products or the provision of services in infringement of consumer protection law, the non-authorised use of copyright protected material" (EP, 2022b, p. 15).

(14) "**Illegal Hate speech** is defined in EU law as the public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin" (Jourová, 2016, p. 1).

(15) "[**Content moderation**] means the activities, automated or not, undertaken by providers of intermediary services aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions" (EP, 2022b, p. 146).

(16) An **intermediary service** is any service involved in the process of information transmission, (temporary) storage, and/or its modification; i.e., "services known as 'mere conduit', 'caching' and 'hosting' services" (EP, 2022b, p. 9).

TITLE II ENSURING A UNION APPROACH

Relevance for deepfakes

Deepfakes thrive in the online information ecosystem that today is characterised by cross-border digital exchanges and dynamics. A Union approach is necessary to prevent regulatory fragmentation and to better protect internal businesses and consumers.

Article 3. Horizontal cooperation

(1) "Public sector bodies and Union institutions, agencies and bodies shall cooperate and assist one another" (EC, 2022c, p. 52).

(2) Signatories of the Strengthened Code of Practice on Disinformation (SCoPoD) are obliged to have good rapport and cooperation both among industry actors and with public institutions, such as to drive best practices in content moderation for purposes of mitigating disinformation. The Union further encourages an adhesion to international fact-checking standards. The adopted content moderation practices ought to be based on practices from fact-checking organisations and scientific research. Signatories therefore ought to maintain regular exchange with fact-checking organisations, both cross-border and cross-platform, to "share information on the tactical migration of known actors of misinformation, disinformation and information manipulation across different platforms" (EC, 2022b, p. 18).

Article 4. A Union approach and a single market

(1) Member states shall "designate at least one authority

with the task to supervise and enforce this Regulation[.] Those authorities should also act in complete independence from private and public bodies, without the obligation or possibility to seek or receive instructions, including from the government" (EP, 2022b, p. 103 & 106).

(2) To allow for the untangling of complex disinformation cases that would involve multiple countries, "an independent advisory group at Union level" (EC, 2022b, p. 39) shall be set up.

(3) "[T]he Commission should have access to any relevant documents, data and information necessary to open and conduct investigations and to monitor the compliance [of the online platforms], irrespective of who possesses the documents, data or information in question, and regardless of their form or format, their storage medium, or the precise place where they are stored" (EP, 2022b, p. 129).

(4) To ensure a single market for AI systems and a Union approach to the regulation of providers of digital services, the European Artificial Intelligence Board and European Board for Digital Services shall "provide advice and assistance to the Commission" (EC, 2021c, p. 72).

TITLE III ONLINE ADVERTISING OBLIGATIONS FOR PROVIDERS OF PLATFORM SERVICES

Relevance for deepfakes

Deepfakes circulate on and attain their target audiences via the particularities of online platforms, whose economic model is importantly shaped by online advertising models.

Article 5. Transparency in online advertising practices

(1) The following stakeholders that are signatories of the SCoPoD shall "significantly improve the scrutiny of advertisement placements, notably in order to reduce revenues of the purveyors of Disinformation[:] advertisers and agencies who are involved in the purchasing of advertising space; publishers and platforms who are involved in the selling of advertising space and approval of advertising campaigns; advertising technology companies who are involved in the targeting or selection of advertising space and/or content and verification reporting; auditing bodies who are involved [in] the accreditation of services ranging from targeting to reporting" (EC, 2022b, p. 4). "[They] will develop, deploy, and enforce appropriate and tailored advertising policies that address the misuse of their advertising systems for propagating harmful Disinformation in advertising messages and in the promotion of content" (p. 7).

(2) Users shall be informed by the online platforms about the targeting mechanisms underlying displayed ads by "providing meaningful explanations of the logic used to that end, including when this is based on profiling" (EP, 2022b, p. 62).

(3) "Very large online platforms [shall] ensure public access to repositories of advertisements presented on their online interfaces [through application programming interfaces] to facilitate supervision and research" (EP, 2022b, p. 89), "until one year after the advertisement was presented for the last time" (p. 219).

Article 6. Transparency in online advertising practices for purposes of political campaigning

(1) "[T]o ensure additional protection of personal data when it is used in the context of targeting political advertising, in full compliance with the [General Data Protection Regulation] and other relevant laws, in particular with regard to acquiring valid consent where required" (EC, 2022b, p. 10), signatories of the SCoPoD will ensure political or issue ads are labelled adequately and that the labelling techniques are published in a user-friendly and transparent way. The information shall include technical details about the labelling and the recommender systems. Signatories shall provide publicly accessible political ad repositories, updated live if possible. These shall include "relevant information for each ad such as the identification of the sponsor; the dates the ad ran for; the total amount spent on the ad; the number of impressions delivered; the audience criteria used to determine recipients; the demographics and number of recipients who saw the ad; and the geographical areas the ad was seen in" (p. 13).

(2) Access to these repositories, via application programming interfaces and other interfaces, also serves for research. Therefore, these interfaces are built by the service providers but with the aid of researchers to be tailored to the needs of the latter.

Article 7. Open and fair online advertising

(1) "[T]o ensure contestability and fairness [of online advertising practices]" (EP, 2022a, p. 7), "[t]ransparency obligations [shall] require gatekeepers to provide advertisers and publishers to whom they supply online advertising services, when requested, with free of charge information that allows both sides to understand the price paid for each of the different online advertising services provided as part of the relevant advertising value chain" (p. 37). This information should include "the method with which each of the prices and remunerations are calculated" (p. 37). Gatekeepers shall also "provide advertisers and publishers, when requested, with free of charge access to the gatekeepers' performance measuring tools and the data, including aggregated and non-aggregated data, necessary for advertisers, authorised third parties such as advertising agencies acting on behalf of a company placing advertising, as well as for publishers to carry out their own independent verification of the provision" (p. 47).

(2) "Gatekeepers [shall] be required to provide access, on fair, reasonable and non-discriminatory terms, to […] ranking, query, click and view data in relation to free and paid search generated by consumers on online search engine services to other undertakings providing such services, so that those third-party undertakings can optimise their services and contest the relevant core platform services. Such access should also be given to third parties contracted by a search engine provider, who are acting as processors of this data for that search engine. When providing access […], a gatekeeper should ensure the protection of the personal data of end users, including against possible re-identification risks, by appropriate means, such as anonymisation of such personal data, without substantially degrading the quality or usefulness of the data." (EP, 2022a, p. 50).

(3) "Ensuring an adequate level of transparency of profiling practices employed by gatekeepers, including, but not limited to, profiling, facilitates contestability of core platform services. […] Enhanced transparency should allow other undertakings providing core platform services to differentiate themselves better through the use of superior privacy guarantees. To ensure a minimum level of effectiveness of this transparency obligation, gatekeepers should at least provide an independently audited description of the basis upon which profiling is performed, including whether personal data and data derived from user activity in line with [the General Data Protection Regulation] is relied on, the processing applied, the purpose for which the profile is prepared and eventually used, the duration of the profiling, the impact of such profiling on the gatekeeper's services, and the steps taken to effectively enable end users to be aware of the relevant use of such profiling, as well as steps to seek their consent or provide them with the possibility of denying or withdrawing consent" (EP, 2022a, p. 63).

(4) The Commission shall be informed of any intended merger and acquisition of a gatekeeper with another "[provider of] core platform services, or any other services in the digital sector or enable the collection of data" (EP, 2022a, p. 129).

Article 8. Ensuring an unhampered access to and switching between Internet services

(1) "[G]atekeepers [shall] not unduly restrict end users in choosing the undertaking providing their internet access service" (EP, 2022a, p. 44).

(2) Gatekeepers shall not hamper an end user that wishes to delete the software or any applications.

TITLE IV OBLIGATIONS FOR ARTIFICIAL INTELLIGENCE SYSTEMS AND PRACTICES

Relevance for deepfakes

Deepfakes are synthetic media generated through machine-learning. Their existence relies on AI systems.

Article 9. Transparency of AI systems

(1) "Transparency obligations […] apply for systems that […] generate or manipulate content ('deep fakes')" (EC, 2021c, p. 14). "Providers shall ensure that AI systems […] are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use" (p. 69). "Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated[, unless] where […] it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties" (69). Providers and users are exempted of this obligation wherever the AI system is "authorised by law to detect, prevent, investigate and prosecute criminal offences" (p. 69).

(2) AI systems are allowed to use personal data as stipulated per the General Data Protection Regulation. Meaning that in most cases the use is prohibited unless explicitly consented to by the end user.

(3) Users of high-risk AI systems "should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate" (EC, 2021c, p. 30).

(4) Member States shall implement "AI regulatory sandboxes and other measures to reduce the regulatory burden and to support Small and Medium-Sized Enterprises […] and start-ups" (EC, 2021c, p. 3). "Member States shall [also] organise specific awareness raising activities about the application of [requirements for high-risk AI systems] tailored to the needs of the small-scale providers and users" (p. 71).

Article 10. AI systems training models risk assessments and traceability, and consumer trust

(1) For high-risk AI systems, "[t]raining, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system. They should also have the appropriate statistical properties, including as regards the persons or groups of persons on which the high-risk AI system is intended to be used [as well as] the features, characteristics or elements that are particular to the specific geographical, behavioural or functional setting or context within which the AI system is intended to be used" (EC, 2021c, p. 29).

(2) "[In the public interest,] to provide trustful, accountable and non-discriminatory access to high quality data for the training, validation and testing of AI systems [and to facilitate] data sharing between businesses and with government[, the Commission establishes] European common data spaces" (EC, 2021c, p. 29).

(3) "In order to ensure a high level of trustworthiness of high-risk AI systems, those systems should be subject to a conformity assessment prior to their placing on the market or putting into service" (EC, 2021c, p. 32).

(4) Providers of high-risk AI systems shall "ensure the bias monitoring, detection and correction" (EC, 2021c, p. 29).

(5) High-risk AI systems are obliged to have a human oversight.

(6) "AI regulatory sandboxes […] shall provide a controlled environment that facilitates the development, testing and validation of innovative AI systems [before their implementation]" (EC, 2021c, p. 69).

(7) Providers of high-risk AI systems are required to comply to a reporting obligation on the AI technical limitations and capabilities as regards "the system, algorithms, data, training, testing and validation processes used" (EC, 2021c, p. 30). Providers of high-risk AI systems shall have "a post-market monitoring system […] to ensure that the possible risks emerging from AI systems which continue to 'learn' after being placed on the market or put into

service" (p. 36). "A risk management system shall be established, implemented, documented and maintained" (p. 46).

(8) Very large platforms ought to audit their algorithms through risk assessment programmes and report to the appointed member state authority at least once a year. Very large platforms shall do additional such risk assessments in relation to: "(a) the dissemination of illegal content through their services; (b) any actual or foreseeable negative effects for the exercise of fundamental rights, in particular the fundamental rights to human dignity, respect for private and family life, the protection of personal data, freedom of expression and information, including the freedom and pluralism of the media, the prohibition of discrimination, the rights of the child and consumer protection; (c) any actual or foreseeable negative effects on civic discourse and electoral processes, and public security; [(d)] any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health, minors and serious negative consequences to the person's physical and mental well-being" (EP, 2022b, p. 203-204). The very large platforms are responsible for the setting up of mitigation means for these risks.

(9) Providers of high-risk AI systems shall "ensure a level of cybersecurity [and prevent any vulnerabilities from being exploited by cyberattacks such as data poisoning and adversarial attacks against training models]" (EC, 2021c, p. 30). They "shall report any serious incident or any malfunctioning of those systems which constitutes a breach of obligations under Union law intended to protect fundamental rights to the market surveillance authorities of the Member States" (p. 75).

(10) "High-risk AI systems shall be designed and developed with [recognised standard] capabilities enabling the automatic recording of events ('logs') while the high-risk AI systems [are] operating. […] The logging capabilities shall ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system" (EC, 2021c, p. 50).

TITLE V OBLIGATIONS FOR PROVIDERS OF DATA PROCESSING SERVICES

Relevance for deepfakes

Deepfakes and data are connected in two important ways. Firstly, deepfakes rely on data for their creation. Secondly, deepfakes rely on data for their targeted dissemination.

Article 11. Data access to users

(1) "Products shall be designed and manufactured, and related services shall be provided, in such a manner that data generated by their use are, by default, easily, securely and, where relevant and appropriate, directly accessible to the user" (EC, 2022c, p. 40). The user shall be provided with "the following information […] in a clear and comprehensible format: (a) [with] the data likely to be generated by the use of the product or related service; […] (c) how the user may access those data; (d) whether [a third party will be allowed to use the data and why]; (e) whether the seller, renter or lessor is the data holder and, if not, the identity of the data holder [and how the user may access the data]" (p. 40-41). "It should be as easy for the user to refuse or discontinue access by the third party to the data as it is for the user to authorise access" (p. 25).

(2) "the *sui generis* right [from the Database Directive that ensures a copyright to database authors] does not apply to databases containing data obtained from or generated by the use of a product or a related service" (EC, 2022c, p. 61).

(3) The user shall be granted access to its data upon "simple request through electronic means where technically feasible" (EC, 2022c, p. 41). A user may not use it for purposes of market competition, but it may provide the data "to a third party offering an aftermarket service that may be in competition with a service provided by the data holder" (p. 23). Only the user is entitled to provide data "generated by the use of a product or related service" (p. 24) to a third party. Gatekeepers are not "an eligible third party" (p. 42).

(4) "The data holder may apply appropriate technical protection measures […] to prevent unauthorised

access to the data[.] Such technical protection measures shall not be used as a means to hinder the user's right to effectively provide data to third parties" (EC, 2022c, p. 46).

(5) "A third party shall process the data made available to it […] only for the purposes and under the conditions agreed with the user, and subject to the rights of the data subject insofar as personal data are concerned, and shall delete the data when they are no longer necessary" (EC, 2022c, p. 43).

(6) "The third party shall not (b) use the data it receives for the profiling of natural persons [(i.e., for any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict […] that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements: EU, 2016, p. 33)], unless it is necessary to provide the service requested by the user; (c) make the data available it receives to another third party, in raw, aggregated or derived form, unless this is necessary to provide the service requested by the user; (d) make the data available it receives to an undertaking providing core platform services for which one or more of such services have been designated as a gatekeeper" (EC, 2022c, p. 43).

Article 12. Data access to data recipients to foster a fair market behaviour

(1) A data recipient is what was identified as third party pursuant Article 9. It follows that "an obligation to make data available to a data recipient shall not oblige the disclosure of trade secrets" (EC, 2022c, p. 44).

(2) "Any compensation agreed between a data holder and a data recipient for making data available shall be reasonable" (EC, 2022c, p. 44).

(3) "A contractual term is unfair if it is of such a nature that its use grossly deviates from good commercial practice in data access and use, contrary to good faith and fair dealing" (EC, 2022c, p. 47). "[It] is unfair […] if its object or effect is to: (a) exclude or limit the liability of the party that unilaterally imposed the term for intentional acts or gross negligence; (b) exclude the remedies available to the party upon whom the term has been unilaterally imposed in case of non-performance of contractual obligations or the liability of the party that unilaterally imposed the term in case of breach of those obligations; (c) give the party that unilaterally imposed the term the exclusive right to determine whether the data supplied are in conformity with the contract or to interpret any term of the contract" (p. 47).

Article 13. Improved interoperability for data reuse

(1) "In line with the [General Data Protection Regulation] and the Free Flow of Non-Personal Data Regulation that enable consumers and businesses to process personal and non-personal data anywhere they want in the Union, the cross-border processing of data within the Union is essential for conducting business in the internal market" (EC, 2022c, p. 8). "High quality and interoperable data from different domains increase competitiveness and innovation and ensure sustainable economic growth" (p. 17). "Barriers to data sharing prevent an optimal allocation of data to the benefit of society" (p. 17). Therefore, "portability of digital assets between different data processing services that cover the same service type [shall be enhanced by open] interoperability specifications and European standards for the interoperability of data processing services" (p. 56).

(2) "[S]witching between cloud and edge services" (EC, 2022c, p. 3), "while maintaining a minimum functionality of service" (p. 32), shall be facilitated. This includes data processing services of infrastructure as a service, software as a service, and platform as a service. Therefore, "[o]perators of data spaces shall comply with, the following essential requirements to facilitate interoperability of data, data sharing mechanisms and services: (a) the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty shall be sufficiently described to allow the recipient to find, access and use the data; (b) the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists shall be described in a publicly available and consistent manner; (c) the technical means to access the data, such as

application programming interfaces, and their terms of use and quality of service shall be sufficiently described to enable automatic access and transmission of data between parties, including continuously or in real-time in a machine-readable format; (d) the means to enable the interoperability of smart contracts within their services and activities shall be provided" (p. 55-56).

Article 14. Data access to public sector bodies

(1) "[D]ata requests made by public sector bodies and by Union institutions, agencies or bodies [in the case of public emergencies] should be made public […] by the entity requesting the data" (EC, 2022c, p. 30). These public sector bodies shall only use the data for the purposes specified, in compliance with "the rights and freedoms of data subjects[, and shall] destroy the data as soon as they are no longer necessary" (p. 51).

(2) "Data made available to respond to a public emergency […] shall be provided free of charge" (EC, 2022c, p. 51).

Article 15. Data access to third countries

(1) "Any decision or judgment of a court or tribunal and any decision of an administrative authority of a third country requiring a provider of data processing services to transfer from or give access to non-personal data within the scope of this Regulation held in the Union may only be recognised or enforceable in any manner if based on an international agreement" (EC, 2022c, p. 54). "In the absence of international agreements[,] transfer or access should only be allowed if [the reasons for the request have been verified and validated]" (p. 34).

Article 16. Dispute settlement

(1) Member states shall certify or set up a certified dispute settlement body (or multiple bodies) that is "impartial and independent[, and] has the necessary expertise in relation to the determination of fair, reasonable and non-discriminatory terms for and the transparent manner of making data available" (EC, 2022c, p. 45).

TITLE VI CONTENT MODERATION OBLIGATIONS FOR PLATFORMS

Relevance for deepfakes

Online content moderation practices are in direct relation with deepfakes by way of their dissemination across the online information ecosystem.

Article 17. Providers of core platform services and data cross-combination

(1) "The gatekeeper shall not: (a) process, for the purpose of providing online advertising services, personal data of end users using services of third-parties that make use of core platform services of the gatekeeper; (b) combine personal data from the relevant core platform service with personal data from other core platform services or from any other services provided by the gatekeeper or with personal data from third-party services; (c) cross-use personal data from the relevant core platform service in other services provided separately by the gatekeeper, including other core platform services, and vice-versa; and (d) sign in end users to other services of the gatekeeper in order to combine personal data" (EP, 2022a, p. 100), unless specifically consented to by the end user.

(2) Gatekeepers "should enable end users to freely choose to opt-in to such data processing and sign-in practices by offering a less personalised but equivalent alternative, and without making the use of the core platform service or certain functionalities thereof conditional upon the end user's consent" (EP, 2022a, p. 29). "The less personalised alternative should not be different or of degraded quality compared to the service provided to the end users who provide consent, unless a degradation of quality is a direct consequence of the gatekeeper not being able to process such personal data or signing in end users to a service. Not giving consent should not be more difficult than giving consent. When the gatekeeper requests consent, it should proactively present a user-friendly solution to the end user to provide, modify or withdraw consent in an explicit, clear and straightforward manner. In particular, consent should be given by a clear affirmative action

or statement establishing a freely given, specific, informed and unambiguous indication of agreement by the end user" (p. 30).

Article 18. Very large platforms and a user-based design of online recommender systems

(1) "Providers of online platforms that use recommender systems shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters. […] They shall include, at least: (a) the criteria which are most significant in determining the information suggested to the recipient of the service; (b) the reasons for the relative importance of those parameters" (EP, 2022b, p. 193).

(2) "[P]roviders of online platforms shall also make directly and easily accessible from the specific section of the online platform's online interface where the information is being prioritised a functionality allowing the recipient of the service to select and to modify at any time their preferred option" (EP, 2022b, p. 193).

Article 19. Enabling and mediating the participation of the user to debunk illegal content

(1) "Providers of hosting services shall put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content" (EP, 2022b, p. 166).

(2) "Signatories [of the SCoPoD] will develop technology solutions to help users check authenticity or identify the provenance or source of digital content, such as new tools or protocols or new open technical standards for content provenance" (EC, 2022b, p. 21). Signatories will allow users to report cases of disinformation by creating reporting features for the users to "flag harmful false and/or misleading information that violates Signatories' policies or terms of service [and signatories will] ensure that this functionality is duly protected from human or machine-based abuse (e.g., the tactic of 'mass-flagging' to silence other voices)" (p. 25).

(3) "Providers of online platforms shall ensure that their internal complaint-handling systems are easy to access, user-friendly and enable and facilitate the submission of sufficiently precise and adequately substantiated complaints" (EP, 2022b, p. 175).

Article 20. Platform accountability for the moderation of illegal content

(1) Online platforms are required to have "easy to access, user-friendly" notice and action "mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content" (EP, 2022b, p. 166).

(2) Very large platforms are required to have an internal compliance department to ensure their compliance to present regulation.

(3) Very large platforms are encouraged to develop and internally defined code of conduct for their users in relation to allowed content.

(4) The voluntary development and implementation of industry-led technological innovations "[to detect and combat] intentional online disinformation spread" (EC, 2021a, p. 145) and monitor content virality, including recommender systems, shall be promoted. Indeed, "[an] effective and consistent application of the obligations in this Regulation [...] may require implementation through technological means, [and it is therefore] important to promote voluntary standards covering certain technical procedures, where the industry can help develop standardised means to support providers of intermediary services in complying with this Regulation" (EP, 2022b, p. 97). "[S]takeholders tackling this issue in the Member States" shall be supported and connected (EC, 2021a, p. 145).

(5) Very large platforms are encouraged to have crisis protocols. "[A] crisis should be considered to occur when extraordinary circumstances occur that can lead to a serious threat to public security or public health in the Union or significant parts thereof. Such crises could result from armed conflicts or acts of terrorism, including emerging conflicts or acts of terrorism, natural disasters such as earthquakes and hurricanes, as well as from pandemics and other serious cross-border threats to public health. The Commission

should be able to require […] service providers to initiate a [voluntary] crisis response as a matter of urgency. Measures that the service provider may identify and consider applying may include, for example, adapting content moderation processes and increasing the resources dedicated to content moderation, adapting terms and conditions, relevant algorithmic systems and advertising systems, further intensifying cooperation with trusted flaggers, taking awareness-raising measures and promoting trusted information and adapting the design of their online interfaces. The necessary requirements should be provided for to ensure that such measures are taken within a very short time frame and that the crisis response mechanism is only used where, and to the extent that, this is strictly necessary and any measures taken under this mechanism are effective and proportionate, taking due account of the rights and legitimate interests of all parties concerned. The use of the mechanism should be without prejudice to the other provisions of this Regulation, such as those on risk assessments and mitigation measures and the enforcement thereof and those on crisis protocols" (EP, 2022b, p. 85).

(6) Providers of intermediary services have "[n]o general obligation to monitor the information which [they] transmit or store, nor actively to seek facts or circumstances indicating illegal activity" (EP, 2022b, p. 151). However, they shall "establish a single point of contact for recipients of services, which allows rapid, direct and efficient communication in particular by easily accessible means [and] which do not solely rely on automated tools" (p. 39).

(7) The adhesion of service providers to global standards on content moderation (fact-checking practices) is encouraged. ►Stronger standards of online content moderation practices should be developed to "agree on ethical and professional codes of conduct, or norms and behaviour" (EP, 2021, p. 23).◄

(8) ►"[T]he decision-making authority of platforms to decide unilaterally on the legality and harmfulness of content [shall be limited]" (EP, 2021, p. 63).◄

Article 21. Transparency of content moderation practices and reporting obligations

(1) Very large online platforms are to be transparent towards the end users of prevention and detection systems in their algorithmic proceedings in content moderation practices.

(2) "Providers of intermediary services shall include information on any restrictions that they impose in relation to the use of their service in respect of information provided by the recipients of the service, in their terms and conditions. That information shall include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making, and human review. as well as rules of procedure of their internal complaint handling system. It shall be set out in clear, plain, intelligible, user friendly and unambiguous language and shall be publicly available in an easily accessible and machine-readable format" (EP, 2022b, p. 161).

(3) "Where a provider of hosting services becomes aware of any information giving rise to a suspicion that a criminal offence involving a threat to the life or safety of a person or persons has taken place, is taking place or is likely to take place, it shall promptly inform the law enforcement or judicial authorities of the Member State or Member States concerned [or Europol] of its suspicion and provide all relevant information available" (EP, 2022b, p. 172).

(4) All online platforms shall report annually on content moderation to the member state where it is established.

Article 22. Transparency of fact-checking practices

(1) "[F]act-checking organisations need to be verifiably independent from partisan institutions and transparent in their finances, organisation and methodology; as well as consistently and continuously dedicated to fact-checking either as verified signatories of the International Fact-checking Network Code of Principles (IFCN), members of the European Digital Media Observatory (EDMO)'s network of fact-checkers, or of the future Code of Professional Integrity for Independent European fact-checking organisations" (EC, 2022b, p. 31).

(2) Signatories of the SCoPoD are responsible for the labelling and rating of the fact-checking organisation

they cooperate with. These collaborations shall be publicly available and shall contain a justification for the choice of cooperation and the involved financial contribution to those fact-checking organisations.

(3) Online platforms ought to implement a system of certified trusted flaggers with entities external to the company that cannot be individuals, and "that have demonstrated […] that they have particular expertise and competence in tackling illegal content, and that they work in a diligent, accurate and objective manner. To avoid diminishing the added value of such mechanism, the overall number of trusted flaggers awarded in accordance with this Regulation should be limited. […] Such entities can be public in nature, such as, for terrorist content, internet referral units of national law enforcement authorities or of the European Union Agency for Law Enforcement Cooperation ('Europol') or they can be non-governmental organisations and private or semi-public bodies" (EP, 2022b, p. 54-55).

(4) Signatories of the SCoPoD shall forbid users that violate their policies and terms of service in regard to the "harmful false and/or misleading information" (EC, 2022b, p. 25). Signatories shall report these violations and "will provide, through meaningful metrics capable of catering for the performance of their products, policies, processes (including recommender systems), or other systemic approaches as relevant to [the mitigation of the viral spread of harmful Disinformation] an estimation of the effectiveness of such measures, such as the reduction of the prevalence, views, or impressions of Disinformation and/or the increase in visibility of authoritative information. Insofar as possible, [signatories] will highlight the causal effects of those measures" (p. 20).

(5) "Signatories [of the SCoPoD] providing trustworthiness indicators will ensure that information sources are being reviewed in a transparent, apolitical, unbiased, and independent manner, applying fully disclosed criteria equally to all sources and allowing independent audits by independent regulatory authorities or other competent bodies" (EC, 2022b, p. 23).

Article 23. Access to platform data for research purposes

(1) Signatories of the SCoPoD are required "to cooperate with [an] independent third-party body", also through funding, "to enable sharing of personal data necessary to undertake research on Disinformation" (EC, 2022b, p. 28). "[W]ithout waiting for the independent third-party body to be fully set up[, signatories] commit to engage in pilot programs towards sharing data with vetted researchers for the purpose of investigating Disinformation" (p. 29). Signatories shall provide open access to "non-personal and anonymised" platform data for research (p. 27). Research entities allowed to access this data include "civil society organisations whose primary goal is to conduct scientific research on a not-for-profit basis [and which shall comply] with relevant sector-related ethical and methodological best practices (as laid down, for example, in the EDMO proposal for a Code of Conduct on Access to Platform Data)" (p. 26).

Article 24. High-quality journalism and diversity

(1) ►"Continue to invest in a pluralistic media landscape and high-quality journalism" (EP, 2021, p. 66).◄

Article 25. Online media literacy

(1) Member states shall organise awareness campaigns and literacy programmes for both experts and the general population. Media literacy shall be supported to enhance critical thinking and ensure a critical engagement with online content and online media.

(2) ►The EU shall boost efforts in "knowledge and tech transfer to developing countries" (EP, 2021, p. 62).◄

R E F E R E N C E S

Aaronson, S. A. (2021). *Could trade agreements help address the wicked problem of cross-border disinformation?*. CIGI Papers, 255. https://www.cigionline.org/publications/could-trade-agreements-help-address-the-wicked-problem-of-cross-border-disinformation/

AFP. (2022, February 14). *Deepfake democracy: South Korean candidate goes virtual for votes*. France24. https://www.france24.com/en/live-news/20220214-deepfake-democracy-south-korean-candidate-goes-virtual-for-votes

Ahmed, S. (2006). Chapter 2: sexual orientation. In *Queer phenomenology: orientations, objects, others* (pp. 65-108). Duke University Press.

Ahmed, S. (2022). Disinformation sharing thrives with fear of missing out among low cognitive news users: a cross-national examination of intentional sharing of deep fakes. *Journal of Broadcasting & Electronic Media*, *66*(1), 89-109. https://doi.org/10.1080/08838151.2022.2034826

Ajder, H. (2020, July 9). Tracer newsletter #58 (09/07/20)-Deepfake threat intelligence: a statistics snapshot from June 2020. *Medium*. https://medium.com/sensity/tracer-newsletter-58-09-07-20-deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020-1af7052e724a

Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2019). *The state of deepfakes: landscape, threats, and impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Ajunwa, I. (2018, April 16). The rise of platform authoritarianism. *Medium*. https://medium.com/aclu/the-rise-of-platform-authoritarianism-d838729e8ad6

Al-Saqaf, W. (2019). A blockchain-based fact-checking registry: enhancing trust in the factcheckers. *Conference for Truth and Trust Online 2019*. https://doi.org/10.36370/tto.2019.6

Arqoub, O. A., Elega, A. A., Özad, B. E., Dwikat, H., & Oloyede, F. A. (2020). Mapping the scholarship of fake news research: a systematic review. *Journalism Practice*, *16*(1), 56-86. https://doi.org/10.1080/17512786.2020.1805791

Ayad, M., Harrasy, A., & Abdullah A, M. (2022, June 14). *Under-moderated, unhinged and ubiquitous: Al-Shabaab and the Islamic State networks on Facebook*. Institute for Strategic Dialogue. https://www.isdglobal.org/isd-publications/under-moderated-unhinged-and-ubiquitous-al-shabaab-and-the-islamic-state-networks-on-facebook/

Azerikatoa Ayounga, D., Naazi-Ale Baadaa, F., & Bugre, C. (2022). Curbing fake news: a qualitative study of the readiness of academic librarians in Ghana. *International Information & Library Review*, 1-14. https://doi.org/10.1080/10572317.2022.2046438

Badiei, F., & Fidler, B. (2021). The would-be technocracy: evaluating efforts to direct and control social change with Internet protocol design. *Journal of Information Policy*, *11*, 376-402.

https://doi.org/10.5325/jinfopoli.11.2021.0376

Barad, K. (2007). *Meeting the universe halfway: quantum physics and the entanglement of matter and meaning*. Duke University Press.

Bateman, J. (2020). Deepfakes and synthetic media in the financial system: assessing threat scenarios. Introduction & scenarios targeting individuals. *Carnegie Endowment for International Peace*, pp. 3-5 & 9-14. https://www.jstor.org/stable/resrep25783.7

Bayer, J., Holznagel, B., Lubianiec, K., Pintea, A., Schmitt, J. B., 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*. European Parliament. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/04-13/EXPO_STU2021653633_EN.pdf

BBC. (2020a, November 30). *Australia demands China apologise for posting 'repugnant' fake image*. BBC. https://www.bbc.com/news/world-australia-55126569

BBC. (2020b, December 23). *Deepfake queen to deliver Channel 4 Christmas message*. BBC. https://www.bbc.com/news/technology-55424730

Bengani, P., Stray, J., & Thorburn, L. (2022, August 8). *A menu of recommender transparency options*. Tech Policy Press. https://techpolicy.press/a-menu-of-recommender-transparency-options/

Berkman Klein Center for Internet & Society. (2022, July 21). *Toward best practices around online content removal requests*. Harvard University. https://cyber.harvard.edu/story/2022-07/toward-best-practices-around-online-content-removal-requests

Bietti, E. (2022, January 28). *How the free software and the IP wars of the 1990s and 2000s presaged today's toxic, concentrated Internet*. Promarket. https://www.promarket.org/2022/01/28/digital-platforms-regulation-free-software-ip-wars-concentration-internet/

Bollmann, H.-S. (2022, June 30). *The EU's regulatory awakening? Hack-and-leak operations in the new EU code on disinformation*. Humboldt Institute for Internet and Digitalisation. https://www.hiig.de/en/hack-and-leak-eu/

Bond, S. (2022, March 27). *That smiling LinkedIn profile face might be a computer-generated fake*. npr. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?t=1649008706058&t=1649146137414

Bowers, J., & Zittrain, J. (2021, June 21). *Internet entropy*. Lawfare. https://www.lawfareblog.com/internet-entropy

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*(January), 499-515. https://doi.org/10.1146/annurev-psych-010419-050807

Breakstone, J, Smith, M., Wineburg, S., Rapaport, A., Carleton, J., Garland, M., & Saavedra, A. (2021). Students' civic online reasoning: a national portrait. *Center for Economic and Social Research*. https://dx.doi.org/10.2139/ssrn.3816075

von der Burchard, H. (2018, May 21). *Belgian socialist party circulates 'deep fake' Donald Trump*

video. Politico. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

Byman, D. & Joshi, A. (2020). Preventing the next social-media genocide. *Survival: Global Politics and Strategy*, *62*(6), 125-152. https://doi.org/10.1080/00396338.2020.1851097

Callon, M. (1984). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc bay. *The Sociological Review*, *32*(1_suppl), 196-233. https://doi.org/10.1111%2Fj.1467-954X.1984.tb00113.x

Castells, M. (2004). Informationalism, networks, and the network society: at theoretical blueprint. In *The network society: a cross-cultural perspective* (pp. 3-45). Edward Elgar. https://annenberg.usc.edu/sites/default/files/2015/04/28/Informationalism%2C%20Networks%20and%20the%20Network%20Society.pdf

Cavaliere, P. (2021). From journalistic ethics to fact-checking practices: defining the standards of content governance in the fight against disinformation. *Journal of Media Law*, *12*(2), 133-165. https://doi.org/10.1080/17577632.2020.1869486

Center for Information, Technology, and Public Life. (2022). *Call for papers: what comes after disinformation studies?*. CITAP. https://citap.unc.edu/ica-preconference-2022/

Chesney, B., & Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753-1819. https://doi.org/10.15779/Z38RV0D15J

Chung, Y.-L., & Vidgen, B. (2022). *Counterspeech: a better way of tackling online hate?*. The Alan Turing Institute. https://www.turing.ac.uk/blog/counterspeech-better-way-tackling-online-hate?_cldee=z0ghsz7UZICUWUlQNDByoB3Wi1qFWf1DekFFarMS1fr7OT_67nKBOgBtIwARs16S&recipientid=contact-ec666d8cb022ec11b6e60022484234ee-c7b82cf428444b66a3f86b482ac86749&esid=ee322393-c50c-ed11-b83e-000d3ad587ce

Clyde, A. (2022, February 17). *Algorithmic systems designed to reduce polarization could hurt democracy, not help it*. Tech Policy Press. https://techpolicy.press/algorithmic-systems-designed-to-reduce-polarization-could-hurt-democracy-not-help-it/

Cover, R. (2022). Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. *Continuum: Journal of Media & Cultural Studies*, *36*(4), 609-621. https://doi.org/10.1080/10304312.2022.2084039

Crain, M. (2022, August 3). *How capitalism—not a few bad actors—destroyed the Internet*. Boston Review. https://bostonreview.net/articles/how-capitalism-not-a-few-bad-actors-destroyed-the-internet/?mc_cid=77d48df6f6&mc_eid=039d2594ab

De Blasio, E., & Selva, D. (2021). Who is responsible for disinformation? European approaches to social platforms' accountability in the post-truth era. *American Behavioral Scientist*, *65*(6), 825-846. https://doi.org/10.1177%2F0002764221989784

Diakopoulos, N., & Johnson, D. (2020). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, (June).

https://doi.org/10.1177%2F1461444820925811

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. *The International Journal of Press/Politics*, *26*(1), 69-91. https://doi.org/10.1177%2F1940161220944364

Donovan, J. (2021, February 3). Just infrastructures speaker series at UIUC [Live online videoconference].

douek, e. (2022). Content moderation as administration [forthcoming]. *Harvard Law Review*, *136*. https://dx.doi.org/10.2139/ssrn.4005326

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729-745. https://doi.org/10.1080/1369118X.2018.1428656

Dumit, J. (2014). Writing the implosion: teaching the world one thing at a time. *Cultural Anthropology*, *29*(2), 344-362. https://doi.org/10.14506/ca29.2.09

Etlinger, S. (2019, October 28). *What's so difficult about social media platform governance?*. Centre for International Governance Innovation. https://www.cigionline.org/articles/whats-so-difficult-about-social-media-platform-governance/

European Audiovisual Observatory. (2020). *Artificial intelligence in the audiovisual sector*. Council of Europe. https://search.coe.int/observatory/Pages/result_details.aspx?ObjectId=0900001680a11e0b

European Commission. (n.d.). *6 Commission priorities for 2019-24*. European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024_en

European Commission. (2020a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European democracy action plan*. European Commission. https://ec.europa.eu/info/sites/default/files/edap_communication.pdf

European Commission. (2020b). *Proposal for a regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)*. European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en#documents

European Commission. (2021a). *Annex to the Commission Implementing Decision on the financing of the Digital Europe Programme and the adoption of the multiannual work programme for 2021-2022*. European Commission. https://ec.europa.eu/newsroom/repository/document/2021-46/C_2021_7914_1_EN_annexe_acte_autonome_cp_part1_v3_x3qnsqH6g4B4JabSGBy9UatCRc8_81099.pdf

European Commission. (2021b, September 15). *Proposal for a decision of the European Parliament and of the Council establishing the 2030 Policy Programme "Path to the Digital Decade"*. European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-decision-establishing-2030-policy-programme-path-digital-decade

European Commission. (2021c). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Commission. (2022a, April 23). *Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-welcomes-political-agreement-rules-ensuring-safe-and-accountable

European Commission. (2022b). *The strengthened code of practice on disinformation 2022*. European Commission. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

European Commission. (2022c). *Proposal for a decision of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)*. European Commission. https://op.europa.eu/en/publication-detail/-/publication/2c5d30ae-961f-11ec-b4e4-01aa75ed71a1/language-en/format-PDF

European Commission. (2022d, May 4). *European Digital Rights and Principles*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/digital-principles

European Parliament. (2021). *Tackling deepfakes in European policy*. Panel for the Future of Science and Technology. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

European Parliament. (2022a). *Regulation (EU) 2022/... of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)*. Council of the European Union. https://www.europarl.europa.eu/news/en/press-room/20220315IPR25504/deal-on-digital-markets-act-ensuring-fair-competition-and-more-choice-for-users, https://www.consilium.europa.eu/media/56086/st08722-xx22.pdf

European Parliament. (2022b). *European Parliament legislative resolution of 5 July 2022 on the 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 (COM(2020)0825 – C9-0418/2020 – 2020/0361(COD))*. Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CONSIL%3AST_10967_2022_INIT&qid=1664162501895, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_10967_2022_INIT&qid=1664162501895&from=EN

European Parliament. (2022c). *Report on foreign interference in all democratic processes in the European Union, including disinformation 2020/2268(INI)*. European Parliament.

https://www.europarl.europa.eu/doceo/document/A-9-2022-0022_EN.pdf

European Union. (2012a). *Charter of fundamental rights of the European Union*. Official Journal of the European Union.

European Union. (2016). *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)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Union. (2017). *Regulation (EU) 2017/1128 of the European Parliament and of the Council of 14 June 2017 on cross-border portability of online content services in the internal marketText with EEA relevance*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2017/1128/oj

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, (August). https://doi.org/10.1007/s13347-020-00419-2

Fischer, S. (2022, July 28). *Scoop: Meta officially cuts funding for U.S. news publishers*. Axios. https://www.axios.com/2022/07/28/meta-publishers-news-funding-cut

Fraga-Lamas, P., & Fernández-Caramés, T. M. (2019). Fake news, disinformation, and deepfakes: leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. *arXiv.org*. https://arxiv.org/abs/1904.05386

Fukuyama, F., Richman, B., Goel, A., Katz, R. R., Melamed, A. D., & Schaake, M. (2020). *Report of the working group on platform scale*. Stanford University. https://fsi.stanford.edu/publication/report-working-group-platform-scale

Gabriel, R. (2021, March 15). *Myth and the mind: saturated with rites and symbols, psychology feeds a deep human need once nourished by mythology*. Aeon. https://aeon.co/essays/how-psychology-fills-the-gap-from-the-disenchantment-of-the-world?utm_source=Aeon+Newsletter&utm_campaign=8e09d211b4-EMAIL_CAMPAIGN_2021_03_18_05_37&utm_medium=email&utm_term=0_411a82e59d-8e09d211b4-72046972

GAO (2020). *Science & Tech Spotlight: deepfakes* [Brochure]. GAO. https://www.gao.gov/pdf/product/704774

Gamble, C. N., Hanan, J. S., & Nail, T. (2019). What is new materialism?. *Journal of the Theoretical Humanities*, *24*(6), 111-134. https://doi.org/10.1080/0969725X.2019.1684704

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*(4), 617-637. https://doi.org/10.1111/jcom.12038

Gensing, P. (2020, September 3). *Auf dem Weg in eine alternative Realität?*. tagesschau. https://www.tagesschau.de/faktenfinder/hintergrund/deep-fakes-101.html

Giansiracusa, N. (2021). *How algorithms create and prevent fake news: exploring the impacts of social media, deepfakes, GPT-3, and more*. Apress.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 1-13. https://doi.org/10.1177%2F20563051221117552

Gladstone, B. (2021, October 15). *Does social media turn nice people into trolls?* [On the Media podcast]. WNYC Studios. https://www.wnycstudios.org/podcasts/otm/segments/does-social-media-turn-nice-people-trolls-on-the-media

Goh, S., & Soon, C. (2019). Governing the information ecosystem: Southeast Asia's fight against political deceit. *Public Integrity*, *21*(5), 523-536. https://doi.org/10.1080/10999922.2019.1603046

Goldberg, S. C. (2010). Introduction. In *Relying on others: an essay in epistemology* (pp. 1-9). Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780199593248.001.0001

Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, *20*(9), 1389-1407. https://doi.org/10.1080/1369118X.2017.1329334

Hackl, C. (2021, June 11). *Andrew Yang turns himself into an avatar and campaigns in the metaverse*. Forbes. https://www.forbes.com/sites/cathyhackl/2021/06/11/andrew-yang-turns-himself-into-an-avatar-and-campaigns-in-the-metaverse/

Hameleers, M., van der Meer, Toni G. L. A., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society*, *8*(3), 1-12. https://doi.org/10.1177/20563051221116346

Hancock, J. T., Woodworth, M. T., & Goorha, S. (2010). See No Evil: The effect of communication medium and motivation on deception detection. *Group Decision and Negotiation*, *19*, 327-343. https://doi.org/10.1007/s10726-009-9169-7

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Murias Munoz, M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *The Proceedings of the National Academy of Sciences*, *118*(50), 1-3. https://doi.org/10.1073/pnas.2116310118

Hannah. (2020, September 28). *New phishing scams detected targeting political opinions*. IT Security Guru. https://www.itsecurityguru.org/2020/09/28/new-phishing-scams-detected-targeting-political-opinions/

Harding, S. (2016). Latin American decolonial social studies of scientific knowledge: alliances and tensions. *Science, Technology, & Human Values*, *41*(6), 1063-1087. https://doi.org/10.1177/0162243916656465

Harris, K. R. (2022). Real fakes: the epistemology of online misinformation. *Philosophy & Technology*, *35*(83), 1-24. https://doi.org/10.1007/s13347-022-00581-9

Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts.

Institute of Electrical and Electronics Engineers, *7*(March), 41596-41606. https://doi.org/10.1109/ACCESS.2019.2905689

Hendrix, J. (2021, November 19). *Researchers: video may be more believable than text, but not necessarily more persuasive*. Tech Policy Press. https://techpolicy.press/researchers-video-may-be-more-believable-than-text-but-not-necessarily-more-persuasive/

Henley, J. (2020, January 29). *How Finland starts its fight against fake news in primary schools*. The Guardian. https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news

Herasimenka, A., Bright, J., Knuutila, A., & Howard, P. N. (2022). Misinformation and professional news on largely unmoderated platforms: the case of telegram. *Journal of Information Technology & Politics*. https://doi.org/10.1080/19331681.2022.2076272

Hinds, S. (2019). *The European Union approach to disinformation and misinformation: the case of the 2019 European Parliament elections* [Master's Thesis, Université de Strasbourg]. Global Campus Europe: EMA. https://doi.org/20.500.11825/1103

Holroyd, M., & Khatsenkova, S. (2022, August 17). *Why the sale of guns on Facebook in Iraq is so 'rife'*. Euronews. https://www.euronews.com/2022/08/17/why-the-sale-of-guns-on-facebook-in-iraq-is-so-rife

Institute for Strategic Dialogue, & Global Disinformation Index. (2021, August 1). *The business of hate: bankrolling bigotry in Germany and the online funding of hate groups*. Global Disinformation Index. https://www.disinformationindex.org/research/2021-8-1-the-business-of-hate-bankrolling-bigotry-in-germany-and-the-online-funding-of-hate-groups/

Institute for Strategic Dialogue. (2022, July 27). *Jiore Craig: the fight against disinformation cannot solely focus on content, but must address the systems at play*. Institute for Strategic Dialogue. https://www.isdglobal.org/isd-events/jiore-craig-the-fight-against-disinformation-cannot-solely-focus-on-content-but-must-address-the-systems-at-play/

Jarvis, J. (2021, December 21). Disinformation is not *the* problem. *BuzzMachine*. https://buzzmachine.com/2021/12/19/disinformation-is-not-the-problem/

Jourová, V. (2016). *Code of conduct—Illegal online hate speech: questions and answers*. European Commission. https://ec.europa.eu/info/sites/default/files/code_of_conduct_hate_speech_en.pdf

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: survey, battleground, and horizon. *International Journal of Computer Vision*, *130*(May), 1678-1734. https://doi.org/10.1007/s11263-022-01606-8

Kalpokas, I., & Kalpokiene, J. (2022). Chapter 7 Regulation: public, private, autonomous? & Chapter 8 Broader implications: politics and digital posthumanism. In *Deepfakes: a realistic assessment of potentials, risks, and policy regulation* (pp. 65-72 & 73-84). Springer International Publishing AG.

Kelly, N. (2022, July 18). *The next-generation news consumer is older than you think and wants more video*. Digital Content Next. https://digitalcontentnext.org/blog/2022/07/18/the-next-generation-

news-consumer-is-older-than-you-think-and-wants-more-video/?utm_source=DCN+InContext+Newsletter&utm_campaign=98049f4a27-incontext-22_07_21&utm_medium=email&utm_term=0_7a1e151592-98049f4a27-399302011

Kim, E.-S. (2020). Deep learning and principal–agent problems of algorithmic governance: the new materialism perspective. *Technology in Society*, *63*(101378), 1-9. https://doi.org/10.1016/j.techsoc.2020.101378

Krasodomski-Jones, A. (2016). *Talking to ourselves? Political debate online and the echo chamber effect*. Demos. https://www.demos.co.uk/wp-content/uploads/2017/02/Echo-Chambers-final-version.pdf

Kreiss, D. (2021). *Polarization isn't America's biggest problem—or Facebook's*. Wired. https://www.wired.com/story/polarization-isnt-americas-biggest-problem-or-facebooks/

Kuchay, B. (2020, December 11). *EU NGO report uncovers Indian disinformation campaign*. Al Jazeera. https://www.aljazeera.com/news/2020/12/11/eu-ngo-report-uncovers-a-15-year-disinformation-campaign

Kuhlman, S., & Rip, A. (2018). Next-generation innovation policy and grand challenges. *Science and Public Policy*, *45*(4), 448-454. https://doi.org/10.1093/scipol/scy011

Kuo, R., & Marwick, A. (2021, August 12). Critical disinformation studies: history, power, and politics. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(4), 1-12. https://doi.org/10.37016/mr-2020-76

Kwok, A. O. J., & Koh, S. G. M. (2020). Deepfake: a social construction of technology perspective. *Current Issues in Tourism*, *24*(13), 1798-1802. https://doi.org/10.1080/13683500.2020.1738357

László, M., Narr, G. E., Hillman, V., Couldry, N., & Newman, R. (2022). 4 ways the new EU digital acts fall short and how to remedy it. *Medium*. https://medium.com/@gregerwinnarr/4-ways-the-new-eu-digital-acts-fall-short-and-how-to-remedy-it-d16b681a88bc

Law, J. (1994). Agency, deletion and relational materialism. In *Organizing modernity* (pp. 100-104). Blackwell.

Law, J. (2006). *Pinboards and books: juxtaposing, learning and materiality*. HeterogeneitiesDOTnet: John Law's STS Web Page. http://www.heterogeneities.net/publications/Law2006PinboardsAndBooks.pdf

Lecomte, J. (2021). *Désinformation, fake news: pourquoi on y adhère et comment s'en prémunir ?*. Philosophie, médias et société. https://www.philomedia.be/desinformation-fake-news-pourquoi-on-y-adhere-et-comment-sen-premunir/

Lenoir, T. (2022, August 1). *Reconsidering the fight against disinformation*. Tech Policy Press. https://techpolicy.press/reconsidering-the-fight-against-disinformation/

Levi-Faur, D. (2011). Regulation and regulatory governance. In D. Levi-Faur (Ed.), *Handbook on the politics of regulation* (pp. 3-21). Edward Elgar Publishing.

Lewis, K., Gonzalez, M., & Kaufman, J. (2011). Social selection and peer influence in an online social

network. Proceedings of the National Academy of Sciences (PNAS), 109(1), 68-72. https://doi.org/10.1073/pnas.1109739109

Life Itself Labs. (n.d.). *Making Sense of Web3*. Life Itself. https://web3.lifeitself.us/about

LinkedIn. (2022). *About us: statistics*. LinkedIn. https://news.linkedin.com/about-us#Statistics

Logic. (2021, December 25). (dis)Info studies: André Brock, Jr. on why people do what they do on the Internet. *Logic*, *15*(Beacons). https://logicmag.io/beacons/dis-info-studies-andre-brock-jr/

Lyons, K. (2020, February 18). *An Indian politician used AI to translate his speech into other languages to reach more voters*. The Verge. https://www.theverge.com/2020/2/18/21142782/india-politician-deepfakes-ai-elections

do Mar Pereira, M. (2022, April 28). *The affective life of knowledge production: conversations on embodied and disembodied voices* [Lecture]. Forschungsplattform GAIN - Gender: Ambivalent In_Visibilities, University of Vienna, Austria. https://gain.univie.ac.at/news/detailansicht/news/maria-do-mar-pereira-phd-msc-the-affective-life-of-knowledge-production-conversations-on-embodie/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail&cHash=d25e32919c377edf7eda94a651491b36

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2019). Epistemic spillovers: learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition*, *188*, 74-84. https://doi.org/10.1016/j.cognition.2018.10.003

Marwick, A., Kuo, R., Cameron, S. J. & Weigel, M. (2021). *Critical disinformation studies: a syllabus*. Center for Information, Technology, & Public Life (CITAP). https://citap.unc.edu/critical-disinfo

Masnick, M. (2022a, March 11). *Performative conservatives are mad that a search engine wants to downrank disinformation*. Techdirt. https://www.techdirt.com/2022/03/11/performative-conservatives-are-mad-that-a-search-engine-wants-to-downrank-disinformation/

Masnick, M. (2022b, May 26). *Senator Gillibrand says we don't have to regulate speech, just misinfo. Who wants to tell her?*. Techdirt. https://www.techdirt.com/2022/05/26/senator-gillibrand-says-we-dont-have-to-regulate-speech-just-misinfo-who-wants-to-tell-her/

Masnick, M. (2022c, August 3). *As expected, Facebook is no longer interested in paying news orgs to post news on Facebook that no one wants*. Techdirt. https://www.techdirt.com/2022/08/03/as-expected-facebook-is-no-longer-interested-in-paying-news-orgs-to-post-news-on-facebook-that-no-one-wants/

McKay, S., & Tenove, C. (2020). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, *74*(3), 703-717. https://doi.org/10.1177%2F1065912920938143

Miller, P. (2013). Postmaterialism and social movements. In D. A. Snow, D. della Porta, B. Klandermans, & D. McAdam (Eds.), *The Wiley-Blackwell encyclopedia of social and political movement* (pp. 1-4). Blackwell Publishing. https://doi.org/10.1002/9780470674871.wbespm165

Mol, A., & Law, J. (2002). Complexities: an introduction. In J. Law & A. Mol (Eds.), *Complexities* (pp. 1-22). Duke University Press.

Monnet, S. (2020, November 16). *Internet protocols and human rights: interplay or interdependence?*. Diplo. https://www.diplomacy.edu/blog/internet-protocols-and-human-rights-interplay-or-interdependence/

Neo, R. (2021). The international discourses and governance of fake news. *Global Policy*, *12*(2), 214-228. https://doi.org/10.1111/1758-5899.12958

Newman, J. (2022, March 18). *It's a cheapfake! Experts laugh off Kremlin misinformation attempt as amateurish 'deepfake' video of Zelensky 'surrendering' is posted by hackers – and spotted almost immediately*. Daily Mail. https://www.dailymail.co.uk/news/article-10625935/Experts-laugh-Kremlins-amateurish-deepfake-video-Zelensky-surrendering.html

Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R. K. (n.d.). *Reuters Institute digital news report 2020*. Reuters Institute. https://www.digitalnewsreport.org/

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv.org*. https://arxiv.org/abs/1909.11573

Nguyễn, S., Kuo, R., Reddi, M., Li, L., & Moran, R. E. (2022). Studying mis- and disinformation in Asian diasporic communities: the need for critical transnational research beyond Anglocentrism. *Harvard Kennedy School Misinformation Review*, *3*(2), 1-12. https://doi.org/10.37016/mr-2020-95

Nicholas, G. (2022, July 13). *Shadowbanning: sorting fact from fiction*. Tech Policy Press. https://techpolicy.press/shadowbanning-sorting-fact-from-fiction/

Nonprofit Quarterly. (2022). *Changing the subject: boards as social movement spaces*. Nonprofit Quarterly. https://store.nonprofitquarterly.org/products/changing-the-subject-boards-as-social-movement-spaces?utm_medium=email&_hsmi=220265941&_hsenc=p2ANqtz-8zOfI1BZrc1QsvOxVkkPKV9g4MnOWxGEEDy8BBAi702LCCzPNU1wedNtF9KdILMmAhZ6DOVrMpxQnmnotLKT1KSKjIbt5iyAYdSomUcgCaE2dFzhg&utm_content=220265941&utm_source=hs_email

Ó Fathaigh, R., Dobber, T., Zuiderveen Borgesius, F., & Shires, J. (2021). Microtargeted propaganda by foreign actors: an interdisciplinary exploration. *Maastricht Journal of European and Comparative Law*, *28*(6), 856-877. https://doi.org/10.1177%2F1023263X211042471

O'Neill, S. J., & Smith, N. (2014). Climate change and visual imagery. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(1), 73-87. https://doi.org/10.1002/wcc.249

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999-1015. https://doi.org/10.1017/S0003055421000290

Ovadya, A. (2019, June 14). *Deepfake myths: common misconceptions about synthetic media*. Alliance for Securing Democracy, German Marshall Fund of the United States.

https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/

Ovadya, A. (2021). *Towards platform democracy: policymaking beyond corporate CEOs and partisan pressure*. Belfer Center for Science and International Affairs. https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure

Oyěwùmí, O. (1997). Visualizing the body: Western theories and African subjects. In *The invention of women: making an African sense of Western gender discourses* (pp. 1-30). University of Minnesota Press.

P2P Models. (n.d.). *Towards a new collaborative economy. Decentralizing power and value using blockchain*. P2P Models. https://p2pmodels.eu/

Paris, B. (2020). The Internet of futures past: values trajectories of networking protocol projects. *Science, Technology, & Human Values*, *46*(5), 1021-1047. https://doi.org/10.1177%2F0162243920974083

Paris, B. (2021). Configuring fakes: digitized bodies, the politics of evidence, and agency. *Social Media + Society*, *7*(4), 1-13. https://doi.org/10.1177%2F20563051211062919

Paris, B., & Donovan, J. (2019, September 18). *Deepfakes and cheap fakes: the manipulation of audio and visual evidence*. Data & Society. https://datasociety.net/library/deepfakes-and-cheap-fakes/

Pérez-Bustos, T., Sánchez-Aldana, E., & Chocontá-Piraquive, A. (2019). Textile material metaphors to describe feminist textile activisms: From threading yarn, to knitting, to weaving politics. *Textile*, *17*(4), 368-377.

Pfefferkorn, R. (2020). "Deepfakes" in the courtroom. *Boston University Public Interest Law Journal*, *29*(2), 245-276.

Plume d'histoire. (2015, November 5). *Marie-Antoinette victime de pamphlets érotico-obscènes*. Plume d'histoire. https://plume-dhistoire.fr/marie-antoinette-victime-pamphlets-erotico-obscenes/

Radsch, C. (2022, July 6). *Artificial intelligence and disinformation: state-aligned information operations and the distortion of the public sphere*. Organization for Security and Co-operation in Europe. https://www.osce.org/representative-on-freedom-of-media/522166

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906-917. https://doi.org/10.1016/j.ejor.2020.09.020

Relph, E. (2021). Digital disorientation and place. *Memory Studies*, *14*(3), 572-577. https://doi.org/10.1177%2F17506980211010694

Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics & Social Philosophy*, *22*(2), 143-161. https://doi.org/10.26556/jesp.v22i2.1628

Romero-Vicente, A. (2022, May 4). *"Crypto-funding" to disinform*. EU DisinfoLab. https://www.disinfo.eu/publications/crypto-funding-to-disinform/

Rosenberg, S. (2022, July 25). *Sunset of the social network*. Axios. https://www.axios.com/2022/07/25/sunset-social-network-facebook-tiktok

Ross Arguedas, A., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022, January 19). *Echo chambers, filter bubbles, and polarisation: a literature review*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review

Roth, A. (2021, April 22). *European MPs targeted by deepfake video calls imitating Russian opposition*. The Guardian. https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition?CMP=Share_iOSApp_Other

de Ruiter, A. (2021, December 8). *Why we should rethink our moral intuitions about deepfakes*. Psyche. https://psyche.co/ideas/why-we-should-rethink-our-moral-intuitions-about-deepfakes

Sancho, D., Eira, M., & Klayn, A. (2021). *Malicious uses and abuses of artificial intelligence (AI)*. Europol EC3. https://rm.coe.int/0900001680a4892f

Saurwein, F., & Spencer-Smith, C. (2020). Combating disinformation on social media: multilevel fovernance and distributed accountability in Europe. *Digital Journalism*, *8*, 820-841. https://doi.org/10.1080/21670811.2020.1765401

de Seta, G. (2021). Huanlian, or changing faces: deepfakes on Chinese digital media platforms. *Convergence: The International Journal of Research into New Media Technologies*, *27*(4), 935-953. https://doi.org/10.1177%2F13548565211030185

Sher, G. (2019). A Wild West of the mind. *Australasian Journal of Philosophy*, *97*(3), 483-496. https://doi.org/10.1080/00048402.2018.1490326

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., & Michalski, M. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *Lecture Notes in Computer Science book series*, *11037*, 1-11. https://doi.org/10.1007/978-3-030-00536-8_1

Silbey, J., & Hartzog, W. (2019). The upside of deep fakes. *Maryland Law Review*, *78*(4), 960-966. https://digitalcommons.law.umaryland.edu/mlr/vol78/iss4/8/

Smith, A. (2021, April 19). *Nvidia is building a giant virtual 'metaverse' of the world, with 'digital twins' of cars, cities, and people*. Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/nvidia-virtual-metaverse-world-b1833707.html

Star, S. L. (1990). Power, technology and the phenomenology of conventions: on being allergic to onions. *The Sociological Review*, *38*(2/S), 26-56.

Statista. (2022). *Number of monthly active Facebook users worldwide as of 2nd quarter 2022*. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Stechemesser, A., Levermann, A., & Wenz, L. (2022). Temperature impacts on hate speech online: evidence from 4 billion geolocated tweets from the USA. *Lancet Planet Health*, *6*(9), E714-E725. https://doi.org/10.1016/S2542-5196(22)00173-5

Strathern, M. (1991). Prosthetic extensions. In *Partial connections* (pp. 105-119). Rowman & Littlefield Publishers.

Strathern, M. (2002). On space and depth. In J. Law & A. Mol (Eds.), *Complexities* (pp. 88-115). Duke University Press.

Stray, J. (2012, July 11). *Are we stuck in filter bubbles? Here are five potential paths out*. NiemanLab. https://www.niemanlab.org/2012/07/are-we-stuck-in-filter-bubbles-here-are-five-potential-paths-out/

Svabo, C., & Bønnelycke, J. (2020). Knowledge catcher: on the performative agency of scholarly forms. *PARtake: The Journal of Performance as Research*, *3*(1). https://doi.org/10.33011/partake.v3i1.477

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

techDetector. (2021, March 26). *Anti-filter bubble algorithm*. techDetector. https://techdetector.de/applications/anti-filter-bubble-algorithm

The Economist. (2019, October 22). *Could deepfakes weaken democracy?* [YouTube video]. The Economist. https://www.youtube.com/watch?v=_m2dRDQEC1A

Tommy Genesis. (2021). A woman is a god [Song]. On *goldilocks x*. Downtown Records.

Tufekci, Z. (2014). Big questions for social media big data: representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014 [forthcoming]. https://arxiv.org/abs/1403.7400

Ugland, E. (2019). Expanding media law and policy education: confronting power, defining freedom, awakening participation. *Communication Law and Policy*, *24*(2), 271-306. https://doi.org/10.1080/10811680.2019.1586407

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, (February), 1-13. https://doi.org/10.1177/2056305120903408

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K., & Tucker, J. A. (2020). Political psychology in the digital (mis)information age: A model of news belief and sharing. PsyArXiv. https://doi.org/10.31234/osf.io/u5yts

Vauss, D., et al. (n.d.). *Nerds Against Humanity* [Facebook page]. Facebook. Retrieved March 31, 2022, from https://www.facebook.com/NerdsAgainstHumanity2018/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

de Vries, K. (2022). Let the robot speak! AI-generated speech and freedom of expression. In *YSEC Yearbook of Socio-Economic Constitutions* (pp. 1-23). Springer, Cham. https://doi.org/10.1007/16495_2021_38#DOI

Wæver, O., & Buzan, B. (2020). Racism and responsibility – The critical limits of deepfake methodology in security studies: a reply to Howell and Richter-Montpetit. *Security Dialogue*, *00*(0), 1-9. https://doi.org/10.1177%2F0967010620916153

Walsh, J. P. (2020). Social media and moral panics: assessing the effects of technological change on societal reaction. *International Journal of Cultural Studies*, *23*(6), 840-859. https://doi.org/10.1177%2F1367877920912257

Williams, I. K. (2022, April 6). *Can AI-driven voice analysis help identify mental disorders?*. The Indian Express. https://indianexpress.com/article/technology/tech-news-technology/can-ai-driven-voice-analysis-help-identify-mental-disorders-7855252/

Winter, R., & Salter, A. (2019). DeepFakes: uncovering hardcore open source on GitHub. *Porn Studies*, *7*(4), 382-397. https://doi.org/10.1080/23268743.2019.1642794

Wong, J. C. (2021, April 12). *Revealed: the Facebook loophole that lets world leaders deceive and harass their citizens*. The Guardian. https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation

Wood, T., & Porter, E. (2019). The elusive backfire effect: mass attitudes' steadfast factual adherence. *Political Behavior*, *41*, 135-163. https://doi.org/10.1007/s11109-018-9443-y

Yadlin-Segal, A., & Oppenheim, Y. (2020). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence: The International Journal of Research into New Media Technologies*, *27*(1), 36-51. https://doi.org/10.1177%2F1354856520923963

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

Zilles, C. (2020, July 31). *If social media companies are publishers and not platforms, that changes everything*. Social Media Headquarters. https://socialmediahq.com/if-social-media-companies-are-publishers-and-not-platforms-that-changes-everything/

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human border at the frontier of power*. Profile Books.

L I S T O F F I G U R E S

A B S T R A C T

Whatever the times and whatever the regime, disinformation that sows distrust in politics and public institutions has always been a sensitive topic. Today, we become increasingly reliant on online media for our consumption of news. The extent of this online information ecosystem makes that political disinformation can have broader reach at faster rates. Adding deepfake technology to the mix—where deep learning allows the creation of fake images, videos, audio files, or texts that are seemingly authentic—leads some to believe democracies to be at the mercy of deepfaked political disinformation. However, research shows that such fatalistic conception is reductive of reality. Further research thereby points at the need for a theoretical reconceptualisation of these deepfakes in order to re-align their conception with their empirical reality. To provide such theoretical contribution, Karen Barad's new materialist development serves to conceptually revisit the deepfake. A revisitation argued to redress the shortcomings of the popular fatalist conception. Most notably, this revisitation shakes the widespread ideal of the clearcut fact-fake demarcation to its core. An ideal nevertheless characterising one of EU's prime efforts to counter deepfakes. The Baradian reconceptualisation of the deepfake is thus further used to analyse the current EU policy approach to deepfakes. A critical examination that in turn serves for the writing of a policy proposal; a proposal not only based on the Baradian reconceptualisation but also on an empirical exploration of the ways through which the deepfake materialises in our society. This proposal—of which the key contribution is to appreciate the netizen in its capacity for autonomous judgment—is argued to redress the absence of a justified basis in the current EU policy approach given that the Baradian-inspired proposal is based on a conception of the deepfake that is more aligned with its observed reality.

Z U S A M M E N F A S S U N G

Unabhängig von der Zeit und dem Regime sind Desinformationen, die zu Misstrauen in der Politik und öffentlichen Institutionen führen, immer ein heikles Thema gewesen. Heutzutage sind wir beim Konsum von Nachrichten zunehmend auf Online-Medien angewiesen. Das Ausmaß dieses Online-Informations Ökosystems führt dazu, dass politische Desinformationen eine größere Reichweite haben können und schneller verbreitet werdennn. Wenn dann noch die Deepfake-Technologie hinzukommt—bei der mit Hilfe von Deep Learning gefälschte Bilder, Videos, Audiodateien oder Texte erstellt werden können, die scheinbar authentisch sind—glauben viele, dass Demokratien gefälschten politischen Desinformationen ausgeliefert sind. Die Forschung zeigt jedoch, dass eine solche fatalistische Sichtweise die Realität verzerrt. Weitere Forschungen weisen daher auf die Notwendigkeit einer theoretischen Rekonzeptualisierung dieser Deepfakes hin, um ihre Konzeption wieder mehr mit der empirischen Realität in Einklang zu bringen. Um einen solchen theoretischen Beitrag zu leisten, dient Karen Barads neue materialistische Entwicklung dazu, das Deepfake konzeptionell zu überdenken. Eine Neubetrachtung, die die Unzulänglichkeiten der populären fatalistischen Konzeption beheben soll. Vor allem erschüttert diese Revision das weit verbreitete Ideal der eindeutigen Fakt-Falsch-Abgrenzung in seinen Grundfesten. Ein Ideal, das jedoch eine der wichtigsten Bemühungen der EU zur Bekämpfung von Deepfakes kennzeichnet. Die baradianische Rekonzeptualisierung des Deepfakes wird daher weiter genutzt, um den aktuellen politischen Ansatz der EU gegenüber Deepfakes zu analysieren. Eine kritische Untersuchung, die wiederum dazu dient, einen politischen Vorschlag zu verfassen; einen Vorschlag, der nicht nur auf der baradianischen Rekonzeptualisierung basiert, sondern auch auf einer empirischen Untersuchung der Art und Weise, wie sich das Deepfake in unserer Gesellschaft materialisiert. Dieser Vorschlag, dessen Hauptbeitrag darin besteht, den Netizen in seiner Fähigkeit zur autonomen Urteilsbildung zu würdigen, soll das Fehlen einer gerechtfertigten Grundlage für den derzeitigen politischen Ansatz der EU beheben, da der von Baradian inspirierte Vorschlag auf einer Konzeption des Deepfakes beruht, die besser mit der beobachteten Realität übereinstimmt.