UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

WestminsterResearch
http://www.westminster.ac.uk/westminsterresearch

From Social Media to Deepfakes: Participatory human rights witnessing and advocacy using audiovisual media, incorporating the emerging impacts of deceptive AI and technologies for authenticity and trust (2007-22)

Gregory, S.

This is a PhD thesis awarded by the University of Westminster.

© Mr Samuel Gregory, 2024.

https://doi.org/10.34737/w955w

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

From Social Media to Deepfakes: Participatory human rights witnessing and advocacy using audiovisual media, incorporating the emerging impacts of deceptive AI and technologies for authenticity and trust (2007-22)

Sam Gregory

PhD 2024

From Social Media to Deepfakes: Participatory human rights witnessing and advocacy using audiovisual media, incorporating the emerging impacts of deceptive AI and technologies for authenticity and trust (2007-22)

Sam Gregory

A thesis submitted in partial fulfilment of the requirements of the University of Westminster for the degree of Doctor of Philosophy by Published Work

March 2024

ABSTRACT

As a researcher and professional human rights worker, my publications and research between 2012-22 generated knowledge and influence on three critical questions: What were significant methodological shifts in human rights tactics and strategies based on audiovisual media? For professional practitioners and everyday activists, what was constant and what changed in ethical and practical challenges around visibility/obscurity, trust and authenticity, and the impact of witnessing amidst media volume? How do emerging technological infrastructure and systems - particularly around media manipulation, deepfakes, authenticity and trust, and artificial intelligence - impact practices and dilemmas, and how can they be shaped from a global human rights perspective? My methods include participatory and action research, field-mapping, expert interviews and convening, speculative/futures-based approaches and specific fieldwork projects. I map, produce case studies, and provide insights into novel forms of human rights participation, documentation and advocacy, including participatory fact-finding and open-source intelligence, as well as new forms of ‘distant witnessing’ via live-streaming, remix and evidentiary documentation at critical moments of their evolution. Emerging forms shape and are shaped by existing advocacy paradigms including smart narrowcasting, as well as by professionalisation trends and the ‘forensic turn’. Their reception cannot be separated from broader questions around trust in media and content and present challenging evaluatory questions around impact. Research findings confirm the consistency of challenges around activists navigating choices around visibility and obscurity of self, presence and content, as well as escalating challenges of confirming trust in content and being found amid volume. ‘Distant witnessing’ involving the active participation of remote activists and co-presence-based strategies, investigated via theorisation and case study, provides an opportunity for more equitable mediated witnessing. Yet experimental action research shows the challenges of self-surveillance. Early global research on deepfakes and authenticity infrastructure closely integrates diverse perspectives in comparative contexts with technical

investigation to ground emerging phenomena in existing interdisciplinary knowledge and propose research and action. The close integration of participatory research with engagement in standard-setting, as well as reference design development via technical artefacts, indicates how research can directly impact these emerging technologies. Future opportunities for research focus on the implications of generative AI, evolving remix activism, and XR.

Acknowledgements

First, I thank my supervisors Professor Graham Meikle and Dr. Andrea Medrado for their support and guidance. This thesis partly draws on collaborations with Gabi Ivens, Professor Tanya Notley, and Andrew Lowenthal - I thank them too.

So much of my research and work draws on the deep collaborative partnership I have had with my colleagues at WITNESS and in the broader human rights, technology and AI world, as well as learning and action alongside the communities we support. Thank you and solidarity.

I'm grateful to key intellectual mentors and cheerleaders whose influence, unseen and explicit, lies within these pages - Alistair 'Slab' Slabczynski and Professor Patty Zimmerman.

My parents probably thought this would be a thesis on 11th-century Crusades history, written 20 years ago, but I took a different path - I still thank them.

This thesis is ultimately dedicated with love to Larry, who put me first many times and ensured I could do so much.

Author's declaration

I declare that all the material contained in this thesis is my own work.

1: AUTOBIOGRAPHICAL CONTEXT AND RESEARCH JOURNEY UNDERLYING THE PORTFOLIO

The publications, public outputs, and research in this portfolio reflect twenty-five years of professional and academic experience as a leading practitioner and reflexive researcher at the intersection of human rights, digital media and activism. Publications from 2012-2022 comprise the portfolio, including peer-reviewed articles, book chapters, public writing, and whitepapers (listed in Section 2). They broadly reflect research, learning, and developments within the period 2007-22, with a stronger focus on post-2012. All have external impacts derived from the research and findings. My synthesis obliquely draws on a body of cited work with impact pre-dating 2012. This prior work shares similar research questions, which I reference where appropriate. The knowledge production here is also interrelated with the production of technical artefacts - for example, reference designs for technology tools – reflecting norms or ethical principles identified in papers.

For over fifteen years, including the synthesis period, I led the programmatic and foresight work of a leading human rights group working in this area of research. WITNESS is a global human rights network focused on the use of video and technology in human rights work. My leadership included directing specific programs on emerging technology and tactical innovations in media activism. These have provided me with a critical, grounded understanding of how individuals and communities use digital media and communications in social change work and the challenges experienced at a grassroots level and inherent in emergent technical infrastructure. In addition, it allowed me to observe both specific actions within my own organisation and a broader sphere of human rights activism. I also led specific foresight-oriented projects - for example, on preserving visual anonymity in video, on live-streaming, on technology for asserting authenticity, and on global preparation for deepfakes and AI-based audiovisual manipulation approaches - that began with particular research questions. These projects

were oriented to improving practice and utilising action research findings to directly impact my field of practice.

Additional perspectives inform my research and writing. For nine years I taught a graduate course at the Harvard Kennedy School on digital video and media for human rights advocacy, and engaged with a broad range of mid-career students with expertise in related fields who informed my thinking through the iterative practice of teaching and discussion. I have also played roles in critical public-service institutions that grapple with questions in this field. At the Partnership on AI, a multi-stakeholder entity focused on the responsible use of AI, I served as co-Chair of their expert group on Social and Societal Influence of AI/AI and Media and at a global effort on establishing technical standards for media authenticity, the Coalition for Content Provenance and Authenticity, I co-chaired its Threats and Harms Taskforce. I was a member of the Technology Advisory Board of the International Criminal Court. Additionally, I have participated in external advisory groups to Twitter, Facebook, and TikTok focused on trust and safety online, specific political events, and emerging technologies.

As I noted above, I have combined my research journey with ongoing work as a leading practitioner in my field. There are several consequences of this.

Firstly, I utilise a range of qualitative methodologies, including direct observation (including of projects in which I play a direct role), participant convening and interviews, expert meetings and reviews, field scanning, specific fieldwork and technical development. These methodologies guide research that is then incorporated into impact-driven work and external public advocacy.

Secondly, I do my work from a position of privileged proximity and access as a staff person at a global human rights organisation, working directly with communities of human rights practice and engaging with colleagues in direct partnership or engagement with human rights defenders. This positionality and consequent subjectivity inform my research.

Thirdly, I have frequently engaged with journals and book projects that are looking for direct reflections and analysis of practice as well as a range of writing genres. The OUP Journal of Human Rights Practice, to which I have contributed multiple times, is an example. In line with these venues, this often means a direction to my analysis that explicitly reflects on my own professional work or that is formatted as a discursive essay. I also engage with publications looking for grounded speculations or provocations based on my research. The methodologies of futures analysis I first learned as a Fellow at the Institute for the Future in 2013 inform this type of publication.

Additionally, my work has been explicitly interdisciplinary in nature. I draw on human rights as an academic field and as a field of practitioner-oriented research (my writing almost exactly tracks the existence of the OUP's Journal of Human Rights Practice, which aligns in affinity with much of my work), as well as communication and media studies, film studies and journalism. I have not drawn as extensively on STS as I would like to do in the future, although I increasingly see the value of this discipline to the infrastructural questions I discuss in Section 5. In this synthesis, I emphasise the value I have perceived in this interdisciplinary practice: this includes how I make connections between different disciplinary practices and translate this into original insights with impact in my field of research and work.

This synthesis explores the primary through-line of my research and writing around the evolution of human rights witnessing, advocacy and activism using video and social media from 2007-2022. Section 2 outlines the publications I draw on to map the core contributions to knowledge that I make by mapping the topographies of my emerging field, by building on others' existing theories and approaches, by adding new insights, by analysing recurring assumptions to test whether they continue to hold true and by developing new paradigms for understanding the actions of people and organisations engaged in human rights witnessing and activism using digital audiovisual tools.

Sections 3, 4 and 5 explore three interrelated Research Questions that I address in my portfolio in relation to the period 2007-22:

- How have human rights witnessing, advocacy and activism tactics and strategies using video and social media evolved, including in relation to existing and emerging forms of human rights communication? (Section 3)
- How have human rights professionals, journalists, and ordinary people participating in activism addressed recurring ethical and practical dilemmas around authenticity and trust, visibility and obscurity, and volume of media? (Section 4)
- How have broader technological infrastructure and systems - particularly evolutions in systems and technologies for falsifying content and, conversely, enhancing trust in audiovisual media - conditioned possibilities for human rights activism by professionals and non-professionals and impacted practices and ethical questions? (Section 5)

In each of Sections 3,4, and 5, I identify insights, coherence and trends that emerge across the timeline and scope of my publications and show how these created original knowledge relevant to my research area and field of practice. In each section, I explore how my work was influenced by emerging contextual literature, ideas, and technical approaches and assess how it evolved.

I bring a strong global and applied perspective to my work and a commitment to the broader impact of inclusive research on society and policy. For this reason, in this synthesis, I also note the evolution of how my research has reflected the evolution of my understanding of my subjectivity as a professional human rights activist and as a researcher and how questions of equity alongside meaningful agency of research participants in decision-making on research and advocacy topics are integral to this approach.

In Section 6, I identify critical ways my work has contributed to knowledge and impact. I discuss the overall contribution of my research to knowledge in my field and how, across the publications, I map, share insights, introduce new concepts and update ‘taken-for-granted’ truths based on evolving practices. In Section 7, I go on to assess impact in both academic settings as well as on policy, public debate, practice and technical standards. I finish, in Section 8, with key future directions for research and practice, as well as exploring my journey as a researcher in order to identify and highlight conceptual limitations or gaps in my own work.

2. PUBLICATIONS IN THE PORTFOLIO

The following publications (in chronological order and accompanied by a brief 50-word synopsis) are included in this synthesis.

These publications reflect the research trajectory I outlined above - including its interdisciplinary nature, my reflexive and practitioner-centred approach, and my commitment to influencing and engaging in the public sphere. In this light, publications include all three categories of work that the University of Westminster regulations identify as suitable for a PhD By Published Work, including - 'Books and Book Chapters', Refereed Journal Papers' and Other Media/Other Public Outputs that 'represent a contribution to research in the academic subject concerned'. In this third category, I include critical research contributions oriented to public debate and developed in the form of industry research and public standards, as well as solicited provocations or speculations about the field grounded in my research and expertise.

I expand on their originality, coherence and contribution in the Research Questions sections (Sections 3, 4, 5) below.

The Participatory Panopticon and Human Rights: WITNESS's Experience Supporting Video Advocacy and Future Possibilities (*'Participatory Panopticon'* used as reference shorthand in this synthesis)
Book Chapter in 'Sensible Politics: The Visual Culture of Nongovernmental Activism', October 2012. *A survey of key lessons learned and approaches in the use of video in advocacy and of emerging human rights video ecosystems, combined with an observational topography of emerging ethical questions with new creators and formats and a set of speculations on implications for human rights advocacy.*

Human Rights Made Visible: New Dimensions to Anonymity, Consent and Intentionality *("Human Rights Made Visible")*

Book Chapter in ‘Sensible Politics: The Visual Culture of Nongovernmental Activism’, October 2012. *A human rights-grounded analysis of key questions of privacy, anonymity and consent in the context of shifts in online video platforms and social media.*

Kony 2012 Through a Prism of Video Advocacy Practices and Trends *(“Kony 2012”)*
Journal of Human Rights Practice, November 2012. *An analysis of the viral video ‘Kony 2012’ through the lens of existing human rights and digital media advocacy practices, highlighting consistency and deviation from established approaches and theorising implications.*

Technology and citizen witnessing: navigating the friction between dual desires for visibility and obscurity *(“Technology and citizen witnessing”)*
The Fibreculture Journal, December 2015. *Discussion of the key friction in human rights activism between needs for visibility and needs for anonymity, grounded in observations from professional practice.*

Ubiquitous witnesses: who creates the evidence and the live (d) experience of human rights violations? *(“Ubiquitous witnesses”)*
Information, Communication & Society 18 (11), 1378-1392, 2015. *Theoretical and practical mapping of two trends in witnessing using digital and social media: filming with ‘evidentiary’ intentions and using live-streaming technologies. Introduction of concepts of distant witnessing and co-presence in a human rights context.*

Human rights in an age of distant witnesses: remixed lives, reincarnated images and live-streamed co-presence *(“Human rights in an age of distant witnesses”)*
Book Chapter, *Image Operations: Visual Media and Political Conflict*, 184-196, 2016. *Mapping, review and case study-based analysis on emerging human rights witnessing practices around remix, curation, video mis-contextualization, and livestreaming, and how these practices relate to, or challenge, norms and expectations in the field. Note: The final section of*

this publication includes some content substantively similar to the earlier published article, 'Ubiquitous witnesses'.

Video for Change: Creating and Measuring Ethical Impact (33% contribution, see attached confirmation) *("Video for Change")*
T Notley, S Gregory, A Lowenthal
Journal of Human Rights Practice 9 (2), 223-246, 2017. *Collaborative action research into practices of measuring impact in diverse forms of digital media activism, including discussions of challenges with emerging practices (focus of author's contribution).*

Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening, June 2018: Summary of Discussions and Next Step Recommendations *("Mal-uses of Deepfakes")*
Public Output. WITNESS. 2018. *A report articulating a public research and action agenda from the first cross-disciplinary global expert gathering on deepfakes.*

Cameras everywhere revisited: how digital technologies and social media aid and inhibit human rights documentation and advocacy *("Cameras everywhere revisited")*
Journal of Human Rights Practice 11 (2), 373-392, August 2019. *Ten years on from a previous publication (Gregory 2010), the paper reviews key trends in human rights witnessing. It then turns to key challenges and identification of future trends related to questions of volume, safety and trust, and the role of platforms, infrastructure and governments. The paper provides insights into how these shifts relate to existing norms and practices in the field.*

Ticks or It Didn't Happen: Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia (50% contribution, see attached confirmation) *("Ticks or It Didn't Happen" or "Ticks")*
G Ivens, S Gregory. WITNESS. 2019. *Co-authored expert report, influential in shaping emerging technical and normative responses to misinformation known as 'authenticity infrastructure' including public impact on the Content*

Authenticity Initiative whitepaper (Parsons et al. 2020), global technical standards of the Coalition for Content Provenance and Authenticity (Coalition for Content Provenance and Authenticity n/d a, n/d b) and regulatory debates.

Live-streaming for frontline and distant witnessing: A case study exploring mediated human rights experience, immersive witnessing, action, and solidarity in the Mobil-Eyes Us Project *("Live-streaming for witnessing")*
NECSUS European Journal of Media Studies, Spring 2021. *Analysis of a live-streaming project in Brazil focused on research questions around the nature of immersive witnessing, relationships of 'mediating distant suffering', and strategies for confronting human rights denial. Findings contextualised within recurring portfolio questions of participation, safety and security and local activism.*

Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism *("Deepfakes responses")*
Journalism, March 2022. *Expert convening and research-based publication identifying how frontline witnessing and civic journalism are impacted by the reality of, the rhetoric about, and proposed solutions for, deepfakes. Paper includes further qualitative-research based discussion on authenticity infrastructure solutions highlighted in other publications.*

3. RESEARCH THREAD 1: How have human rights witnessing, advocacy and activism tactics and strategies using video and social media evolved, including in relation to existing and emerging forms of human rights communication?

3.1 Introduction: **In this analysis of the first of three interrelated threads running through my research, I map insights into how professionals and non-professionals participated in human rights activism using audiovisual media in a period between 2007 and 2022 discussed in the publications**. In addition to mapping, my research in this area builds an understanding of how this participation relates to both existing forms of human rights communication, such as documentation and accountability reporting as well as emerging formats reliant on open-source intelligence ("OSINT") or participatory fact-finding. I look at how practices in a changing activist and participatory media landscape relate to existing advocacy strategies, such as the targeting of narratives to specific audiences ('smart narrowcasting').

I go on to discuss how 'distant witnesses' (activists and others actively participating in human rights audiovisual witnessing at a distance) engage in remix, aggregation, curation or live-streamed engagement and the role of evidentiary documentation by 'citizen witnesses'. These areas are particular elements of the participatory human rights video ecosystem I explore in *Ubiquitous witnesses*, *Human Rights In An Age of Distant Witnesses* and *Cameras everywhere revisited*. I look at how **this work has contributed to a number of conceptual frameworks to understand how the engagement of both 'distant witnesses' and 'first responders'** challenges the power dynamics of established forms of solidarity, mediated witnessing and direct witnessing and our understanding of denial and 'self-expressive' witnessing.

I go on to explore the implications of evolutions in tactics and practices for definitions and understandings of impact, as well as how I have integrated an understanding of the relationship between journalism and human rights.

3.2 A key thread through these publications is the nature of increasing participation in witnessing. My contribution to the discussion in this area must be placed in the context of existing practices of frontline witnessing, activism and documentation for human rights, as well as distant witnessing and organised advocacy practices. Contextually, my work sits within the context of growth in participation using video and social media in formal human rights activism, along with other forms of 'citizen' or 'civilian' witnessing and participatory activism and scholarship on these efforts.

Human rights practices evolved significantly during the period of research. Within this synthesis, there is not adequate space to provide more than a summary, but in this landscape, direct first-person practices of human rights witnessing range in approach. Activists and first-responders participate in the collection of evidentiary video that can function as potential proof of violence (as discussed in *Ubiquitous Witnesses*, and also Human Rights Center 2014), often overlapping to citizen journalism, citizen witnessing and participatory newsgathering taking place within a journalistic ecosystem (Allan 2013). The participatory fact-finding paradigm of broadening participation in human rights advocacy has been explored by Land (Land 2009 and 2016). Connective witnessing acts to maintain, generate and engage ongoing publics (Mortensen 2015, Tufekci 2017) while particular domains of human rights and activist witnessing draw their strength and practice from specific traditions of resistance with longer histories —for example, "Black witnessing" (Richardson 2020).

Frontline witnesses and activists also interact with a field of advocates, activists and other distant witnesses. These analysts, journalists, investigators, debunkers, and verifiers of user-generated content act

themselves with “a range of expectations, purposes, accountability, and processes” (*Deepfakes responses*). New practices and communities have emerged around this, such as OSINT work for human rights, where people at a distance identify, verify, contextualise and use social media and other documentation (Dubberley, Koenig and Murray 2020). Other related practices include the broad collection of data via “mass archiving” approaches, human rights collectives collecting and making sense of comprehensive archives of violence and mediating the footage for publics (Deutch and Para 2020, Ristovska 2021), aggregative data witnessing practices (Gray 2019), ethical curation of witnessing acts and “forensic architecture” (Weizman 2017). There has also been a rejuvenation, amplification and further corporate co-opting of existing oppositional advocacy forms such as video remix and culture-jamming and the development of new mediums such as consumer live-streaming (*Live-streaming for witnessing*).

3.3 Participants in the expanding ICT-enabled human rights activism universe I articulate above utilise a range of advocacy and storytelling techniques. A focus through my published work is on mapping these novel approaches and identifying their relation to existing and emergent advocacy strategies from within ‘traditional’ human rights advocacy as well as parallel and broader fields of social activism.

In *Participatory Panopticon* I look at existing advocacy strategies, including ‘smart narrowcasting’ in which advocates make calculated media appeals to specific audiences (a concept I articulated and explored in earlier work, including Gregory 2006, 2010; Gregory et al. 2005). How do existing advocacy practices and principles such as audience targeting, mitigating sensationalism, and a commitment to ethical witnessing practices relate to emerging online-centric storytelling techniques and the prioritisation of the sensational on online platforms? My analysis identifies shifts by drawing on existing theory and reference points of transnational advocacy (Keck and Sikkink 1998, Bob 2005) along with contemporary case studies. In *Participatory Panopticon* I also map, drawing on participatory activism

(Jenkins 2006a), the impact of emerging DIWO (Doing It With Others) practices that draw on the wisdom, size, presence and energy of crowds. My forward-looking assessment of potential developments in the field was prescient in identifying the emergence of OSINT, spatial and participatory approaches from 2012 onwards.

3.4 Storytelling and advocacy shifts can also be understood through specific analysis of case studies, discussion of technology-enabled formats such as live-streaming and via field surveying and speculative work grounded in signals and analysis of trends. In *Kony 2012* I use the case study of the eponymous video as an example of unexpected viral video success and explore its relation to participatory activism concepts such as drillability and spreadability (Jenkins 2009, Mittel 2009) as well as to formulations of advocacy video (Gregory et al. 2005, Gregory 2006). In *Ubiquitous Witnesses* I map storytelling formats in live-streaming and in 'non'-storytelling formats of video gathered as potential evidence. Then, based on practitioner interviews and discussions in Brazil and elsewhere, I draw connections to various media formats, including radio. I build on this understanding in the *Live-streaming for witnessing* analysis of the Mobil-Eyes Us research, where I look at concrete strategies for mitigating denial of atrocity (Cohen 2001, Seu 2011), for countering the centring of violence and placing it in a broader matrix of life, and for understanding livestreaming within diversified witnessing practices, struggling for attention and grappling in storytelling with issues of visibility, risk and trust. In *Cameras everywhere revisited* I frame the issue of volume (see further discussion under Research Thread 2/Section 4) as a critical question that conditions a range of new storytelling and advocacy practices and the technologies that inform them - including the growing prevalence and necessity for OSINT approaches, 'video as evidence' field growth, and techniques of spatial analysis, polyvocal storytelling and the use of AI. This question of volume also draws upon the questions of diversification of practice and professionalisation of practice raised by Land (Land 2009, 2016) and Ristovska (Ristovska 2001). I also draw on informed analysis of signals and trends to identify, as in *Participatory Panopticon*, how human rights

defenders will likely need to change their narrative and advocacy strategies - drawing on smart narrowcasting and the positives of networked authenticity in online communities but also engaging with the risks of 'firehose of falsehood' attacks on human rights narratives and the challenges of compassion fatigue compounded by volume of content.

3.5 My work studying the who and how of participatory video activism for human rights contributes to discussion around mediated witnessing and to new conceptualisations of 'distant witnessing' and the establishment of the concept of 'co-presence' in a human rights context. Early work in this portfolio primarily focuses on conceptualisations of ethical witnessing that derive from the literature on witnessing (Peters 2001 and others) and then applies these to emerging practices such as remix, aggregation, curation and live-streaming (*Participatory Panopticon*; *Human Rights In the Age of Distant Witnesse*s; *Live-streaming for witnessing*). In *Ubiquitous witnesses* I introduce the concepts of 'distant witnessing' and also 'co-presence' as frameworks to understand existing conceptualisations of witnessing focused on active collaboration, immersive experience and solidarity between 'viewers' and frontline activists, based on the affordances of live-streaming and the possibility of synchronous interactions between the audience and the frontline witness. Within *Live-streaming for witnessing* I explore how this live-streamed and co-present distant witnessing takes place via tools (including Mobil-Eyes Us, Periscope, Facebook Live) in the context of practices and via the actions of community-based activists in Brazil.

In these papers, I particularly engage with how distant witnessing relates to theories of mediated witnessing in a more participatory media context (Orgad and Seu 2014, Richardson 2020), discussions of denial as a phenomenon in human rights activism (Cohen 2001, Seu 2011), commodity activism and trends in 'self-expressive' activism (Chouliaraki 2006, 2013), and how to manage issues of immersive witnessing and 'improper distance' (Chouliaraki 2015). Issues of denial were central to my research and practice before the publication period (see

Gregory et al. 2005, Gregory 2006). However, I increasingly centre concerns around mediated suffering. My contribution to the discussion is to ground an evolving understanding of mediated witnessing and 'mediated suffering' in specific human rights contexts of live-streaming and immersive witnessing (informing subsequent work by others in this area, see citations on *Ubiquitous Witnesses*), as well as to complement recent work by Ong (Ong 2015 and 2019) and others that push back on analyses of mediated 'suffering' that centre suffering over other experiences, and exclude proximate audiences of local activism and attention from consideration.

3.6 The concept of remix as integral to human rights activism with video, and subsequently considered as a potential act of 'distant witnessing' is central to my work from *Participatory Panopticon* onwards. In *Participatory Panopticon* I discuss the findings from two initial remix experimentation processes at WITNESS oriented towards broader publics and towards policymakers, and in *Human Rights In An Age of Distant Witnesses*, I look at more distributed single video remix approaches as well as curation approaches occurring in the early 2010s. I apply a distant witnessing framework to remixers and aggregators and reflect on this as an act of political participation (following Hinegardner 2009). Based on participant observation and case studies, I also start to conceptualise an understanding of 'reincarnated images' (what I would later term 'shallowfakes'), drawing on how videos and images are repurposed from one country or context to another. As I discuss further below, in the *Video for Change* paper, my co-authors and I identify remix video activism as one of the most ethically complicated forms of video for change, noting the challenges of moving away from source material, originator and consent.

3.7 Alongside the conceptualisation of distant witnessing and remix activism I note in the sections above, one dynamic that I analyse is the growth of the concept of 'video as evidence'. I am an early participant-observer in this area, participating in developing protocols and practices, a process that began in earnest during the Syrian conflict, as well as developing tools in this area, such as ProofMode. This evidentiary focus

reflects an evolution from my work before 2012. Indeed, in the *Participatory Panopticon* paper, I largely downplay a focus on video as potential evidence as an advocacy and documentation strategy. I increase my emphasis in subsequent publications. As I draw on my colleagues' work at WITNESS on video as evidence (Matheson 2015), my concern in analysing the potential collection of video as evidence is often to understand it in the context of decisions by witnesses about managing decisions around immediacy and long-term usage and visibility and protection (see further under Section 4). From *Ubiquitous Witnesses* onwards, this work is also in the context of the professionalisation discussion (as articulated by Ristovska 2021), which highlights the dynamics of emerging norms and expectations around the 'video as evidence' field and the specificities of how this constrains and enables activists. Although this is not a heavy focus in the work I present in this synthesis, the role of evidence-gathering norms as a way to set unhelpful and unrealistic professional boundaries is a recurring backdrop in my research and a recurring concern in my work.

3.8 In *Ticks or It Didn't Happen* I indicate how a **shift to authenticity infrastructure and a growing prevalence of deepfakes compounds a similar professionalising momentum around image analysis and media forensics** and highlights the power dynamics inherent in this given the dominance of law enforcement in existing media forensics work, and the role this may give to technologists over lawyers or ordinary human rights activists. The work in *Ticks or It Didn't Happen* and in *Deepfakes responses* also reflects the complications of forensic proof as a complement to evidentiary documentation practices and the contribution of an epistemic undermining of video (echoing Chesney and Citron 2019, Rini 2020) to driving particular practices of evidentiary video (e.g. multi-source or spatial mapping), to motivating particular technological approaches (e.g. authenticity infrastructure), and reinforcing perpetrator denial strategies such as the so-called Liar's Dividend (Chesney and Citron 2019). One area I will explore further in the future is the relation of all these phenomena to the broader so-called 'forensic turn' (following Anstett and Dreyfus 2015) in human rights investigations and how this trend, exemplified in projects such as Bellingcat

and Forensic Architecture, impacts a more diversified human rights witnessing field, and the expectations of what passes as real or truthful, or can be validated as such.

3.9 A corollary of these expanding strategies is the necessity to interrogate their impact, and what impact measurement strategies are effective. In *Video for Change* - the work in this portfolio solely focused on this area - my co-authors and I push back on tendencies in related fields, such as impact film, to rely on final outputs and metrics over process and 'impact pathways', and we argue for cross-disciplinary 'thick data' (Wang 2013) over numbers. One section on impact frameworks, where I focused my contributions, highlights how the diversification of the field and the range of participants and 'video for change' approaches noted above complicate impact measurements, particularly when we also apply lenses of ethics and professionalism. Citizen witnessing, aggregation, and remix confound known ethical questions in human rights witnessing and also take place on corporate-controlled platforms. Witnessing by ordinary people usually lacks a planned impact pathway, video curation often abrogates the intention of the creator and neglects risk, and perpetrator-shot video compromises all ideas of consent and informed participation.

3.10 On a broader level of discussion on 'impact', given the timeframe and research context, **my work is also in dialogue with the hype and corresponding research and public backlash to naivete around the role of social media in protests, social organising and activism, and in resistance to binary framings around social media's potential or actual impact.** It instead navigates a more intermediate, nuanced position. At the beginning of my research period, *Participatory Panopticon* avoided the naive liberation technology focus of the Arab Spring era, and *Kony 2012* made a contrarian argument for understanding the intentions and impacts of that contentious advocacy video. In *Cameras everywhere revisited*, eight years on from *Participatory Panopticon*, conversely I draw on field knowledge and survey the literature and case studies to assess that the pessimism around social media-mediated activism is not entirely consonant with the evidence.

3.11 Concluding this Section on the evolving how (and who) of practice, **the growing intersection of human rights activism using technology and citizen journalism with journalism, declining trust in the media and a growing 'radical distrust' in witnessing characterise my work from *Ubiquitous Witnesses* onwards**. Within my earlier work (including before the Published Works discussed in this synthesis) I largely excluded potential intersections of theorising and research between journalism and human rights activism. These intersections came into greater focus for me with others' work conceptualising citizen journalism and 'citizen witnessing' (Allan 2013), and my work after *Ubiquitous Witnesses* draws on these intersections in terms of research subjects, understanding my findings and potential approaches for translating insight into impact.

From *Mal-uses of Deepfakes* onwards, my research process, including in *Deepfakes responses*, and *Live-streaming for witnessing*, has explicitly included journalists, with a focus on citizen journalists and civic journalists as expert resources. One reason is that within the professionalisation pressures noted above, both human rights defenders and journalists engage with similar professional logics of reasserting control over processes such as verification (McPherson 2015, Hermida 2015).

In addition, as noted in Section 5 on technology and technology infrastructure, any consideration of the technology infrastructure impacting human rights defenders must consider journalism as an adjacent sector. Consequently, the *Ticks or It Didn't Happen* and *Deepfakes responses* work identifies both mainstream journalism and news outlets, in addition to citizen journalism and local journalism globally, as key protagonists in the dilemmas and participants in research.

4: RESEARCH THREAD 2: How have human rights professionals, journalists, and ordinary people participating in activism addressed recurring ethical and practical dilemmas around authenticity and trust, visibility and obscurity, and volume of media?

4.1 Introduction: **A consistent thread in my work is mapping the recurring dilemmas central to broader, more distributed human rights witnessing and activism using video, social media and technology and how these evolve or remain constant over time.** In Section 3, I explored how existing and emerging human rights approaches and ecosystems evolved. In this section, again as a reflexive practitioner and researcher, I review how participants grapple with underlying and recurring ethical and efficacy questions and how this raises research questions or requires refinement and reconceptualisation of theorisation or practices in response.

The particular recurring themes I engage with include how activists using digital media balance privacy, consent and visibility; how they grapple with the increasing volume of online and audiovisual media; and how they reconcile the need to prove trust, authenticity and evidentiary value.

4.2 A focus of my research has been around how activists and others manage the dilemmas of visibility and obscurity and of ephemerality and permanence. My work contributes a specific focus on the intersection of visual material and human rights contexts while building on broader academic discussions of the nature of social media (such as Marwick and boyd on context collapse, 2011), as well as sector-specific discussions around risk analysis for human rights defenders (e.g. Ganesh et al. 2016).

I approach questions of audiovisual visibility and obscurity from multiple perspectives in order to find new insights into this recurring dilemma in contemporary human rights work. In *Human Rights Made Visible* I articulate an understanding and analysis of visual anonymity grounded explicitly in human rights norms and practices of ethical witnessing and assess potential steps to address needs within practice and infrastructure. Methodologically, the research in the *Human Rights Made Visible* paper informed the parallel development of media artefacts to illustrate key issues and research outcomes and to use for advocacy towards tech platforms. One such media artefact, ObscuraCam, was the first publicly available tool focused on visual anonymisation for mobile images and video and was used as a reference design to advocate for the introduction on YouTube of a blurring tool available in-platform.

Activists make both active and forced choices around visibility. In *Technology and citizen witnessing* and then in *Ubiquitous Witnesses* I explain one core friction that relates to the discussion of witnessing approaches noted above, and that recurs across the research areas in the synthesis. This friction is between the synchronous activism value and the asynchronous evidentiary value of documentation. Human rights defenders must make contingent decisions around visibility and obscurity that are not constant and where their content is subject to context collapse. In both *Technology and citizen witnessing* and the subsequent live-streaming specific research of *Livestreaming for witnessing* I highlight how activists' risk assessment occurs, but is constrained by the tools available, their literacy on those tools and the structural conditions of platform governance and broader surveillance. In *Livestreaming for witnessing* I explore how, in a specific scenario in Brazil, activists make decisions on avoiding 'self-surveillance' (Kavada and Treré 2019) from constant self-documentation, boundary management and surveillance realism. As I discuss further below in Section 5 on technology platforms, such understanding of the contingent and uncertain nature of activist participation in these audiovisual platforms and with these audiovisual technologies is also mirrored in my own usage of a range of ambivalent and non-binary terms that other scholars have

conceptualised to describe the overarching surveillance architecture in which human rights witnesses as a subset of society participate (including 'participatory panopticon' per Cascio 2005, 'sousveillance' per Mann, Nolan and Wellman 2003).

4.2 In keeping with a growing research focus on broader technical infrastructure and government regulation as a precondition for understanding activism, from *Cameras everywhere revisited* onwards I bring an increasing focus on the role of infrastructure and data surveillance in visibility and obscurity. This emphasis on understanding visibility and obscurity, surveillance and data mining is a core focus of the *Ticks or It Didn't Happen* report by Ivens and myself. In Dilemma 3 of *Ticks,* we consider the effects of authenticity infrastructure on chilling and enhancing voice. This analysis also recurs in the expert consultations that form the research basis of *Deepfakes responses'* analysis of civil society and human rights defender concerns about deepfakes. Returning to *Ticks*, in Dilemma 14 (a chapter by an outside author supervised by the primary authors), the technical concept of blockchain is used to explore issues of mutability, visibility and permanence. The impact of the *Ticks* analysis is explored further in the Impact section below but includes its impact on key industry white papers and technical standards. The Content Authenticity Initiative whitepaper (Parsons et al. 2020) and the Guiding Principles and Technical Standards for the Coalition for Content Provenance and Authenticity (Coalition for Content Provenance and Authenticity n/d a, n/d b) include key principles around privacy and anonymity, prioritise human rights activists with privacy concerns as primary workflows, and build on illustrative use cases centred on global human rights and journalism needs.

4.3 The impact of the volume of digitally mediated and online audiovisual content is a recurring contour of my research. **What is the relationship of volume as a characteristic of much of the communication environment to both the effective and ethical practice of human rights activism?** In the topographical essay work in this synthesis - in *Participatory Panopticon* and *Cameras everywhere revisited* - I map this

volume as having positive and negative consequences in both the broad environment and specific circumstances. I identify its implications based on the field surveying work and the grounded perspective of the network at WITNESS. In *Cameras everywhere revisited,* I ground volume in the theorisation of more diverse stories and a tilt into 'participatory human rights fact-finding' (Land 2009 and 2016) and into citizen witnessing (Allan 2013). I indicate how this provides both more opportunities for diverse advocacy and networked authenticity of local stories, but also the risk of loss of critical stories, exclusions of those that are less compelling audiovisual material and the risks of 'pics or it didn't happen'. I connect the reality of audiovisual volume to the prevalence of emerging OSINT, geospatial and 'forensic architecture' strategies. In *Human Rights Made Visible* I articulate an ongoing dilemma addressed in the professional training materials that I engage with and supervise at WITNESS, around how the audiovisual content can create harm when released, even when it has no impact on the desired terms of the protagonist human rights defenders because of its obscurity amid volume. I return to this dilemma in the *Livestreaming* research, where live-streams result in self-surveillance that is counterproductive to the aims of the local activists and lacks impact on the terms they seek.

4.4 Volume is usually transmuted into questions of 'scale' when the focus is on technology platforms and their consideration of emerging technologies and mitigation strategies for harms. This understanding of the challenge of 'scale' for platforms informs the work I focus on in Section 5 of this synthesis. In *Ticks or It Didn't Happen,* as well as in *Deepfakes responses* and in my contributions to Bontcheva et al, 2020, I centre an understanding of how scale informs platform decisions on implementing authenticity infrastructure, on utilising tools of deepfake detection and how they must rely in both cases on technical adjuncts to support human decision-making, and the implications of this automation. Here, I build on a broad range of academic discussions on automated content moderation (Gillespie 2018, Roberts 2019, Jaloud et al. 2019) and apply it to this particular area of interest.

4.5 As in other research areas, I look at what norms, professional and amateur practices, tools, and infrastructure emerge as responses to pressures on authenticity and trust. As noted before, my early work de-emphasized the evidentiary value of video. However, in line with the growing discussion around authenticity and trust in online and open–source video, this changed by the time I was working on *Ubiquitous witnesses*. This shift is reflective of the growth of a more explicit human rights 'video as evidence' field (Human Rights Center 2014, Matheson 2015). In *Ubiquitous witnesses* I indicate the combination of skills, tools and authenticity approaches needed to engage in the nascent 'video as evidence' professional practice field.

Authenticity and trust concerns are also driven by the growing pressure on the epistemic value of the more diversified set of news and human rights information sources with which my work and research engages and also reflect the broader discussion of declining trust in media and trusted sources. *Cameras everywhere revisited* provides a field perspective on the real and manufactured trust crisis. *Ticks* builds on this in a specific context, raising specific questions about the risks in binding trust to social media platforms, the problems of the 'forensic' or sceptical mindset, and how certain approaches normalise techno-centric epistemic foundations around who and what we trust (e.g. in the use of blockchain). The needs for skills, tools and infrastructure enter sharper focus in the *Mal-uses of Deepfakes* report. *Deepfakes responses* confirms these explicit understandings via research into the risks anticipated by key human rights and civic journalism ecosystem participants, who live within a climate where civilian and civic witnessing is already confronted by radical doubt and distrust.

4.6 Metadata has a complicated public reputation. **An interesting observation tracking the trajectory of this work in relation to trust and authenticity is to note how it engages with the idea of metadata over a period from 2015 to 2022** including the negative connotations of metadata that predominated in the wake of the Snowden revelations (as highlighted in

Technology and citizen witnessing), the move to a possible ‘metadata for good’ in WITNESS’s own work and its co-design on the tool ProofMode, and then the re-normalization of metadata as a response to misinformation and disinformation fears that underlies the initiatives analysed in *Ticks or It Didn’t Happen* such as controlled capture apps and authenticity infrastructure.

4.7 A critical research question for me is the consistency of concerns over time. Reviewing this thread of research I find that concerns around safety, security, visual anonymity, efficacy and trust remain constant as factors yet evolve in practice. In *Participatory Panopticon* I articulate consent, safety and efficacy as key recurring concerns identified in a topographic survey. In *Human Rights Made Visible* I connect these ethical strategies to actual practice and forward-looking proposals for how to reinforce these underlying principles with the force of human rights norms and the actualisation of technology, law and practices that perpetuate them. In specific concrete scenarios, I then review these concerns in *Live-streaming* as they apply to live-streaming and co-presence strategies in the Mobil-Eyes Us program in Brazil. Here, participants had to make day-to-day decisions about their interventions in an environment characterised by volume and consider the safety implications of their actions for themselves and their communities.

Cameras everywhere revisited returns to these key ongoing questions of safety and security, trust and credibility, and content volume. However it reads them more explicitly in intersection with platform power and digital authoritarianism. I discuss this evolution of my thinking further in Section 5 below on technology platforms, reflecting a growing public awareness and academic discussion on platforms' mediating and coercive power. The key insight I make in *Cameras everywhere revisited* is that the broad issues named above have remained consistent in mapping the topography but have changed in scope and scale. They are now also being more effectively weaponised against human rights defenders by opponents. In *Ticks or It Didn’t Happen*, Ivens and myself and in *Deepfakes responses* I apply these

recurring dilemmas to activists' and journalists' desired interventions into the emerging authenticity infrastructure and related responses to deepfakes.

4.8 A final recurring contribution from my work is to emphasise the need for contextualisation and historicisation. As discussed in Section 3 on the evolving practices of witnessing, I first apply this to techniques and approaches in *Participatory Panopticon*. In other work, I draw on existing human rights norms and standards (*Human Rights Made Visible*), reflecting existing established human rights law and indicating the viability of existing human rights-based approaches in contrast to a Silicon Valley laissez-faire or Chinese digital authoritarianism (in *Cameras everywhere revisited*, Bontcheva et al. 2020) This application of human rights law to online social media platforms is best contextualised in the work of former UN Special Rapporteur on Freedom of Expression David Kaye's reports and the growing field of business and human rights that followed the creation of the UN Guiding Principles on Business and Human Rights. I also endeavour to place a hyperbolised new phenomenon, such as deepfakes or authenticity infrastructure, into a broader context of information disorder, journalism, state suppression of civil society and existing attack strategies on journalists and human rights defenders. I also connect synthetic media and responses to existing expertises and fields, including journalism and OSINT, and existing long-standing experiences of human rights defenders and vulnerable populations (*Mal-uses of Deepfakes*, *Ticks or It Didn't Happen* and *Deepfakes responses*).

5: RESEARCH THREAD 3: How have broader technological infrastructure and systems - particularly evolutions in systems and technologies for falsifying content and, conversely, enhancing trust in audiovisual media - conditioned possibilities for human rights activism by professionals and non-professionals and impacted practices and ethical questions?

5.1 Technologies impact practices and critical voices are excluded from agency over the development and operationalisation of these technologies.

I have increasingly focused within academic publications as well as in practitioner-oriented and technology white papers and research reports on how **existing online platforms, as well as emerging technologies and technical infrastructure, intersect as active protagonists with existing and nascent ecosystems and with practices of mediated human rights video and image-making.**

The contextual role of content moderation and platform policy has been widely explored in work including Youmans and York 2012, Gillespie 2018, Roberts 2019, and Jaloud et al. 2019, among others. While algorithms, social media company policy and the affordances of particular platforms impact all users, participants and broader societies, civic journalists and human rights defenders globally are among the most vulnerable participants. They are also critical platform users from a public interest perspective. Consequently, it is essential to assess how these contexts impact the ability of these individuals and organisations within a diversifying human rights communication and organising field to effectively and ethically witness, report and advocate and to pinpoint ongoing dilemmas participants face.

One part of my research contribution focuses on how users act in consciousness of constraints they cannot control, both broadly within human

rights work but also in the evolving map of specific sub-genres and approaches I articulated in earlier sections, such as live-streaming and remix activism (*Human rights in the age of distant witnesses,* as well as prior work Gregory and Losh 2012*)*.

However, activists and human rights defenders also look to be key protagonists in developing protocols and participants in design. They are also end-users and, in addition, targets of emerging technologies. This engagement also often needs to extend beyond platforms to other technologies. **Via deep field grounding, expert interviews and structured convening, my research contributions identify critical dilemmas facing human rights practice, specifically in the design of emerging technologies,** including live-streaming, authenticity infrastructure, deepfakes and AI-based media creation and manipulation. I then identify how to build infrastructure and develop emerging technologies responsive to these dilemmas and centring the needs of a global and diverse range of human rights communicators.

The **research and commentary in this Section 5 is deeply grounded in the earlier Section 3 and 4 discussions on the *how* (and who) of human rights practice and the recurring ethical and normative dilemmas of visibility/privacy, trust and presence amid volume.** In this research and publication area, of the three I discuss in this synthesis, I particularly bridge between academic publications*,* synthesis, and research work intended to influence the field, alongside explicit contributions to critical technical standard-setting*.*

5.2 Human rights defenders are frequently compromised yet critical public interest participants in private online spaces (Zuckerman 2010)*.* **Human rights activists' participation in commercial spaces reflects both adaptive practices at the community level as well as largely insuperable infrastructure barriers.** A through-contribution within my work

is an evolving understanding of the nature of activist participation in commercial spaces and their increasing consciousness of how they operate within these constraints. Contextually, my work is in dialogue with the range of discussions around reconciling more widespread media production and social media participation with broad surveillance and sousveillance on platforms and elsewhere.

5.2 In the earliest work in the portfolio, in the discursive essay, *Participatory Panopticon,* I conduct a field analysis of the state of online video and human rights, drawing on insider research on the challenges of running an independent online video site at WITNESS ("The Hub"). *Participatory Panopticon*, *Human Rights Made Visible*, and *Kony 2012* demonstrate **how activists and defenders attempt to reconcile concepts of human rights advocacy – such as targeting a specific audience and crafting narratives – and core ethical dilemmas, such as consent and ethical witnessing, with the demands of operating in the 'vaudevillian'** (Jenkins 2006b) **characteristics of online venues, such as early YouTube.** My primary framework for understanding these contexts builds on understandings of participatory activism as explored by Jenkins 2006a, and early analyses of the nature of the online space and specific venues such as YouTube (Burgess and Green 2009, Zuckerman 2010). A research and theoretical gap - somewhat reflective of the field at the time - is a more limited theorisation of the role of platforms as content moderators. This area of work is more prominently reflected in later work in my portfolio, reflecting the further growth of a field of academic discussion on content moderation (including Gillespie 2018, Roberts 2019, and Jaloud et al. 2019) and the capacity to build on these understandings and approaches. In the *Mal-uses of Deepfakes responses* paper and in *Cameras everywhere revisited*, I revisit the dilemmas of public activism in private spaces that I first discussed in *Participatory Panopticon* (drawing on Zuckerman 2010) and the context of platforms as invisible and visible gatekeepers. In *Human Rights In An Age of Distant Witnesses* (drawing on Gregory and Zimmerman 2011 and Gregory and Losh 2012) I explore the contingency of remix activism on commercial platforms via case studies, and how users anticipate this contingency and

constraint on a platform like YouTube. My work contributes to a further understanding of how human rights actors view and understand their role in these commercial spaces, building on the broader scholarship noted above.

5.3 Live-streaming and co-present activism is a particular focus of my analysis for understanding this intersection of practices and infrastructure. In *Ubiquitous Witnesses and Live-streaming for witnessing* I draw on case studies and field survey-based work to provide a topography of emerging human rights livestreaming approaches and their relationship to mediated witnessing. I use a futures-based research methodology in *Ubiquitous Witnesses* to identify risks and opportunities at the intersection of practice and technology infrastructure, including connecting audiovisual technologies to parallel developments in technologies of task deployment that enable people to engage in a coordinated manner on shared needs as well as to concepts of co-presence more often discussed in relation to immersive and virtual environments. I continue on in *Live-streaming for witnessing* to explore these practices in a specific empirical context of the experience of favela-based activists in Brazil. This experimental project further highlights how issues of surveillance and sousveillance are addressed in the specific circumstances of Brazil and live-streaming practices- without the live-streaming 'self-surveillance' evident in the Occupy movements (Kavada and Treré 2019), and with cautious boundary management and surveillance realism by participants, reflecting the visibility/obscurity dynamics discussed in Section 4.

5.4 Infrastructure is determinative for many actual and potential human rights usages of technologies (Section 3), as well as for the ability to mitigate safety, security, visibility and trust issues (as identified in Section 4). Commercial infrastructure is not always an optional choice for activists. In *Cameras everywhere revisited*, I contextualise a field understanding (Lim 2017, Tufkekci 2017, Kayyali 2018, Xiao Mina 2019) around the nature of platforms as critical infrastructure for quotidian communication, community organising and advocacy that most activists cannot easily discard in response to public clamour to #DeleteFacebook.

This understanding informs the research on emerging technology that could be incorporated into platforms and infrastructure.

5.5 The questions of what activists can and cannot obtain from critical infrastructure form the context of the research questions in my work in *Ticks or It Didn't Happen*, *Cameras everywhere revisited* and *Deepfakes responses*. In *Mal-uses of Deepfakes*, the intention is to translate these dilemmas into concrete proposals and impact upon an emerging area of challenge in activism and information-sharing in private spaces, namely deepfakes - for example, identifying content detection, authenticity and moderation approaches to these phenomena.

Ticks or It Didn't Happen and my work on deepfakes address the conceptualisation of 'misinformation' and 'disinformation' as societal phenomena with potential technical solutions and the implications of these 'solutions' for marginalised communities and human rights defenders globally as well as for grassroots practice in the field of digital media and human rights. I highlight a critical need to historicise these developments within broader understandings of media manipulation and power and within an intentional conceptualisation of the potential harms to vulnerable groups arising from not only neglect in participation in design but also forced inclusion in emergent infrastructure.

5.6 In *Ticks*, myself and co-author Ivens **make a critical contribution to understanding emerging authenticity infrastructure,** researching a nascent area of technology that was then just about to enter the mainstream and using expert interviews, research and analysis to consider how recurring assumptions around privacy, security and trust apply in this area, and to identify key dilemmas that inform technology infrastructure development. These dilemmas include those highlighted in Section 4 as recurring normative questions and assumptions that must be assessed and tested to see if they continue to hold true - around visibility and obscurity, surveillance, and credibility of content in a broader ecosystem of content and communication volume.

Ivens and myself also specifically address platform control of emerging authenticity infrastructure, adding new insights concerning this area of existing discussion. Our focus is on how platforms will mediate trust in new ways, their ability to 'lock-in' users via their ability to do verification and new considerations of how content moderation and appeal occur when a 'verification' layer is also part of this decision-making. In both areas of recurring ethical questions and the specific context of platform power, we then identify approaches drawn from research and consultation to address these concerns.

5.7 In *Mal-uses of Deepfakes* and *Deepfakes responses*, research and convening grounds the first critical work to identify the need for proactive design with equity and inclusion in responding to deepfakes and to concretely propose insights on how to do this. Based on a series of expert consultations globally, I further expand on the role of 'invisible' infrastructure in the *Deepfakes responses* paper and how civil society concerns must be incorporated. I identify platform roles in handling content moderation and the particular fault lines around how they manage remix and satire that will be most challenging in the context of AI-manipulated media, as well as raise the research questions on how they will manage the authenticity infrastructure dilemmas first identified and outlined in the *Ticks or It Didn't Happen* paper. I identify the implications for human rights defenders in *Deepfakes responses* of the gaps in access to detection tools, of forced inclusion in authenticity infrastructure, of the 'ratchet effect' of new infrastructure and point to the issues of AI and its additional challenges with auditing and transparency around how it functions.

***5.8 Deepfakes responses* and *Ticks or It Didn't Happen* reflect an attempt within my work to engage more explicitly with issues of inclusion and exclusion in tech design and to translate this into external impact on policy and technical standards.** My work and research design in *Participatory Panopticon* and *Human Rights Made Visible* has an under-theoreticised understanding of the absence of diversified global input

and consideration in platform decision-making. This under-theorisation also reflects a broader field of advocacy and research that was yet to develop in this space (a development that I describe in *Cameras everywhere revisited*). These gaps also reflect limitations in my own theoretical background (for example, in relation to STS) and an evolving process of recognising my subjectivity and positionality in relation to my professional work and research.

In my work from the *Mal-uses of Deepfakes responses* project onwards, I more explicitly focus on inclusion - global, intersectional, early in processes and ongoing - as a critical dimension of technology and human rights. I draw on research methodologies for consultation around public technology issues, such as the University of Washington's Diverse Voices methodology (Young et al. 2019) and justice-oriented design practices (Benjamin 2019). This focus on inclusion enables the foregrounding of issues around deepfakes and authenticity infrastructure identified by globally diverse and vulnerable communities. *Ticks* centres global concerns and considerations of marginalised groups and resulted in the inclusion of critical issues reflective of concerns heard in the research process in the conceptualisation and technical standards being developed globally for authenticity infrastructure (Parsons et al. 2020*,* Coalition for Content Provenance and Authenticity n/d b). Similarly, the inclusive consultation and research outlined in *Deepfakes responses* enabled critical input on the development of tools and policies in this area (further outlined below in terms of impact in Sections 6 and 7)

6. IMPACT ASSESSMENT: CONTRIBUTION TO KNOWLEDGE

6.1 Between 2007-22, the human rights field saw a growth in participation in both formal human rights work and other forms of civilian witnessing, facilitated by the accessibility and utilisation of video and social media. New practices emerged including human rights-centred OSINT to discover, verify and prove crimes, and curation on platforms. Existing practices of strategic advocacy evolved, and there was a new iteration on cultural advocacy forms such as video remix and the utilisation of novel consumer formats such as live-streaming. Throughout the period, practitioners and activists faced ongoing challenges, evolving with societal and technical shifts, around what and how to trust media and content, around choices of personal and media visibility and obscurity and around how to make their content meaningful amidst escalating volume. Activists using digital media had to balance privacy, consent and visibility; consider how to reconcile the need to prove authenticity, guarantee trust and enhance evidentiary value; and challenge established power dynamics of existing forms of solidarity and mediated witnessing.

6.2 The publications within this PhD by Published Work synthesis collectively provide a contribution to knowledge during this decade of significant evolution and continuity in the field of human rights practice, professionalised and otherwise. They open up new fields for further research and provide new insights from fieldwork to build on existing theories and paradigms. They introduce new models, paradigms, conceptual frameworks and test frameworks for distant witnessing in practice. They show that “taken for granted” truths or assumptions about professional practice, strategy, and underlying ethical concerns are not substantiated by contemporary evidence or have significantly evolved. I place research grounded in specific scenarios and activist contexts within a broader geopolitics of increased digital surveillance, platform power, rising authoritarianism, and global problem framings such as ‘misinformation’ and ‘disinformation’. I outline each of these contributions below.

6.3 This work has received **a range of citations covering a range of disciplinary fields and including prominent journals** such as *Information, Communication & Society* combined with a number of contributions to key journals specific to the human rights space, such as the *Journal of Human Rights Practice*. All citations are listed in the Appendix.

6.4 The research has also directly impacted policy, practice, and critical and emerging technical standards and approaches, which I discuss in Section 7.

6.5 Mapping the field: As a practitioner-researcher working within a rapidly evolving field of practice, publications including *Participatory Panopticon* and *Cameras everywhere revisited* map emerging topographies of professional and civilian activism using audiovisual media at two specific moments: in the early days of mass online video and participatory human rights practices; and at a subsequent moment of challenge and frustration, constrained by platform power and (digital) authoritarianism. In each case, they effectively reflect on continuity and change in key concerns, issues and approaches, drawing on informed field scanning, professional practice, and experience from my work context.

6.6 Evolution of existing paradigms and assumed truths: Reflecting the nature of evolutions and constants, the research looks at how assumptions taken for granted evolve, and how existing models and paradigms can be improved: for example, around smart narrowcasting and effective human rights advocacy (*Participatory Panopticon*), around remix and curation, as well as advocacy strategy (*Human rights in an age of distant witnesses*), the efficacy of live-streaming in complex human rights situations (*Live-streaming for witnessing*), and the nature of anonymity and privacy in an increasingly visual era (*Human rights made visible, Technology and citizen witnessing*). *Deepfakes responses* provides new insights beyond the hype on how to understand a highly publicised phenomenon. *Video for Change* shows the

limitations of existing models for assessing impact when considering civilian video, remix and perpetrator video.

6.7 Insights into emerging areas, grounding them in existing context and demonstrating the need for multidisciplinary and case study-based work: The work within *Mal-uses of Deepfakes*, *Deepfakes responses* and *Ticks Or It Didn't Happen* provides insights into the emerging technical phenomena of deepfakes and authenticity infrastructure grounded in fieldwork and expert consultation, as well as case studies. They provide a map for potential research, policy and advocacy actions, grounded in the perspectives of human rights defenders and journalists and a human rights approach. They consider the implications of emerging technical infrastructure for the ability of individuals and organisations within a diversifying human rights communication and organising field to effectively and ethically witness, advocate and report. The contribution of my research rests in the close integration of understanding of critical dilemmas in human rights and civic journalism practice - identified via convening and research - with the theorisation of how to build infrastructure and develop emerging technologies with responsiveness to these dilemmas and centring the needs of a global and diverse range of human rights communicators. These perspectives were not present in the literature on these areas at the time.

Based on in-depth convening and research work, these publications highlight the complexity of these 'problems' and their 'solutions' and the need for a multidisciplinary response and an integration of professional practice. In each case, the research draws on contextual and inter-disciplinary knowledge to historicise, re-contextualize and deepen understanding of emerging fields around deepfakes and authenticity infrastructure by connecting them to existing scholarship and research on similar phenomena - e.g. verification, OSINT, authenticity, misinformation and disinformation - as well as to lived experience and community experience of related technologies and problems.

6.8 Original data from fieldwork to expand understanding and theory: Specific case studies in *Live-streaming for witnessing* and *Kony 2012* as well as those in *Human rights in an age of distant witnesses* add progressively to the understanding of professional and amateur practice of human rights work with video and social media. In *Live-streaming for witnessing*, I applied in practice a new conceptual framework of co-presence and of 'distant witnessing' first proposed in *Ubiquitous Witnesses*. The action research helped generate understanding on how this distant witnessing and co-presence model worked in relation to engagement, mediated witnessing, denial and surveillance. From a theoretical understanding of these areas, it introduced original data from fieldwork to better inform work on immersive and co-present witnessing and the relationship of these new practices to existing understandings of mediated witnessing and spectatorship.

6.7 New insights and new conceptualisations: The research in *Ubiquitous witnesses* introduces new conceptualisations around witnessing ('distant witnessing') and looks at the existing witnessing literature around the two poles of live-streaming and evidentiary documentation. Building on data from case study analysis, a research scan of WITNESS experience on 'video as evidence' and speculative analysis via research and interviews on distant witnessing, the article provides new insights into these emerging areas from the fieldwork and research.

A key output from my ongoing process of field-mapping, identifying shifts, and field-testing is the concept of active 'distant witnessing'. Academic literature has focused extensively on distant witnesses in the context of trauma, natural disasters and terrorism incidents (Howie 2015) and on understanding mediated witnessing at a distance (Peters 2001, Orgad and Seu 2014, Ong 2015 and 2019). In *Ubiquitous Witnesses* and onwards, I identified a category of media activism mainly centred on live-streaming, remix action and 'forensic' analysis such as OSINT, and placed it in a witnessing framework, naming characteristics of 'distant witnessing' and correlating it with co-presence, active engagement with the witnessing texts, and collaborative action. Distant and frontline witnessing concepts were

integrated into others' analysis of the field, including via citation to *Ubiquitous Witnesses* (see above) as well as in work by Richardson (2020) and Martini (2018).

6.8 Novel artefacts to answer and illustrate research questions within professional practice and public advocacy: The research projects discussed in this synthesis have, a number of times, been created in tandem with reference designs for technology and prototypes that propose answers to research questions in a real-world case study. Examples of these include the development with collaborators at the activist collective, Guardian Project, of ObscuraCam (a tool for visual anonymity developed in conjunction with research around choices on visibility/obscurity for activists), ProofMode (a reference design for authenticity infrastructure), and Mobil-Eyes Us (a co-presence and distant witnessing tool developed and discussed in the *Live-streaming for witnessing* paper). I have seen how physical artefacts, tools and reference designs illustrate research findings and contribute both to public advocacy on related issues and utilisation for impact.

7: IMPACT AND IMPLICATIONS FOR POLICY AND PRACTICE

7.1 Impact on public discussion, policy and professional field practice

The works identified here on deepfakes and authenticity infrastructure have significantly influenced public and policy discussions in these areas. A strength of my research has been its integration and bridging between academic publications, related synthesis and research work intended to influence the field, and contributions to critical technical standard settings. This research's impact on practice includes field influence on other practitioners in the field and in broader related fields such as journalism, technology development, and mis/disinformation.

The *Mal-uses of Deepfakes* research identified a series of research and action areas that have subsequently helped shape the field of concerns and solutions around deepfakes and related generative AI tools. These include a focus on ways to understand media provenance and shaping a category of infrastructure known as 'authenticity infrastructure' and similar terms (including via the subsequent *Ticks or It Didn't Happe*n report). Other concerns foregrounded in public discussions via the research also include the need for global and inclusive threat modelling incorporating particular attention to marginalised populations and attention to the issues of satire/humour.

The *Mal-uses of Deepfakes* and *Deepfakes responses* research on deepfakes and appropriate responses has widely informed subsequent media discussion (see Google News listing here), as well as public discussion on issues including the accessibility of detection infrastructure for deepfakes and the trade-offs inherent between access, inclusion/exclusion at a global scale and utility. The *Mal-uses of Deepfakes* and *Deepfakes responses* research has informed editorials in the Washington Post (Washington Post 2019), as well as multiple Congressional witness testimonies and briefings by myself and others (Clark 2019, U.S. House Oversight Committee 2022, Gregory 2023 b and Gregory 2023 c, U.S.

House Oversight Committee 2023, U.S. Senate Commerce Committee 2023). The research in *Ticks or It Didn't Happe*n has informed reports, including the US Federal Trade Commission report to Congress on *Combatting Online Harms Through Innovation* (Federal Trade Commission 2022), the Congressional testimonies noted above, submissions to UN Human Rights Council and Special Rapporteur mechanisms and inquiries in the UK House of Lords, as well as dialogue with lawmakers on particular legislation in 2023-4 focused on provenance and authenticity of AI and human-generated media.

The research's direct impact on practice also includes the implementation of findings in my professional practice at WITNESS, a global human rights organisation operating currently with more than 50 team members across twelve countries. All publications identified in this portfolio have had an influence on the programmatic decision-making of WITNESS, including

- *Participatory Panopticon* (vis-a-vis tactical and strategic decisions on online video)
- *Human Rights Made Visible* (tools and tactical investment in visual anonymity)
- *Technology and citizen witnessing* (investments in tools to secure authenticity)
- *Ubiquitous Witnesses* (strategic decisions to do research and experimentation work on distant witnessing and co-presence)
- *Human Rights in the Age of Distant Witnesses* (strategic decisions to do research and experimentation work within WITNESS Media Lab)
- *Mal-uses of Deepfakes* (input to shape a globally-leading 'Prepare, Don't Panic' initiative focused on inclusive preparation for deepfakes)
- *Cameras everywhere revisited* (focus in strategic plan on volume, security, authenticity as key guiding principles)
- *Ticks or It Didn't Happen* (direct advocacy leading to impact articulated in this Section on commercial and multi-stakeholder authenticity infrastructure initiatives and standards)
- *Live-streaming for witnessing* (informed decision making on continued approaches to diverse forms of distant witnessing at WITNESS)

- *Deepfakes responses* (articulated a continuing roadmap for an inclusive plan to counter both deepfakes and potential harms from solutions).

7.2 Impact on technical approaches and standards, and on commercial and independent tools

Two prominent initiatives focused on content authenticity and provenance have been heavily influenced by the research in *Ticks or It Didn't Happen* and *Mal-uses of Deepfakes*.

The Content Authenticity Initiative whitepaper (Parsons et al., 2020) is the foundational whitepaper of an eponymous initiative including hundreds of media, technology and academic groups working on issues in this area. The *Ticks or It Didn't Happen* report was shared widely with senior leaders in founding actors including Adobe, Microsoft and the BBC, and the key findings were presented at the initial conference launch for the initiative. Subsequently, key findings reflecting the dilemmas in the research were integrated into the whitepaper - including a focus on privacy, accessible tech, global applicability and minimising harms and risks. Guiding Principles directly reflecting the *Ticks or It Didn't Happen* report include 2.2. Privacy 2.3 Global Audience and Applicability, 2.7 Simplicity and Cost Burden, and 2.9 Misuse, with its focus on review for 'ability to be abused and cause unintended harm, threats to human rights, or disproportionate risk to vulnerable groups globally'. Additionally, the whitepaper explicitly includes, as expected users, both human rights defenders and professional and citizen journalists in high-risk environments and highlights Human Rights Activists as one of three critical workflows, and one needing a focus on privacy, anonymity and redaction.

The Coalition for Content Provenance and Authenticity (C2PA) Specifications (Coalition for Content Provenance and Authenticity n/d b), a global technical standard for shared development of authenticity and provenance approaches, also draws on the research from *Ticks or It Didn't Happen*, and

incorporates related principles into their approach. The C2PA also conducted further extensive threats and harms analysis grounded in the dilemmas identified in the *Ticks* report and incorporated global focus groups and analysis of potential technical and normative solutions. These findings further consolidated the validity of the findings in the *Ticks* report.

Research on authenticity infrastructure, including *Technology and citizen witnessing* and *Ticks or It Didn't Happen,* has also informed the development of a leading authenticity tool, ProofMode.

Research on visual anonymity conducted in *Human Rights Made Visible, Technology and citizen witnessing* introduces a novel dimension to discussions of anonymity and privacy. The necessity of incorporating functionalities for visual privacy and anonymity into consumer tools informed decisions by YouTube to incorporate a 'blurring' functionality into the most widely-used video-sharing platform globally.

8: FUTURE DIRECTIONS, PREVIOUS LIMITATIONS - ISSUES, METHODOLOGIES AND PROJECTS

8.1 Introduction: Some areas in this synthesis are at significant inflection points in external society and my own research in 2023/24. In this section, I highlight critical areas of future exploration around the implications of generative AI, the evolution of platform and technical infrastructure and the renewal of remix practice both in deepfake/synthetic media and prevalent social media practices. I go on to highlight methodological limitations and disciplinary gaps in the portfolio work and how I see these relating to future directions. I conclude by noting how my own situated personal subjectivity and professional subjectivity have influenced my research methodology and future directions.

8.2 A first future research direction I am exploring in research and practitioner contexts is how to reinforce the integrity of human rights documentation given the increasing versatility of synthetic media creation and the increasing accessibility (within defined parameters of class, wealth and global positionality) of advanced media production and editing tools. After the dates of this synthesis, an initial publication in this area is *Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI* (Gregory 2023).

The publications in this portfolio were produced in advance of the increased accessibility of so-called multimodal ‘Generative AI’ tools (based on Large Language Models and diffusion approaches to machine learning) for text-to-image, text-to-video, image-to-image, image-to-video as well as video-to-video creation and other modalities. These generative AI tools exacerbate the dynamics articulated in the *Mal-uses of Deepfakes* and *Deepfakes responses* papers related to global inclusion in understanding potential harms. They likely increase intersectional risks to vulnerable populations and the potential for undermining critical accounts with the ‘liar’s dividend’ and easier plausible deniability of evidentiary content.

8.3 Work on detection approaches and authenticity infrastructure in the generative AI context now needs to engage further with the research questions identified in the *Deepfakes responses* paper around centring global threats and priorities and ensuring equitable access and related capacity to detection and other technical solutions. In *Deepfakes responses* and the *Ticks or It Didn't Happen* report, myself and co-authors highlight key risks in authenticity infrastructure - responses to which have been incorporated into some of the emergent approaches, as detailed in the Impact section above. However, as I identify in *Deepfakes responses*, we need 'qualitative and quantitative' research on how to shape the increasingly prevalent authenticity infrastructure in response to, at minimum, known existing harms, as well as to understand better the impacts of audience reception of authenticity claims. The concerns I highlighted in *Cameras everywhere revisited* around platform power and governments' increasing sophistication in both democratic regulation and digital authoritarianism also emphasise the importance of the need for further research highlighted in *Deepfakes responses* around how these emergent technical infrastructures are being or risk being 'co-opted within platforms as well as political regimes and their free expression-suppressing response to "fake news".

I note a limitation of my previous work is its engagement with significant work in both digital media studies more broadly and in STS that could inform this work on emerging infrastructure and norms in response to generative AI and human rights. STS approaches could both inform the underlying theory related to the publications in Section 5 of my portfolio and future directions in this work and help me conceptualise how to move research principles and findings into policy advocacy and external impact.

8.4 A future research question will be around further grounding advancing modes of media synthesis and manipulation in broader and additional frameworks alongside mis/disinformation. This area includes understanding how creator cultures, including those creating narratives for human rights advocacy, adopt synthetic media technologies for imaginative and

campaigning purposes and how, in these use cases, harms and risks to vulnerable groups are addressed.

8.5 Early work in this synthesis looked at remix activism and was heavily influenced by pioneering frameworks of participatory activism from Jenkins (Jenkins 2006a). An area of future exploration is how participatory remix activism - particularly as it has evolved since the work under Research Thread 1/Section 3 - relates to deepfakes and synthetic media. In *Deepfakes responses* and in related research work supervised at WITNESS, including Ajder and Glick 2021, I have considered the use of deepfake remix and satire as a form that complicates both the commercial platform and regulatory governance of malicious deepfakes. However, a further conceptualization of deepfakes as inherently a remix because of the nature of training data and their creation process (as discussed in Meikle 2022 and in current discussions of generative AI training data sets) will benefit from additional research, conceptualisation and advocacy. Similarly, I will explore the question of remix using the more accessible tools of generative AI such as text-to-image and in-painting and out-painting, and the intersection of these generative AI tools with existing remix-oriented platforms such as TikTok (and the rich scholarship around these platforms including Abidin, Divon, Jaramillo et al. 2022, and the TikTok Cultures Research Network).

8.6 My work included an early practical exploration of the concept of ‘distant witnessing’. Distant witnessing, as explored in the context of live-streaming, also carries strong resonances in the immersive, augmented reality (AR) and virtual reality (VR) approaches (often termed XR in aggregate) that form part of the substance of ‘metaverse’ discussions. I have explored this in work outside the portfolio (for example, Gregory 2016a). Recurring issues for activists identified in this portfolio – including managing within a context of volume, understanding and mitigating evolving security concerns, and navigating trust and authenticity - take on additional dimensions in these new XR contexts. A key area of further research via case studies and participatory approaches would be understanding how issues of mediation, spectatorship and effective action are addressed in these environments.

Similarly, there is research to be done on how the recurring research threads of Sections 3,4 and 5 relate to these XR technologies - namely, the how (and who) of video activist practice and how it evolves in response to existing advocacy strategies and novel approaches, the attention to recurring ethical and normative issues, and the responsiveness of technical infrastructure and emerging technologies to these concerns.

8.7 An area of critical reflection throughout developing this synthesis has been upon my growth as a researcher. As an outcome of this progression, I have been able to articulate novel research directions and identify commonalities across time (e.g. around recurring ethical concerns), conceptualise potential terms for specific concepts (e.g. ‘distant witnessing’ in human rights contexts), and produce substantive research outputs that shape broader fields of discussion, public policy, industry investment and impact (e.g. work on authenticity infrastructure and deepfakes).

I identify some specific areas of previous limitations in my research approach and of future growth. An approach I have frequently used in both research and presentation is the use of speculations and provocations. This usage reflects both journal and conference requests for this type of approach and my positionality as a practitioner with a public voice in my field. An advantage of these approaches is that they provoke attention and secure public impact around the research presented, as evinced in the responses to work on concepts including distant witnessing. However, one limitation of my integration of this approach, particularly evidenced in my earlier papers, is in terms of presenting clear research methodologies. In the first publications in this Portfolio (notably *Participatory Panopticon*), my research methodology is more opaque than necessary - both regarding the underlying approach and in terms of how to do responsible speculative work. From 2013, I used an underlying approach to speculation and provocation based on assessing the field, conducting participant observation as a practitioner, and utilising versions of a futures methodology developed by the Institute for the Future. Moving forward, I would like to apply more rigorous futures and speculative thinking methodologies in the abovementioned areas.

In terms of my research practices and approaches, I continue to explore how I can deepen my inclusive, participant-grounded and centred analysis. This exploration also reflects an awareness of my subjectivity as an able-bodied queer cis-gendered white man with a national origin in the Global North, as well as my situated professional subjectivity as a professional staff member of a human rights non-governmental organisation. Steps forward include a further articulation of how design justice applies particularly to emerging technologies and infrastructure (Design Justice Network as well as Costanza-Chock 2020) as well as further iterating the work I already do grounded in the Diverse Voices methodology (Young et al, 2020) as well as justice-oriented frameworks (Benjamin 2019). It also includes drawing on the rich, engaged scholarship around AI harms (including Buolamwimi & Gebru 2018 and subsequent, Noble 2018, Bender et al. 2021) that provides valuable intersections to work on deepfakes, generative AI and authenticity infrastructure.

APPENDIX: Academic citations

Ubiquitous witnesses: who creates the evidence and the live (d) experience of human rights violations?
Information, Communication & Society 2015.
Citations: 65 Impact Factor: 4.22 (2022)

Kony 2012 Through a Prism of Video Advocacy Practices and Trends
Journal of Human Rights Practice 2012.
Citations: 53 Impact Factor: 0.852 (2022)

Cameras everywhere revisited: how digital technologies and social media aid and inhibit human rights documentation and advocacy
Journal of Human Rights Practice 2019.
Citations: 32 Impact Factor: 0.852 (2022)

Human Rights Made Visible: New Dimensions to Anonymity, Consent and Intentionality In 'Sensible Politics: The Visual Culture of Nongovernmental Activism' 2012.
Citations: 22

The Participatory Panopticon and Human Rights: WITNESS's Experience Supporting Video Advocacy and Future Possibilities
In 'Sensible Politics: The Visual Culture of Nongovernmental Activism' 2012.
Citations: 16

The Content Authenticity Initiative: Setting the Standard for Digital Content Attribution (link) (contributing author)
Adobe, 2020
Citations: 9

Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism

Journalism 2022

Citations: 12 Impact Factor: 2.9 (2023)

Technology and citizen witnessing: navigating the friction between dual desires for visibility and obscurity

The Fibreculture Journal 2015.

Citations: 9

Human rights in an age of distant witnesses: remixed lives, reincarnated images and live-streamed co-presence

Book Chapter, *Image Operations: Visual Media and Political Conflict*, 2016.

Citations: 6

Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening, June 2018: Summary of Discussions and Next Step Recommendations WITNESS, 2018.

Citations: 3

Video for Change: Creating and Measuring Ethical Impact

T Notley, S Gregory, A Lowenthal

Journal of Human Rights Practice 2017.

Citations: 2 Impact Factor: 0.852 (2022)

Live-streaming for frontline and distant witnessing: A case study exploring mediated human rights experience, immersive witnessing, action, and solidarity in the Mobil-Eyes Us Project

NECSUS European Journal of Media Studies, 2021.

Citations: 1

LIST OF REFERENCES

Abidin, C. https://wishcrys.com/academic-publications/ [Accessed January 31, 2024].

Ajder, H. and Glick, J. (2021). *Just Joking: Deepfakes, Satire and the Politics of Synthetic Media*. WITNESS/Co-Creation Studio at MIT Open Doc Lab. Available at: https://cocreationstudio.mit.edu/just-joking/ [Accessed January 31, 2024].

Allan, S. (2013). *Citizen Witnessing: Revisioning Journalism in Times of Crisis*. Oxford, England: Polity Press.

Anstett, E., and Dreyfus, J. (2015). Introduction: Why Exhume? Why Identify? In: Anstett, E. and Dreyfus, J. (eds.) *Human Remains and Identification: Mass Violence, Genocide, and the 'Forensic Turn'*. Manchester: Manchester University Press.

Bender, E.M., Gebru, T., McMillan-Major, A. and Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610-623.

Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Oxford, England: Polity Press.

Bob, C. (2005). *The Marketing of Rebellion: Insurgents, Media, and International Activism*. New York:Cambridge University Press.

Bontcheva, K., Posetti, J., Teyssou, D., Meyer, T., Gregory, S., Hanot, C. and Maynard, D. (2020). *Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression.* UNESCO Broadband Commission Report, UNESCO. Available at: https://en.unesco.org/publications/balanceact [Accessed January 31, 2024].

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77-91.

Burgess, J. and Green, J. (2009). *YouTube: Online Video and Participatory Culture.* Oxford, England: Polity Press.

Coalition for Content Provenance and Authenticity (C2PA) (N/d a). *Coalition for Content Provenance and Authenticity Guiding Principles*. C2PA. Available from: https://c2pa.org/principles/ [Accessed January 31, 2024].

Coalition for Content Provenance and Authenticity (C2PA) (N/d b). *Coalition for Content Provenance and Authenticity Specifications*. C2PA. Available from: https://c2pa.org/specifications/specifications/1.3/index.html [Accessed January 31, 2024].

Cascio, J. (2005). The Rise of the Participatory Panopticon. *Worldchanging*. Available at: http://www.openthefuture.com/wcarchive/2005/05/the_rise_of_the_participatory.html [Accessed January 31, 2024].

Cascio, J. (2013). https://twitter.com/cascio/status/364112024818556928 [Accessed January 31, 2024].

Chesney, R. and Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security". 107 *California Law Review* 1753. Berkeley, California: University of California Berkeley School of Law.

Chouliaraki, L. (2006). *The spectatorship of suffering*. Thousand Oaks, CA: SAGE Publications.

Chouliaraki, L. (2013). *The ironic spectator: Solidarity in the age of post-humanitarianism*. Oxford, England: Polity Press.

Chouliaraki, L. (2015). Digital witnessing in conflict zones: the politics of remediation. *Information, Communication & Society* 18(11), 1362–1377.

Clark, J. (2019) Written Testimony of Jack Clark Policy Director OpenAI at hearing on “The National Security Challenges of Artificial Intelligence, Manipulated Media, and ‘Deep Fakes’” before the House Permanent Select Committee on Intelligence June 13th, 2019. Available from: https://democrats-intelligence.house.gov/uploadedfiles/clark_deepfakes_sfr.pdf [Accessed January 31, 2024].

Cohen, S. (2001). *States of Denial : Knowing about Atrocities and Suffering*. Oxford, England: Polity.

Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need.* London, England: MIT Press.

Deutch, J. and Para, N. (2020). Targeted mass archiving of open source information: a case study. In: Dubberley S, Koenig, A. and Murray, D. (eds.), *Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability.* Oxford, England: Oxford University Press, 165-184.

Divon, T. https://tomdivon.com/PUBLICATIONS [Accessed January 31, 2024].

Dubberley, S., Koenig, A., and Murray, D. (eds.) (2020) *Digital Witness: Using Open Source Information for Human Rights Documentation, Advocacy and Accountability*. Oxford, England: Oxford University Press.

Federal Trade Commission (US). (2022) *Combatting Online Harms Through Innovation: FTC Report to Congress.*

Ganesh, M., Deutch, J. and Schulte, J. (2016) *Privacy, visibility, anonymity: dilemmas in tech use by marginalised communitie*s. Report, Tactical Technology Collective. Available from: https://xyz.informationactivism.org/en/dilemmas-tech-use-by-marginalised-communities/ [Accessed January 31, 2024].

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

Gray, J. (2019). Data witnessing: attending to injustice with data in Amnesty International's decoders project. *Information, Communication & Society* 22(7), 971–991.

Gregory, S., Caldwell, G., Avni, R., and Harding, T. (eds) (2005) *Video for Change: A Guide for Advocacy and Activism*. London: Pluto Press.

Gregory, S. (2006). Transnational Storytelling: Human Rights, WITNESS, and Video Advocacy. *American Anthropologist* Volume 108, Issue 1: March 2006, 195-204.

Gregory, S. (2010). Cameras Everywhere: Ubiquitous Video Documentation of Human Rights and Considerations of Safety, Security, Dignity and Consent. J*ournal of Human Rights Practice* 2(2), 191–207.

Gregory, S. and Zimmerman, P. (2011). Speculations on the Virtual and Viral Witness to Human Rights Crises. *Mediascape* Winter 2011. Available from: https://web.archive.org/web/20180427084755id_/http://www.tft.ucla.edu/mediascape/Winter2011_HumanRights.pdf [Accessed January 31, 2024].

Gregory, S. (2012). The Participatory Panopticon and Human Rights: WITNESS's Experience Supporting Video Advocacy. In: McLagan, M. and McKee, Y. (eds.) *Sensible Politics: Visual Cultures of Nongovernmental Politics*. Cambridge, MA: MIT Press.

Gregory, S. (2012). Human Rights Made Visible: New Dimensions to Anonymity, Consent and Intentionality. In M. McLagan and Y. McKee (eds), *Sensible Politics: The Visual Culture of Nongovernmental Activism*, 551–61. Cambridge, MA: MIT Press.

Gregory, S. (2012). Kony 2012 Through A Prism of Video Advocacy Practices and Trends. *Journal of Human Rights Practice* 4 (3), 1–6.

Gregory, S., and Losh. E. (2012). Remixing Human Rights: Rethinking Civic Expression, Representation and Personal Security in Online Video. *First Monday* 17(8). Available from: http://firstmonday.org/ojs/index.php/fm/article/view/4104/ 3279 [Accessed January 31, 2024].

Gregory, S. (2015). Technology and Citizen Witnessing: Navigating the Friction between Dual Desires for Visibility and Obscurity. *Fiberculture Journal* (26). Available from: https://twentysix.fibreculturejournal.org/fcjmesh-005-technology-and-citizen-witnessing-navigating-the-friction-between-dual-desires-for-visibility-and-obscurity/ [Accessed January 31, 2024].

Gregory, S. (2015) Ubiquitous Witnesses: Who Creates the Evidence and the Live(d) Experience of Human Rights Violations? *Information, Communication and Society* 18(11),1378–92.

Gregory, S. (2016a). Immersive Witnessing: From Empathy and Outrage to Action. *WITNESS Blog*. Available from: https://blog.witness.org/2016/08/immersive-witnessing-from-empathy-and-outrage-to-action [Accessed January 31, 2024].

Gregory, S. (2016b). Human rights in an age of distant witnesses: remixed lives, reincarnated images and live-streamed co-presence. In J. Eder and C. Klonk (eds). *Image Operations: Visual Media and Political Conflict*, Manchester: Manchester University Press.

Gregory, S. (2019). Cameras Everywhere Revisited: How Digital Technologies and Social Media Aid and Inhibit Human Rights Documentation and Advocacy. *Journal of Human Rights Practice* 11(2), 373–92.

Gregory, S. (2021). Live-streaming for frontline and distant witnessing: A case study exploring mediated human rights experience, immersive witnessing, action, and solidarity in the Mobil-Eyes Us project. *NECSUS_European Journal of Media Studies*. #Solidarity, Jg. 10 (2021-06-06), 1, 145-171.

Gregory, S. (2022). Deepfakes, Misinformation and Disinformation and Authenticity Infrastructure Responses: Impacts on Frontline Witnessing, Distant Witnessing, and Civic Journalism. *Journalism* 23(3), 708–29.

Gregory, S. (2023). Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI. *Journal of Human Rights Practice* 16(1): Advance Publication.

Gregory, S. (2023b). Written Testimony of Sam Gregory at U.S.Senate Commerce Committee, Subcommittee for Consumer Protection, Product Safety and Data Security (2023). *The Need for Transparency in AI*. September 12. Available from: https://www.commerce.senate.gov/services/files/DAD2163A-EF02-41B5-B7BA-2BA8B568C977 [Accessed January 31, 2024].

Gregory, S. (2023c). Written Testimony of Sam Gregory at U.S. House Oversight Subcommittee Hearing on Cybersecurity, Information Technology, and Government Innovation on *Advances in Deepfake Technology.* November 8. Available from: https://oversight.house.gov/wp-content/uploads/2023/11/Sam-Gregory-House-Oversight-Committee-Advances-in-Deepfake-Technology-November-2023.pdf [Accessed January 31, 2024]..

Hermida, A. (2015). Nothing but the truth. In: Carlson, M. and Lewis, S.C. (eds.), *Boundaries of Journalism Professionalism, Practices and Participation*. London, England: Routledge, 37–50.

Hinegardner, L. (2009). 'Action, Organization and Documentary Film: Beyond a Communications Model of Human Rights Videos.' *Visual Anthropology Review* 25(2), 172–85.

Howie, L. (2015). Witnessing terrorism. *Journal of Sociology*, Volume 51: 3.

Human Rights Center, University of California - Berkeley School of Law (2014). *FIRST RESPONDERS: An International workshop on Collecting and Analyzing Evidence of International Crimes.* Berkeley: UC Berkeley.

Ivens, G. and Gregory, S. (2019). *Ticks or It Didn't Happen: Key Dilemmas in Building Authenticity Infrastructure for Multimedia*. Report, WITNESS. Available from: https://lab.witness.org/ticks-or-it-didnt-happen/ [Accessed January 31, 2024].

Jaloud, A., Rahman, A., Al Khatib, H., Kayyali, D. and York, J. (2019). *Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content*. Report. Electronic Frontier Foundation, WITNESS and Syrian Archive. Available from: https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content [Accessed January 31, 2024].

Jaramillo-Dent, D., Alencar, A., and Asadchy, Y. (2022). #Migrantes on TikTok: Exploring Platformed Belongings, *International Journal of Communication* 16 (2022).

Jenkins, H. (2006a). *Convergence Culture: Where Old and New Media Collide*. New York, NY: New York University Press.

Jenkins, H. (2006b). YouTube and the Vaudevillian Aesthetic. *Confessions of an Aca-Fan,* *http://*www.henryjenkins.org/2006/11/youtube_and_the_vaudeville_aes.html [Accessed January 31, 2024].

Jenkins, H., (2009). The Revenge of the Origami Unicorn: Seven Principles of Transmedia Storytelling (Well, Two Actually. Five More on Friday). *Confessions of an Aca-Fan*. 12 December. Available from: http://henryjenkins.org/2009/12/the_revenge_of_the_origami_uni.html [Accessed January 31, 2024].

Kavada, A. and Treré, E. (2019). 'Live Democracy and Its Tensions: Making Sense of Livestreaming in the 15M and Occupy', *Information, Communication & Society*, 23, no. 12, July 2019, 1787-1804.

Kayyali, D. (2018). Delete Facebook? Not Just Yet. *WITNESS*. Available from: https://witness.org/delete-facebook- not-just-yet [Accessed January 31, 2024]

Keck, M. and Sikkink, K. (1998). *Activists beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

Land, M. (2009). Peer producing human rights. *Alberta Law Review*, Vol. 46, No. 4, 2009*; NYLS Legal Studies Research Paper* No. 09/10 #12 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1712021.

Land, M. (2016). 'Democratizing Human Rights Fact-Finding'. In: P. Alston and S. Knuckey (eds), *The Transformation of Human Rights Fact-Finding*. Oxford: Oxford University Press, 399–424.

Lim, M. (2017). Digital Media and Malaysia's Electoral Reform Movement. In: Berenschot, W., Schulte Nordholt, H., and Bakker, L. (eds.) *Citizenship and Democratization in Southeast Asia*. Leiden: Brill, 211–37.

Marwick, A. E., and boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133.

Mann, S., Nolan, J., and Wellman, B. (2002). Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance & Society*, Vol.1 No. 3.

Martini, M. (2018). Online Distant Witnessing and Live-Streaming Activism: Emerging Differences in the Activation of Networked Publics, *New Media & Society*, 20, no. 11, April 2018, 4035-4055.

Matheson, K. (2015). Basic practices: Capturing, storing & sharing video evidence. *WITNESS.* New York, NY: WITNESS. Available from:

http://library.witness.org/product/video-evidence-basic-practices-capturing-storing-sharing/ [Accessed January 31, 2024].

McPherson, E. (2015). Digital human rights reporting by civilian witnesses: Surmounting the Verification Barrier. In: Rebecca, A.L. (ed), *Produsing Theory in a Digital World 2.0The Intersection of Audiences and Production in Contemporary Theory* Volume 2. New York, NY: Peter Lang.

Meikle, G. (2022). *Deepfakes.* Oxford, England: Polity Press.

Mittel, J. (2009). To Spread or to Drill? 25 February. *Just TV*. Available from: http://justtv.wordpress.com/ 2009/02/25/to-spread-or-to-drill [Accessed January 31, 2024].

Mortensen, M. (2015). 'Connective Witnessing: Reconfiguring the Relationship between the Individual and the Collective', *Information, Communication & Society*, 18, no. 11, July 2015, 1393-1406.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: New York University Press.

Notley, T., Gregory, S., and Lowenthal, A. (2017). Video for Change: Creating and Measuring Ethical Impact. *Journal of Human Rights Practice,* 2017, 1–24.

Ong, J. (2015). ''Witnessing distant and proximal suffering within a zone of danger: Lay moralities of media audiences in the Philippines', *International Communication Gazette*, 2015: 1-16.

Ong, J. (2019). 'Toward an Ordinary Ethics of Mediated Humanitarianism: An Agenda for Ethnography', *International Journal of Cultural Studies*, 22, no. 4, February 2019, 481-498.

Orgad, S and Seu, I. (2014). 'The mediation of humanitarianism: Toward a research framework', *Communication, Culture & Critique*, 7, 2014, 6-36.

Parsons, A., Rosenthol, I., Scouten, E., et al. (2020). *The Content Authenticity Initiative: Setting the standard for digital content attribution.*

White paper, Content Authenticity Initiative, USA, August. Available at: https://contentauthenticity.org/how-it-works [Accessed January 31, 2024].

Peters, J. D., (2001). 'Witnessing.' *Media, Culture & Society* 23(6), 707–23.

Richardson, A. (2020). *Bearing witness while Black: African Americans, smartphones, and the new protest #journalism*. New York, NY: Oxford University Press.

Roberts, S.T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media.* New Haven: Yale University Press.

Rini, R. (2020). Deepfakes and the Epistemic Backstop *Philosophers' Imprint* 20(24), 1–16.

Ristovska, S. (2021). *Seeing Human Rights: Video Activism as a Proxy Profession*. Cambridge, MA: MIT Press.

Seu, I.B. (2011). "Shoot the Messenger": Dynamics of Positioning and Denial in Response to Human Rights Appeals. *Journal of Human Rights Practice* 3(2), 139–161.

Sherman, T. (n.d.). *Vernacular Video*. Available from: https://org.noemalab.eu/sections/ideas/ideas_articles/pdf/sherman_vernacular_video.pdf [Accessed January 31, 2024].

Treré, E. (2018). *Hybrid media activism: Ecologies, imaginaries, algorithms*. United Kingdom: Taylor & Francis.

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest.* New Haven, CT: Yale University Press.

U.S. House Oversight Committee Roundtable (2022). *On the Frontline: Responding to the Threat of Election Misinformation*. August 11. Available from: https://www.youtube.com/live/rBQFyzM2eok?si=w3dc0OCEYebImOAd [Accessed January 31, 2024].

U.S. House Oversight Committee, Subcommittee on Cybersecurity, Information Technology, and Government Innovation (2023). *Advances in Deepfake Technology,* November 8. Available from: https://oversight.house.gov/hearing/advances-in-deepfake-technology-2/ [Accessed January 31, 2024].

U.S.Senate Commerce Committee, Subcommittee for Consumer Protection, Product Safety and Data Security (2023). *The Need for Transparency in AI*. September 12. Available from: https://www.commerce.senate.gov/2023/9/the-need-for-transparency-in-artificial-intelligence [Accessed January 31, 2024].

Wang. T. (2013). *Big Data Needs Thick Data. Ethnography Matters*. Available from: http://ethnographymatters.net/2013/05/13/big-data-needs-thick-data [Accessed January 31, 2024].

Washington Post Editorial Board. (2019) Deepfakes are dangerous - and they target a huge weakness. *Washington Post,* June 6. Available from: https://www.washingtonpost.com/opinions/deepfakes-are-dangerous--and-they-target-a-huge-weakness/2019/06/16/d3bdbf08-8ed2-11e9-b08e-cfd89bd36d4e_story.html. [Accessed January 31, 2024].

Weizman, E. (2017). *Forensic Architecture: Violence at the Threshold of Detectability.* S.L.: Zone Books.

Xiao Mina, A. (2019). *Memes to Movements: How the World's Most Viral Media is Changing Social Protest and Power.* Boston, MA: Beacon Press.

Youmans, W. and York, J. (2012). Social Media and the Activist Toolkit: User Agreements, Corporate Interests, and the Information Infrastructure of Modern Social Movements. *Journal of Communication* 62(2), 315–329.

Young, M., Magassa, L. and Friedman, B. (2019). Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21(2), 89–103.

Zuckerman, E.. (2008). Public Spaces, Private Infrastructure - Open Video Conference. October 1, 2010. *My Heart's in Accra blog.* Available from: https://ethanzuckerman.com/2010/10/01/public-spaces-private-infrastructure-open-video-conference/ [Accessed January 31, 2024].