Check for updates

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–19
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517231219243
journals.sagepub.com/home/bds
Sage

Sharenting and social media properties: Exploring vicarious data harms and sociotechnical mitigations

Pamela Ugwudike[1], Silke Roth[1], Anita Lavorgna[2], Stuart E. Middleton[3], Natalie Djohari[1], Morena Tartari[4] and Arpan Mandal[3]

Abstract

In this paper, we demonstrate how social media technologies can co-produce data-related harms unless preventative measures are instituted. To this end, we draw on a passive ethnography of a public Facebook group in the UK practicing sharenting which occurs when parents and guardians post sensitive and identifying information about children in their care on social media. Theoretically, we draw on the ‘harm translation’ concept from digital criminology and the ‘seductions of crime’ perspective from cultural criminology. Further we analyse documents on the operations of Facebook's content filtering algorithms published by Meta (Facebook's parent company). With insights from these sources, we demonstrate how platform technologies go beyond facilitation to the inadvertent co-production of harm via embedded mediative properties that shape user perception and action. We show that, in the specific context of sharenting, the properties *invite* rather than simply facilitate the practice and can also *invite* subsequent misuses of child-centric data. Through our analysis of these dynamics, we set out an empirical basis for challenging reductive depictions of social media technologies as solely facilitative of human action including harmful conduct. We also outline our vision to integrate insights from the analysis into a new sociotechnical harm prevention framework informed by Natural Language Processing approaches.

Keywords

Sharenting, harm, Facebook, social media platforms, natural language processing

Introduction

Sharenting is the practice of parents and guardians posting child-centric data such as stories and videos of their own children on social media, often without consent (Brosch, 2018). Some have used the term ‘oversharenting’ to conceptualise ‘high frequency’ sharenting (e.g. Klucarova and Hasford, 2023). In this paper, we argue that regardless of frequency, all forms of sharenting are potentially harmful, and preventative measures are required to protect the affected children. Associated data harms include privacy violations (Brosch, 2018), identity crimes and fraud (Williams-Ceci et al., 2021), contaminated digital and online identities (Bezáková et al., 2021), as well as various cyberharms (Bezáková et al., 2021).

News media reports (e.g. Pierre, 2022) and academic studies (e.g. Potter and Barnes, 2021) exploring the risks and harms of the practice, often focus on the motives and agency of sharenters. Limited attention is paid to the perhaps inadvertent role of social media platforms in co-producing both the practice and adverse data-related outcomes. In this paper, our aim is to unravel the dynamics of this co-production and consider remedial strategies that can help protect children from the intentional and unintentional harms of sharenting. We therefore have two objectives. The first is to explore and understand how sharenters and the platforms co-produce the data harms associated with

[1] Department of Sociology, Social Policy and Criminology, University of Southampton, Southampton, UK
[2] Department of Political and Social Sciences, University of Bologna, Bologna, Italy
[3] School of Electronics and Computer Science, University of Southampton, Southampton, UK
[4] Department of Philosophy, Sociology, Pedagogy, and Applied Psychology, University of Padua, Padua, Italy

Corresponding author:
Pamela Ugwudike, Department of Sociology, Social Policy and Criminology, University of Southampton, Southampton, UK.
Email: P.Ugwudike@soton.ac.uk

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

sharenting. The second is to propose a sociotechnical Natural Language processing (NLP) framework for preventing such harms.

To fulfil our first objective, we draw on the data from our passive ethnography of open Facebook groups practicing sharenting in the UK. We manually extract posts from one of the groups for qualitative content analysis. To address our second objective which is to explore the co-productive role of social media platforms, we perform a qualitative content analysis of documents published by Meta (Facebook's parent company). We focus on documents detailing the operations of the platform's content filtering algorithms which influence both the distribution of, and access to, user generated content.

We analyse the documents, and the selected Facebook posts, with reference to the concept of 'harm translation' as developed by Wood and colleagues (2023) within digital criminology. We also draw on the 'seductions of crime'[1] perspective in cultural criminology by Katz (1988).

Both criminological perspectives are useful for exploring how technology properties embedded in social media platforms can co-produce with users, what one of us (Ugwudike) conceptualises as vicarious data harms (VDHs). These are harms that affect people whose sensitive or identifying data are disclosed voluntarily by others (e.g. sharenters) on social media platforms.

Through our analyses of how sharenters and the platforms with which they interact co-produce VDHs, we move away from the tendency of existing studies to focus on sharenters' motives and actions whilst ignoring the co-productive capacity of platform technologies. The studies seem to adopt an approach that complements Actor-Network Theory (ANT) (Latour, 2005), which similarly de-emphasises the role of technology properties in mediating forms of use and co-producing outcomes (see also Wood et al., 2023).

To address our final objective which focuses on remedial strategies, we propose a sociotechnical NLP framework for preventing sharenting harms. We demonstrate how social media group administrators can use the framework to moderate posts and issue relevant warnings, or deploy other preventative measures to protect children.

Sharenting and vicarious data harms

Insights from the extant literature on sharenting suggest that the practice exposes affected children to data harms (e.g. Lavorgna et al. 2022). Data in this context is defined as 'any information relating to an identified or identifiable natural person' (European Union, 2016). We conceptualise the previously mentioned harms associated with sharenting, as VDHs. The concept of VDHs refers to online harms affecting those whose sensitive or identifying data are disclosed voluntarily by others on social media platforms. For example, sharenters post child-centric data and inadvertently expose their children to privacy violations and other harms. Such harms are vicarious in the sense that the victims pay the price for the risky data practices of others engaging with social media platforms. The victims are also exposed to further victimisation by those who subsequently misappropriate the shared data.

In the context of sharenting, the concept of VDHs usefully denotes both (1) the voluntariness of data disclosure, and (2) the non-participation of affected children who ultimately bear the risks and harms. Applying the concept of VDHs to the study of sharenting specifically allows us to reflect on the powerlessness of the children exposed to risks or even victimised by the practice. They are denied informed consent rights, which undermines their autonomy. In jurisdictions such as the UK, parents exercise digital custodianship rights over children. Paradoxically, the UK is a signatory to the United Nations Convention on the Rights of the Child (UNCRC), which in Article 12 grants children the right to exercise their agency and autonomy in matters that affect their lives. Article 16 also provides that:

> Every child has the right to privacy. The law should protect the child's private, family and home life, including protecting children from unlawful attacks that harm their reputation.

Sharenting practice runs contrary to these provisions. Through the practice, sensitive and identifying child-centric data are shared on social media platforms without the consent of the affected children and with adverse implications for them. As we discuss in the next section, the focus of sharenting research has been on the actions and agency of sharenters. But the technical elements (the role of technology properties) also warrant attention.

Some conceptualise the technology properties as 'affordances', which, on social media platforms, include artefacts such as the 'like' or 'share' features that offer opportunities and possibilities for various forms of use (Bucher and Helmond 2017). That said, the meaning of affordances is contested and discipline specific. As such, a detailed elaboration of the concept is beyond the scope of this paper. Therefore, we use the term 'technology properties' which we explore at length in the paper, to conceptualise the often-ignored features of social media platforms which co-produce both sharenting and associated harms.

Focusing on users and ignoring the invitational role of technology properties

In digital environments, practices such as sharenting involve the use of technology properties embedded in social media platforms. On Facebook, examples include artefacts such as the 'comment' and 'share' features. They enable the sharing of photos and both audio-visual and textual child-centric material. As such, they *facilitate* usage. But, to prevent online harms such as those associated with sharenting, attention must also be paid to other technology properties which go beyond facilitation to *invite* harmful conduct.

Nevertheless, the multidisciplinary research on data harms associated with sharenting often depicts the social medial platforms where it is practiced as primarily *facilitative* and portrays the sharenters' actions as causative. More attention is typically paid to the motives and agency of the sharenters. Opportunities to forge social networks and access support and advice are commonly cited examples (e.g. Haslam et al. 2017). Others include lucrative brand endorsements featuring children (Abidin, 2017) and parents' pursuit of self-representation through their children (Holiday et al. 2022).

The studies do provide very useful insights on the data-related risks and harms of sharenting. But the emphasis tends to be on the reductive notion of technological facilitation, with causation often attributed to sharenters and to other users who subsequently misappropriate the shared data. As such, the literature on sharenting harms specifically, seems to reflect the tenets of ANT (Latour, 2005). The theory has been used to explain how technologies and users, as co-actants, operate within a digital network. But ANT appears to depict the role of technologies as merely facilitative. The premise seems to be that no actant dominates the other in any deterministic manner.

ANT usefully avoids the problem of determinism by recognising that both human and technological actants can co-produce harm. However, the theory conflates the capabilities and agency of both. Recognising this, Wood and Colleagues (2023) note that, 'Actor-network theory […] acknowledges that technologies can contribute to producing violent events but struggles to meaningfully distinguish between the different kinds of contributions technological and human agents can make'.

As Wood and colleagues (2023) demonstrate in their analysis of how technologies can invite harm, ANT does not address the capacity of technologies to play a more active mediative role in harm production. They argue that 'actor-network theory cannot account for the specific mediatory role technologies can play' in influencing human perception, conduct, and experience. ANT's depiction of human-technology interaction can discourage critical scrutiny of how a technology goes beyond facilitation to *invite* motivated users to actualise harmful conduct (Wood et al., 2023). In this scenario, technologies play a mediative role that is difficult to capture through the lens of ANT. Instead, the theory appears to elide the structural conditions of technology use by reductively minimising the power and agentive asymmetry inherent in interactions between powerful technologies such as social media platforms and their users.

Platform power: Going beyond facilitation to invitation

Social media platforms can facilitate high levels of interactivity, connectivity, and content creation. As such, they empower users such as sharenters to access many benefits, from building social networks (Brosch, 2018) to securing lucrative brand endorsement contracts (Abidin, 2017). But the platforms have also been described as holders of 'algorithmic power' (Bucher, 2012) capable of going beyond facilitating access to such benefits, to *inviting* particular conduct.

That aside, the structural conditions of technology use are such that, with the properties embedded in them by their designers, technologies can exercise greater agency than users. For example, with content filtering technologies, social media platforms can actively manipulate users' choices and actions by inviting or encouraging them to engage with amplified and recommended posts. This goes beyond facilitating usage, to inviting engagement with specific content. In the context of sharenting, two concepts that are useful for exploring the role of social media platforms in influencing users' actions through invitations and co-producing associated harms are 'harm translation' and 'seductions of harm'.

Harm translation

Wood and colleagues (2023) from the field of digital criminology use the concept of 'harm translation' to explain how technologies invite (rather than simply facilitate) harmful forms of use. Such solicitation occurs via embedded *inviting* properties. Harm translation itself occurs when technology users translate *inviting properties* into harmful action using 'affordances-in-use' (e.g. *facilitative artefacts* such as the 'like', 'comment', and 'share' buttons) which facilitate usage. According to Wood and colleagues (2023), 'a technology's affordances-in-use […] capture an actor's use of a technology to carry out certain (harmful) actions'. They allow users to deploy a technology 'as an instrument of action'.

The concept of 'harm translation' allows us to unpack the capacity of powerful platform technologies to literally circumscribe user agency. It is a concept that usefully differentiates between the inviting and facilitative artefacts of social media platforms. From the 'harm translation' perspective, platform technologies, for example, do not simply facilitate sharenting. Through invitations that *motivate or encourage* user action, they mediate the actions of sharenters as well as the other users who subsequently misuse the child-centric data.

The notion of translation originates from Verbeek's (2005) postphenomenological theory of technology and draws attention to the role of user perception and agency. When invited by technology properties to use a technology in specific ways, users interpret the invitations. They reflect on the various ways in which they may be deployed. Users who interpret invitations as routes to their desired harmful goals will engage in harm translation.

We recognise that not all users can or do appropriate technology invitations for harmful purposes. The impact of technology invitations depends on their alignment with the target actor's interests and capabilities. Further, technology invitations differ across different social media platforms (Tufekci, 2017). In addition, even when they are invitational, they can be constraining. Recommender systems, for example, invite users to interact with specific content. At the same time, they constrain interaction by restricting users' access to other content. In this way, social media platforms can circumscribe user agency. These dynamics reveal the unequal structural conditions of usage, particularly the power asymmetry between the platforms and users.

Applied to sharenting practice, the concept of 'harm translation' allows us to further reflect on how technology properties can operate as inviting properties and initialise the harm production chain. They can influence users' perceptions about possible forms of harmful use. In contrast, facilitative artefacts allow users to use features provided by the technology to realise harmful goals.

Seductions of harm

Beyond the concept of the harm translation concept developed within digital criminology, Katz's (1988) 'seductions of crime' perspective from cultural criminology is also useful for exploring how social media platforms can invite harmful conduct. Katz introduced the perspective to explain the factors that solicit and encourage deviance. The perspective views transgressive conduct including crime as a form of cultural expression and ascribes causality to factors such as affective benefits (e.g. excitement, thrills, frustrations, and anger) that make such conduct meaningful to perpetrators. It is a perspective that implicitly points to the capacity of technology properties to, not only facilitate, but also invite and precipitate VDHs. The properties do so through 'seductions' that influence users' perceptions about potential affective benefits. For Katz (1988), harmful conduct can be perceived as euphoric or emotionally beneficial by perpetrators and this explains why they welcome opportunities to engage in such conduct.

Recent studies have explored how seductions of crime operate in digital networks. For example, in their study of how the internet invites young people to engage in transgressive behaviour including hacking, Goldsmith and Wall (2022) drew partly on Katz's (1988) work. They used the notion of 'affordances of seduction' to demonstrate how the properties of 'technological environments' (such as the internet and social media platforms) motivate users to move from legal to illegal forms of use.

With reference to Cooper's (2000) work on the addictive quality of online harms, Goldsmith and Wall (2022) cited accessibility, affordability, and anonymity as examples of 'affordances of seduction' capable of inviting harmful conduct. They are invitational technology properties that are seductive to transgressors including hackers pursuing thrills and other affective benefits.

We broaden the concept of seductions of crime by using the term seductions of 'harm' instead, to accommodate forms of behaviour that are legal but harmful. Indeed, Katz (1988) who pioneered the seductions of crime perspective did incorporate in his analysis, transgressions which may be legal but risky and potentially harmful. Sharenting involving the disclosure of legally appropriate content is legal and is most likely often done in good faith. But studies describing its risks and harms suggest that the practice falls within the range of behaviours that Katz (1988) described as transgressive products of emotional seductions. Sharenting has been depicted as an 'uplifting', 'tantalising and compulsive pursuit' (Martindale, 2014), hinting at its affective dimensions. It is also portrayed as a means of combating social isolation (Brosch, 2018), forging useful networks, and accessing both information and support (Haslam et al., 2017). All these can enhance the addictiveness, thrills, and broader emotional attractions of the practice.

In sum, the seductions of harm perspective and the concept of harm translation are insights from criminology which expand our understanding of the often-overlooked role of social media technologies in producing the VDHs associated with sharenting. Our study of sharenting builds on the insights to offer a starting point for thinking generally about the ways in which the properties of technologies go beyond facilitation to play an active mediative role in harm causation by inviting (or in other words, encouraging and motivating) harm. We also propose an NLP remedial framework.

An NLP approach to harm prevention

Our proposal for an NLP remedial strategy is motivated principally by the ever-increasing volume of online discussion content on social media that requires moderation, and the corresponding increase in pressure on social media moderators to check this content and ensure violations are handled appropriately. There is a lot of related work on AI for content moderation, also known as algorithmic moderation (e.g. Gorwa et al., 2020), in related areas including harassment (Van Hee et al., 2018), cyberbullying (UNICEF n.d.), and hate speech (Davidson et al., 2017). But our approach is new in that it focuses on a sociotechnical NLP methodology which allows users to incrementally improve the NLP model focus over time. This human-in-the-loop approach is important because moderation is a subjective process requiring human judgement. It would be risky to delegate entirely moderation decisions to an AI algorithm as demonstrated by recent social media failures around AI content moderation (Hamilton 2022). As such, we propose using NLP models at the pre-moderation triage stage, which usually occurs at the start of a moderation session.

Moderation triage is where a batch of posts is sorted in terms of risk level (e.g. level of risk that sharenting violations are present) and urgency (e.g. child sexual abuse or other harms representing an immediate threat to life). Moderators can sort posts in triage risk order to tackle priority ones first and make decisions regarding the depth of analysis needed for individual posts at different risk levels. In our sociotechnical NLP methodology, the final moderation decisions are always made by the human moderators. But our NLP algorithms help moderators make these decisions as efficiently as possible.

As already noted, we do acknowledge that there are alternative strategies for using AI to support moderation. These include prevention-based methods focusing on user education. They operate via real-time content analysis to alert users to their risks prior to posting a message. Recommender systems that monitor user streams for regular risky behaviour and then recommend intervention material such as help websites can also be used. These other methods usually train AI without human-in-the-loop methods. But they can scale well and can as such complement our proposed sociotechnical NLP remedial strategy. Other offline measures such as awareness raising campaigns and strategies that can improve the digital literacy of social media users are also potentially useful.

In the remaining sections, we describe how we selected and analysed both the Facebook posts and Meta's documents. We also present our findings. Subsequently, we discuss how insights from our analysis can be used to develop the proposed NLP framework.

Methodology

We explored how sharenters and technology properties co-produce VDHs. To analyse the sharenters' role, we conducted a virtual passive ethnography involving non-participant observations of open/public Facebook communities of parents disclosing information about themselves, their children, and their families from January to April 2022. We then used qualitative content analysis method to analyse a selection of posts. To explore the role of technology properties, we also used qualitative content analysis method to analyse a selection of documents published by Meta.

Method 1: Passive ethnography and qualitative analysis of Facebook posts

Our passive ethnography focused on open Facebook groups. The platform has long been identified as a popular site for sharenting (Marsali et al., 2016). Further, to a greater extent than platforms such as Twitter and image-based platforms (e.g. Instagram), Facebook provides opportunities to form insular groups that are amenable to passive digital ethnography.

Purposive sampling. Since we had an exemplificatory aim, we decided to adopt a purposive sampling method to select relevant groups for passive ethnography. The purposive sampling method is a well-recognised sampling technique in qualitative research (e.g. Campbell et al. 2020). Following discussions with our project partners and advisory board members, we searched Facebook for open groups. For maximum variation, we focused on groups populated by communities of parents discussing diverse topics (e.g. parenting in general, health, hobbies, legal matters). We hypothesised that such groups would display various types, modes, and frequencies of sharenting. We then purposively sampled a selection of the groups and selected those who had the highest number of members and were more active in terms of posts and comments.

The passive ethnography. Our passive ethnography involved both observing and annotating users' conversations and behaviours over 4 months, using a detailed observation grid agreed by four researchers in the project team to avoid bias. The grid (Table 1) set out aspects of the digital field (the open Facebook groups) we sought to observe and usefully ensured that the ethnographic observation was systematic to reduce bias.

With the grid, we observed the full flux of posts during the period indicated (January to April 2022), initially focusing on the more recent posts until data saturation was reached and then going back to repeat the process. This lasted for a period of 4 months. Our ethnographic observation notes and memos revealed that most of the posts shared in the groups had the same characteristics, and in each observation period, we reached data saturation quickly.

This paper focuses on one of the groups we found which had, as its main topic, divorce-related legal advice. The search terms '(parent OR mother OR father) AND (legal OR divorce OR separation) AND (UK)' led us to several groups including the one on which we focus in this paper. The group is an English-speaking open/public Facebook community[2] of 1600 members. For the purposes of this contribution, we selected this specific group because of the sensitivity, publicness, and high frequency of posts which included disclosures of child custody and maintenance arrangements. To provide in-depth insights into the type of sharenting being practiced by the group members, we manually extracted 18 posts from the group for qualitative content analysis. The posts were specifically reflective of the type of sharenting taking place amongst group members as observed during the passive ethnography.

Table 1. The observation grid.

Code	Subcode (examples)
Platform	Facebook, Instagram, Twitter, TikTok
Date	Post date
Sharenter	Mother, Father, Teacher, Carer
Sharenter info	Demographics and child-centric disclosures
Sharenter motivation	Advice, networking, support, other
Children	Gender and number of children
Children's info	Other demographic information
Information shared and characteristics	Images (photos or videos) and/or text
Groups' topic and activities	Divorce/legal issues), health issues, leisure
Type of sharenting	Deliberate, incidental, unintentional
Comments	Group members or strangers
Sharenting in comments	Sharer (e.g. author, others) & information
Additional information on sharenting	Sensitive and/or identifying information
Members' comments about sharenting	Risks, harms, motives
Audience's comments about sharenting	Risks, harms, motives
Moderators' comments about sharenting	Risks, harms, motives
Offline interactions with group members	For advice, networking, support, other
Sharenting jargon used	Expressions denoting unique digital culture
Social context	E.g. parental separation, family problems
Frequency of posts	Number of posts by each sharenter
Community/group rules	E.g. disclosure of sensitive information
Sample posts	Extracts (Text posted by group members)
Miscellaneous	Other relevant information

Extracting the 18 posts allowed us to perform the in-depth analysis that underpins qualitative content analysis (see also Clark et al. 2021) and develop insights that can provide a starting point for broader debate and future research. Varis (2016) emphasises the informational and insightful utility of in-depth qualitative analysis involving a small number of social media posts, after a period of ethnographic observation. Further, other criminological studies have conducted qualitative analysis of social media data as we did. Schneider and Trottier (2012), for example, explored the impact of Facebook on the 2011 Vancouver riot. They analysed Facebook posts and demonstrated how the platform exposed the rioters to potential prosecution by publicising their images and posts.

Analysis of Facebook posts. We conducted in-depth, qualitative content analysis of the selected Facebook posts. As described by Clark and colleagues (2021), this method involves the systematic categorisation of a body of texts, usually with a coding frame (Table 2).

Neuendorf (2017) notes that the aim of qualitative context analysis is to unravel patterns in the data. We applied the approach by uploading the posts to Nvivo where we used codes to label segments of the data, and group the mutual codes (referring to the same phenomenon) into categories. The categories were in essence, overarching themes relevant to the topic (Fereday and Muir- Cochrane, 2006) of how the actions of sharenters co-produce VDHs.

Method 2: Analysis of documents – Meta's reports

Sampling. For our analysis of how technology properties co-produce the VDHs with the sharenters, seven publicly available documents published by Meta (Facebook's parent company) on the topic of algorithmic content filtering (Table 3) were reviewed.

The documents were selected through searches of the company's publications available on Meta's website (https://ai.facebook.com/) and Google search engine. We used the following search terms: 'Facebook personalisation', OR 'Facebook recommender systems', OR 'Facebook ranking system'**.** These produced relevant documents on Meta's site. Only articles focusing on content filtering since 2018 were selected to ensure that the documents were up to date given the rapidity of technological evolution and change in recent years.

The selected documents are best described as 'official documents from private sources' and are as such suitable for qualitative content analysis (Clark et al., 2021). Bowen (2009) argues that the documents are useful complementary sources of data. Reflecting this, the ones we selected provided vital information on how platform technologies *invite* and *facilitate* the VDHs of sharenting.

Other studies have similarly included reviews of organisational documents in their data collection strategy, with some undertaking the review as part of an ethnographic study (e.g. Angers and Machtmes (2005). This type of data triangulation which combines data from various sources (e.g. documents and digital ethnographic data),

Table 2. Coding frame for Facebook posts.

Code	Subcode (examples)
Platform	Facebook
Date	Post date
Sharenter	Mother, Father, Teacher, Carer, Other
Other Sharenter info	Demographics and child-centric disclosures
Children	Gender and number of children
Children's info	Other demographic information
Information shared and characteristics	Content of actual text and images
Groups' topic and activities	Divorce/legal issues
Type of sharenting	unintentional vs deliberate
Comments	Group members or strangers
Sharenting in comments	Sharer (e.g. author, others) & information
Community/group rules	E.g. disclosure of sensitive information

Table 3. The documents.

Documents	Data analysed
Meta (2018) Bringing People Closer Together. https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/	How Facebook connects users to posts from friends and family
Meta AI (2019a) DLRM: An advanced, open-source deep-learning recommendation model. https://ai.facebook.com/blog/dlrm-an-advanced-open-source-deep-learning-recommendation-model/	Features of a deep-learning recommendation model (DLRM)
Meta (2019b) Updates to Video Ranking https://about.fb.com/news/2019/05/updates-to-video-ranking/	How videos are disseminated to support content creators and help users find personalised videos.
Meta (2020) Personalized Advertising and Privacy Are Not at Odds. https://about.fb.com/news/2020/12/personalized-advertising-and-privacy-are-not-at-odds/	How user actions are used to personalise the content they see.
Meta (2021) How Does News Feed Predict What You Want to See? https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/	How Facebook's machine learning ranking system pushes content to users' New Feeds.
Meta (2022) New Ways to Customize Your Facebook Feed. https://about.fb.com/news/2022/10/new-ways-to-customize-your-facebook-feed/	How users can customise and personalise their News Feed to show specific types of posts.
Meta (2023) How AI Influences What You See on Facebook and Instagram. https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/	How Facebook uses AI systems to determine the content users discover.

facilitates 'a deeper and fuller understanding' of the topic (Bowen, 2009), in our case, the ways in which social media technology and users (sharenters) can co-produce VDHs.

Analysis of documents. The documents published by Meta were uploaded to Nvivo, and again, a coding framework (Table 4) was used to perform manual qualitative content analysis and ensure that coding and categorisation were systematic.

This time, the focus was on categorising the patterns in the data that explain how technology properties configure the contexts of sharenting and co-produce VDHs.

Methodological limitations

The purposive sampling method we adopted for the digital ethnography and our focus on a limited number of Facebook posts limit the study's generalisability. Regarding our review of Meta's documents, we acknowledge the lack of 'representativeness' as a potential methodological limitation (see Clark et al., 2021) given our purposive sampling method. Yet, our research is credible since we adopted a systematic approach to data collection and analysis. But in qualitative research, credibility also refers to the question of whether the selected data are accurate and free of distortion (Scott 1990). We recognise that Meta's publications, including the documents we selected,

Table 4. Coding frame for analysing the documents.

Code	Subcode (content filtering technologies)
Visible properties	• Visible technological artefacts (e.g. 'like', 'search', and 'comment' features) which facilitate usage.
Invisible (hidden) properties	• Recommender systems for personalisation. • Recommendation metrics (e.g. usage signals). • Amplification systems for higher visibility. • Amplification metrics (e.g. popularity signals).

do not fully disclose the architecture and operations of technology properties such Facebook algorithms. The models are proprietary and often unamenable to external scrutiny.

Lack of transparency and explainability are well-documented ethical issues affecting such technologies (Ugwudike 2022). Knowledge of the precise workings or operations of the models are as such limited to the unverifiable information provided by technology companies themselves. Further, organisational documents are not necessarily 'windows into social and organizational realities. [The documents represent] a distinct level of "reality" in their own right [or] particular versions of reality' (Clark et al., 2021). They reveal the contexts or background of their production. For example, Meta's depiction of content filtering as beneficial to users conjures up the image of an organisation committed to transparency and positive user experience. But studies exploring the profit model of social media companies generally point to commercial motives, in this case, promoting the high levels of user engagement required for lucrative data production (Kopf 2020). In sum, given the methodological limitations, what we offer here is a theoretically and empirically informed starting point for understanding how technologies can co-produce data harms via embedded inviting and facilitative properties.

Ethical considerations

Our passive digital ethnography poses ethical implications, primarily lack of informed consent, as well as privacy and confidentiality considerations (Cera, 2023). As such, we took steps to comply with best practice for digital research and data security by following the ethical guidelines and codes developed the Association of Internet Researchers (AOIR 2019), now described as 'the industry standard' for researchers deploying the digital ethnographic method (Cera, 2023).

In line with their provisions, we anonymised all social media data at the point of extraction. For this, we deleted all personal identifiers (all names and personal details). Further, we did not profile users according to sensitive attributes or personal data of any kind and we did not extract any images. We entered all the data in Excel files and stored them securely in the University's secure Research Filestore, with a dedicated project SharePoint area. This can only be accessed by the project team and is a university drive protected by BitLocker Device Encryption.

Further, this paper does not quote any of the extracted social media data since posts can be traced back to users via internet search engines and other means, defeating our anonymisation aims.

These anonymisation efforts fully comply with the ethical provisions of the previously mentioned organisations. Our methodology and overall research design were also approved by the Research Ethics Committee of the University of Southampton ethics committee (Number: 66341). Finally, the anonymised social media data have been stored in the *UK Data Service's* secure repository to enable replication and support future research. The Service's secure infrastructure will host the data securely and provide controlled access via a strict request/approval process.

Findings

Codes and categories from the analysis of Meta's documents

One of the categories that emerged from our qualitative content analysis of Meta's documents was 'invisible properties', and the embedded codes[3] were models (39)[4], system/s (37), ranking (32), predictions (28), personalise/personalisation (24), and recommendations (16). These all relate to technology properties that push personalised content to users' news feeds and invite user engagement. Later, we demonstrate how they co-produce VDHs. Another category identified from our analysis of Meta's documents was 'visible properties' and the embedded codes were share (22), follow (12), and comment (9); both the category and codes refer to the visible technological artefacts that facilitate usage.

Codes and categories from the analysis of Facebook posts

A category that emerged from the analysis of Facebook posts was 'sensitive posts'. The embedded codes[5] were Court (denoting court hearings including *sub judice* cases) (39)[6], 'contact' (mostly acrimonious child contact arrangements) (25), children's characteristics (gender and age) (8), 'live' (7), and school (7). Most references to 'contact', 'live', and 'school' provided information about the location of affected children. No images of the children were shared directly in the group page. But in 12 cases, the profile pages of the group members did provide open and unrestricted access to sensitive and identifying child-centric such as images of children and locational information.

A second category from the analysis of Facebook posts was 'emotive posts'. These refer to the pejorative terms with which members of the group described their estranged partners' actions. Embedded codes were 'abuse' (includes financial, domestic, drug and alcohol, and emotional abuse) (12), 'manipulation' (6), 'narcissist/narcissism/narc' (6), and 'molestation' (particularly with reference to non-molestations orders filed against estranged partners) (6). The emotive posts violated the group's rules which instructed users to refrain from posting emotive content such as slanderous and other disparaging remarks about ex-partners. The posts, including claims of abuse, also disclosed children's traumatic experiences, and as we demonstrate later, exposed them to VDHs. Later, we also show how our proposed NLP framework can support content moderation in such cases.

Meanwhile, we developed the categories that emerged from both our analysis of Facebook posts ('sensitive posts' and 'emotive posts') and Meta's documents ('invisible properties' and 'visible properties') by exploring how they intersect (along with their dimensions/embedded codes). This produced three themes which illustrate the co-production of VDHs, and they are *Initiating VDHs co-production, co-producing primary VDHs, and accelerating secondary VDHs co-production.* Below, we demonstrate how intersections between the categories from our analysis of Facebook posts and Meta's documents produced these themes.

Initiating VDH co-production. 'Invisible properties', which as mentioned earlier is a category that emerged from our analysis of Meta's documents, initiate the production of child-centric data such as those present in the 'sensitive posts' and 'emotive posts' created by sharenters. In doing so, the invisible properties expose affected children to VDHs. The category, 'Invisible properties' and its embedded codes (models, system/s, ranking, predictions, personalise/personalisation, and recommendation), all refer to hidden AI models in the form of recommender systems that personalise content. They provide a personalised gateway to suggested groups, posts, and even search results with which sharenters are most likely to engage based on their predicted preferences. Engaging with the content suggested by the invisible properties lead to the production of 'sensitive posts' and 'emotive posts' which go on to produce VDHs.

Invisible properties: Recommendation systems. Facebook has long used invisible properties such as recommender systems to analyse users' browsing histories and predict their preferences in order to personalise content and motivate user engagement with suggested posts. In one of their publications, Meta notes that, 'when you take an action on Facebook, such as following a Page or liking a post, we use that information to personalize your experience' (Meta, 2020). Content personalisation is as such based on predicted user preferences. According to Meta (2023), Facebook uses data-driven AI models for these predictive analytics: 'our AI systems predict how valuable a piece of content might be to you, so we can show it to you sooner'.

Meta appears to have become somewhat transparent about its recommenders. In 2019, they released an open-source 'Deep-Learning Recommender Model' (DLRM) that combines collaborative filtering techniques with approaches based on predictive analytics (Meta, 2019a). The latter operate by 'predicting what you're most likely to be interested in or engage with. These predictions are based on a variety of factors, including what and whom you've followed, liked, or engaged with recently' (Meta, 2021).

Invisible properties: Amplification systems. Amplification models, which are also invisible properties, operate as ranking algorithms that heighten the visibility of popular content to stimulate user interest and engagement (Meta AI 2019b). In several publications, Meta has published updates on how the platform ranks posts to determine eligibility for amplification and wider reach. In 2018, the company remarked that, 'today we use signals like how many people react to, comment on or share posts to determine how high they appear in News Feed… We will also prioritize posts from friends and family over public content' (Meta, 2018). In a 2019 update, the company focused on the amplification of content posted by video creators and media companies (Meta, 2019b). The company introduced ranking algorithms that 'further prioritize original videos that people seek out'. In other words, the aim is to amplify the most popular videos.

Amplification and recommendation systems as initiators of VDHs. How do invisible technology properties in the form of amplification and recommender systems initiate the production of child-centric data such as the sensitive and emotive posts created by the sharenters? As the foregoing suggests, our review of the documents published by Meta revealed that the systems push personalised content to users' news feeds via artefacts that users see in their news feed. We conceptualised the artefacts and the underpinning invisible properties (recommender and amplification models) as *initiators* of VDHs, and examples are provided below in Figure 1. They alert users to the personalised content and initialise user engagement.

In the case of sharenters, the artefacts include 'suggested groups' which the recommender algorithms consider relevant to their interests. Other examples include 'search results' which direct them to groups and other content that are also personalised on the basis of their predicted preferences. From a harm translation perspective, the artefacts (see Figure 1) and underpinning algorithms are *inviting* technology properties. They form part of what Wood and colleagues (2023) describe as 'persuasive technologies designed to 'hook users'. They are initiators of VDHs in that they invite forms of use such as sharenting which can produce harm. Viewed from the 'seductions of harm' perspective, they constitute 'seductions' that hold the promise of affective benefits such as the emotional gains of engaging with a supportive social network. As such, they are capable of influencing users' perceptions and incentivising transgressive behaviour (in this case, sharenting leading to VDHs).

The 'facilitators'. The initiators such as those described above and reproduced in the dotted square of Figure 2 below are insufficient for the 'harm translation' that produces VDHs. Once the initiators (powered by invisible properties such as amplification and recommender algorithms) instil in users the perception that the platform can be deployed for sharenting, motivated users must adopt visible properties to actualise the practice. 'Visible properties' is another category that emerged from our analysis of Meta's documents and the embedded codes were: share, follow, and comment. As noted earlier, they refer to visible artefacts and we described them as facilitators.

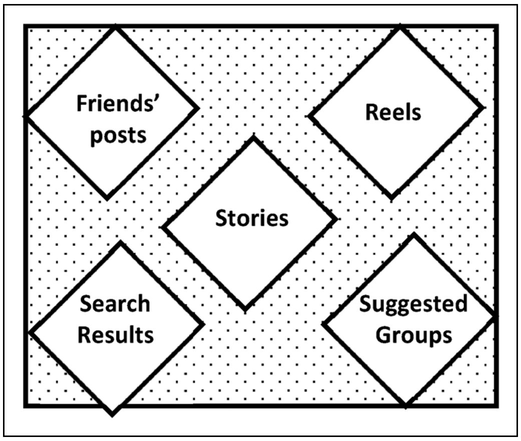

Figure 1. Diagrammatic representation of initiators.

Facilitators (see the outer lined square above) enable content creation and other forms of human agency and action.

Co-producing primary VDHs

Together with the initiators and facilitators, the sharenters in our study inadvertently (and with no observable malicious intent) co-produced what we conceptualised as primary VDHs. These are harms that are directly linked to the interactions between the technology properties (visible and invisible properties) and the sharenters' actions (posting sensitive and emotive content). Our analysis of both Meta's documents and the sharenters' Facebook posts revealed that through the process of harm translation, the sharenters responded to the invitations from the initiators (innermost dotted square) by using facilitators (lined square) to disclose child-centric data in the form of the sensitive and emotive posts mentioned earlier. Together, the posts exposed affected children to primary VDHs (outermost dark grey square of Figure 3 below). The sensitive and emotive posts shared by the parents, violated the privacy of affected children. Apart from that, the revelation of traumatic divorce/separation-related incidents in the lives of the children could cause them psychological distress. Such posts can also contaminate their digital identities and cause embarrassment.

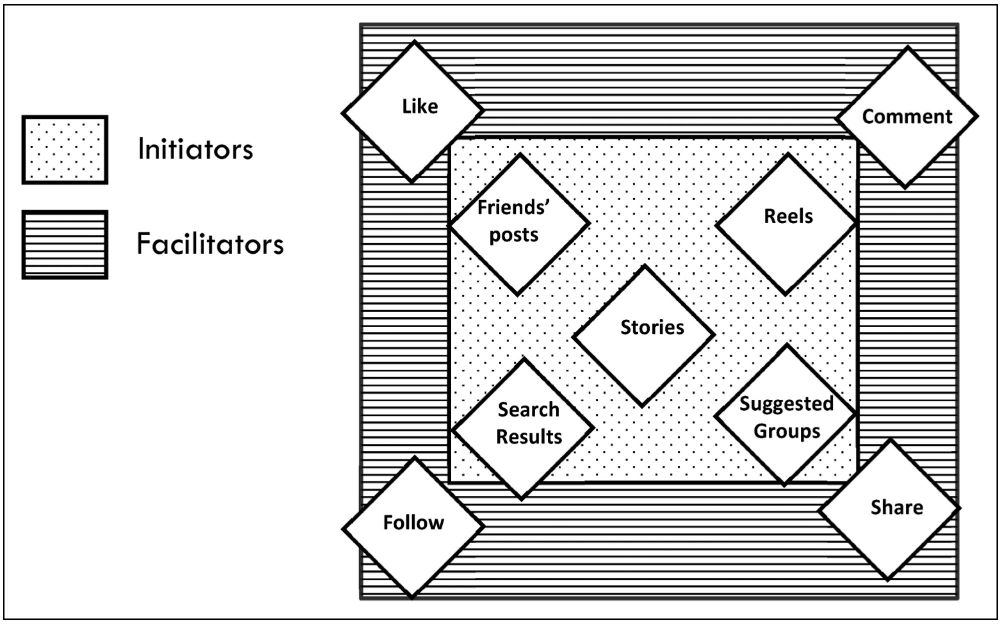

Figure 2. The initiators and facilitators.

Accelerating VDH co-production

In the next phase of the VDH production process, the previously mentioned recommender and amplification systems again operate as invisible inviting properties. They manifest themselves through various artefacts which appear in users' news feeds. At the beginning of the trajectory, the artefacts and underpinning invisible properties operate as 'initiators' and invite the production of sensitive and emotive posts resulting in primary VDHs. In this current phase, they become 'accelerators' of secondary VDHs. Unlike primary VDHs, secondary VDHs are the harms directly linked to the interactions between technology properties and other users who, with malicious intent, subsequently access and misappropriate the child-centric data disclosed by sharenters.

Meta's publications indicate that optimised for data personalisation, recommender and amplification algorithms will push the child-centric content shared by the parents we observed to the news feed of targeted users. These will be the users whose browsing histories such as the content they like or share, denote an interest in child-centric data (see the outermost black square in Figure 4 below). As noted earlier, one of the documents we reviewed makes it clear that, 'When you take an action on Facebook, such as following a Page or liking a post, we use that information to personalize your experience' (2020).

Secondary VDHs occur when other users perceive or interpret child-centric data pushed to their news feed by recommender and amplification systems, as invitations to engage with the data and harm children. As the harm translation perspective suggests, in response, the perpetrators

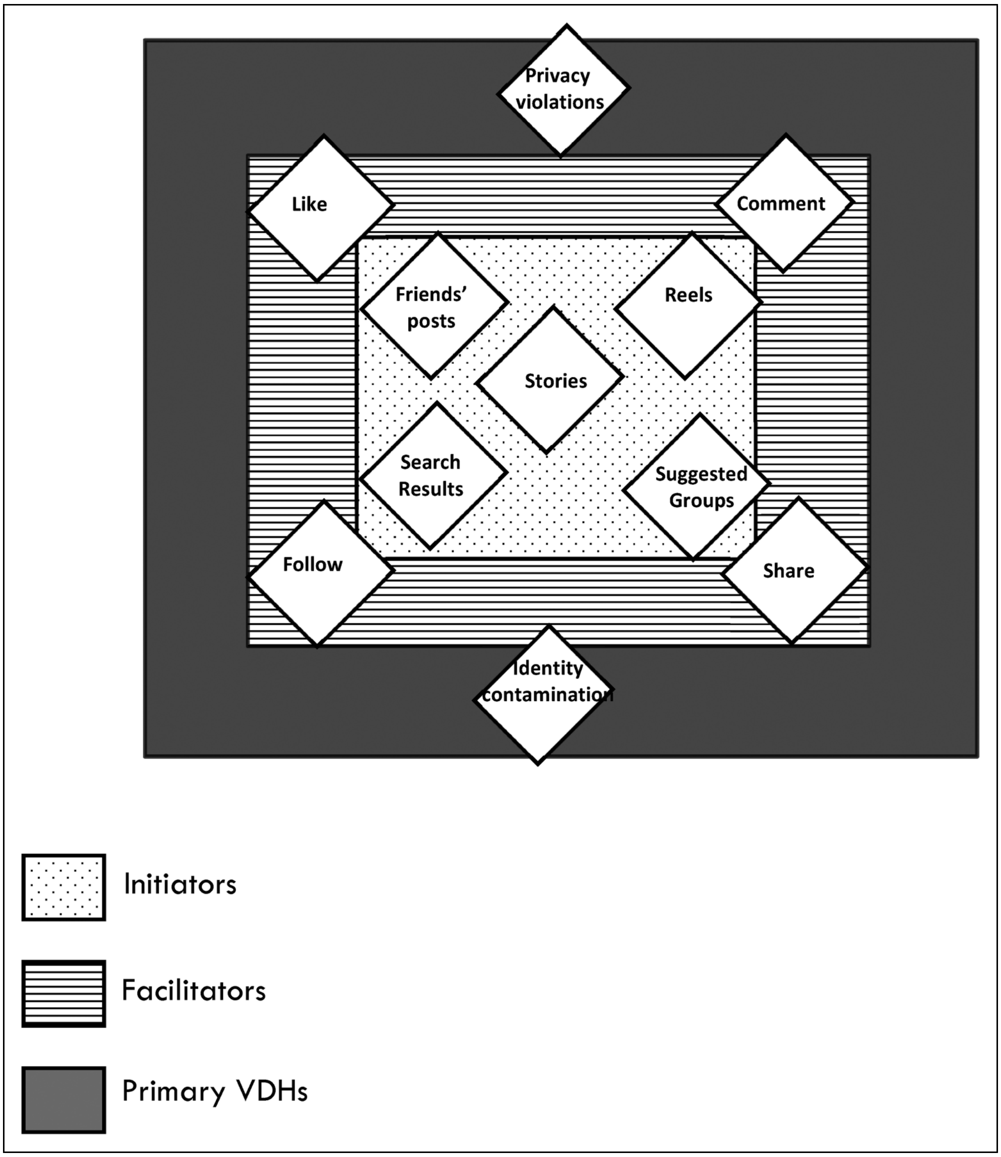

Figure 3. Primary VDHs co-produced by technology properties and sharenters.

adopt facilitative artefacts (such as those in the lined square of Figure 4 above) which enable them to misuse the data.

Therefore, in the absence of harm prevention measures such as the NLP framework that we describe later, the initiators (dotted square) and facilitative artefacts (lined square) will again provide the required pathway to harm. Figure 5 below summarises the process of co-producing VDHs.

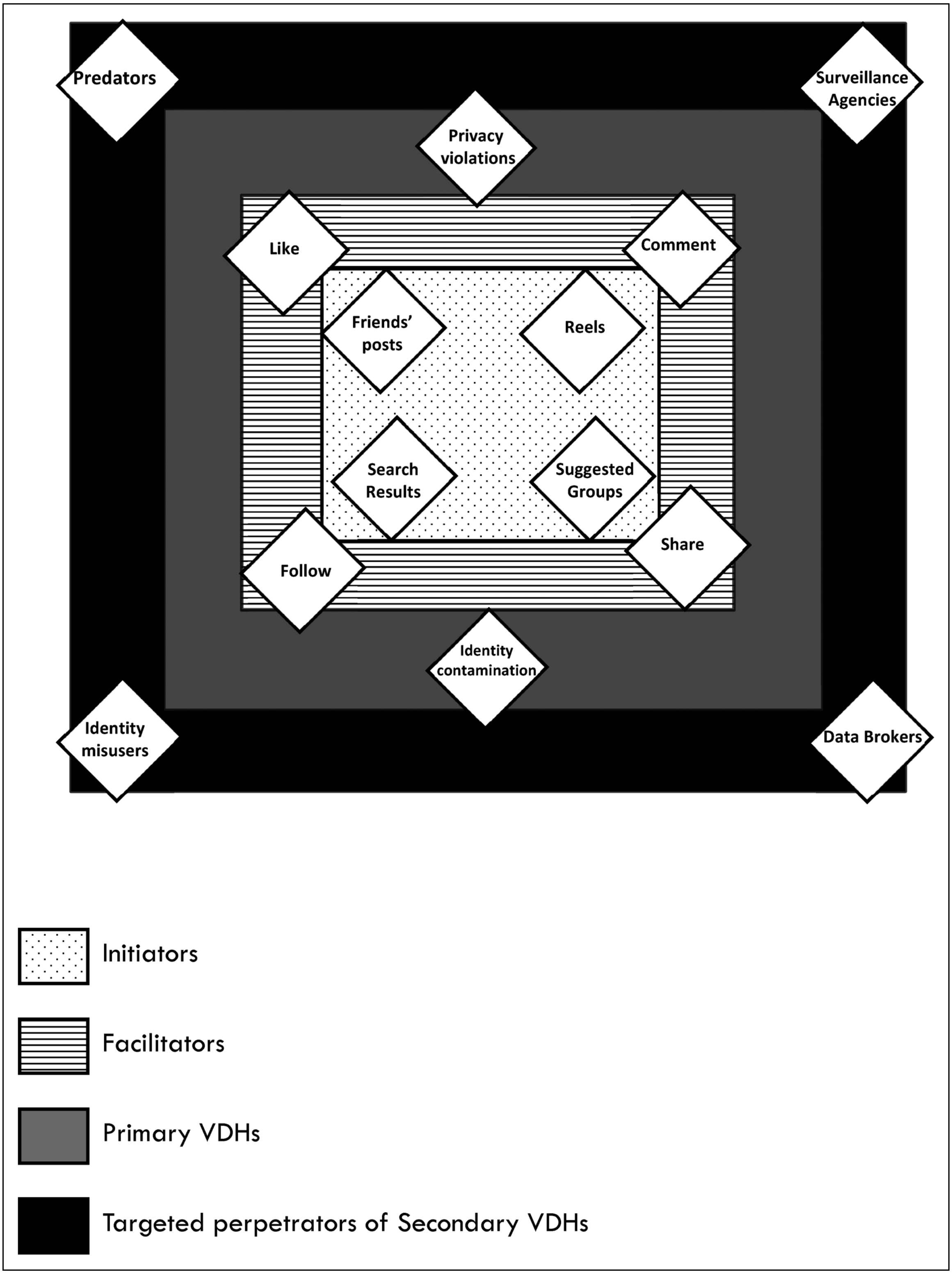

Figure 4. Targeting perpetrators of secondary VDHs.

Figure 5 shows that although the sharenters we observed did not provide information about their children's routines and activities, the child-centric data that they did share via sensitive posts and emotive posts, for example, offered insights into their children's lives. Such content can expose the children to secondary VDHs such as those in the outermost grey square of Figure 5 above. Predators at whom the content is targeted via inviting properties

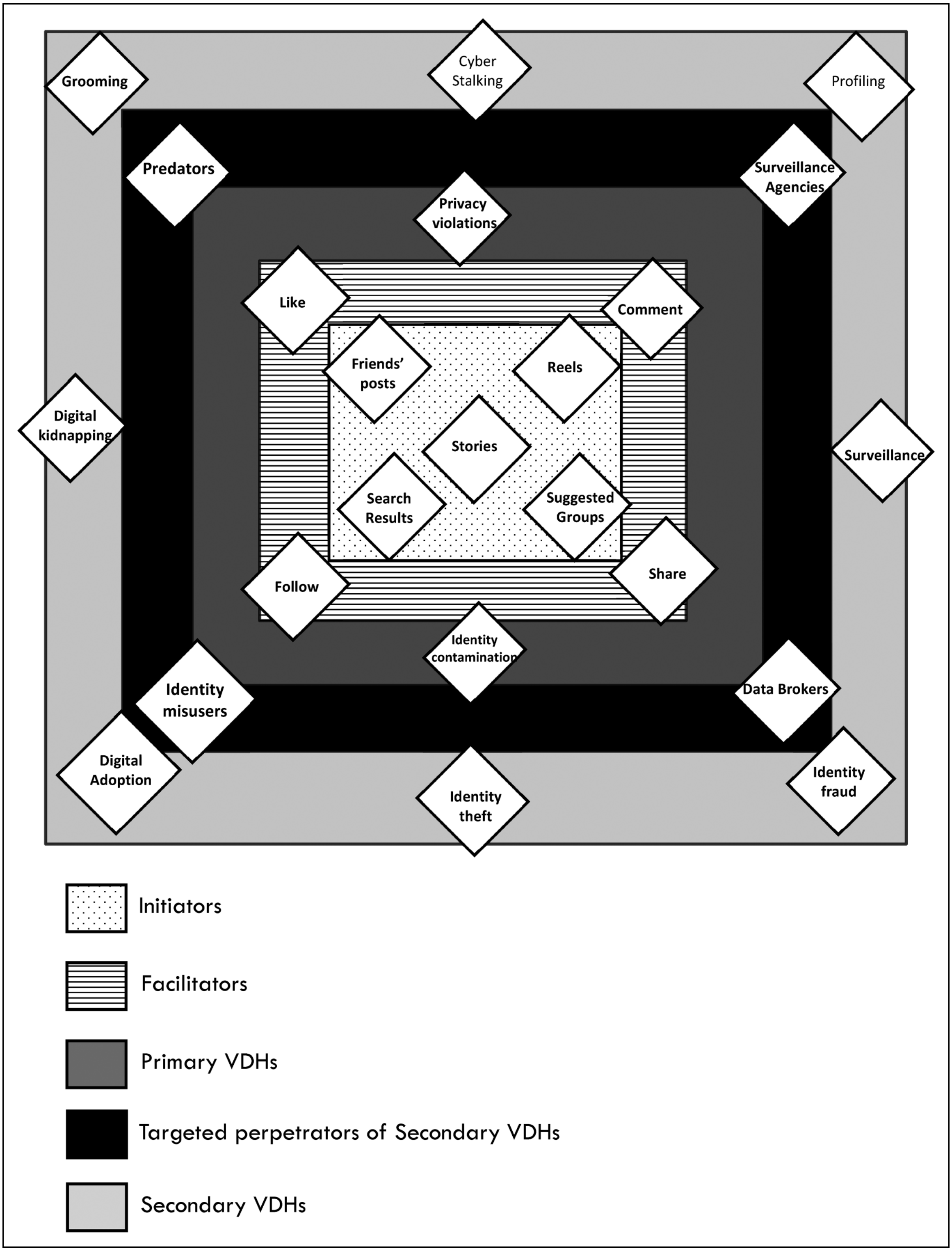

Figure 5. The co-production of VDHs.

(see inner dotted square of Figure 5 above) could consider whether accepting the invitation provides opportunities to prey on affected children due to their apparent vulnerability.

The predators may then engage in harm translation by deploying facilitative artefacts such as those in the innermost lined square which enable such harmful conduct. The children can also be exposed to profiling by companies using child-centric data for targeted advertisements. Child protection and social welfare agencies, known to particularly target and surveil children and families from deprived groups who are also often labelled as 'at risk' (Edwards and Ugwudike 2023), may access such data for profiling and related activities. Additionally, the data can be open to misappropriation (e.g. usage without appropriate consent) by law enforcement agencies relying on digital data for various intelligence, profiling, and surveillance activities.

We do acknowledge that VDH co-production does not necessarily display the linearity depicted here. Further, algorithmic models underpinning initiators and facilitators are proprietary, unamenable to external scrutiny, and their operations can become too complex for even the developers themselves to unravel. But we do provide insights based on relevant theories, a digital ethnography, and an analysis of documents published by a major social media company. In doing so, we offer a theoretically and empirically informed starting point. Others can build on our study to develop further understanding of how technology properties and users co-produce data harms.

An NLP remedial strategy

Here, we outline our vision of a new sociotechnical harm prevention framework-informed NLP approaches.

Our sociotechnical NLP methodology is shown as an information flow diagram in Figure 6. This methodology follows a human-in-the-loop AI (Middleton et al., 2022) cyclic workflow. It focuses on the use of few-shot NLP models (Schick and Schütze, 2021) to assist human sensemaking (Middleton et al., 2020) of intelligence extracted from large volumes of forum posts. The posts are gathered for the purpose of identifying content which exhibit group policy violations such as the sensitive and emotive posts shared in the Facebook group we studied.

Each cycle represents a moderation or criminological analysis session which we expect will last about one hour depending on the number of posts to be reviewed each cycle. The NLP violation classifier model will be applied to a corpus of forum posts, allowing risky posts such as sensitive and emotive child-centric data to be flagged and inspected by human experts. Posts deemed in violation of group policy can then trigger appropriate interventions such as activating the forum's reporting features or sending educational information to the authors. Initially, the NLP violation classifier model is trained using a small few-shot dataset of manually annotated example posts such as those extracted for the qualitative content analysis presented in this paper.

At the end of each session, some of the more difficult and ambiguous example posts are annotated, appended to the initial few-shot training examples, and the NLP models are re-trained overnight, ready for the next session. This is an example of a human-in-the-loop AI deep active learning (Ren et al., 2021) method to incrementally re-train and improve NLP models over time.

The few-shot NLP model itself will follow a state-of-the-art Pattern-Extraction Training (PET) regime which has been proven highly effective for various text classification tasks (Schick and Schütze, 2021). This model training approach is shown in Figure 7. It will use a Pre-trained Language Model (PLM) such as DeBERTa (He et al.,

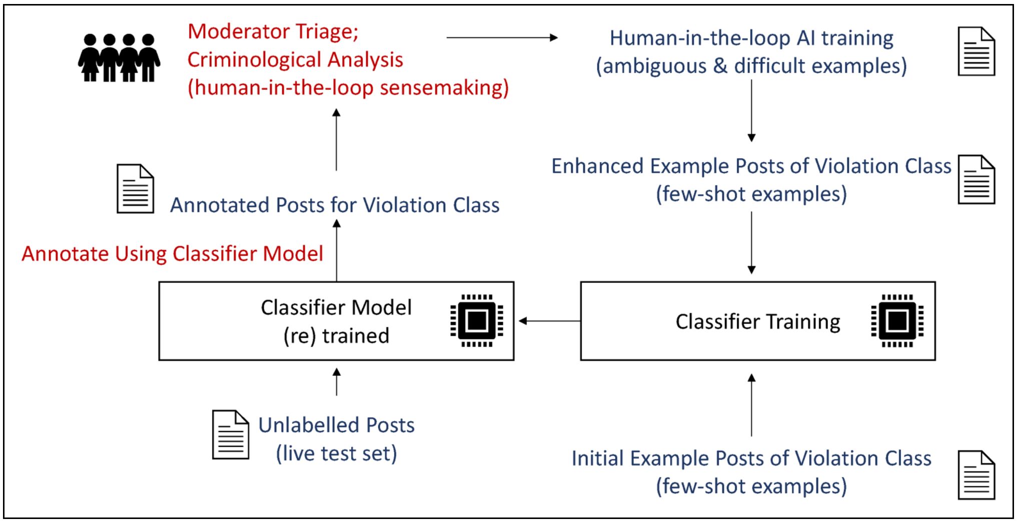
Moderator Triage;
Criminological Analysis
(human-in-the-loop sensemaking)
Human-in-the-loop AI training
(ambiguous & difficult examples)
Enhanced Example Posts of Violation Class
(few-shot examples)
Annotated Posts for Violation Class
Annotate Using Classifier Model
Classifier Model
(re) trained
Classifier Training
Unlabelled Posts
(live test set)
Initial Example Posts of Violation Class
(few-shot examples)

Figure 6. Sociotechnical NLP harm prevention framework.

2023), two to five language model-specific patterns, and 18 few-shot example posts. The training creates a PET model for each pattern, which are then used as an ensemble to noisily annotate the full unlabelled corpus usually consisting of 1000s of forum posts. This noisily annotated corpus is then used to train the final NLP classifier model.

We will initially focus on classifying coarse-grained violation classes, such as sensitive child-centric posts such as those that reveal location information. However, our few-shot NLP models only need small numbers of example posts to train. As such, it should be feasible for training many models that use more finer-grained classes, such as posts discussing children's location on a particular case or exposure to emotional abuse by a particular person type such as an ex-partner. In the coming phases of the ProTechThem project, we plan to further explore and test how this, in addition to the more generic coarse-grained violation classes, can support our sociotechnical NLP harm prevention framework.

Discussion

Our findings regarding the co-production of VDHs associated with sharenting highlight the important role of technology properties. This has been largely overlooked by the extant literature on sharenting risks and harms (e.g. Bezáková et al., 2021; Holiday et al., 2022). As such, we provide new insights and advance current knowledge in the field. Our findings also expand understanding of how technology properties play mediative roles in the context of human-computer interaction. We show that on social media platforms, for example, invisible technology properties channel to each user, the content predicted to stimulate their interest and engagement based on their usage history (Meta, 2019b). As such, they mediate human perception and action by inviting users such as the sharenters we studied, to engage with such content. In our study, we conceptualised them as initiators of primary VDHs by sharenters and accelerators of secondary VDHs perpetrated by others.

Our findings expand current knowledge about invisible or hidden technology properties which invite various forms of use. Although they are hidden, they manifest themselves through technological operations that provide evidence of their existence. On the Facebook platform, for example, our study shows that they manifest themselves through the presence of specific artefacts such as 'suggested groups' and 'search results' which are visible in a user's news feed. The invisible properties are also quite persuasive and invitational in that they encourage or invite users to discover and engage with specific (personalised) content. Without adequate preventative measures, such inviting properties can encourage and motivate harmful conduct unintended by the designers.

Our findings also show that although technology properties such as recommenders and amplifiers invite VDHs by operating as 'initiators' and 'accelerators', facilitative artefacts are required to finalise the process of VDH production. As such, our findings distinguish between the mediative and facilitative role of technologies in harm production. We

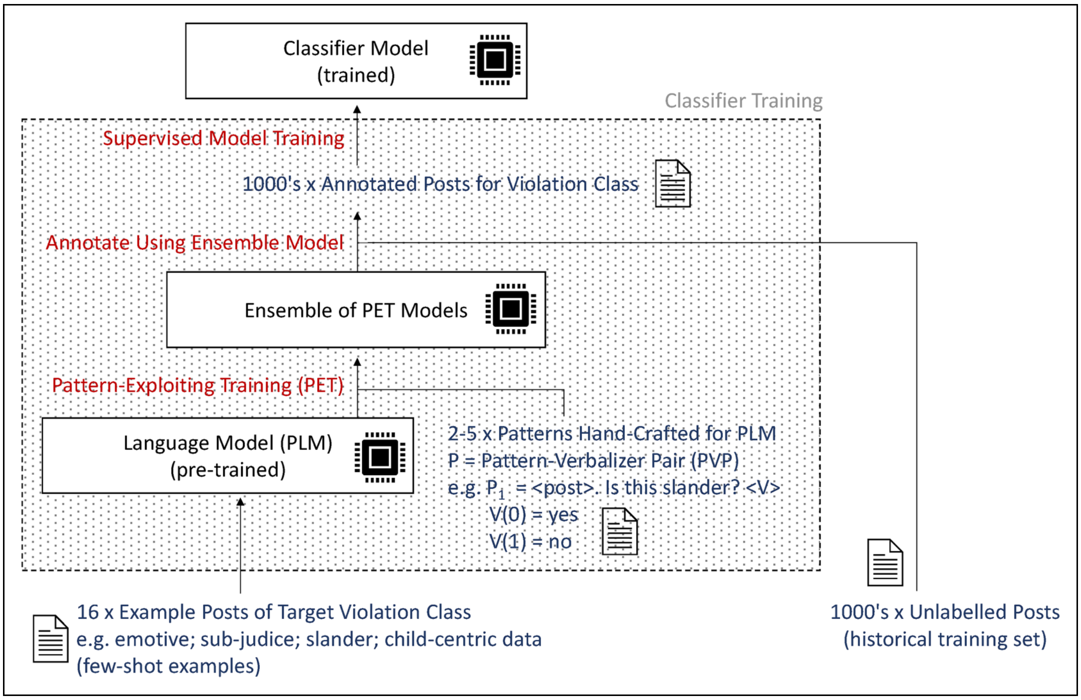
Classifier Model
(trained)
Classifier Training
Supervised Model Training
1000's x Annotated Posts for Violation Class
Annotate Using Ensemble Model
Ensemble of PET Models
Pattern-Exploiting Training (PET)
Language Model (PLM)
(pre-trained)
2-5 x Patterns Hand-Crafted for PLM
P = Pattern-Verbalizer Pair (PVP)
e.g. P_1 = <post>. Is this slander? <V>
V(0) = yes
V(1) = no
16 x Example Posts of Target Violation Class
e.g. emotive; sub-judice; slander; child-centric data
(few-shot examples)
1000's x Unlabelled Posts
(historical training set)

Figure 7. Few-shot NLP violation classifier using Pattern-Exploiting Training (PET).

demonstrate that invitational technology properties (initiators and accelerators) are mediative and distinct from facilitative artefacts. In their criminological analysis of intersections between technology properties and harm translation, Wood and Colleagues (2023, emphasis added) observe that, the inviting properties of a technology such as a social media platform are 'what a technology encourages users to do'. But facilitative artefacts are different. They are 'what a technology *allows its users to do*'. Reinforcing this, our findings distinguish between the two and highlight the distinct roles that inviting properties play beyond facilitation. We also propose an NLP model that can disrupt the harm production process and protect children. The framework reflects our understanding that the ANT approach inherent in current studies of sharenting is limited. The studies ignore or minimise the mediative capacity of technology properties and emphasise the motives and actions of sharenters. In contrast, our NLP approach follows the premise of the 'harm translation' and 'seductions of harm' perspectives which is that technology properties can inadvertently mediate harm through invitations and seductions respectively. In response, the NLP framework is optimised to detect risky sharenting content before it becomes subject to the mediative actions of technology properties (such as recommender and amplification systems) which initiate and accelerate secondary VDHs.

Conclusion

In this contribution, our aim has been to shed light on how technologies can, through various properties co-produce data harms with users such as sharenters. We have also outlined our vision for a new sociotechnical harm prevention framework informed NLP approaches. Our analysis of the operations of technology properties shows that although they may be integrated into design logics or features in good faith to enhance user experience, they can produce unintended consequences. They are amenable to misappropriation and can mediate the process of VDH production via harm translation processes. This occurs when inviting properties, for example, instil in a user the perception that the technology in question can be deployed for harmful conduct and the user adopts facilitative artefacts that allow them to actualise such an outcome.

What this demonstrates is that studying technology properties allows us to observe how design logics and features influence perception and usage (see also Wood et al., 2023). As Bucher and Helmond (2017) observe in relation to properties embedded in technology design and capable of offering various action possibilities, 'power is placed in the hands of designers who have the power to enable and constrain certain action possibilities through their design choices'. Recognising this, Norman's (1988) concept of 'perceived affordance' enjoins designers to pay careful attention to the ways in which hidden properties in the form of recommendations and amplification can inadvertently alert harm perpetrators and others to the presence of child-centric data in a social media site and invite unintended forms of use.

Our analysis has also produced new concepts that expand the multidisciplinary fields of human-computer interaction and sharenting, bridging gaps and providing pathways for both distinct areas of study to speak to each other in ways that can inform future research and advance knowledge. With the concept of *invisible properties*, researchers can integrate ethical considerations pertaining to the lack of transparency associated with proprietary data-driven machine learning models such as opaque recommender and amplification systems, into their analysis of sharenting risks and harms.

Both systems are proprietary models, and as already noted, such models are largely unamenable to external scrutiny. Some describe them as selective models that privilege and enhance the visibility of certain posts over others, distorting information for political reasons, profit-related imperatives, and other unknown purposes (e.g. Shin et al., 2022). Their opacity poses ethical challenges since they obfuscate the ways in which platforms channel information to users including potential data harm perpetrators. But they are arguably embedded in the architecture of social media platforms to enhance user friendly personalisation (Meta, 2022).

Although they can invite harmful usage unintended by the designers, the hidden properties are useful for platform companies keen to foster the high levels of use necessary for lucrative user data production. This points to tensions between profit-focused imperatives and user protection. The former is more concerned with ensuring that invisible proprietary algorithms can maximise both user engagement and associated profit. In contrast, user protection requires a commitment to promoting transparent models and remediating systemic vulnerabilities such as properties that invite harmful forms of use.

Added to the notions of *initiators* and *facilitators* introduced in this paper, the new concepts of *primary* and *secondary VDHs* further expand current understandings of how technology properties and potentially harmful forms of use (e.g. sharenting) intersect. We have shown how primary VDHs emerge from the actions of those who share the sensitive data of others without consent, exposing them to risks and harms. Secondary VDHs, on the other hand, emerge when others misappropriate the data.

With respect to future policy development, our findings highlight several points in the harm causation trajectory where preventative techniques can be introduced to forestall the effects of inviting properties. Social media platforms can, for example, target harm prevention measures at the initial phase when *initiators* invite sharenters who use *facilitators* to produce sensitive child-centric data, resulting in the primary VDHs that we observed.

Applying the NLP measures we propose would go some way towards interrupting primary VDH co-production (with sharenters). It can also help prevent secondary VDH co-production (with others, e.g. predators). This type of VDH occurs when visible and invisible technology properties acting as accelerators, progress the ham production trajectory. They do so by inviting others such as predators whose user profiles indicate an interest in child-centric data to engage with the content, exposing affected children to further (secondary) data harms. Child protection services and social welfare agencies known to surveil children and families from deprived groups (Edwards and Ugwudike 2023) and law enforcement agencies using social media data for surveillance (e.g. Trottier 2017).

Platform technologies are inherently sociotechnical and comprise not only social artefacts but also technical dimensions. Therefore, remedial strategies must address both the social dynamics (perceptions and actions of users) and technical elements (the operations of invisible and visible technology properties). The NLP framework we proposed is an example of a sociotechnical approach that can mitigate VDHs.

Although we conceptualise invisible and visible technology properties as initiators and facilitators of VDH production respectively, we do not propose that they should be abandoned. We recognise that even if profit motives inspire their design, they can enhance user engagement and opportunities to access numerous benefits. Examples include information, social interaction, and other positive outcomes that can improve human wellbeing. Sharenting, for example, can involve sharing child-centric data for networking, informational, and other beneficial outcomes (Haslam et al., 2017). Our primary contention here is that the properties can produce unintended harms, and preventative measures must be introduced for user protection.

Further, we certainly do acknowledge that technology designers cannot anticipate all the potential uses of their technologies, and embedded properties cannot discern user's moral or amoral intentions. But it is incumbent on the designers to consider the possible harmful uses their products can afford. As a guiding design principle, designers should ask the question: what forms of use can this product *afford* users? Our study shows that both visible and invisible properties should be considered for a comprehensive understanding of the specific role of technologies in the co-production of data harms.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the UK's Economics and Social Research Council (ESRC) project 'ProTechThem: Building Awareness for Safer and Technology-Savvy Sharenting' (grant number: ES/V011278/1).

ORCID iDs

Pamela Ugwudike https://orcid.org/0000-0002-1084-7796
Stuart E. Middleton https://orcid.org/0000-0001-8305-8176
Natalie Djohari https://orcid.org/0000-0002-7636-2863

Notes

1. We use the broader term – seductions of 'harm' – to accommodate forms of behaviour that are not 'crimes' *per se* but can be harmful (e.g. sharenting).
2. The community was based in the UK, used English as language, and its topic concerned legal issues such divorce, separation, and children custody. Further details of the community are provided later.
3. These are in vivo codes (verbatim terms extracted from the data), and they featured in the top 10 frequently used words relevant to the topic.
4. The numbers in brackets refer to the number of times the term appeared in the documents we analysed.
5. These are in vivo codes (verbatim terms extracted from the data), and they featured in the top 10 frequently used words relevant to the topic.
6. The numbers in brackets refer to the number of times the code appeared in the posts we selected for analysis.

References

Abidin C (2017) #Familygoals: Family influencers, calibrated amateurism, and justifying young digital labor. *Social Media + Society* 3(2): 1–15.

Angers J and Machtmes K (2005) An ethnographic-case study of beliefs, context factors, and practices of teachers integrating technology. *The Qualitative Report* 10(4): 771–794.

AOIR (2019) Internet Research: Ethical Guidelines 3.0 (2019) https://aoir.org/reports/ethics3.pdf.

Bezáková Z and Švec M (2021) Security risks of sharing content based on minors by their family members on social media in times of technology interference. *Media Literacy and Academic Research* 4(1): 53–69.

Bezáková Z and Švec M (2021) Security risks of sharing content based on minors by their family members on social media in times of technology interference. *Media Literacy and Academic Research* 4(1): 53–69.

Bowen GA (2009) Document analysis as a qualitative research method. *Qualitative Research Journal* 9(2): 27–40.

Brosch A (2018) Sharenting: Why do parents violate their children's privacy? *The New Educational Review* 54(4): 75–85.

Bucher T (2012) Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society* 14(7): 1164–1180.

Bucher T and Helmond A (2017) The affordances of social Media platforms. In: Burgess J and Poell T (eds) *The SAGE Handbook of Social media*. London: Sage, 233–253.

Campbell S, Greenwood M, Prior S, et al. (2020) Purposive sampling: Complex or simple? Research case examples. *Journal of Research in Nursing* 25(8): 652–661.

Cera M (2023) Digital ethnography: Ethics through the case of QAnon. *Frontiers in Sociological Theory* 8: 1–15. https://doi.org/10.3389/fsoc.2023.1119531

Clark T, Foster L, Sloan L, et al. (2021) *Bryman's social research methods*, Sixth Edition Oxford: oxford University Press.

Cooper A (2000) Cybersex and sexual compulsivity: The dark side of the force. *Sexual Addiction & Compulsivity* 7: 1–3.

Davidson T, Warmsley D, Macy M, et al. (2017) Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media* 11(1): 512–515.

Edwards R and Ugwudike P (2023) *Governing families: problematising technologies in social welfare and criminal justice*. Abingdon: Routledge .

European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union* L119: 1–88.

Fereday J and Muir-Cochrane E (2006) Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods* 5(1): 80–92.

Goldsmith A and Wall DS (2022) The seductions of cybercrime: Adolescence and the thrills of digital transgression. *European Journal of Criminology* 19: 98–117.

Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1): 1–15. https://doi.org/10.1177/2053951719897945.

Hamilton IA (2022) AI is not smart enough to solve Meta's content-policing problems, whistle-blowers say. https://www.businessinsider.com/meta-facebook-ai-cannot-solve-moderation-frances-haugen-daniel-motaung-2022-6?r=US&IR=T.

Haslam DM, Tee A and Baker S (2017) The use of social media as a mechanism of social support in parents. *Journal of Child and Family Studies* 26(7): 2026–2037.

He P, Gao J and Chen W (2023) DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, The Eleventh International Conference on Learning Representations. Available at: https://openreview.net/forum?id=sE7-XhLxHA.

Holiday S, Norman MS and Densley RL (2022) Sharenting and the extended self: Self-representation in parents' Instagram presentations of their children. *Popular Communication* 20(1): 1–15.

Katz J (1988) *Seductions of Crime*. New York: Basic Books.

Klucarova S and Hasford J (2023) The oversharenting paradox: When frequent parental sharing negatively affects observers' desire to affiliate with parents. *Current Psychology* 42(8): 6419–6428.

Kopf S (2020) "Rewarding good creators": Corporate social media discourse on monetization schemes for content creators. *Social Media+ Society* 6(4): 2056305120969877.

Latour B (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford.: OUP.

Lavorgna A, Tartari M and Ugwudike P (2022) Criminogenic and harm-enabling features of social media platforms: The case of sharenting practices. *European Journal of Criminology* 20(3): 1037–1060. DOI: 10.1177/147737082211316.

Marasli M, Suhendan E, Nh Yilmazturk, et al (2016) Parents' shares on social networking sites about their children: Sharenting, . *The Anthropologist* 242: 399–406. DOI: 10.1080/09720073.2016.11892031.

Martindale S (2014) Obsessive 'sharenting' could be more than digital narcissism. *The Conversation*. Available at: https://theconversation.com/obsessive-sharenting-could-be-more-than-digital-narcissism-30331.

Mata (2020) *Personalized Advertising and Privacy Are Not at Odds*. https://about.fb.com/news/2020/12/personalizedadvertising-and-privacy-are-not-at-odds/

Meta (2018) Bringing People Closer Together. https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/

Meta (2021) How Does News Feed Predict What You Want to See? https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

Meta AI (2019a) DLRM: An advanced, open-source deep-learning recommendation model. https://ai.facebook.com/blog/dlrm-an-advanced-open-source-deep-learning-recommendation-model/

Meta (2019b) Updates to Video Rankinghttps://about.fb.com/news/2019/05/updates-to-video-ranking/

Meta (2023) How AI Influences What You See on Facebook and Instagram. https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/

Middleton SE, Lavorgna A, Neumann G, et al. (2020) Information Extraction from the Long Tail: A Socio-Technical AI Approach for Criminology Investigations into the Online Illegal Plant Trade. In: 12th ACM Conference on Web Science.

Middleton SE, Letouzé E, Hossaini A, et al. (2022) Trust, regulation, and human-in-the-loop AI: Within the European region. *Communications of the ACM* 65(4): 64–68.

Neuendorf KA (2017) *The content analysis guidebook*. Sage Publications.

Norman DA (1988) *The Psychology of Everyday Things*. New York: Basic Books.

Pierre R (2022) 'Sharenting' your child's struggles online is abuse. *The Independent*. Available at: https://www.independent.co.uk/voices/sharenting-parenting-social-media-consent-b2063895.html.

Potter A, Barnes R (2021) The 'sharent' trap: parenting in the digital age and a child's right to privacy. In: Holloway D, Wilson MA, Murcia K, et al (eds) *Young children's rights in a digital world: Play, design and practice*. Cham: Springer, 283–297.

Ren P, Xiao Y, Chang X, et al. (2021) A survey of deep active learning. *ACM Comput. Surv* 54(9): art.180.

Schick T and Schütze H (2021) Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (ECAL-2021).

Schneider C and Trottier D (2012) The 2011 Vancouver riot and the role of Facebook in crowd-sourced policing. *BC Studies: The British Columbian Quarterly* 175: 57–72.

Scott J (1990) *A Matter of Record*. Cambridge: Polity.

Shin D, Hameleers M, Park YJ, et al. (2022) Countering algorithmic bias and disinformation and effectively harnessing the power of AI in Media. *Journalism & Mass Communication Quarterly* 99(4): 887–907.

Trottier D (2017) Fear of contact': Police surveillance through social networks. *European Journal of Cultural and Political*

Sociology 4(4): 457–477. DOI: 10.1080/23254823.2017.1333442.

Tufekci Z (2017) *Twitter and Tear Gas. The Power and Fragility of Networked Protest*. New Haven: YUP.

Ugwudike P (2022) Predictive algorithms in justice systems and the limits of tech-reformism. *International Journal for Crime, Justice, and Social Democracy* 11(1): 85–99.

Van Hee C, Jacobs G, Emmery C, et al. (2018) Automatic detection of cyberbullying in social media text. *PLoS One* 13(10): 1–22.

Varis P (2016) Digital Ethnography. In: Georgakopoulou A and Spilioti T (eds) *The Routledge Handbook of Language and Digital Communication* (55-68) Abingdon: Routledge.

Verbeek PP (2005) *What things do: Philosophical reflections on technology, agency, and design*. Pennsylvania State University Press.

Williams-Ceci S., Grose G. E., Pinch A. C., etal (2021) Combating sharenting: Interventions to alter parents' attitudes toward posting about their children online. *Computers in Human Behaviour* 125: 106939. DOI: 10.1016/j.chb.2021.106939.

Wood MA, Mitchell M, Pervan F, et al. (2023) Inviting, Affording and Translating Harm: Understanding the Role of Technological Mediation in Technology-Facilitated Violence. *British Journal of Criminology*. 63 (6) 1623-1624. https://doi.org/10.1093/bjc/azac095.