LSE Media and Communications

Media@LSE MSc Dissertation Series

Editors: Bart Cammaerts, Saumyadeep Mandal and Hao Wang

A 'CANARY IN THE COALMINE' FOR SYNTHETIC MEDIA REGULATION

The Emerging Threat of Deepfake Image Abuse

OLIVIA OTTS

Published by Media@LSE, London School of Economics and Political Science ("LSE"), Houghton Street, London WC2A 2AE. The LSE is a School of the University of London. It is a Charity and is incorporated in England as a company limited by guarantee under the Companies Act (Reg number 70527).

ABSTRACT

This research aims to reveal how policy actors conceptualise and rank their main concerns for the potential online harms of synthetic media technologies - and with what underpinning rationales. It also asks how these actors conceptualise gender-based online harms compared with other forms of harm, and with what justifications. Proposals for synthetic media regulation in liberal-democratic contexts are argued to fit roughly into two camps. One primarily fears their use in influence operations and disinformation - and their risk to the information environment required for democratic deliberation. Another group, smaller than the prior, commonly quotes one (now admittedly outdated) statistic – that 96% of all deepfakes are pornographic depictions of female bodies (Ajder et al., 2019). One might argue that preventing collective harms, such as the discreditation of election outcomes, should rationally take precedence over more 'individualised' harms, such as deepfake image abuse. However, after ten expert interviews with policymakers, OSINT analysts, and technologists active in the synthetic media space, I have developed a more nuanced view. Conceiving synthetic media regulation as a balancing act between two forms of harm ignores one important outcome from these interviews: seemingly individualised image abuse is in fact enmeshed within - and a 'canary in the coalmine' (P3) for - the core epistemic threat of synthetic media itself. When women in power are discredited or shamed into silence by their political adversaries – when their democratic rights are violated - the implications of their plight are necessarily collective. The potential consequences of synthetic image abuse - her shame, her silence, or her death – are not simply individual. Moving forward, regulators are encouraged to critically consider the interconnected ethical concerns – and epistemic underpinnings – of these rising online harms.

INTRODUCTION

> ...if we try to roll this all into one and we forget about the problem that already exists, then women are just gonna be left behind again - and it's gonna be about the economic harm, and it's gonna be about the national security harm again - instead of, like, this very fundamental right to privacy and integrity and bodily autonomy that's being taken away from people (P3)

Open-source intelligence (OSINT) researchers have investigated the use of synthetic media in information operations since the early stages of their development (see: Ajder et al., 2019). These researchers' first-hand perspectives have the advantage of being embedded in real-time and evolving cases of online disinformation. One of the most significant contributions of open-source research to the understanding of synthetic media is the report (Ibid) that claimed 'non-consensual deepfake pornography *(...)* accounted for 96% of the total deepfake videos online' (n.p.). This statistic, years after its publication, still begs the question - if deepfake image abuse is a prevalent online harm, how is its prevention weighed against other, potentially graver dangers that the advancement of synthetic media may bring? Some theorists have warned from early on of synthetic media's potential risk to national security (Chesney and Citron, 2018) and the functioning of democracies (Pawelec, 2022). Others have noted its potential utility in large-scale disinformation campaigns (Jeong et al., 2021). Given these large-scale 'collective' risks, it is reasonable to wonder if deepfake image abuse does not naturally float to the top of a policymaker's to-do list.

Emerging policy coalitions around synthetic media appear to mirror those around the broader liberal-democratic discourse of AI regulation. Two distinct factions – those who warn of superintelligence and world-ending catastrophes (see Roose, 2023), and those who argue for a focus on harms in the here and now (see O'Neil, 2023) appear to conflict in their core conceptions of what online harms regulation is for. Scant literature analyses these discourses in relation to synthetic media, leaving much of what is published divided between 'hard' geopolitical studies of deepfakes (see: Byman, 2023) and 'soft' narratives of their impacts on women's lives (see: Burgess, 2020). There are some exceptions (see: Chesney and Citron, 2018 – who combine the two), but these are rare. A more critical consideration of how gendered

online harms – for example, shaming a female politician through deepfake pornography (*Editorial*, 2020; Mackintosh and Gupta, 2020) - intersect with hard geopolitical considerations, is needed.

As of now, the United States' deepfake policymaking process is cautiously underway. Lawmakers have proposed legislation (ex. Deepfake Task Force Act S.2559 (117th Congress, 2022)) to mitigate the risks of deepfakes to the information ecosystem. However, the U.S. Special Ops Command has also submitted a procurement request (SOCOM, 2023) to consider employing deepfake technologies in its operations. As these 'adversarial dynamics' (Leibowicz et al., 2021: 1) develop, is the regulation of synthetic media a balancing act between the potential for large-scale epistemic harms and the reality of frequent, individualised ones? This dissertation aims to present a new perspective on how gendered online harms directly interrelate with grander national security and democratic harms. It also aims to explore how experts in the synthetic media space juggle collective and individual forms of harm. The wider aim of this research study is to deepen the existing understanding of how policy coalitions are developing around synthetic media.

The cornerstone of this research aim is the analysis of ten qualitative expert interviews with policymakers, OSINT analysts, technologists, and a victim of deepfake image abuse. The interviews cover a wide range of themes – ethical considerations of the usage of deepfakes in war, the utility of OSINT analysis in policy topic prioritisation, and more. The core interest underpinning these themes is their treatment in comparison with - and their relationship to - gendered harms. First, relevant debates on epistemic and democratic theories and their relationship to synthetic media will be overviewed. Then, the methodological intent and process of expert interviewing will be explained in detail. Interviews will then be analysed and discussed critically within a broader conversation on the regulatory prioritisation of online harms.

THEORETICAL CHAPTER

This dissertation focuses on a series of core concepts and theories as they relate to emerging debates on the regulation of synthetic media. These concepts – epistemic security

and deliberative democracy; simulacrums and the 'hyperreal' (Baudrillard, 1983); information operations and gendered harms; and OSINT investigations and policy coalitions – all connect to two core concerns. First - how are online (and offline) harms from synthetic media being conceptualised in the lens of a wider social contract? Second – how are these varied harms being prioritised in liberal-democratic policy discourses? Each of these concepts will be defined, connected to recent synthetic media developments, and critically analysed as to the debates they raise in the relevant literature.

A few key terms require definition before delving into the bulk of the literature review. First, synthetic media refers to 'source media (that) has been manipulated or synthesized using AI techniques' (Liebowicz et al., 2021). All deepfakes are a form of synthetic media, but not all synthetic media is a deepfake. The definition of deepfakes themselves has been murky for some time – and contested. In a recent comprehensive literature review of their definition, Whittaker et al. (2023) proposed the following useful definition of a deepfake, as:

> …synthetic media generated using artificial intelligence and deep learning technology which produce realistic yet fake representations of people undertaking actions or saying words in the form of video, image, or audio content.

I would expand upon this definition to note that deepfakes may also depict non-human content in a way that is usable for disinformation purposes – a false satellite image of a military base, as a hypothetical example. It is not possible here to explore the origins of deepfakes in depth, but some context is necessary. The term *deepfake* is generally credited to a user who created r/deepfakes on Reddit in November 2017 to disseminate artificial pornography of female celebrities (Cole, 2018). However, deepfake-adjacent technology first arose two decades earlier through the 1997 *Video Rewrite* program, which altered mouth movements to allow for the semi-realistic insertion of another audio track (Bregler et. al, 1997). Video manipulation technologies are not a recent development – what is so new, and so concerning, is the capacity to make them hyper-realistic – a problem that will be explored in detail later.

Epistemic Security and Deliberative Democracy

A theoretical lens that has been frequently used in analyses of the rise of synthetic media is that of epistemic security. This refers to epistemology – the study of knowledge – and the degree to which knowledge itself is secure. Epistemic security is defined by the Alan Turing Institute (Seger et al., 2020: 11) as follows:

> …a holistic umbrella for investigations into the processes by which societies produce, distribute, evaluate and assimilate information, and into threats that restrict access to information, or undermine our ability to evaluate information veracity or information source reliability

The Alan Turing Institute argues that 'an epistemically secure society (is) one that reliably averts threats to the processes by which reliable information is produced, distributed, acquired and assessed within the society' (Seger et. al, 2020: 2). This concept is relevant in the context of the threat of geopolitical influence operations, but also to the foundation of the democratic process more generally. A few illustrations are useful to outline this risk. First, imagine a military decision-maker views a video of a president asking their troops to surrender (See Allyn, 2022 for a real Zelenskyy example). To what extent is this military official's visual and aural experience still a reliable basis upon which to pass judgement – and how integral is their decision to trust or not trust the images they see? The basis upon which previously commonplace spaces of visual legitimacy– broadcast news reports, for example - has arguably become increasingly shaky. As synthetic media gains the potential to blur the line between realistic representation and constructed fiction, the relevance of debates around epistemic security increases.

The maintenance of epistemic security has also been argued in the literature as core to the functioning of democratic decision-making structures (see: Seger et al., 2020). A shared knowledge base has been argued to be a prerequisite for a society to make informed decisions on its own governance (see: Cohen, 2002 for a discussion of the Rawlsian view). The deterioration of this knowledge base – through risks such as the 'manipulation of elections', 'lowering *(of)* trust in institutions and authorities', and the 'undermining *(of)* journalism and information' (Waldemarsson, 2020) have been the focus of much concern in the literature. Despite a convincing argument and a wide swath of possible harms discussed, this report does not address any gender-specific vulnerabilities to deepfakes, or any political implications of

image abuse itself. It is important to note that some analyses of the subject have addressed deepfake pornography alongside national security and election concerns (Chesney and Citron, 2018; European Science-Media Hub, 2021). Some, from early on, have also explicitly discussed the balance between considering 'election chaos' harms and image abuse concerns that primarily target women (Meaker, 2019). However, there is a relative paucity in the literature on how geopolitical and gendered harms interrelate and exacerbate each other. This gap - and its potential policy implications - is a core basis of the later analysis, and should be kept in mind throughout this literature review. Before discussing this in more detail, philosophical theories that underpin the core threats of synthetic media will be presented as a backbone for understanding current debates on deepfake harms prioritisation.

Simulacrums and the Hyperreal

The French sociologist Baudrillard's seminal *Simulacra and Simulation* (Baudrillard, 1983, trans. Glaser) has been noted by multiple authors for its utility in academic critiques of synthetic media (notably Wagner and Blewer, 2019). Baudrillard's concept of the 'hyperreal' - 'the generation by models of a real without origin or reality' (Baudrillard, 1983: 3) - is particularly relevant here. For example, if an ML (machine learning) model is trained on a dataset of Piet Mondrian images and then outputs its own version based on this dataset, it does have an *origin* in that artist's work. However, its reality as a Mondrian piece is debatable – one could argue that it lacks any real authorship. Similarly, models used to create realistic AI portraits, such as the aptly named *this-person-does-not-exist.com* (2023), are not creating real people – although their datasets are based on real faces. They are arguably amalgamations of once-real images and ML interpretations of them. Synthetic media, although created in its current form decades after Baudrillard's hyperreal, is argued by Waite (2023, n.p.) to be a fitting example of the concept.

Baudrillard also presents the concept of the 'simulacrum'. As defined by Ecclesiastes - 'The simulacrum is never what hides the truth - it is truth that hides the fact that there is none.' (Baudrillard, 1983: 3). Baudrillard's multi-step conception is relevant to the understanding of synthetic media as a transformation of existing datasets and knowledge bases:

> Such would be the successive phases of the image: it is the reflection of a profound reality; it masks and denatures a profound reality; it masks the absence of a profound reality; it has no relation to any reality whatsoever; it is its own pure simulacrum (Baudrillard, 1983, trans. Glaser: 6)

As seen through the lens of Baudrillard's theory, synthetic media wears not only on the contours that denote what is perceptible and what is reasonable. It also arguably wears on the core epistemic value (or lack thereof) embedded within objects that a person perceives as knowledge-bearing. It can be conceptualised, in essence, as an ignorance-creation loop.

On the topic of ignorance, a proposed area of study that contrasts with the previous overview of epistemology is that of *agnotology*, defined by Proctor and Schiebinger (2008) as 'the study of ignorance making' (p. 5). This, in his view, can arise from 'secrecy, stupidity, apathy, censorship, disinformation, faith, and forgetfulness' (Ibid: 9). Ignorance of the *falseness* of synthetic image, video, or audio content will be argued to arguably correlate with many of the prior causes Proctor and Schiebinger (2008) present. When questioning the epistemic status and origins of a piece of synthetic media, I argue that it is essential to consider whether the *production* of ignorance is an intended result. As an example - a deepfake meant to reassure citizens of the 'realness' of an event that never came to pass, especially in the context of geopolitical conflicts, can be a clear example of the automated production of ignorance for strategic gain.

Multiple theorists have since brought the topic of ignorance-creation into conversation with the effects of modern platforms and algorithms. RAND published *Truth Decay: A Threat to Policymaking and Democracy* (Kavanagh and Rich, 2018) on the causes, impacts, and solutions to the titular phenomenon. The 'most damaging effects' of this 'truth decay' process, they argue, are the 'erosion of civil discourse, political paralysis at the federal and state level, individual disengagement from political and civic life, *(and)* uncertainty in national policy' (p. 3). This is relevant in understanding policymakers' fears of what malign synthetic media can inflict on an information ecosystem - and on fellow citizens. This theoretical connection, and its relationship to the gendered harms of deepfakes, will be crucial in later analyses.

Gradual Incorporations of Gendered Harms

Deepfakes were considered through a national security lens soon after they came into the public lexicon. Early research on the security implications of deepfakes comes from Chesney and Citron (2018). A particularly accurate prediction made in this piece was that 'The capacity to generate persuasive deep fakes will not stay in the hands of either technologically sophisticated or responsible actors.' Now, with the rise of publicly accessible synthetic image creation servers (see Discord, 2023) - deepfake creation did, as Chesney and Citron predicted, 'democratize' (p. 1762) - at least in terms of accessibility. The aforementioned synthetic media-creation channel, at the time of writing, has fourteen million registrants (Discord, 2023). Over a year after Chesney and Citron's report, the Dutch cyber-security company *Deeptrace* published *The State of Deepfakes* (Ajder et. al, 2019). This study covered the expected bases of a general analysis of deepfakes - their commodification, market impacts, implications for cybersecurity, and the like. However, they also included a distinct section on 'deepfake pornography' (Ibid: 6) - and did not limit this section to 'individualised' harms alone. Quoting Danielle Citron, the report states:

> Deepfake technology is being weaponized against women by inserting their faces into porn. It is terrifying, embarrassing, demeaning, and silencing. Deepfake sex videos say to individuals that their bodies are not their own and can make it difficult to stay online, get or keep a job, and feel safe (Ibid: 6)

The final section of this statement represents a small but important turn in academic literature's coverage of online harms from synthetic media – the argument that deepfake pornography has firm collective implications. Citron's focus on the challenges of maintaining digital connectivity and meaningful employment in the face of synthetic image abuse highlights a wider concern – that of women's' societal participation. A case study within this piece noted that a monetised app used to create this content, 'DeepNude', was taken down after being flooded with requests. However, they note, its 'software continues to be independently repackaged and distributed through various online channels, such as opensource repositories and torrenting websites' (Ibid: 8). This is described as allowing software to 'spread and mutate like a virus' (Ibid: 8) - further complicating cybersecurity and regulatory approaches.

Beyond the technical difficulty of tracking or removing content and content-creation tools, impacts on individuals' lives are argued by other authors to be similarly long-lasting. A CNN report (Mackintosh and Gupta, 2020) noted that 'India's youngest parliamentarian, Chandrani Murmu' had her 'face superimposed onto an obscene video, before she was elected last year' (n.p.). One piece of Indian news coverage on this issue noted 'If Indian politics is a negotiation with entrenched power structures and multiple bigotries', the experiences of Murmu and a fellow female leader show 'how easy it is to turn on the tap of prejudice to discredit women in public life' (*Editorial*, 2020, n.p.). A recent study by Pawalec (2022) is an important exception to the trend in the literature of separating gendered harms and other forms of harm. In using deliberative democratic theory, they argue that the existence of pornographic deepfakes 'discourages certain societal groups, in particular women, from participating in the public sphere—aggravating existing discrimination' (p. 19). This perspective, although relatively uncommon in the literature, goes some way to beginning to bridge the gap between individual and collective notions of harm.

OSINT Data and Policy Coalitions

One group of policy actors who have arguably shaped the discourse on gendered harms within synthetic media is the OSINT analyst community. As previously noted, this is a wide moniker for researchers – both hobbyists and professionals – who work in 'collecting and analysing information gathered from open sources to produce actionable intelligence' (European Data Portal, 2022, n.p.). Beyond expected uses, such as tracking deepfake videos in the Russia-Ukraine war (see Gleicher, 2022) – they have had a unique impact on the developing understanding of gender and information operations. An analysis by SensityAI, a 'deepfake detection company' (Burgess, 2020, n.p.) reported in 2020 that 'at least 104,000 women' on Telegram have been 'targeted' via a deepfake pornography bot. Hard statistics such as this one have received considerable exposure in news coverage of synthetic media – and helped to construct narratives for the loose policy coalitions emerging in the area. OSINT analysts' own perspectives on this debate matter because their reports on deepfakes can be (and have been) used as evidence in national-level policy proposals on synthetic media. For an example of this, see how the frequently quoted '96% of deepfakes' statistic was used in a report discussing US

regulatory proposals on the technology – including the previously mentioned *Deepfake Task Force Act* (Jeong et al., 2021, n.p.). OSINT analysts' potential influence on the development of policy coalitions should not be underestimated – and forms a key aspect of this study's practical focus.

Ethical considerations behind any attempt at regulatory intervention on synthetic media harms are incredibly complex. As argued above, these efforts involve much deliberation between which harms are individual and which are collective – leaving a significant conceptual gap to be filled. The technology's uses in conflict scenarios and individual abuses – as it will be demonstrated later - frequently overlap. These boundaries and grey areas connect to much disagreement between policy actors in the space. Regulatory approaches to this technology also arguably require a position on its epistemic threats, which are impossible to address sufficiently here, but important to consider regardless. Additionally, OSINT analysts as a disparate group have been and will be argued as relevant to policy actors' prioritisations of synthetic media harms. However, their work, as it will be shown, has not come without criticism for its frequently renegade nature. The aim of this research is not – and cannot – be to sufficiently address all core issues within this varied group of debates. My aim is rather to understand how these debates are grappled with by certain policy actors in practice – and to support the nascent understanding of policy coalitions being developed from these core themes.

CONCEPTUAL FRAMEWORK

This dissertation's approach, as noted above, borrows from multiple disciplines – epistemics (see: Kavanagh and Rich, 2018; Seger et al., 2020), media theory (see: Ajder and Glick, 2021), deliberative democratic theory (see: Chambers, 2003; Pawelec, 2022), and gender studies (see: Meaker, 2019). Epistemic theories have led me to be interested in how policy actors conceptualise vague and large-scale harms to knowledge bases. Deliberative democratic theories have led me to consider how policy actors who address synthetic media view the social contract between a regulatory body and its citizens. Finally, perspectives from gender studies have prompted a particular interest in how online harms to women and feminine-

presenting people through synthetic media are seen as individual or societal in nature, and why.

This variety of approaches is taken because online harms themselves are a multidisciplinary – and multi-causal, policy problem. In discussing research relevant to proposals for deepfake regulation, this study adopts a specific definition of what 'policy' refers to. Policy as it is used here 'is the result of deliberate choice between alternative ways of proceeding' which 'results in action (or deliberate inaction)' (Oxford Policy Engagement Support, Guidance Note 2, n.p.). This definition leaves open the possibility of influence from a range of people – referred to as 'policy actors' (Ibid). There are a variety of policy actors involved in synthetic media discussions – regulators, think tank researchers, victims of online harms, and OSINT analysts alike. All these categories of actors will be interviewed for this study.

A key aspect of debates on online harms is whether they are societal/collective or individual in nature (see Smuha, 2021) – and whether this differentiator is even productive. Smuha (2021), writing on AI governance, defines individual harm as that which 'occurs when one or more interests of an individual are wrongfully thwarted' (p. 5), and collective harm as that which 'occurs when one or more interests of a collective or group of individuals are wrongfully thwarted' (p. 5). This differentiation may seem clear as it is written by Smuha, but its application in practice is not. These contested categorisations are argued here to be relevant to every stage of the policymaking process – and to the final result of harm-prioritisation by regulatory agents. The intended conceptual contribution is an expanded understanding of the rationalisations behind individual and collective harm-labelling in the context of synthetic media. Gendered aspects of online harms perpetrated through synthetic media appear to require a more nuanced judgement of their potential collective impact. Potential consequences of image abuse, such as damage to the ability to express oneself or reluctance to participate in political discussions, will be critically considered within the context of policymakers' prioritizations of online harms. Although this critical framework is relevant to and embedded within the study's analytical perspective, interviewees will be encouraged to express

unexpected perspectives outside of the framework's constraints - and to counter its presumptions.

RESEARCH OBJECTIVES

For years it has been noted that gendered harms should play a significant role in discussions on synthetic media (ie: Chesney and Citron, 2018). However, I have not noted any studies which have critically mapped the rationale for the prioritisation of these harms within wider epistemic and political conversations around synthetic media. Given this research gap, this study aims to understand how gendered harms are prioritised within the wider regulatory conversation. I hope to contribute to an understanding of a specific, limited, and contemporary debate space between tech and policy elites within liberal-democratic contexts.

Key hypotheses are as follows:

(I) The desire to preserve epistemic security in the online information landscape is expressed as a motivator for policy actors to engage in advocacy relating to synthetic media.

(II) The desire to preserve the feasibility of deliberative democracy is expressed as a motivator for policy actors to engage in advocacy relating to synthetic media.

(III) Gender-specific harms that can be perpetrated via synthetic media technologies are more frequently categorised by policy actors as individualised than collective in nature.

The overall aim is to enhance an understanding of how policy actors justify and rationalise their stances on the varied harms that can be perpetrated through synthetic media. Research questions are:

RQ1: How do policy actors conceptualise and rank their main concerns for synthetic media technologies' current and potential online harms, and with what underpinning rationales?

RQ2: How do policy actors conceptualise gender-based online harms compared with other harms, and what justifications do they provide for any differentiation in their prioritisations?

RESEARCH DESIGN AND METHODOLOGY

This chapter outlines the research design, methodological approach, and form of data analysis used. The method's limitations and ethical considerations are then discussed. Finally, the researchers' reflexive position and influence on the research are critically analysed.

Methodological Rationale

My goal was to speak to influential people who had recently written about, researched, or debated policy stances on synthetic media technologies. I wanted to speak with a generally equal combination of OSINT analysts, synthetic media technologists, policy professionals, and academics/think tank members. The aim – to understand policy actors' conceptualisations and rationalisations of varying online harms within synthetic media policy discourses – required a conversational and customisable approach. This study consists of ten qualitative semi-structured interviews with policy actors from a variety of backgrounds – government, industry, and non-profits alike. Gaskell (2000) notes that one outcome of the interviewing method is 'a 'thick description' of a particular social milieu' - and that is what is intended here. Interviewing was chosen as the most suitable method because subjects' internalised rationales - and the revealing responses that follow – required methodological flexibility.

Originally, the idea of a focus group appeared appealing for its ability to spark dialogue between professionals, but the sensitive topics and frequent redactions that were required made a shared space of dialogue ethically unmanageable. Warren (2001) argues that 'the purpose of qualitative interviewing (and associated fieldwork) is to understand others' meaning making'. In this way, it was necessary to be open to unexpected themes and responses. Gaskell (2000) similarly notes that the method's aim is 'a fine-textured understanding of beliefs, attitudes, values and motivations in relation to the behaviours of people in particular social contexts' (p. 3). Rather than follow a rigid interview guide, a semi-structured approach was chosen so that questions could be customised to participants' fields of expertise and research outputs.

The epistemological approach here is constructionist and is based on a postmodern rationale wherein 'the qualitative research interview appears as a construction site of knowledge' (Brinkmann and Kvale, 2018: 24). This perspective, therefore, opposes any notion that interviewing is an exercise in the extraction of existing knowledge. The conversational aspect of the interviewing method lends itself to a conception that the knowledge produced is a form of narrative construction. Unlike positivist epistemologies, which generally aim to reduce 'influence by the person of the researcher' (Brinkmann and Kvale, 2018: 12), postmodern and constructionist approaches to interviewing see the interrelation of perspectives as key to the knowledge-generation process. This is not to suggest that all influence from the researcher is beneficial – a critical view of the researcher's influence is presented in the section on reflexivity.

Findings from this study apply to a specific temporal context, and to a small group of elites within that primarily liberal-democratic and English-speaking context. Sweeping conclusions on the opinions of policymakers or OSINT analysts as a group cannot be extrapolated from this data. However, insights gained from this research may help to inform studies that interview policy actors in similar contexts. One unexpected problem that arose in the interview process was that some interviewees were willing to disclose details of their work, but not willing for these details to be included in the research itself. This meant that it was important not to allow my knowledge of these disclosures to unnecessarily influence my analysis of existing data, given the interviewees' redaction requests.

Methods and Procedures

This section overviews the sampling and data selection strategy, the development of the interview guide, and the use of thematic analysis on the interview corpus.

Sampling and Data Selection

The sampling process was approached through a combination of convenience and snowball techniques with the hope that multiple forms of outreach might lead to a group with a higher diversity of opinion. First, I reached out to the authors of publications which were most closely related to my topic of interest, based on prior research from my literature review

process. This initial outreach strategy provided me with eight interviewees. Another interviewee was a recommendation from a member of the prior group, and the final interviewee responded to a request that I had posted on a professional forum. I ended sampling once I reached ten interviews due to a level of data saturation and participant variety sufficient for my research aim and time limitations. I was able to include perspectives from four continents, but my interviewees were mostly, like me, English-native speakers from privileged backgrounds. If I were to redo my recruitment process, I would search for interviewees within more regional online groups that were less likely to fall into my outreach pool from the beginning.

During the interview process, there were a handful of instances in which I was asked to redact a specific disclosure. Two of the ten interviewees requested to review my final selection of quotes due to their desire not to be identifiable. Participant 6's organisation required the right to make stylistic edits before their quotes could be included. None of these alterations were substantial in nature. The data itself is best conceptualised as representing a specific, hyper-current conversational sphere of tech and research elites in the English-speaking world. There is much regulatory conversation happening in the EU, China, and elsewhere which is beyond the scope of the study. That limitation is necessary to keep in mind when considering data scope and applicability.

Development of the interview guide

Following Döringer's re-conception of Bogner and Menz' (2009) 'theory-generating expert interview' (2021: 266), the development of the interview guide began with a search for core tensions in the relevant literature. Döringer (Ibid: 266) argues that 'individual perceptions and orientations of experts are seen as essential for shaping social practices in a field of action.' In this vein, the interview guide was constructed to allow interviewees to negotiate their place within these spectrums of opinion, as well as for them to introduce other tensions specific to their expertise. The interview guide was customised for each interviewee based on prior reading of their research output. All questions asked are available in Appendix D – including some that were originally asked to the first few interviewees, but later redacted for repetitiveness or an inability to provoke productive response. Most interviews lasted between

45-60 minutes, although three required a thirty-minute limit. All interviews were recorded remotely and transcribed with a combination of GDPR (General Data Protection Regulation) compliant AI software and subsequent hand-correction.

Thematic analysis

The chosen method of interview data analysis was the thematic approach. As these interviews explored experts' prioritisations of various online harms relating to synthetic media, it was essential to break down interview data into relevant thematic categories for comparison. Döringer (2021: 274, referencing Bogner & Menz, 2009/18) notes that 'experts represent 'a complex interdependence of knowledge and power''. Thematic analysis therefore had to be approached with an awareness of interviewees' motivations and power-positions, and how these may have strategically influenced their prioritisations of certain forms of online harm. A deductive approach was first employed in the literature review to locate core themes. However, thematic analysis is a primarily inductive method, and the sub-themes used in the final analysis were gleaned from knowledge gained in interviews.

The coding process was achieved through a straightforward digital highlighting technique of overarching themes. In cases where a quote was relevant to more than one theme, it was duplicated into each category so that the framework was not unnecessarily prescriptive. Once all overarching themes were coded, a document of each quote within these themes was then further analysed and divided into relevant sub-themes, which are visible in Appendix D. At this stage the commonalities and differences in interviewees' expressed opinions, and their underpinning rationales, were analysed. A multi-step approach to coding allowed for a flexible framework to develop over time. One weakness of the thematic analysis approach was that, as experienced by Guest, MacQueen, and Namey (2012: 26) I was 'simultaneously confronted with a richness of data and real constraints on (my) ability to analyze them.' The volume of data acquired made it challenging but necessary to maintain focus on relevance and a feasible overall scope.

Ethics and Reflexivity

The ethics approval process involved additional REC approvals and data security reports due to the sensitive nature of information operations and image abuse. All interviewees were provided with - and signed – a comprehensive information sheet and consent form. My main concerns were protecting the identities of interviewees and preventing the unnecessary risk of retaliatory deepfake abuse. The latter concern was mentioned by two female interviewees who discussed their public-facing work - and was taken into serious consideration. Redactions of any reasonably identifying aspects of the interviews were made meticulously. The topic of image abuse was approached sensitively. In the case where an interviewee had experienced it themself, I asked at the beginning of our discussion if any topics were off-limits.

Dodgson (2019: 221, quoting Berger, 2015) argues that reflexivity is a 'conscious and deliberate effort to be attuned to one's own reactions to respondents and to the way in which the research account is constructed'. Deepfakes have been a theoretical interest of mine for years now, and this intense interest may have clouded out other relevant forms of synthetic media - such as synthetic audio and data. I also have personal experience with the OSINT community, and to balance this proximity bias, I took care to ask critical questions about their research impact. Overall, as a native English-speaking, educated, white researcher, my advantages and blind spots have influenced my networks of influence and eventual pool of interviewees. In awareness of this, my outreach method included cold-contact methods as well as postings across multiple open-access channels. This mitigation was not a cure-all, and I realise upon reflection that more should have been done to encourage a diverse interview pool.

ANALYSIS AND DISCUSSION: DIVERGING PRIORITISATIONS

This dissertation asks what underpins policy actors' prioritisations of the online harms that arise from the use of synthetic media in influence operations. Do they view some forms of harm as more urgent to act upon than others? Do they differentiate in their prioritisations of large-scale and individualised harms, and if so, what is their justification? From ten interviews conducted with high-influence professionals in the synthetic media space, areas of

significant consensus and disagreement emerged. Major inductive and deductive themes - available in Appendix C - are first examined, with relevant quotations then analysed. Excepting 'gendered harms', every topic area originally expected within the data was discussed by every participant. In addition, each topic was the primary focus of at least one interview within the dataset. Gender-specific harms were passionately focussed on by some participants, and not particularly on the radar for others. This contrast in focus, prioritisation, and rationale between participants and across themes is discussed in detail below.

Fears for Epistemic Security

All interviewees expressed, in one way or another, a concern for the stability of collective knowledge bases when faced with synthetic media. Some interviewees also expressed a wider fear of the deterioration of reality – such as Participant 1's reference to Baudrillard's Simulacra and Simulation (1983, trans. Glaser) - despite the interviewer not mentioning it:

> ***P1:*** *(speaking of their research on deepfakes)* So, you know, very much kind of Baudrillard's simulacrum and that kind of stuff, right?

The epistemically destabilising effect of synthetic media was expressed in one way or another as a concern held by every interviewee in this study. Epistemic concerns are widely discussed in this study on synthetic media, but interestingly, these appear to be less frequently mentioned in other studies that analyse the broader discussion on AI. One broad study focussing on 'rhetorical dynamics in AI' (Imbrie et al., 2021) did not mention truth decay, epistemic security, or any other closely related phenomenon in its overview of dominant AI-related discourses up to 2020. It did, however, share commonalities with other themes detected in this study, such as worries related to AI in conflict. Perhaps this is because that study was released a few years ago when the AI and epistemics discourse was less developed. Another possible explanation is that since it was published before the generative AI boom of the past year or so, concerns around replication and AI realism were less predominant. It isn't clear what the cause of this difference is, but it appears productive to keep in mind.

Truth Decay

The first subtheme of this wider topic is truth decay, defined by Kavanagh and Rich (2018: x-xi) as consisting of four specific trends. The first aspect, 'increasing disagreement about facts and analytical interpretations of facts and data', is reflected in Participant 7's statement that:

> **P7:** ... it feels that it's… there is a harm that comes from directly believing what that synthetic media is, but there is possibly a greater harm which comes from the - kind of - undermining trust in the collective knowledge base and this, you know, the splintering of people into kind of fragmented communities of people that they trust and the creation of information silos.

The idea of 'undermining trust in the collective knowledge base' (P7) is closely reflected in the second aspect of the truth decay phenomenon - 'a blurring of the line between opinion and fact' (Kavanagh and Rich, 2018: xi). In this vein, P2 similarly fears that the public-informing and accountability functions of human rights observers could be weakened:

> **P2:** *(deepfakes are)* going to have all sorts of ramifications from human rights observers' inability to - or undermining their ability to – inform international audiences about human rights abuse because a lot of that is based on images *(…)* at the most basic level, it will undermine the trust of any image that we see.

The third element of truth decay is argued to be 'the increasing relative volume, and resulting influence, of opinion and personal experience over fact' (Kavanagh and Rich, 2018: xi). This change, from the influence of 'fact' to the influence of 'opinion' is similarly expressed in P9's concern for the degrading effect of synthetic media.

> **P9:** But really my view is that anytime something like this is used to harm someone privately (…) that degrades trust in our information ecosystem and what is real, what is fake – and just like with disinformation, the less - I guess - society trusts what information they're seeing, or the more easily they're able to pick and choose what they like to think is true, the more destabilized society becomes, and that has huge national security risks (…) threat*(s)* - to epistemic integrity...'

P9's fear of societal destabilisation is based here on information consumers' proposed tendency to 'pick and choose' in the face of a generally untrustworthy information

environment. There is an interesting tension here between an information consumer's seemingly unobjectionable right to autonomy – to choose what they believe, and when – and the way this autonomy is framed by P9. They may have been referring to information consumers' risks of polarisation, or perhaps to consumers' tendencies towards less 'legitimate' sources of information when trust is lost in mainstream sources – it isn't entirely clear here.

The final element of the proposed truth decay phenomenon is 'declining trust in formerly respected sources of factual information' (Kavanagh and Rich, 2018: xi). The fear that synthetic media would lead to a decline in trust was expressed by many interviewees. A majority were concerned about the large-scale epistemic risks of synthetic media – risks involving conflict, the 'collective knowledge base' (P7), and a reduction in accountability for online and offline harms. Participant 4, a high-level professional at an OSINT-based firm, presented a more moderate viewpoint:

> **P4:** The amount of effort and then the quality level that's required to do it *(deepfakes)* at sort of an illicit level, I think it's still very high. So, I'm not quite as worried now, because even if that technology is out there, I still think it's going to be very hard for *(them)* to pull off like more than once, right?'

Participant 4 expresses the belief that a high barrier to entry, as well as the likelihood that repeated successful operations would be challenging to hide, may both limit the potential harms of synthetic media. So, amongst interviewees, the fear that synthetic media will inevitably lead to truth decay was predominant, but not always expressed without caveats.

Suggested accountability-based solutions

Literature suggests that policy actors disagree on who is burdened with the duty of care in building authenticity infrastructure. Some nations believe that the state should develop these detection requirements and directly outsource to providers (Interesse, 2022), whereas others believe that platforms should bear more responsibility (see: European Science-Media Hub, 2021). Some broadly view detection processes as problematic to begin with (Leibowicz, McGregor, Ovadya, 2021). However, most of this study's respondents suggested accountability-based solutions. One interviewee who works in the synthetic media industry supported the idea of audits, and bluntly suggested moving away from individual

accountability, noting '...You know, reports, audits, et cetera are essential for our industry to navigate in this, you know, very big mess' (P5). Interviewees within private industry all expressed some degree of eagerness for more compliance infrastructure. However, there was tension as to the feasibility of media literacy. Participant 5 argued that marking something as synthetic...

> **P5:** ...would be much more effective than trying to educate, you know people - because again, you have stupid people, you have smart people, you have people that have time, you have people that do not have time...

Their accountability-based suggestion appeared to place the burden of compliance onto verification companies, internal audits, and the like – as opposed to individuals. However, this attitude was not shared consistently across the interviewee pool. Another interviewee with a background in government cautioned against prescriptive source-ranking, noting:

> **P3:** …I'm not into fact-checking. I don't want anybody to rank sources in terms of trustworthiness. I think all of that just drives polarisation deeper. It drives greater mistrust in the organisation that's spreading all that stuff. I want people to be able to make those decisions for themselves

So, although accountability-based solutions were generally offered, there was no clear, singular consensus on whose shoulders accountability itself should fall on. Some interviewees felt that not all individuals could be expected to gain the skills needed to navigate their information environment – and therefore audit mechanisms and tools that tag synthetic media would be more productive. Still others, like P3, were wary of the potentially Orwellian nature - or perception - of prescriptive solutions to defining source legitimacy.

Interviewees who spoke from a concern for gender-based online harms had a distinct perspective on the issue of epistemic security. One noted:

> **P6:** Even if it's very clear - and it's labelled and that this isn't actually a picture of *(own name)* engaging in this material, there's nevertheless that attack on my dignity and my bodily autonomy that you've taken my image and put it in that context without my permission

Labelling synthetic pornography - as an attempt to maintain some epistemic security in the information environment - was not considered sufficient by participant 6. Wider harms – namely the 'attack on my dignity' (P6) were not thought to be resolved from labelling alone. Another participant with a gendered-harms perspective noted that:

> **P3:** …by tackling the women's image abuse issue we're gonna be - maybe not outright solving, but at least testing - for the solution to many of the other AI issues. We're gonna learn how oversight works better, which rules…

In this way, they suggest that women's abuse, as a subset of other epistemic threats being prioritised for regulation, could serve as a test-scape for wider solutions. This harm-prioritisation rationale of addressing one issue as a test for others was relatively unique to interviewees with a gendered-harms perspective.

A core hypothesis of this research is that the desire to preserve epistemic security in the online information landscape motivates policy actors to engage in advocacy on deepfakes. The hypothesis, in part, can be answered from these results. From this group of experts, epistemic security was a significant concern - each participant mentioned epistemic risks in some way or another. For some interviewees, this was the concern. An unresolved but interesting tension within this theme is the disagreement as to which accountability mechanisms – individual or societal - would be more suitable for the prevention of harm. This is not to suggest that regulation requires a firm choice between the two. This is to suggest that more conversations would be beneficial as to what an ethically justifiable approach to these fears looks like – and on whose shoulders accountability should lie. Responses also suggest that general epistemic security approaches – such as image verification – may not be sufficient to address scenarios of gendered harms. Harms relating to image abuse, such as 'that attack on my dignity and my bodily autonomy' (P6) may require interventions beyond the lens of epistemic security and verification alone.

Fears for Deliberative Democracy

Most interviewees expressed concern for the feasibility of democratic deliberation if citizens lack a shared and trusted information base through which to make decisions. This

concern relies on 'a central tenet of all deliberative theory' which 'is that deliberation can change minds and transform opinions' (Chambers, 2003: 318). However, as Chambers notes (Chambers, 2003, referring to Mansbridge 1996, Sunstein 2002), 'Perhaps deliberation sharpens our disagreements, intensifies social competition, and polarizes opinion' (Ibid: 318). Despite the suspicion amongst some theorists over the potential validity of deliberative theories, these ideas, and the threat to their core tenet, hold firm sway amongst most interviewees. Some interviewees shared fears of the development of an environment antithetical to collective governance. Election interference was a common concern:

> **P2:** Creating a photograph of Fauci hugging Trump... I mean, as far as moral violations go, that's pretty on the low end... *(But)* for Russia to create a deep fake that's intended to subvert a US election and submit it like, just days before the election with obvious nefarious and geopolitical intentions...

Interviewees frequently distinguished between degrees of harm that a piece of synthetic media may cause – from relatively harmless satire on one end, to nation-destabilising events on the other, like the Russia example above. A similarly nuanced view of synthetic media in politics was constructed by WITNESS and MIT's Co-creation Studio in their report on satire and synthetic media (see: Ajder and Glick, 2021). This report brought to mind that a blanket fear of synthetic media as used in political contexts may be unproductive, given that satire itself is often used as a form of political expression. This is not to suggest that interviewees did not express a sensitivity to other political uses of deepfakes, but to suggest that their primary focuses on democracy and deepfakes, to the point of potential hyperbole, were on concrete, nation-destabilising threats. Solutions to these threats, as shown below, were generally proposed cautiously, and with caveats.

Pessimism around government interventions

The general attitude to reform was either cautious or, in some instances, relatively fatalistic. This fatalism, in the wider conversation on Western democratic nations, was primarily focussed on the US and its hurdles to effective regulation:

> **P2:** The US is a lost cause *(…)* Yeah, I mean, the legislation can't do *(expletive)*. So, the notion of passing complex regulatory regulation... the political environment here is very limited.

Another interviewee, noting a deepfake legislative proposal in Washington, stated:

> **P9:** …I think she *(Rep. Clarke)* submitted that three times, and it's obviously not been passed *(…)* I'm not very hopeful.

In contrast to the above, some interviewees, primarily in academia, noted that they thought government would be the most effective regulatory actor. This was mentioned to a former federal employee, who replied, laughing: 'Have any of those people ever worked in government before?' (P3). Another concern was that much regulatory progress 'will be rolled back depending on, like, how a bunch of elections go around the world this or next year' (P3). The regulatory environment was conceptualised as volatile in some ways, and rigid and stuck behind in others. A general pessimism was countered by Participant 4's concern that one cannot 'just reliably expect companies to sort of do the ethical or moral thing and just not put it (deepfakes/deepfake models) out there...'. This suggests that participants see regulation as a sort of balancing act between industry and state – with the main point of contention being which is more fit, or more motivated, to act ethically.

Consciousness of unique regional vulnerabilities

Although most discussed these risks in a Western context, some interviewees brought in useful - and generally underreported - perspectives on other regions' vulnerabilities. One technologist noted that Africa will experience 'suffering from a lack of regulation', despite the suggestion that 'AI could play - and synthetic content could play - a huge role there'. Similarly, another interviewee stated:

> **P9:** …while here in the US, we might worry about someone influencing a few votes on an election or sparking Black Lives Matter protests – in other places like in India, these things could spark massacre

This response suggests that regulation may benefit from a nuanced view of the geopolitical impacts of deepfakes, given regional variations in potential destabilisation. Another interviewee noted a connecting concern for how the 'online space' itself is used in the context of women's political participation:

> **P9:** Particularly where you have authoritarian *(…) (or)* deeply socially conservative regimes where women's physical participation is limited in many ways, in many of these contexts the online space is becoming a really important vehicle for women's voice*(s)*

This interviewee's concern appears to be for the relative importance that online expression holds for those who otherwise may not have access to traditional forms of political engagement. Their comment is particularly relevant to synthetic media technologies that create sexual content, considering variations in political freedoms and morality-driven punishments into silence.

Suggestions for increased international cooperation and citizen involvement

Despite relative pessimism in the regulatory space, interviewees did encourage the notion that international agreement was needed, and worth aiming for.

> **P8:** ...either multiple countries simultaneously could influence the big tech companies that are working on those things, or maybe some *(private sector companies could get involved)*. So like internal rule building, right, to influence multiple governments and go with a single front by multiple governments...'

A policymaker who focusses on online harms noted that some international cooperation on online safety has been underway – referring to the 'new Global Online Safety Regulators Network, which took inspiration from the equivalent network of existing privacy regulators' (P6). The mention of this network, although not specifically aimed at synthetic media alone, suggests that this policymaker may be attuned to other actors' desires for international cooperation.

Many interviewees were also keen to note support for the spotlighting of citizens' voices in the policymaking process.

> **P5:** But there is one part that is always missing and it's the users, the civilian society. They are never asked about how *(deepfakes will affect them) (…)* In theory we would like, I mean, to create this triangle of, you know, discussions - but in practice it cannot happen because of the nature of our democracies.

So, participant 5 notes that 'civilian society' is absent from many regulatory discussions but is not particularly optimistic about their potential inclusion. Normative desires for citizen participation did not always line up with the reality of what technologists in industry said on the subject. Participant 5 continued, noting:

> **P5:** I would say that the idea to involve users *(in deepfake discussions)* has been expressed many times and I'm not aware at the moment unfortunately... and that would be a nice surprise to know that it is happening.

This technologist, although not aware of any user-involvement scenarios in discussions around deepfakes, did express a firm sense of responsibility over what his company's technologies may be used for, stating 'We have our responsibility to instruct and make sure that this technology sits well - again - in the social contract' (P5). The idea of instructing, as opposed to involving, was a visible tension amongst interviewees in discussions of citizens' roles in the policymaking process.

Discussions of gender-based political vulnerabilities

Regional vulnerabilities and suggestions for international cooperation are further complicated when viewed through the lens of gendered harms. Participant 9 expressed the concern that deepfake pornography:

> **P9: …**could have a chilling effect, especially on politicians who are… you have to be very concerned with the public image, especially women...

Fears of the 'chilling effect' (P9) showed up with far more frequency in discussions of gender-based harms than general political harms. The same interviewee noted that, in their publication of deepfake research, they feared retaliatory deepfake pornography. They said that risk 'would never stop me from thinking or talking about it' but that they did wonder '...is this something that I'll have to handle - I'll have to deal with, I mean, personally?'. Their statement suggests that the chilling effect – even if braved by this researcher – did come into consideration in their own lived experience. Participant 9 also stated that they feared deepfake image abuse is 'a bellwether for how this *(synthetic media)* could be used to chill free speech more broadly.'

In earlier discussions of the democratic effects of synthetic media, there was a general pessimism around any regulatory interventions. However, interviewees who spoke about gender-based, sexualised online harms expressed a firm desire for something to be done about it. Important to note is that interviewees suggested particular harms may come to those in vulnerable regions, and to women. If those two groups are combined, it is reasonable to be concerned about intersectional harms that combine aspects of both of those vulnerabilities. A core hypothesis within this research is that concerns for the feasibility of deliberative democracy motivate policy actors to engage in policy advocacy on deepfakes. This fear was shared by a vast majority of interviewees. Some approached it from a pessimistic regulatory lens, some viewed it in terms of regional vulnerabilities, and still others analysed it in the context of women's political participation. This section suggests that particular regional and gender-based vulnerabilities would likely benefit from a targeted focus within the wider conversation on the democratic impacts of synthetic media.

Changing norms of influence operations

> **P3:** ...only when it started to get more broad and affect not just women is *(society)* starting to have that conversation...

'Strategic information operations', as defined by Starbird et al. (2019), '...encompass efforts by individuals and groups, including state and non-state actors, to manipulate public opinion and change how people perceive events in the world by intentionally altering the information environment.' A key element of this definition is the intention to manipulate public perceptions – a concern which has shown itself in interviewees' fears around the use of synthetic media in conflict scenarios. Policy actors unsurprisingly disagree on whether or not militaries can justify using deepfakes in their own operations, given that their adversaries are doing so. Some view the USA's participation in a potential deepfake arms race as representative of a 'next generation propaganda aspiration' (Biddle, 2023). Is it better to trust one's own government to take the technological offensive, or is it instead preferable to institute a blanket refusal on safety grounds? Almost all interviewees mentioned a concern for the trajectory of influence campaigns amidst increasingly realistic 'deepfake' technologies. Some also expressed wider fears of changing ethical and practical norms of warfare.

Troubled moral rationalisations of deepfakes in conflict

When asked for their perspectives of the use of deepfakes in warfare, most interviewees stated that it was either inevitable or rational given defence dynamics and competition. Participant 5's statement suggests that the relationship between military apparatuses and AI research made any discussions of their potential banning unproductive:

> **P5:** ...Most of the research that is done on AI comes from the military. So, it means that - whether or not - you will not be able to control and to forbid the use of deepfakes during a conflict.

Another interviewee felt that the social contract that a democratic country has with its citizens makes the idea of deepfakes in warfare challenging for it to justify:

> **P9:** the fact that it *(a redacted democratic country)* does have a very different social contract with its people than, let's say, an authoritarian country, might... I think that the potential damage, let's say public confirmation that they use technology like this - is so much... so far outweighs the possible benefits…

This discomfort with using deepfakes in warfare within a democratic context was also shared by Participant 3, who noted that influence operations such as this add 'fuel to the fire for authoritarians who wish to use 'what-about-ism' in their arguments against the United States and other Western democracies.' However, the generally accepting tone of most reflections on this topic can be summarised by Participant 1's statement that 'Deepfakes are an extension of media manipulation which is as old as time, right? *(...)* This is just a new way of manipulating.'

Another concern around the usage of deepfakes in warfare related to the prosecution of war crimes:

> **P1:** ...when you're potentially doing war crimes prosecution, the ability of defendants to potentially claim that any captured footage or video evidence of them committing those war crimes is (a) deepfake...

This concern connects to the earlier theme, epistemic security. The epistemic harms possible from synthetic media, as rationalised by this interviewee, are not limited to a conflict arena alone. Two main differences to the above consensus are as follows. Participant 10, a technologist, notes that 'there should be a general stigma around it' but that they would prefer

to 'focus on building the differential tech to avoid this because *(of)* the complacency that may set in because you're thinking that no one would use it.' This perspective brought the previously unmentioned idea that complacency could be a side effect of comprehensive regulatory strategies against deepfakes in warfare. If no one thinks they're out there, this interviewee argues, defences may not be put in place.

Another interviewee with knowledge of conflict dynamics tempered the generally myopic focus on synthetic media in warfare, noting:

> **P8:** ... *(Deepfakes are)* not the main concern because there are like far, far worse and at the same time conventional things happening, but it might become a problem

Similarly, one interviewee who specialises in conflict-affected and technologically underdeveloped states, said that in these scenarios '...synthetic media is not the major risk that is facing the information, you know, environment in those contexts' (P7). This served as a reminder that deepfake technologies, given their complexity, are not internationally ubiquitous in conflict scenarios. There was not a universal expression of inevitability as far as deepfakes in conflict were concerned, although most respondents saw their use as rationalizable given competition dynamics. There was rarely a firm consensus expressed around the usage of deepfakes in warfare, and no confident policy coalitions appeared to be developing from what was discussed by this group of experts.

Participants who focus on the gendered harms of synthetic media expressed frustration with the national security community. Participant 3 noted a desire to 'point out kind of the huge blind spot that the policy community, in particular the national security community, has' in conversations around gender-specific harms. Interviewees with this perspective argued that gender and conflict are intimately related. Participant 7 stated 'I think where we're seeing it more predominantly is in this evolution of sexual violence in conflict'. This concern relates to the earlier mention of a female Indian politician who was deepfaked (*Editorial*, 2020) - similarly to 'real' violence, digital sexual violence can, and has, been used as a tactic for silencing women's voices. Participant 3 similarly noted 'I have again been frustrated for a long time about the lack of understanding about how gender is used by hostile states in their influence

operations'. Cultural conceptions of shame, morality and suitability for leadership were all argued in some way to be important when considering synthetic media in conflict.

One expert noted that there have been discussions of the gendered origins of deepfakes, 'but really nothing happening in the defence space' (P9) on that theme. An OSINT analyst who researches conflict scenarios noted that they come across deepfakes 'in geopolitical terms' (P8) but that they are aware image abuse 'is a huge problem and challenge from the policy perspective.' This statement, although understandable, represents a discourse that the above interviewees expressed frustration over – the separation of geopolitical and gendered online harms.

A core hypothesis of this research has been that gender-specific harms are more frequently categorised by policy actors as individualised than as collective in nature. It is not possible, from this small sample size of interviewees, to speak to the wider policy environment. However, it is meaningful to note that from within this sample, a majority discussed to some degree the collective implications of gendered harms. Although some interviewees did not explicitly draw the connection between gendered and geopolitical harms, almost all interviewees expressed an awareness of them. Compared with generally stereotyped gender ratios in conflict studies and national security communities, this awareness may be related to an almost equal gender balance within my interviewee pool. However, it is not possible to be confident that this is the case. Regardless, it appears productive to consider the benefits of women's voices within the broader conversation on synthetic media and conflict scenarios.

Desires for OSINT-Based Literacies

Most interviewees discussed the potential utility of OSINT (open-source intelligence) techniques for synthetic media detection and/or accountability. Many presented examples of OSINT's broader utility in detecting disinformation and influence operations. Interviewee 2 specifically argued in the context of the Russian invasion of Ukraine that:

> **P2:** ...OSINT tools have been critical on a number of different levels *(…) (in outing)* Russian fake content, like, some of their attempts to show Ukrainian mistreatment of prisoners or whatever, right?

The concept of 'outing' these operations was reflected in interviewee 4's belief that OSINT analysts' accountability functions have influenced considerations to deploy deepfakes.

> **P4:** …there's just an arsenal, like, an army of people online that are just going at it and finding things that you could never imagine. And there's like that instant accountability. I never thought that was going to happen. So, I actually love that. *(…)* I do believe that that has impacted the countries' calculus of whether to use them or not *(…)*

Their perception as an 'army of people' (P4) who carry out a digitised form of mass accountability, however, is not always well-received. A professional within an OSINT-based organisation noted that some analysts in the context of the Russian invasion:

> **P9:** …have come out and been like 'ohh yeah, here's my hot take with, like, minimal amount of data, but lots of the very authoritative sounding…

So, the lack of accountability of some more casualised OSINT actors, as opposed to those operating within legitimised networks and organisations, is seen by one insider as a source of concern. However, OSINT professionals generally suggested that, if the investigation is traceable and uses legitimate techniques, it should be taken seriously.

OSINT skill sets, such as image verification, geolocation, and the like, have been described by most interviewees as potentially useful as part of larger media literacy initiatives. However, this majority view was caveated by one interviewee, who noted that '...the onus is also on, like, media companies to sort of validate themselves...' (P4). So, the balance between collective responsibilities and individual responsibilities was yet again a point of contention. OSINT vigilantism, in the context of wider social responsibilities towards the information ecosystem, may understandably be perceived as a simplistic and individual solution to a collective problem. However, one interviewee noted that concrete data and examples can be used for larger-scale change in policy, noting on the regulation of deepfake pornography:

> **P3:** What we're told a lot by tech innovators is 'Ohh, that's really hard', right? But if we show that it *(regulation)* can be done with a problem that already exists, rather than just saying, 'OK, this is what we think would work' and then them pushing back on us, which is what they did *(with)* the Online Safety Bill in the UK, right? *(…)* 'This is too hard. We can't do that. We don't have this data.' *(…)*

In the vein of policy impacts, P1 noted that OSINT sometimes '...feels more like journalism that is telling a story rather than research which is pointing to a natural conclusion for policymakers.' This in some ways contrasts with P3's hope in concrete data's influence – in the sense that P1 does not see much explicit advocacy coming from these data-producing organisations. Generally, OSINT analysis and its utility for broader media literacy goals was one of the most frequently mentioned accountability-based suggestions for harms that can be perpetrated through synthetic media.

One impact of OSINT work that appeared influential to interviewees was its demonstration of the severity of gender-based harms perpetrated by synthetic media. The '96% of deepfakes' (Ajder et al., 2019) statistic, mentioned by multiple interviewees, made an expected reappearance. Participant 1 reinstates the importance of the influence of hard data on their prioritisation, noting:

> **P1:** There are theoretical issues *(…)* around disinformation around cyber security, market manipulation, defamation, all these kinds of things. But they were very much short to medium term issues that hadn't emerged yet, whereas image abuse against women was a real problem right then and there and has become an even bigger problem. So that's how I kind of focused in on that space was... that was what the data said. That's where the impact was being felt, why work was needed.

Participant 3's rationalisation was that 'that was what the data said'. The influence of OSINT research - on top of its relevance as an accountability mechanism as discussed above - appears to also be in actors' own data-led prioritisations of gendered online harms.

Similarly, participant 1 noted 'I've worked with victims and so on. And I again, I can see that that is the biggest harm right now *(...)* millions of women who have been targeted by these tools.' This set of interviewees suggests that OSINT data outputs helped them focus on 'where the impact was being felt' (P1), as opposed to worst-case hypotheticals. This is not to say that other OSINT-discovered harms, such as targeted disinformation, are not occurring – this is to suggest that what interviewees expressed had been occurring most frequently – according to widely quoted statistics – was the image-based online abuse of women. A strong-voiced minority of interviewees consistently underlined their belief in the collective nature of gender-related harms. Some spoke of this in relation to political silencing, others in relation to

its psychological effects. An important implication of these findings is that there appears to be a coalition emerging that desires synthetic media regulation to follow hard data. Relevant now is the question – how will synthetic media's potential gender-based online harms be conceptualised moving forward, and how will they be received by the wider policy-making community? Time will tell, but this research suggests there is reason for optimism.

CONCLUSION

This dissertation has explored what at first appear to be two disjointed priorities in the regulation of synthetic media – preventing collective and individual harms. After ten expert interviews, I believe that the dichotomy between collective and individual online harms in this space is unproductive and lacks nuance. I have argued that conceiving synthetic media regulation as a balancing act between two forms of harm ignores one important reality: seemingly individualised image abuse is actually enmeshed within - and a 'canary in the coalmine' for - the core epistemic threat of synthetic media itself. I borrowed from multiple disciplines in this research trajectory – epistemics, media theory, deliberative democratic theory, and gender studies. From these analyses, I concluded that gendered harms and frequently hypothetical geopolitical ones are, at the core, different manifestations of the same epistemic harm. The participants in this ten-interview research study - OSINT analysts, synthetic media technologists, policy professionals, academics, and think tank members – all brought unique perspectives to the table. The overarching goal has been to paint the picture of a specific liberal-democratic discourse around synthetic media and understand its main points of consensus and contention. Core to this conceptualisation have been theories of epistemic security and deliberative democracy, and how both may face multifaceted harms from synthetic media.

It was hypothesised that a desire to preserve epistemic security in the online information landscape motivates policy actors to engage with discourse on synthetic media's potential online harms. Overall, each interviewee expressed fears of the epistemic risks that synthetic media may bring. Interestingly, interviewees speaking from a gender-related harms view brought an unexpected perspective. They noted that strategies which may be effective

for some epistemic harms may not address other gender-specific ones, like shame and a sense of bodily violation. It was also hypothesised that the desire to preserve the feasibility of deliberative democracy was a key motivator for interviewees' engagements in this policy discourse. Almost all interviewees were concerned about the effects of synthetic media on collective information ecosystems. Many suggested that international cooperation and increased citizen involvement were essential. Intersectional harms, especially in the context of conflict-affected and gender-divided societies, were convincingly argued by participants to require more nuanced attention.

This dissertation also hypothesised that gender-specific harms would be more frequently categorised by policy actors as individualised than collective in nature. A targeted lens on women's vulnerabilities to synthetic media uncovered multiple interrelations between geopolitical harms and gendered harms themselves. Using examples of defamation and abuse faced by female politicians, multiple interviewees drew an explicit connection between the two seemingly disjointed regulatory priorities. Interestingly, many respondents quoted OSINT-based data in their justifications for focussing on gendered online harms – noting in one way or another that they went *where the data led them*. Moving forward, regulators of synthetic media are encouraged to critically consider the interconnected ethical concerns – and epistemic underpinnings – of these rising online harms. One possibility for further research would be to speak with a wide range of women in positions of power who have experienced synthetic image abuse. Doing so may help to understand more complex political impacts of this technology's emerging capabilities and provide integral data for forthcoming regulatory actions.

REFERENCES

117th Congress (2022) S.2559 - *Deepfake Task Force Act*. URL: https://www.congress.gov/bill/117th-congress/senate-bill/2559 [Last consulted 15 August 2023].

Ajder, H., Glick, J. (2021) *Just Joking! Deepfakes, Satire, and the Politics of Synthetic Media. WITNESS, MIT Co-Creation Studio, MIT Open Doc Lab. URL:* https://cocreationstudio.mit.edu/wp-content/uploads/2021/12/JustJoking.pdf [Last consulted 02 August 2023].

Ajder, H., Patrini, G., Cavalli, F., and Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact.* Deeptrace Labs (now SensityAI). URL: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf [Last consulted 21 March 2023].

Allyn, B. (2022) *Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn.* NPR Technology. URL: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia [Last consulted 15 August 2023].

Baudrillard, J. (1983) *Simulacra and Simulation.* Trans. Glaser, S. Ann Arbor: University of Michigan Press. URL: https://0ducks.files.wordpress.com/2014/12/simulacra-and-simulation-by-jean-baudrillard.pdf [Last consulted 15 August 2023].

Biddle, S. (2023) *U.S. Special Forces Want To Use Deepfakes For Psy-ops*. The Intercept, URL: https://theintercept.com/2023/03/06/pentagon-socom-deepfake-propaganda/ [Last consulted 13 August 2023].

Bogner, A., Menz, W. (2009) The Theory-Generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction. In: Bogner, A., Littig, B., Menz, W. (eds) *Interviewing Experts.* Research Methods Series. Palgrave Macmillan, London. URL: https://doi.org/10.1057/9780230244276_3 [Last consulted 13 August 2023].

Bregler, C., Covell, M., Slaney, M (1997) *Video Rewrite: Driving Visual Speech with Audio. Interval Research Corporation. URL:* http://chris.bregler.com/videorewrite/ [Last consulted 21 March 2023].

Brinkmann, S., and Kvale, S. (2018) *Epistemological issues of interviewing*. SAGE Publications Ltd, URL: https://doi.org/10.4135/9781529716665 [Last consulted 15 August 2023].

Burgess, M. (2020) *A deepfake porn bot is being used to abuse thousands of women.* WIRED. URL: https://www.wired.co.uk/article/telegram-deepfakes-deepnude-ai [Last consulted 15 August 2023].

Byman, D., Gao, C., Meserole, C., Subrahmanian, V.S. (2023) *Deepfakes and international conflict.* Brookings Artificial Intelligence and Emerging Technology Initiative. URL: https://www.brookings.edu/articles/deepfakes-and-international-conflict/ [Last consulted 16 August 2023].

Cohen, J. (2002) Deliberation and Democratic Legitimacy. *Debates in Contemporary Political Philosophy.* (eds. Matravers, D., Pike, J.) London, UK: Routledge: 342-360 URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203986820-28/deliberation-democratic-legitimacy-joshua-cohen [Last consulted 15 August 2023].

Cole, S. (2018) We Are Truly F*cked: Everyone Is Making AI-Generated Fake Porn Now. *Vice Motherboard. URL:* https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley [Last consulted 15 August 2023].

Chambers, S. (2003) Deliberative Democratic Theory. *Annual Review of Political Science,* 6 (1) p. 307-326. URL: https://doi.org/10.1146/annurev.polisci.6.121901.085538 [Last consulted 15 August 2023].

Chesney, R. and Citron, D. (2018) Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *107 California Law Review* 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [Last consulted 13 August 2023].

Discord (2023) *Midjourney Server Community Invite.* Discord. URL: https://discord.com/invite/midjourney [Last consulted 15 August 2023].

Dodgson, J. (2019) Reflexivity in Qualitative Research. *Journal of Human Lactation.* 35(2): 220-222. URL: https://doi-org.gate3.library.lse.ac.uk/10.1177/0890334419830990. [Last consulted 15 August 2023].

Döringer, S. (2021) 'The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge, *International Journal of Social Research Methodology*, 24:3, 265-278. URL: https://doi.org/10.1080/13645579.2020.1766777 [Last consulted 15 August 2023].

Editorial (2020) *The sexist playbook.* The Indian Express – Journalism of Courage. URL: https://indianexpress.com/article/opinion/editorials/kamal-nath-item-remark-bjp-congress-rahul-gandhi-bypollls-6811608/ [Last consulted 15 August 2023].

European Data Portal (2022) *Open-Source Intelligence The lessons that OSINT provides to open-data portals.* European Commission. URL: https://data.europa.eu/en/publications/datastories/open-source-intelligence [Last consulted 15 August 2023].

European Science-Media Hub (2021) *Fighting abusive deepfakes: the need for a multi-layered action plan*. European Parliamentary Research Service. URL: https://sciencemediahub.eu/2021/10/13/fighting- abusive-deepfakes-the-need-for-a-multi-layered-action-plan/ [Last consulted 15 August 2023].

Gaskell, G. (2000) Individual and Group Interviewing. *Qualitative Researching with Text, Image and Sound.* (eds. Gaskell and Bauer). SAGE Publications. URL: https://methods.sagepub.com/book/qualitative-researching-with-text-image-and-sound/n3.xml [Last consulted 13 August 2023].

Gleicher, N. (2022) Tweet: *'Earlier today, our teams identified and removed a deepfake video claiming to show President Zelensky issuing a statement he never did'* Twitter. URL: https://twitter.com/ngleicher/status/1504186935291506693?s=20 [Last consulted 15 August 2023].

Guest, G., MacQueen, K. M., & Namey, E. E. (2012) *Planning and preparing the analysis*. SAGE Publications, Inc.. URL: https://doi.org/10.4135/9781483384436 [Last consulted 15 August 2023].

Imbrie, A., Gelles, R., Dunham, J. and Aiken, C. (2021) *Contending Frames Evaluating Rhetorical Dynamics in AI.* Center for Security and Emerging Technology. URL: https://doi.org/10.51593/20210010 [Last consulted 20 March 2023].

Interesse, G. (2022) *China to Regulate Deep Synthesis (Deepfake) Technology Starting 2023. China Briefing. URL:* https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/ [Last consulted 15 August 2023].

Jeong, S., Sturtevant, M., Stephen, K. (2021) *Responding to Deepfakes and Disinformation.* The Regulatory Review. Penn Program on Regulation. URL: https://www.theregreview.org/2021/08/14/saturday- seminar-responding-deepfakes-disinformation/ [Last consulted 15 August 2023].

Kavanagh, J. and Rich, M. (2018) *Truth Decay: A Threat to Policymaking and Democracy.* RAND Corporation, Santa Monica, CA. URL: https://www.rand.org/pubs/research_briefs/RB10002.html.[Last consulted 15 August 2023].

Leibowicz, C., McGregor, S., and Ovadya, A. (2021) The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21).* Association for Computing Machinery, New York, NY, USA, 736–744. URL: https://doi.org/10.1145/3461702.3462584 [Last consulted 16 August 2023].

Mackintosh, E., Gupta, S. (2020) *Troll armies, 'deepfake' porn videos and violent threats. How Twitter became so toxic for India's women politicians.* CNN World. URL: https://edition.cnn.com/2020/01/22/india/india-women-politicians-trolling-amnesty-asequals-intl/index.html [Last consulted 15 August 2023].

Meaker, M. (2019) *It's not male politicians who edited 'deepfake' videos target—it's women turned into pornography.* Prospect Magazine. URL: https://www.prospectmagazine.co.uk/politics/39507/its-not-male-politicians-who-edited-deepfake-videos-targetits-women-turned-into-pornography [Last consulted 15 August 2023].

O'Neill, L. (2023) *These Women Tried to Warn Us About AI.* Rolling Stone. URL: https://www-rolling stone-com.cdn.ampproject.org/c/s/www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/amp/ [Last consulted 15 August 2023].

Oxford Policy Engagement Support (n.d.) Guidance note 2: Understanding the policy process. Using research to engage. URL: https://www.ox.ac.uk/research/using-research-engage/policy-engagement/guidance-policy-engagement-internationally/guidance-note-2-understanding-policy-process [Last consulted 15 August 2023].

Pawelec, M. (2022) Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *DISO* 1 (19). URL: https://doi.org/10.1007/s44206-022-00010-6 [Last consulted 15 August 2023].

Proctor, R. and Schiebinger, L. (eds.) (2008) *Agnotology: The Making and Unmaking of Ignorance*. Stanford, California: Stanford University Press. URL: https://philpapers.org/archive/PROATM.pdf [Last consulted 20 July 2023].

Roose, K. (2023) *Inside the White-Hot Center of A.I. Doomerism.* The New York Times. URL: https://www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html [Last consulted 15 August 2023].

Ryan-Mosley, T. (2023) *An early guide to policymaking on generative AI.* The Technocrat. MIT Technology Review. URL: https://www.technologyreview.com/2023/03/27/1070285/early-guide-policymaking-generative-ai-gpt4/ [Last consulted 30 May 2023].

Seger, E., Avin, S., Pearson, G., Briers, M., Heigeartaigh, S., and Bacon, H. (2020) *Tackling threats to informed decisionmaking in democratic societies: Promoting epistemic security in a technologically-advanced world.* The Alan Turing Institute. URL: https://www.turing.ac.uk/sites/default/ files/ 2020-10/epistemic-security-report_final.pdf [Last consulted 20 July 2023].

Smuha, N. (2021) Beyond the individual: governing AI's societal harm. *Internet Policy Review* 10(3). URL: https://doi.org/10.14763/2021.3.1574 [Last consulted 15 August 2023].

SOCOM (2023) Broad Agency Announcement USSOCOM-BAAST-2020, *Amendment 3 For Technology Development And Advanced Technology Development.* DocumentCloud. URL: https://www.documentcloud.org/documents/23696654-us-socom-procurement-document-announcing-desire-to-utilize-deepfakes [Last consulted 4 May 2023].

Starbird, K., Arif, A., Wilson, T. (2019) Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *PACMHCI.* Vol: CSCW, Article 127. ACM. URL: https://dl.acm.org/doi/pdf/10.1145/3359229. [Last consulted 15 August 2023].

This-Person-Does-Not-Exist.Com. (2023) *Random Face Generator (This Person Does Not Exist).* URL: https://this-person-does-not-exist.com/en [Last consulted 1 May 2023].

Tsukayama, H., McKinney, I., Williams, J. (2019) *Congress Should Not Rush to Regulate Deepfakes.* Electronic Frontier Foundation. URL: https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes [Last consulted 15 August 2023].

Waite, T. (2023) *The treachery of images in the age of AI*. Dazed Digital. URL: https://www.dazeddigital.com/art-photography/article/59695/1/the-treachery-of-images-in-the-age-of-artificial-intelligence-luke-nugent [Last consulted 21 July 2023].

Waldemarsson, C. (2020) *Disinformation, Deepfakes & Democracy - The European response to election interference in the digital age.* Alliance of Democracies Foundation. URL: https://www.allianceofdemocracies.org/wp-content/uploads/2020/04/Disinformation-Deepfakes-Democracy-Waldemarsson-2020.pdf [Last consulted 15 August 2023].

Wagner, T. & Blewer, A. (2019) 'The Word Real Is No Longer Real': Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science*, 3(1), 32-46. URL: https://doi.org/10.1515/opis-2019-0003 [Last consulted 15 August 2023].

Warren, C. (2001) Qualitative Interviewing. *Handbook of Interview Research*. (eds. Gubrium and Holstein). Sage Research Methods. URL: https://methods.sagepub.com/book/handbook-of-interview-research/d7.xml [Last consulted 13 August 2023].

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., Russell-Bennett, R. (2023) Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation*, Volume 125. URL: https://doi.org/10.1016/j.technovation.2023.10278 [Last consulted 15 August 2023]

APPENDICES

APPENDIX A: Participant Profiles

Pseudonym	Career Stage	Position
P1	Early/mid-career	Deepfake expert and independent advisor (various)
P2	Late-career	Behavioural scientist - expertise in disinformation/extremism
P3	Mid-career	VP of an OSINT-based research non-profit
P4	Mid-career	Director-level role at private OSINT-led firm
P5	Mid-career	Technology ethicist at a deepfake creation firm
P6	Mid-career	Manager-level role at an online-safety focused regulator
P7	Late-career	Director of an OSINT-based research non-profit
P8	Mid-career	OSINT analyst at a US-headquartered think tank

P9	Early/mid-career	Fellow/researcher on OSINT/AI for US-based think tanks
P10	Early/mid-career	Data scientist/technologist, focus on misinformation mitigation

There were four female and six male interviewees. One interviewee disclosed being a victim of synthetic media image abuse. This interviewee was, unsurprisingly, female.

APPENDIX B: Thematic Analysis Grid

Theme	Code	Description	Example
epistemic security **Expresses concern for the stability of collective knowledge bases when faced with synthetic media and/or fears the deterioration of shared realities*	Truth decay	Interviewee expresses a current or theoretical fear of societal truth decay or similar epistemic phenomenon	*'... it feels that it's there is a harm that comes from directly believing what that synthetic media is, but there is possibly a greater harm which comes from the, kind of, undermining trust in the collective knowledge base and this, you know, the splintering of people into kind of fragmented communities of people that they trust and the creation of information silos.'* ***(P7)***
	Not a prescriptivist	Interviewee fears expressing moral authority - has discomfort with saying what 'should' be done	*'So I can say that - from a moral perspective - that anything that's intended to deceive, sort of, you know, crosses over into that... morally dark (...) I'm not sure I'm in the business of saying what people ought to or ought not to use.'* ***(P2)***
	Prescriptivist	Interviewee suggests potential solutions related to epistemic harms	*'So we have a kind of verification process that we go through for any piece of footage that we detect through those open source human rights monitoring projects, which - you know - includes tests of authenticity on content'* ***(P7)***

deliberative democracy **Expresses concern for the feasibility of democratic deliberation if citizens lack a shared and trusted info base and/or fears an environment antithetical to collective governance*	'Watchdog' function	Interviewee gives example of OSINT data's utility in policymaking	*'...I think that our investigations, they work as basically cases and case studies so that then the legislators (can use them) to show that 'hey, that's how this thing was used in a bad way to do harm, so we should prevent this thing in the future'.'* ***(P8)***
	Citizen involvement	Interviewee suggests the ethical importance of citizen involvement in policy-making	*'So I do think in those rooms, you know, (the) public should also be present - and I'm not saying tech companies should be excluded, but currently the dynamic is a bit opposite. The tech companies are the ones in the room most of the time.'* ***(P10)***
	International cooperation	Interviewee suggests the need for international cooperation on deepfake policymaking	*'Obviously, we've kind of* (had) *polarisation in the UN Security Council, where it's very difficult to see that you're going to get international consensus* (…) *I feel, like, you know, from a moral perspective - it feels like* (there needs to be agreement on) (…) *what is acceptable and what is not acceptable in certain contexts.'* ***(P7)***
information operations **Expresses concern for the trajectory of influence campaigns amidst increasingly realistic 'deepfake' tech – fears changed ethical and practical norms of warfare*	Moral judgments	Interviewee attempts to frame/justify/construct a judgement of the ethics of deepfakes in war	*'And, you know, one person's opportunities, the other side's threat, right? So for me - I don't think of many things happening in warfare which aren't, you know... like that Biden (deepfake) example would be seen as propaganda by the Russians and as a threat. So for me it's... I think it's... the framing is difficult.'* **(P1)**
	Predictions	Interviewee makes predictions on the usage of deepfakes in war	*'... So I do think that there's - probably at least in the US (…) there are definitely people that are just looking to cause – sow - mayhem on the domestic scale and I think they're just very good at identifying what are those flash points that get people very excited. So I would guess that, yeah,*

			there's probably going to be more use of that - or attempted use - on the domestic side.' ***(P4)***
	Potential risks	Interviewee posits potential/actual downsides of deepfakes in conflict scenarios	*'I mean, not that China doesn't regulate the Internet, you know. Absolutely. It's (…) morally dubious, but effective, but you know they're not gonna probably have the same, you know, sensitivities around use of deepfakes in warfare, right?'* ***(P7)***
	Potential upsides	Interviewee posits potential/actual positives of deepfakes in conflict scenarios	*'...maybe you could use (…) deepfakes or generative content to, for example, make it look like Biden is speaking Ukrainian to Ukrainian people to help bolster their morale or something like this'* ***(P1)***
	Solutions	Interviewee suggests solutions related to potential harms from deepfake info ops	*'So this is a very common problem in cyber attribution: when a nation state attacks (…), the victim nation state(s) - when they make the attribution claim - it's hard to believe them. You need, like, a neutral third party to do that attribution and that's why it's important to start laying those foundations, at least for AI incidents or synthetic media.'* ***(P10)***
gender-specific harms ** Expresses concern for overlooked 'deepfake' harms women and girls may face – fears 'grander' concerns overshadow private ones*	Origins/history	Interviewee mentions the gendered origins of deepfakes/image abuse	*'Obviously given its origins in non consensual intimate image abuse, it was quite clear that there was harms already emerging, and there were lots of potential harms that were being theorised (…) And I kind of went down the pretty unpleasant rabbit hole.'* ***(P1)***
	Overlooked	Interviewee claims that image abuse is an overlooked/important issue	*'...people are still very much dismissing it because (they) think it's entertainment like they think (...) Deep Tom Cruise TikToks, things like that - or you think of pornography. And then again, there's that*

			immediate dismissal of like 'well, it's not a problem, right'' ***(P9)***
	Accountability and awareness	Interviewee brings up the relevance of OSINT as an accountability or awareness mechanism	*'I wasn't going into that piece of work expecting to be like 'ohh, ok, and after this I'm gonna become an advocate for legislative reform on image abuse and laws around digital sexual abuse'. But, the findings of that research* (that) *96% of deepfakes at the time - or deepfake videos at the time - were non consensual pornography (…) I followed the evidence to its conclusions (…)'* ***(P1)***
	Policy-specific	Interviewee makes comments or predictions on policy developments	*'*(There was proposed) *an amendment to the Violence Against Women Reauthorization act (...) - the Republicans didn't want to pass it - so like, I'm not, I don't - especially because it's coming from a member of the minority – and you know - then it would have to get through the Senate as well. I just don't see it going very far. And that's so incredibly frustrating.'* ***(P3)***
	Gender and power	Interviewee argues women in power (or gaining a voice in public spaces) are specifically vulnerable/frequent targets of online harms	*'... particularly where you have authoritarian regimes, or you have deeply socially conservative regimes where women's physical participation is limited in many ways - that in many of these contexts the online space is becoming a really important vehicle for women's voice - and we're seeing, you know, that technology facilitated gender based abuse.'* ***(P7)***
	Personal experience	Interviewee mentions their own experiences or risks of being deepfaked	*'...when I look at like, you know, a deepfake pornographic video of me that doesn't even really resemble me, I'm just kind of like, 'yeah, this is horrible and I'd like it to be taken down', but I don't know if it's worth the effort given all of the huge pile of other things I have to deal with (...)*

			I'm like, so broken at this point that it just... it, it just kind of rolls off...' ***(anonymised)***
	Children	Interviewee mentions child rights/protections or exploitation through deepfakes	*'Instead of using deepfake, you know, to protect those people, they (law enforcement) wanted to become, for example, little kids physically with the voice to attract, you know, paedophile people' (for arrest)* ***(P5)***
	Solutions	Interviewee suggests solutions related to gender-specific harms	*'It's not that they're completely separate, but that by tackling the women's image abuse issue we're gonna be - maybe not outright solving, but at least testing - for the solution to many of the other AI issues. We're gonna learn how oversight works better, which rules that we could make – are, you know - something that's palatable to the American public...'* ***(P3)***
OSINT literacies ** Broadly discusses OSINT techniques relating to synthetic media and/or expresses a desire for improved online information literacies for experts or civilians*	Personal motivations	Interviewee discusses their personal motivations to do OSINT-related work	*'I mean for us this sort of sense that, like, access to information is a fundamental human right, and you know - reliable, trustworthy, comprehensive information is absolutely fundamental to the decisions that people make in their lives. And that's across every sphere, right. (...)'* ***(P7)***
	What is 'impact'?	Interviewee defines or questions what impactful OSINT work is/can be	*'If it's accurate and it's well done, it's going to be picked up by the national media organisations* (…) *I see that happen all the time, like VICE News, Motherboard and things - they'll pick up people like that - like you're describing - and then from there it'll get picked up by, like, the, you know, international news media.'*

			(P4)
	OSINT critiques	Interviewee criticises OSINT techniques, experts, or hobbyists	*'… I don't mean this disparagingly because we do exact same thing, but like armchair investigators who - especially in the Ukraine war - have come out and been like 'ohh yeah, here's my hot take with, like, minimal amount of data, but lots of the very authoritative sounding...' And I think that gets frustrating (...)'* ***(P9)***
	Specific cases/data	Interviewee provides concrete examples of the utility of OSINT data	*'So, for example, you know the study by DeepTrace (…) about the volume of synthetic videos on the Internet and then how 96% were, um, non-consensual pornography. I have seen that stat so often, and that's such a powerful statistic to refer to (…) having concrete examples to point to, I think, are really important (...)'* ***(P9)***
	Information literacies	Interviewee supports teaching OSINT (or general media literacy) skills to the public	*'And yeah, we should absolutely bake in information literacy into the curricula that are being developed - but that's kind of already happening, I mean, librarians know this is a problem, social studies teachers know this is a problem and that effort is occurring. And I think it's much more systematic and it's much more... It's not quite like flipping a switch, right?'* ***(P3)***

APPENDIX C: Interview Guide

**There was significant customisation in the preparation of each interview guide depending on the interviewee's role, experiences, and research output – with many of these identifying the participants. So, to accurately share the breadth of questions asked, I display them in the following format: those asked to everyone (besides two who were on a strict time limit), questions asked to a few but then discarded, and then a long list of questions asked specifically to individuals. I have redacted identifying aspects of the individual-targeted questions.*

Questions asked to everyone (except a few removed for those with thirty-minute time limits):

1. *Great, is it alright if I start recording?*
2. *So to start off - could you outline for me what motivated your decision to get involved in issues around synthetic media?*
3. *Great, so could you outline for me what advocacy in this area has looked like for you? I'm interested in how your priorities have developed, what stakeholders have been productive to approach - any 'wins' in the area – that sort of general overview.*
4. *I'd like to ask - what are the values that have motivated you to engage in advocacy around deepfakes – is it due to concerns for democracy, a desire for a particular political outcome, gender-specific worries, or perhaps something else?*
5. *What role would you say that your form of research plays in a democracy? Do you consider your work to be part of the fourth estate – of journalism and public accountability – or would you categorise it in another way?*
6. *So I'm interested in your viewpoint on the usage of deepfakes in warfare. Do you think that a hard line should be drawn in not allowing for their use, or is it justifiable as defence when a bad actor deploys them first?*
7. *I saw that US Special Ops Command submitted a procurement request to use deepfake technologies in their own operations – what do you feel about this 'race to the bottom' so to speak, to develop deepfake technologies for usage in conflict?*
8. *What is your biggest concern about the usage of deepfakes in warfare?*
9. *Do you think that work by open-source analysts in tracing disinformation – like discovering the origins of the Zelenskyy deepfake for example - has an impact on the policy process around synthetic media – or do those sorts of reports have less policy impact than their writers might hope?*
10. *Do you think anything needs to change at a societal level before policy interventions around synthetic media can really become effective?*

11. *What sorts of negative impacts are you concerned that deepfakes and synthetic media may have in the future?*
12. *Do you think it could be productive to expand the traditional understanding of media literacy to include teaching people basic OSINT techniques in school to detect information origins, or is that a bit overly optimistic?*
13. *Which country or groups of countries are you most optimistic about in terms of their ability to develop ethical policies around synthetic media?*
14. *Is there a specific concern around synthetic media that you intend to make your top priority in the future, or is it not possible to rank one worry above the others?*
15. *Is there anything else you'd like to share before we conclude?*

Questions asked early on but then discarded for repetitiveness or lack of usefulness:

1. *What would you say was your reasoning to prioritise this issue area over others relating to AI?*
2. *How would you define a deepfake?*
3. *So this is a bit of a hypothetical - if you were to be given the task of developing policies around synthetic media tomorrow – would you rather do so within a tech platform, in a government role, or within a civil society organisation?*
4. *Are there any pieces of legislation on deepfakes being proposed in Washington at the moment that you're feeling optimistic about?*
5. *In what ways do you feel your work has had an impact on the policymaking process around deepfakes?*
6. *What are some positive ways that deepfakes may impact the information space?*
7. *I'd like to ask - what are the political values (or worries) that have motivated you to engage in policy advocacy?*

Questions specific to individuals (based on pre-reading of their research):

1. *So with your (redacted) project, what sorts of decision-makers are you and (redacted) hoping to reach, and what sorts of impact are you aiming for?*
2. *So I saw a tweet that noted a (redacted) film has covered (redacted) using OSINT techniques – I'm curious when this sort of content is being created, who are the main audiences or stakeholders (redacted) intends to reach, and is there always an explicit policy motivation underneath the creation of this sort of content?*
3. *I saw in a (redacted) that (redacted) experienced online hate through 'cheapfakes' and that it was likely a (redacted) operation – do you find that the line between politically-motivated deepfake use and gender-specific deepfake use is often blurred – as in, do they tend to occur in unison?*

4. *Could you share any other examples you know of in which foreign actors, like (redacted) potentially in (redacted) case, have specifically used female-targeted deepfakes for geopolitical gain?*
5. *I noted that in (redacted), media literacy strategies as a potential solution. Would that also include teaching basic information investigation skills, like rudimentary OSINT analysis, or would that ideally focus more so on just basic determination of source legitimacy? I suppose I'm asking – how optimistic do you think we should be in terms of the level of these media literacy initiatives?*
6. *I saw in your recent (redacted) that you said (redacted) - this reminded me of recent critiques of AI 'doomerism' and people freaking out about the singularity... these ideas taking away attention from real issues in the here and now. Do you think that it's necessary to take attention from some areas of the AI discourse and funnel it over to women's issues, or is there space for tackling all of these in unison?*
7. *So to start off – could you outline for me what motivated you to work with (redacted), and what the ethics aspect of your role consists of?*
8. *So I saw in one press release that you were referred to as (redacted) – what does that entail?*
9. *So in the (redacted) report that you worked on, I saw that the scope was more on (redacted) than on deepfakes, but from that, what would you say are some of the main risks in general when dealing with automated as opposed to traditional forms of info ops?*
10. *What are some ways in which large language models might be used in conjunction with deepfakes for information operations?*
11. *Could you tell me a bit more about the (redacted) research group you worked on at (redacted)?*
12. *Part of the theory aspect of my research looks at epistemic security and the extent to which influence operations using synthetic media are a threat to the collective knowledge base – do you feel there's something specific about synthetic media that makes it more of an epistemic threat than other tools?*
13. *So I understand one of your areas of expertise is in (redacted) – from your knowledge of this area, how significant is synthetic media in the context of the whole conversation around emerging tech in warfare – is it the main 'hot topic' or area of concern lately, or less prominently discussed than other developments, like autonomous weapons for example?*
14. *Could you outline for me your involvement in the world of OSINT and related strategy? I saw in your bio that you (redacted) and am curious if any of it relates to issues around synthetic media?*
16. *What do people in the more formalised OSINT spaces think of those who practise OSINT online in casual Discord and Reddit communities? Are their reports ever used in think tank publications, for example?*
17. *How often do you see synthetic media and OSINT analysis cross paths in academic research, if at all?*

18. *Have you seen any examples of OSINT investigations informing policy discussions around synthetic media?*
19. *In an article you cowrote for (redacted), I saw that it mentioned (redacted) - when we think about deepfakes as used in warfare, what are some of the ways that this mindset could be applicable?*
20. *To start off - could you outline for me around when deepfakes came to (redacted)'s attention, and how the team's priorities in that area have developed over time?*
21. *Great, could you tell me a bit about what your role has been in that process?*
22. *I'd like to ask - what do you think are the values that have motivated the (redacted) team to do work on deepfakes – is it due to concerns for gender-based harms, disinformation and election security, or perhaps something else?*
23. *Great, so I'm interested in learning which stakeholders have been productive to approach – has (redacted) found platforms to be cooperative in terms of the deepfake-related suggestions that have been put forward?*
24. *I saw on your LinkedIn that a lot of your previous work centred around (redacted) – from your experience, how do you think deepfake regulations should be approached specifically when considering the protection of children?*
25. *Does media literacy form a significant part of (redacted)'s online harm prevention strategy?*
26. *When developing policies around (redacted), has your team used insights found in journalistic 'open-source' investigative reports, or does the data used to make decisions tend to come more so from within established academia and government-adjacent think thanks?*
27. *In what ways do you feel your team's work has had an impact on the global conversation around deepfake policymaking?*
28. *I saw on your LinkedIn that you've worked with (redacted) - do you feel that these sorts of states have some unique vulnerabilities to synthetic media, and if so, what might those vulnerabilities be?*
29. *So to start off – could you outline for me what motivated you to work in the security, intelligence, and OSINT world?*
30. *Great, could you tell me a bit about what you do in your current role?*
31. *Since you wrote that article, what have been some of the main changes you've seen in the world of synthetic media?*
32. *Is the public market for deepfakes still low, or do you think that demand is increasing?*
33. *When a company is concerned about the impact of deepfake scams or propaganda or whatnot on their business, what sorts of companies do they go to for advice – or is there not yet much of a formal B2B security environment for this issue?*
34. *(redacted) - what should a software company, for example, take into consideration when considering open-sourcing some of its tech, given that it can also be used in really unintended ways?*

35. *What would you say that people in the 'official' or sort of mainstream intelligence and security world think about OSINT enthusiasts who sort of do their own rogue reporting outside of an organisation?*
36. *(redacted) – I was wondering, in the security and intelligence spaces that you're in, do people tend to speak about the gender issues of deepfakes or do they tend to stick to their risk, for example, in political propaganda?*
37. *So in the (redacted nation) context, have you seen people be particularly concerned about the use of deepfakes in conflict, or is the issue not as much of a hot topic as other forms of propaganda and disinformation?*
38. *Do you think there should likely be a strict ban on synthetic media in conflict scenarios, or is it more of a moral grey area where some uses may be justifiable?*
39. *So I understand that a lot of OSINT work, especially in conflict scenarios, can really have a mental health impact. Do you feel that enough attention has been given to this risk in your field?*
40. *So I saw on your LinkedIn that you've (redacted) - do you think teaching OSINT and general media literacy is a potential solution to some of the risks of synthetic media in (redacted nation)?*
41. *What are some policies that you feel would be most helpful to have as a way to fight the influence of deepfakes in (redacted nation)?*
42. *So to start off – could you outline for me what motivated you to write (redacted) for (redacted), and what motivated you to get involved in issues around synthetic media more generally?*
43. *So I saw you (redacted) – how do you think (redacted) recruitment might be made more palatable to young people?*
44. *So I understand you focus on (redacted) and radicalization – what are some of your worries around deepfakes as they relate to those areas?*
45. *I think people often don't realise how traumatic online abuse, especially image abuse, is until they've experienced it. How would you say that the harassment you've experienced has informed the way you approach deepfake regulation?*