Project Minerva: Interim Report

Dr Katherine Allen
Megan Hermolle

July 2022

DISCLAIMER

While every effort has been made to ensure that the information contained in this draft report is accurate and up to date, the author(s) cannot accept legal responsibility or liability for any actions taken by readers as a result of any errors or omissions.

Published by:

Institute of Social Justice and Crime
University of Suffolk
Waterfront Building
Ipswich
Suffolk
IP4 1QJ

Contents

Table of figures

Abbreviations

AI	Artificial Intelligence
DA	Domestic Abuse
GBV	Gender-Based Violence
HCI	Human-Computer Interaction
IBSA	Image-Based Sexual Abuse
SWGfL	South West Grid for Learning
TFA	Technology-Facilitated Abuse
TFVAWG	Technology-Facilitated Violence Against Women and Girls
UK	United Kingdom
UoS	University of Suffolk
VAWG	Violence Against Women and Girls

1.1 Introduction & background

Harmful online acts and speech can have far-reaching emotional, psychological, financial, reputational and interpersonal repercussions for victim-survivors (Maple *et al*, 2011; McGlynn *et al,* 2019), and pose a growing concern as we transition into a digitized society (Glitch & EVAW, 2020). Avoidance or withdrawal is no longer a practicable option, as work, commerce, leisure and socializing increasingly require access to online spaces, a transition which has gathered pace during the pandemic. Further, individualising behavioural prescriptions to log off or 'ignore the trolls' elide the fact that online and social media spaces provide a crucial forum for women and girls to exercise their freedom of expression (Glitch & EVAW, 2020).

Project Minerva is designed to address behaviours which harm and/or endanger women and girls and impede their ability to safely navigate online (and offline) spaces, as well as linking patterns of online abuse and offline gendered violence.

The South West Grid for Learning (SWGfL) has commissioned the University of Suffolk (UoS) to conduct the research and evaluation strands of the project, undertaking research to inform the development and evaluation of a bespoke AI tool for women experiencing online abuse. The Minerva tool is intended to equip women subject to online abuse with the knowledge and resources they need to find safety and regain control.

1.2 Research design and objectives

In order to ensure that the development and evaluation of the Minerva tool reflects the needs and concerns of key stakeholders such as VAWG sector professionals and victim-survivors, researchers chose to adopt a participatory research approach involving in-depth qualitative interviews and elements of co-design.

This interim report summarises the key findings of the research stage of the project which extended from March – mid-July 2022. As an interim report, this paper focuses on detailing researchers' findings regarding:

- The forms of online abuse & technology-facilitated violence against women and girls (TFVAWG) experienced by women in the UK, and the impacts associated with these
- The support and reporting options currently available, including barriers, gaps and dysfunctionalities in this system and opportunities for more collaborative working practices
- The needs, expectations and preferences of UK women subject to online abuse/TFVAWG in relation to the Minerva tool
- The potential benefits, risks and challenges associated with the tool, and how these can be monitored and evaluated

Our sources for these interim findings include:

- Survey responses from 148 women with lived experience of online abuse and/or TFVAWG
- Survey responses from 58 professionals whose work involves supporting or engaging victim-survivors of online abuse/TFVAWG

- Interviews with five victim-survivors regarding their experiences of reporting and help seeking
- Interviews with five professionals regarding their experiences of engaging victim-survivors and/or conducting training, research and advocacy work in relation to online abuse/TFVAWG
- Co-design interviews with five victim-survivors regarding what they would like to see in an AI tool
- A scoping review mapping available literature on online abuse and TFVAWG, including: prevalence and harms, definitional questions; solutions and interventions; contextual factors such as rurality; how the Internet of Things (IoT) and other technologies are instrumentalised by perpetrators to surveil and control victims
- Published reports (e.g., SWGfL & Home Office, 2021) and RPH/RHC data shared by SWGfL

This report explores emerging answers to the following research questions:

RQ1. What forms of online abuse and technology-facilitated VAWG are UK women experiencing, and what are the harms associated with these experiences?

RQ2. What do women experiencing online abuse expect/want/need from an AI tool, and what outcomes are they looking for?

RQ3. What remedies are currently available to UK women experiencing online abuse, and what are the gaps and vulnerabilities within existing systems? Where are there opportunities for enhanced collaboration and connectivity, and how can Minerva support and build on these?

RQ4. What are the potential risks associated with developing and implementing such a tool and how can these be pinpointed, tracked and managed?

Based on these answers, researchers frame a set of recommendations, and present a preliminary evaluation framework to assess incoming user feedback and outcomes against projected benchmarks.

2. Key findings

2.1 "*This is something that's going to be with me for the rest of my life*"[1]: tracking abuse and understanding harms

Our survey, interview and scoping review findings indicate that women in the UK and beyond experience a range of forms of online abuse, including as part of a wider pattern of abusive, coercive, controlling and/or harassing offline behaviours. Moreover, there are striking parallels between these findings on online abuse and the wider literature on how VAWG shapes women and girls' lives, including exposure to social and professional harms and reduced access to collective goods and public spaces (Vera-Gray & Kelly, 2020).

Scoping review

Our scoping review uncovered an extensive global evidence base regarding the prevalence and impacts of online and tech-facilitated harm, including cyberstalking (Alexy *et al*., 2005); image-based sexual abuse (Campbell *et al*., 2022); intimate partner cyber-abuse and surveillance (Brem *et al*., 2019); and sexual coercion (Drouin *et al*., 2015).

In terms of prevalence, reviewed studies suggested high levels of both victimisation and perpetration among studied populations (Lenhart et al., 2016). Negative emotional impacts including feelings of fear, anxiety, depression, distrust, and anger were reported across many of the articles amongst other negative effects (Branch *et al.,* 2017; Campbell *et al*., 2020; Dreißing *et al,* 2014). Women were more negatively impacted and more likely to be impacted by online harms than men (Gamez-Guadix et al, 2015), reflecting the gendered nature of online harassment and TFA.

In relation to technology facilitated VAWG and coercive control, included articles identified a range of abusive behaviours by perpetrators, such as breaking and monitoring phones or computers (Belknap et al., 2012), image based sexual abuse (Henry et al.), and sexual coercion (Brown et al., 2021), and the kinds of technology they use to engage in such abuse, for example spyware apps and GPS (Havard & Lefevre, 2020; Eckstein, 2020). This duality – whereby perpetrators are able to control and terrorise victim-survivors through practices of digital exclusion and/or digital surveillance or 'omnipresence' – highlights the complex nature of TFVAWG and online abuse.

Many of the studies examining the role of the Internet of Things (IoT) in TFVAWG were qualitative, indicating that the themes and findings which emerged came from lived experience or experience of supporting women experiencing TFA (Douglas *et al*., 2019). Some of the common findings in this literature included the existence of many different apps that are used for intimate partner surveillance, many made specifically for that purpose, but also numerous which are dual use – apps which

[1] Quote from victim-survivor interviewee, 'Cara' (all interviewees have been assigned pseudonyms to protect their confidentiality)

have ‘legitimate’ purposes such as ‘find my iPhone’, which can be used to monitor or control (Chatterjee *et al*., 2018). Also commonly highlighted was the dual nature of technology – victim-survivors need technology such as phones, email, and social media, to connect with support networks or support services, to collect and collate evidence of abuse, or even as part of a perpetrator’s mandated contact with children (i.e., facetime or zoom); however, abusers use the same technology against them, and victims are often less ‘tech-savvy’ than perpetrators (Leitao, 2019). Similarly, although this is covered more specifically in a later section, it was emphasised that many support services do not have the technological expertise or capacity to deal appropriately with women experiencing TFA (Tanczer *et al.,* 2018). Many recommendations centred around this issue, in addition to proposals for improved and easier to navigate privacy and security solutions, multi-factor authentication for IoT devices, and better tools for storing and retrieving digital evidence of TFA or online abuse.

There were mentions throughout the literature of links between offline and online abuse. For example, Brem et al. (2017) found high levels of cyber-abuse and cyber-monitoring behaviours in men who had been arrested for domestic violence, while Dreißing et al. (2014) found that many victim-survivors of cyberstalking reported transitions from online to offline stalking, or vice-versa, suggesting a commonality between online and offline stalkers. Additionally, Douglas *et al*. (2019) carried out interviews of DA survivors aimed at exploring DA in general. Participants were not asked about TFA specifically, but 83% of women volunteered information about experiencing behaviours such as smartphone coercion, IBSA; social media-abuse, and online harassment, which suggests that victim-survivors consider there to be an intrinsic link between online and offline violence.

Survey findings

Among survey participants, the most commonly experienced forms of abuse included receiving unwanted sexual messages (61%) and cyberstalking or harassment (44.9%). A third of respondents also reported receiving unwanted violent or pornographic content (36.4%), hate speech (33.9%), and experiencing threats or blackmail (33.1%). Other forms of abuse, such as doxing, were reported by only a minority of participants.

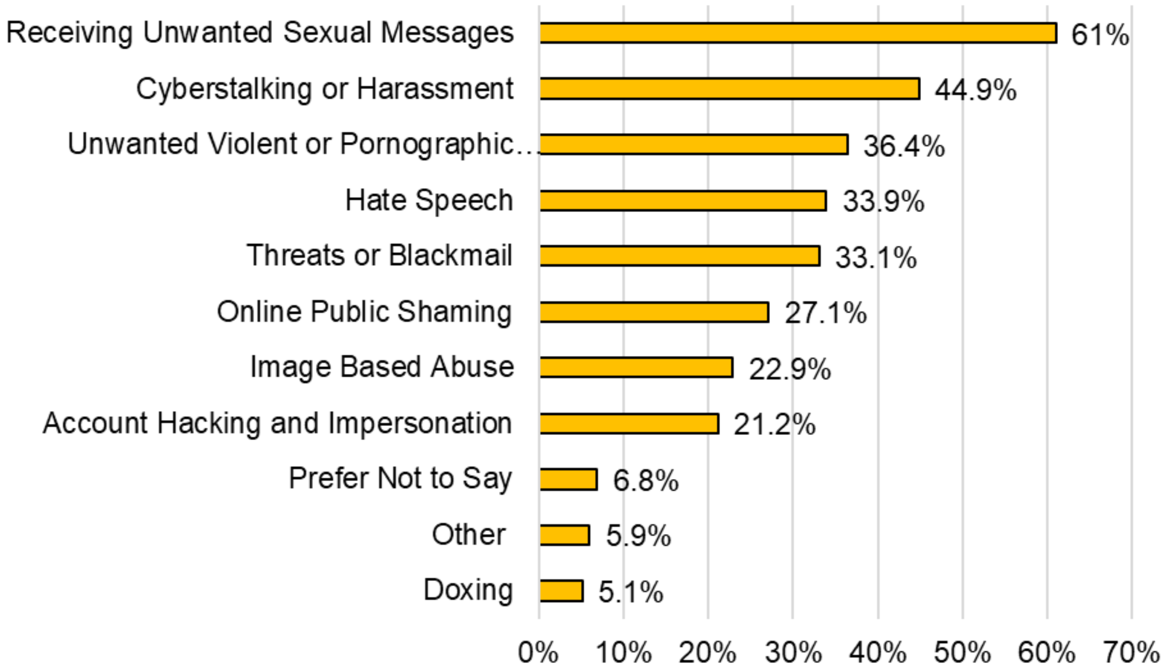


Figure 1: Most frequently experienced forms of online abuse

However, it is important to note that achieving accurate prevalence estimates via self-reporting hinges on clear and accessible terminology. As the scoping review identified, variance in definitions of concepts can inhibit shared understanding and, where classification and definitional issues arise, minoritised groups or those with additional language or literacy needs may be excluded as a result (Blackwell et al., 2017). As open-ended survey responses show that not all victim-survivors who participated were familiar with the term 'doxing', it is plausible that this terminological ambiguity contributed to under-reporting. This finding underlines the importance of accessibility and clarity of language when developing the Minerva tool (a point which was also emphasised by professional and victim-survivor interviewees).

As well as being the most commonly reported forms of abuse, cyberstalking and harassment (21.8%) and receiving unwanted sexual messages (21.8%) were identified as the most enduring forms of abuse experienced (see Figure 2). As stalking and harassment are legally defined as course of conduct offences involving an ongoing pattern of fixated and unwanted behaviour, it is unsurprising that this is reflected in the findings regarding duration. Based on additional insights from interviews and the wider literature, researchers formulated several possible hypotheses regarding the unwanted sexual messages finding, including:

- A series of discrete/unconnected messages from a number of individual perpetrators (e.g., unwanted sexual overtures or cyberflashing by strangers encountered online or via apps)
- A pattern of connected, but not coordinated, messages from a number of perpetrators, potentially associated with prior experiences of image-based sexual abuse (IBSA) and "collector culture", which may include sharing of identifying details and personal information [2]

[2] Collector culture is defined as the posting, collating and trading of intimate images online

- An orchestrated campaign of messages from a number of perpetrators, potentially associated with prior experiences of IBSA and collector culture, or other forms of gender-based violence. Unsolicited sexual messages may form part of a wider harassment campaign, for example during the GamerGate harassment of indie game developer Zoe Quinn (Salter, 2018)

Victim-survivors facing ongoing abuse of this form are likely to have different security and support needs based on which of these hypotheses (if any) fits their individual circumstances; therefore, where possible wider contextual factors regarding the pattern of abuse should be taken into account when tailoring support and signposting options.

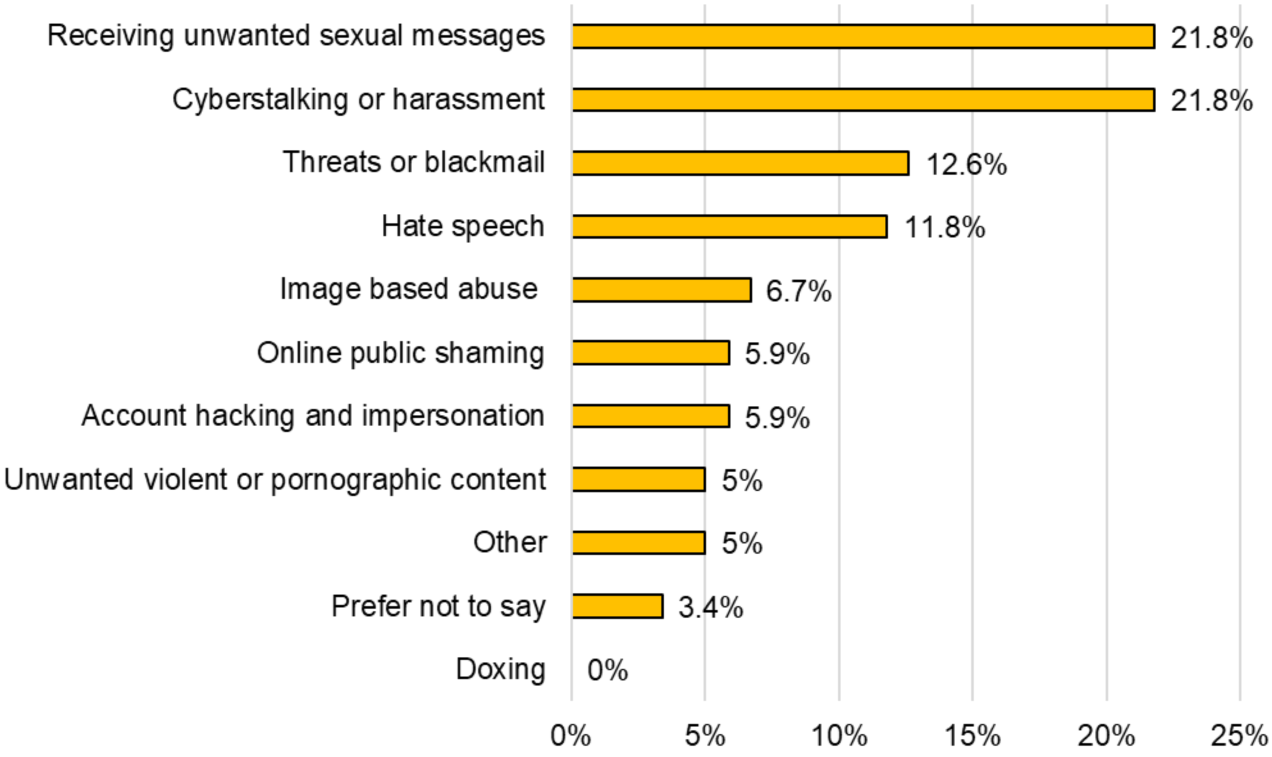


Figure 2: Most enduring form of abuse experienced

There was a polarised distribution of abuse duration, meaning that participants' experiences tended towards both extreme ends of the presented timeline. For example, 21% of respondents experienced abuse for 1-6 days, while 17.6% of respondents experienced abuse for 2+ years, with fewer responses at 6-9 months (2.9%), indicating that respondents either experience shorter-term or longer-term abuse.

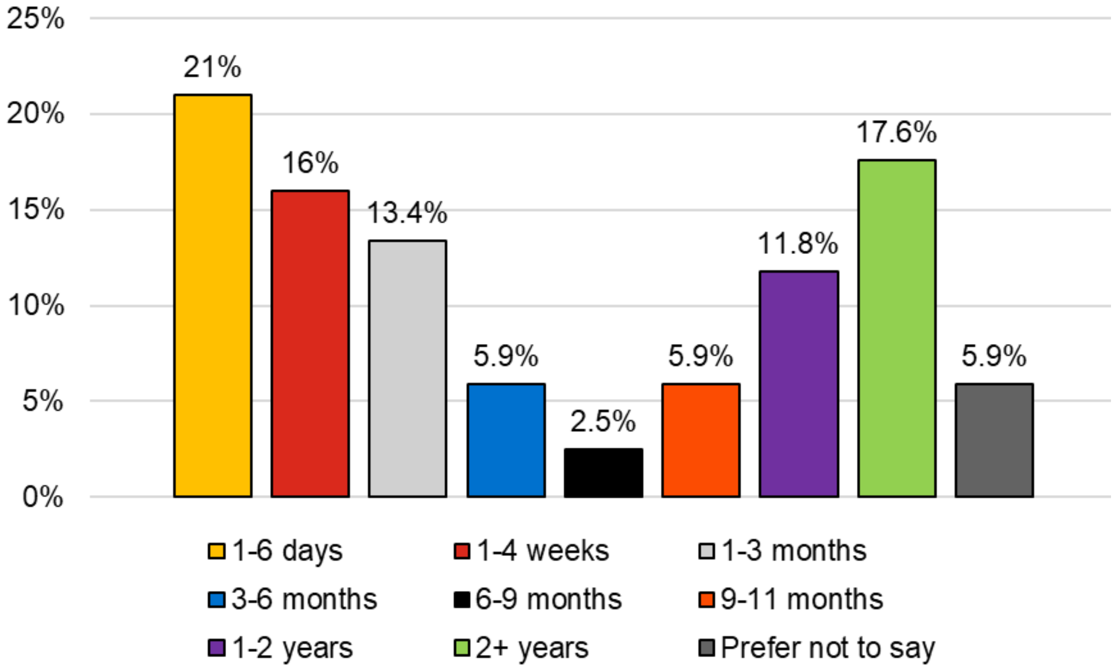


Figure 3: Duration of longest period of abuse

In terms of the victim-survivor's relationship to the abuse perpetrator, over half of respondents indicated that their perpetrator was a stranger (53.8%). The next most frequent response was former partner (22.7%), while the least frequent response was family member (1.7%).

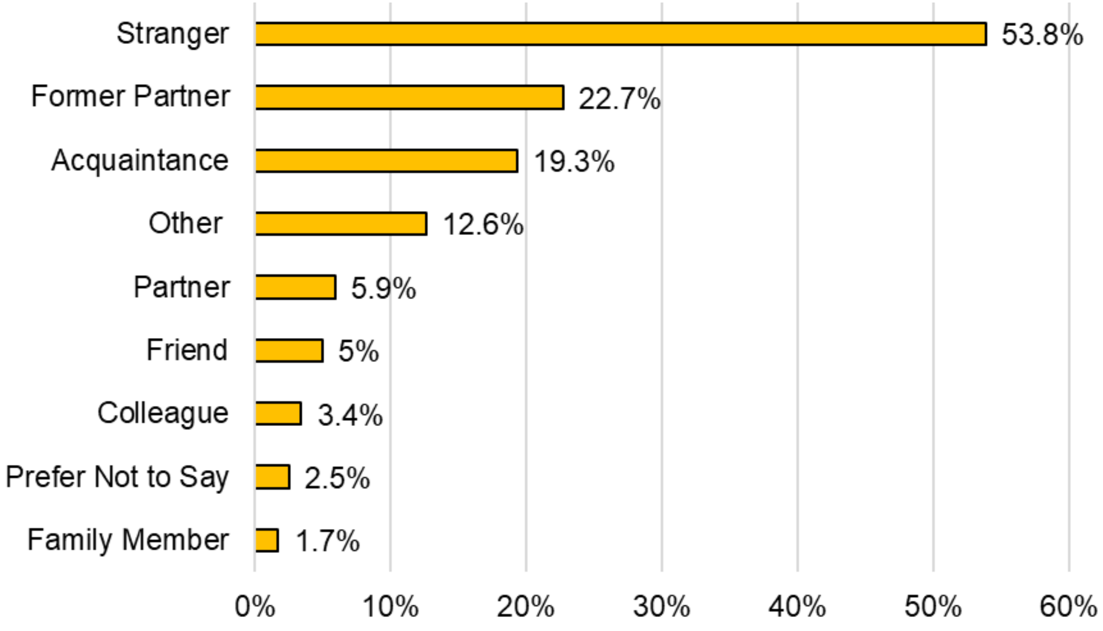


Figure 4: Perpetrator relationship to victim-survivor

The most frequent platforms where participants experienced online abuse were Facebook or Facebook Messenger (44.1%). Instagram (30.5%) and WhatsApp (25.4%) were also commonly cited. 29.7% of respondents also chose ‘other’, indicating that there are very common platforms not mentioned in this survey. Open-ended responses to this question suggest that platforms such as pornography sites, email, text, Skype, YouTube, and Tumblr are also common places where abuse happens. The least common platforms were Reddit (1.7%), LinkedIn (2.5%), and TikTok (2.5%).

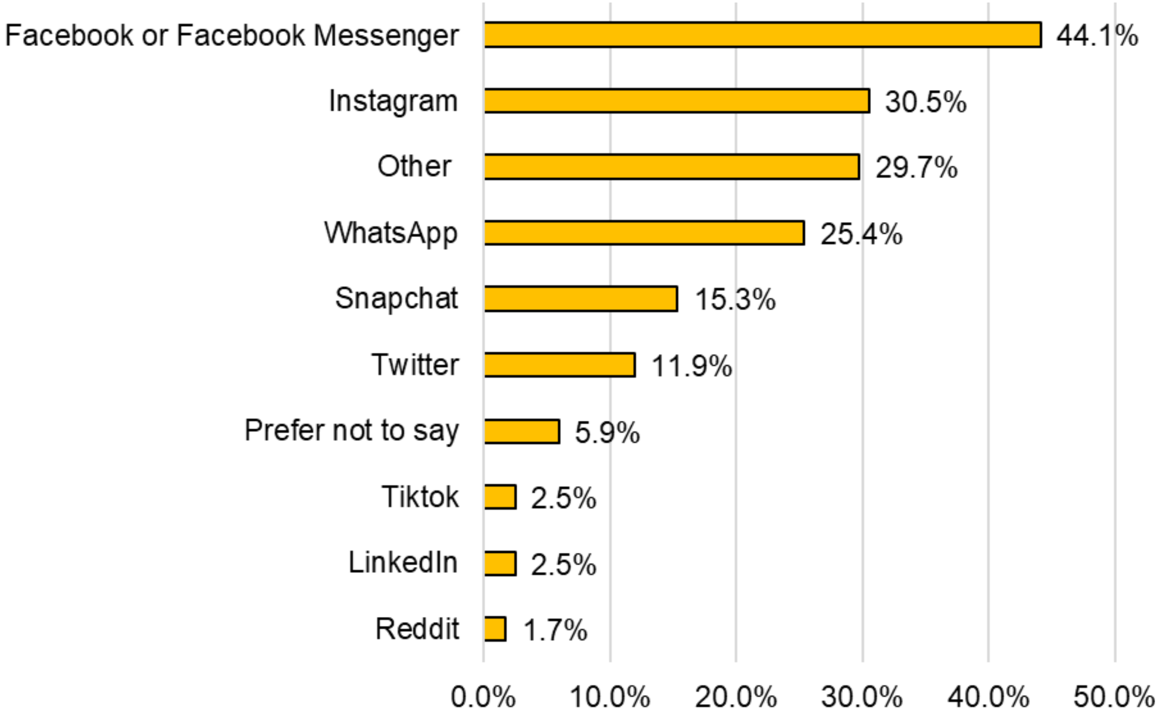


Figure 5: Where the online abuse happened

In terms of the impacts experienced as a result of the abuse, findings suggest that emotional impacts were extensive and varied. Participants most commonly reported feelings of anxiety (62.2%), stress (53.8%) and anger (52.1%). Additionally, many of the emotions in this question scored above 25%, with traumatised (23.5%) and panic attacks (19.3%) representing the lowest frequency items, although this is still very high. 16.8% of respondents specified other emotional impacts, including fear of their home address being shared, suicidal feelings or the reactivation/triggering of previous trauma.

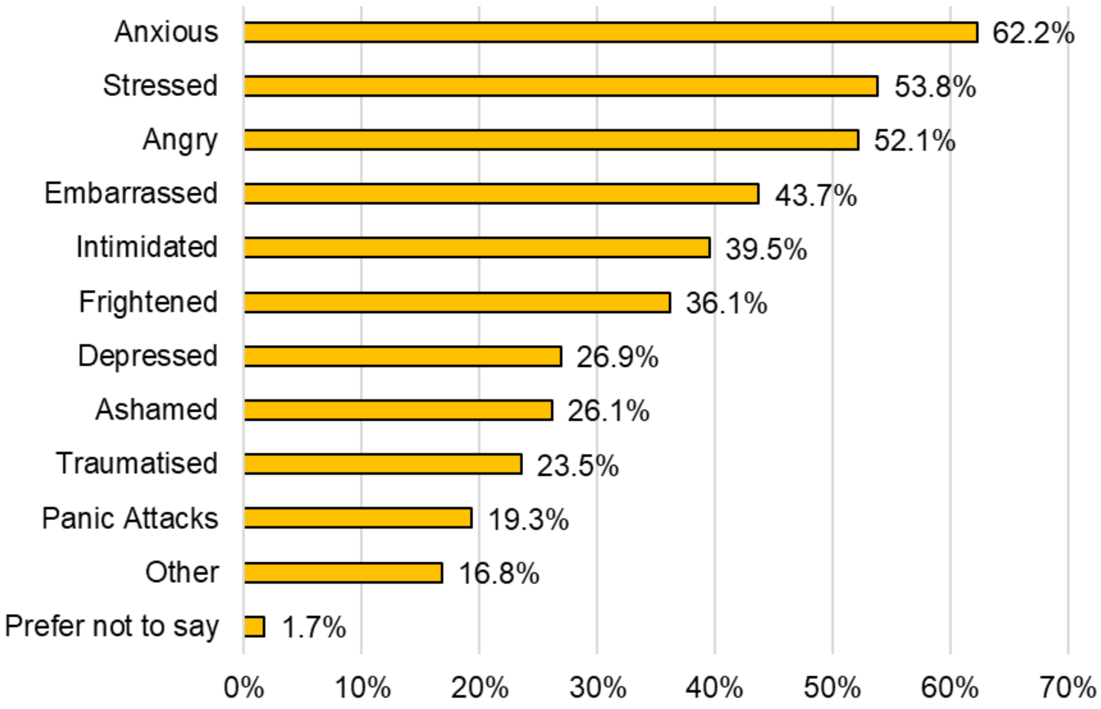


Figure 6: Emotional impacts of abuse

Participants also reported a range of social and professional impacts, including loss of confidence (57.3%) and feeling isolated (37.6%) or bullied (34.2%). Meanwhile, 23.1% of respondents indicated that they lost friends or acquaintances as a result of the abuse. 27.4% of respondents reported experiencing 'Other' social and professional impacts such as worry for the safety of their children, loss of trust in people, and effects on loved ones such as their family and friends. These findings underline the tangible 'real world' impacts associated with online and technology-facilitated abuse.

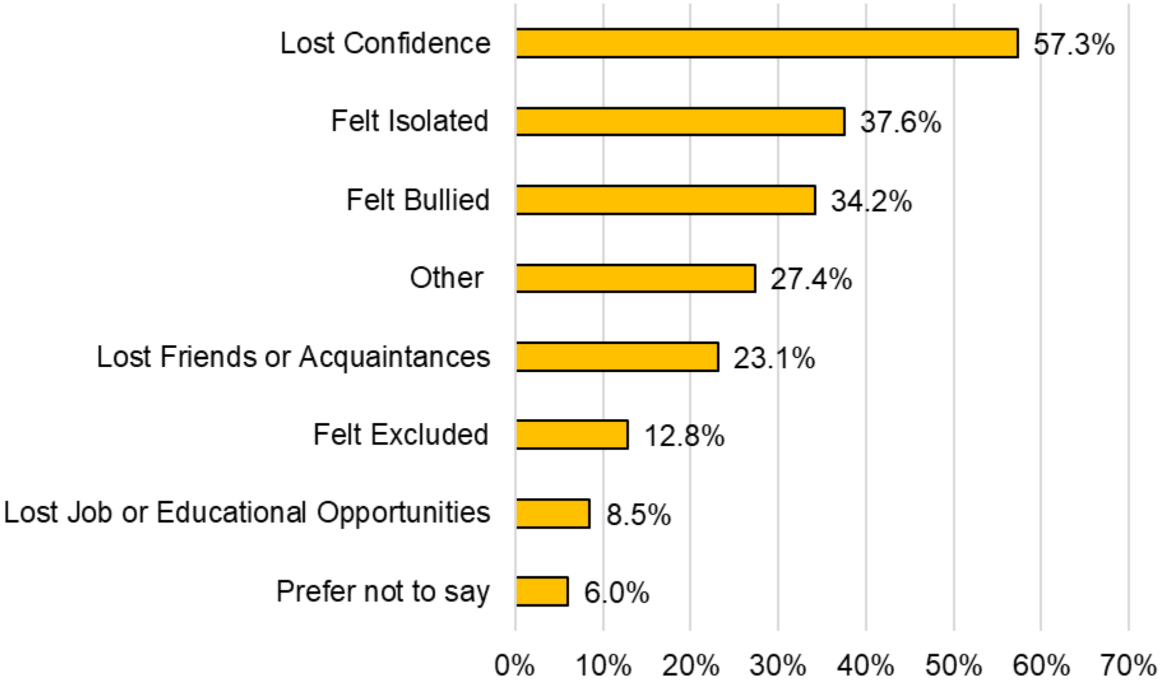


Figure 7: Social and professional impacts of abuse

Regarding the behavioural impacts of online abuse and/or TFVAWG, participants reported limiting their use of online spaces and communication, with more than half of participants (55.9%) stopping or reducing online interactions, 47.5% stopping or reducing their use of social media and 39% expressing themselves less online. While less common, there was also a substantial impact on some participants' offline interactions and relationships, with 32.2% reporting that they expressed themselves less in real life and 24.6% isolating themselves from family and friends.

This chilling effect on women's expression and ability to access to public spaces underlines the "silencing effect" of online abuse and TFVAWG and emphasises that its harmful effects extend far beyond offence (Glitch & EVAW, 2020: 28).

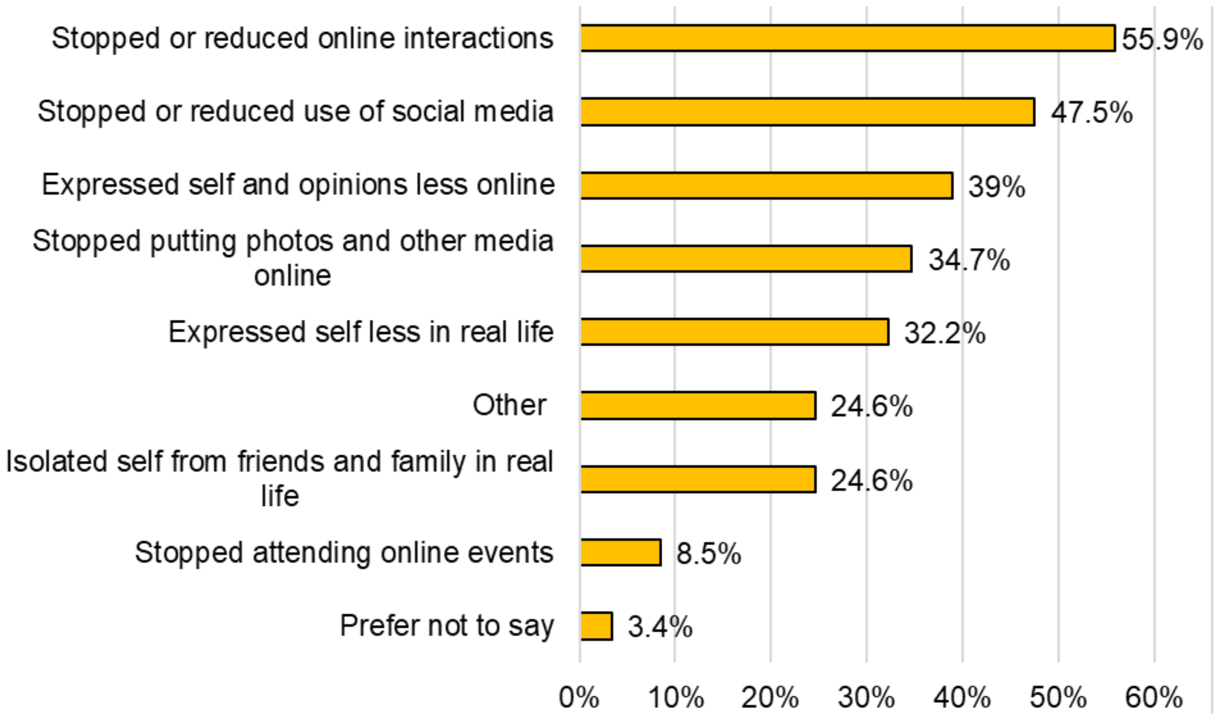


Figure 8: Behavioural impacts of abuse

Correlational analyses

Many significant moderate and some significant weak and strong correlated relationships were found to exist between survey items. Due to the nature of the analysis and the way the data was collected, we can only look at the existence and strength of association and cannot draw any conclusions about the direction of the relationships.

In terms of forms of abuse and how they relate to the victim-survivor's relationship to the perpetrator, threats or blackmail were found to have a relationship to family member (weak) and former partner (moderate). Cyberstalking or harassment were found to have moderate associations with stranger and former partner, while image-based abuse had a moderate relationship to partner. Unwanted sexual messages had a weak relationship with stranger and a moderate relationship with ‘other’ (qualitative responses to this indicated elected officials, and external professionals). Hate speech and family member also had a weak relationship.

Forms of abuse and emotional impact were found to have many moderate relationships. Threats or abuse and feeling ashamed, traumatised and having panic attacks each had moderate relationships, and cyberstalking had moderate associations with feeling frightened, anxious, and having panic attacks, Similarly, image-based abuse was moderately associated with feeling frightened, embarrassed, depressed, traumatised, and having panic attacks; and strongly associated with feeling ashamed. Unwanted sexual messages were moderately associated with feeling embarrassed, while doxing had a moderate relationship to feeling intimidated. Online public shaming was moderately related to feeling ashamed, stressed, traumatised, having panic attacks, and ‘other’ (qualitative responses indicated include upset, annoyed, and body conscious). ‘Prefer not to say’ was moderately related to embarrassment and anxiety.

Forms of abuse and professional and social impact also had several similar relationships: threats or blackmail were moderately related to losing friends or acquaintances and jobs or educational opportunities, and image-based abuse had moderate relationships with losing confidence, feeling isolated, and losing friends or acquaintances. There was also a moderate relationship between receiving unwanted sexual messages and feeling bullied; and hate speech and feeling bullied. Doxing was moderately associated with losing jobs or educational opportunities. There were several impacts associated with online public shaming: feeling isolated, excluded, bullied, and losing jobs or educational opportunities were all moderately associated, while losing friends or acquaintances were relatively strongly associated. Cyberstalking or harassment was moderately associated with feeling bullied, losing friends or acquaintances, and losing jobs or educational opportunities.

In terms of forms of abuse and changes to victim-survivors' behaviour as a result of online abuse, there were several associations, all moderate strength. Threats or blackmail was associated with self-isolation, while image-based abuse was related to the respondent expressing themself less online and expressing themself in real life. Similarly, receiving unwanted pornographic or violent content was associated with expressing themself less online and putting less photos or media online. Online public shaming was related to self-isolation, expressing themself less in real life, reducing online interactions, and stopping attending online events. Finally, cyberstalking was associated with putting less photos and media online.

The final correlation analyses which produced significant findings was whether certain forms of abuse have any relationship to other forms of abuse. There were several moderate associations: threats or blackmail was associated with image-based abuse and online shaming, while image-based abuse was found to be related to receiving unwanted sexual messages, receiving unwanted violent or pornographic content, and online public shaming. Hate speech, doxing, and cyberstalking were also all associated with online public shaming.

Correlational analyses were additionally carried out to determine whether there were relationships between rurality and duration, gaps in support, and reporting, and no significant associations were found.

Interviews

While the focus of the follow-up interviews conducted with victim-survivors was their experiences of reporting and help seeking rather than the abuse they had experienced, these interviews also shed light on some the long-term impacts of online abuse and/or TFVAWG.

In addition to the emotional, social and professional impacts reported by survey participants, interviewees’ accounts suggest that the legacies of online abuse can be both long lasting and unpredictable, an intermittent but intrusive presence threading through victim-survivors' lives:

You know, this is something that's going to be with me for the rest of my life […] As for the reporting, yeah, I mean, a lot of them are taken down, and the [RPH], they do give me bits of information on how many they report and the percentage that are removed. I think there's a 90% removal rate at the moment. But I asked for a recent report off them for the police and yesterday they messaged me to say this month alone the images had been shared hundreds of times on certain sites. And this is what the police are looking into, because, you know, fake Facebook profiles keep being made as well, which are then messaging people I know. […] They've also, whoever, in the past years, has got photos of me from Facebook. You know, so I've had friends and strangers messaging me through messenger or you know, someone has even emailed my work email, and they've seen photos of me.

'Cara', victim-survivor

Another interviewee recalled how ongoing TFVAWG in the wake of a coercive and controlling relationship significantly altered the trajectory of her life, prompting a move and change of career:

I was going through really prolonged post-separation abuse with my [child']s dad, which had been going on since [several years ago]. And it still wasn't over and it was ongoing and various things sort of came to a head and eventually there was a lot of- there was massive damage to my career and I had to move. I was kind of harassed to the point where, you know, I just couldn't stay in the area where I could do my previous career, which was in [technology] and which I'd got a [qualification] in. So I was just like hounded out of that.

'Elle', victim-survivor

A third interviewee described how her sense of regained safety following a period of persistent online harassment by a former partner came down to chance:

The only thing that was actually making me feel safe was the fact that this guy legally could not leave the country [he was living in] because the borders were closed [due to the pandemic]. But that wasn't a reflection on... you know, services that I might have contacted, or me and things in and out of my control. It's just the way things were.

'Maeve', victim-survivor

These accounts point to a loss of control over major aspects of one's life as a correlate of online abuse, which may account for the predominance of anxiety among survey participants. This loss (or theft) of one's ability to share or withhold personal information as one sees fit, to decide major life events based on one's own preferences, aligns with characterisations of trauma as a loss of meaning associated with the subversion of foundational beliefs about the world, such as the ability to control one's own destiny (Janoff-Bulman, 1992). This felt loss of control – and the importance of restoring a sense of agency - was a theme that would re-emerge during co-design interviews, underlining the importance of victim-survivor consultation for understanding user journeys and values.

2.2 “*With the police, they don't believe you basically*”[3]: victim-survivor journeys to safety and support

In addition to the scoping review, which identified a dearth of targeted interventions for those experiencing online abuse or TFVAWG and a lack of technological knowledge among VAWG services, survey and interview findings suggest that, despite pockets of good practice and collaborative working, there are systemic gaps and barriers that may prevent victim-survivors from getting timely help and support, particularly within UK law and the criminal justice system. These gaps include legislative loopholes that enable some perpetrators to evade accountability and uninformed – and at times victim-blaming - police responses that deter reporting and make it harder for victim-survivors to access justice, as well as widescale resource and training issues.

The findings also pointed to cultural barriers such as the normalisation or minimisation of some forms of online abuse/TFVAWG, stigma and victim-blaming.

Scoping review

Many of the tech and online solutions in the identified literature were aimed at women experiencing intimate partner violence offline (Ford-Gilboe et al., 2020; Glass et al., 2017). Most of the literature includes evaluating or reviewing specific interventions, such as tech safety and education programs (Finn & Atkinson, 2009) and tailored safety decision aids (Hegarty et al., 2019; Koziol-McLain et al., 2015). One article carried out a thorough review of all possible technological solutions to IPV, including Internet of Things devices, wearable devices to monitor vitals in case of violence, and emergency measures such as smart alarms (Rodriguez-Rodriguez et al., 2019). In terms of outcomes, the interventions broadly reported improvements for measures such as PTSD, depression, levels of violence and decisional conflict.

It is worth noting that one identified solution was specifically aimed at victims of online harassment (Blackwell et al., 2017). This involved the evaluation of HeartMob, a by-and-for online platform for women experiencing online harassment. Recommendations arising from this study included the emphasis that care needs to be taken in classifying online harassment by tech companies and social media sites, and that any classifications are intersectional and do not exclude minoritised groups. An additional recommendation was a user-led process.

A common theme throughout the portion of the literature which dealt with domestic abuse support services was that frontline services and staff supporting women experiencing TFA or online abuse are struggling with a lack of expertise in the technology used in perpetrating abuse. For example, the professionals in Freed et

[3] Quote from victim-survivor interviewee ‘Elle’

al.’s (2017) study emphasised that they lacked sufficient knowledge to help service users, and the only advice they could give them regarding TFA was very basic and not very actionable. Similarly, the findings of several linked research projects by Tanczer et al. (2018) indicated that support services are not adequately equipped to respond to TFA, and especially abuse related to IoT devices., and Tanczer et al. (2021) pointed out that the entire VAWG sector is outrun by the fast pace of technology with limited capacity posing a significant issue.

Many of the recommendations made in reviewed articles were aimed at VAWG sector support services, the criminal justice system, or social media platforms, with a very small proportion aimed at refining any technological solutions for women experiencing TFA and online abuse. However, some useful general recommendations from the scoping review can be applied to Project Minerva: for example, the emphasis on co-design or by-and-for, and survivor-led approaches and solutions which should include ease of setting digital security and privacy and language inclusivity; intersectionality and consideration of marginalised and minoritised groups; clear definitions of TFA related terminology; and a specific recommendation for tech experts to work with legal professionals to develop new techniques for collecting legally valid digital evidence (Freed et al., 2017).

Survey

Victim-survivor respondents most frequently indicated reporting their abuse to family or friends (44.9%). In terms of formal reporting avenues, social media moderators or admins were the most common (28.8%), followed by police (25.4%). No respondents chose Crimestoppers or the Rape Crisis Helpline. Additionally, almost a quarter of respondents (24.6%) indicated that they had never reported what had happened to anybody.

These findings are in line with the research literature on other forms of GBV such as intimate partner violence, which suggests high rates of disclosures to informal support sources such as family or friends, with informal support networks being the option of first resort for many victim-survivors (Sylaska & Edwards, 2014). This finding also chimes with feedback during the co-design interviews, where several participants suggested providing safe access to peer support or ‘success stories’ from those with similar lived experiences as one of the Minerva tool’s functions.

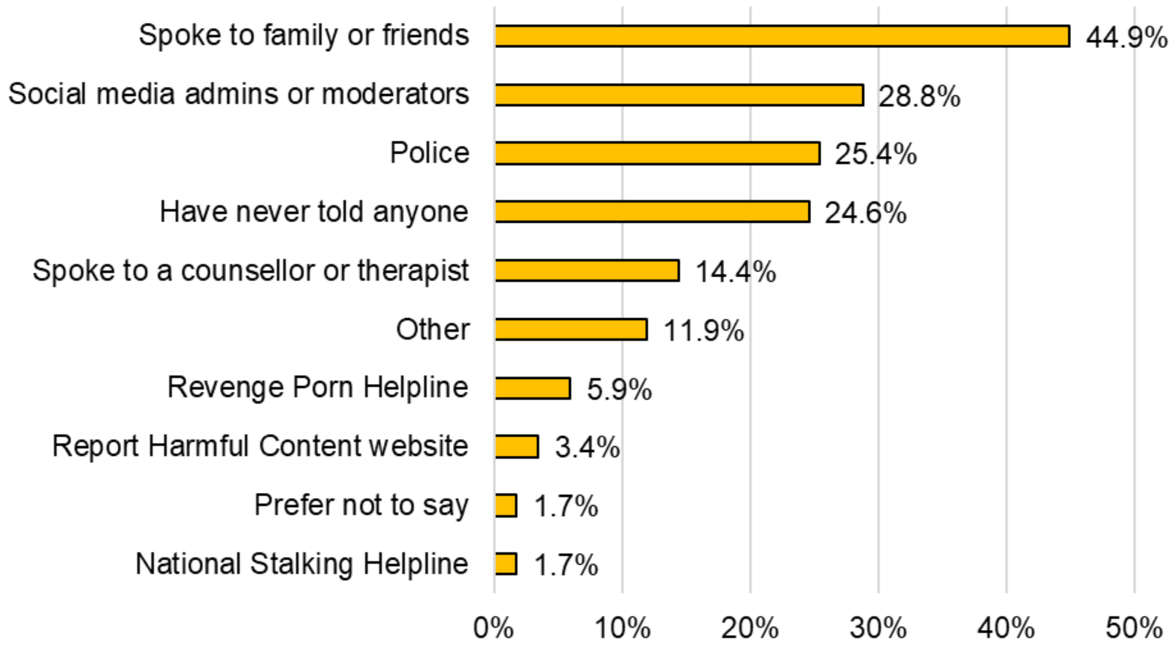


Figure 9: Where the victim-survivor reported abuse

As a result of reporting their abuse, respondents indicated a variety of positive and negative emotional responses, including feeling heard (35.3%), followed by feeling supported (32.9%) and feeling more in control (27.1%). However, there were also high levels of negative impacts of reporting: 24.7% felt no action was taken to stop abuse, 23.5% did not feel taken seriously, and 22.4% did not feel heard. Additionally, the lowest frequency item in this question aside from ‘Prefer not to say’ was Felt Safer (14.1%). 20% of respondents also chose ‘Other’, with open-ended responses citing a mixture of positive and negative impacts, stating that whether they felt heard or not varied based on which the platform they reported to, or stating that the answer is more complex than any combination of the options given.

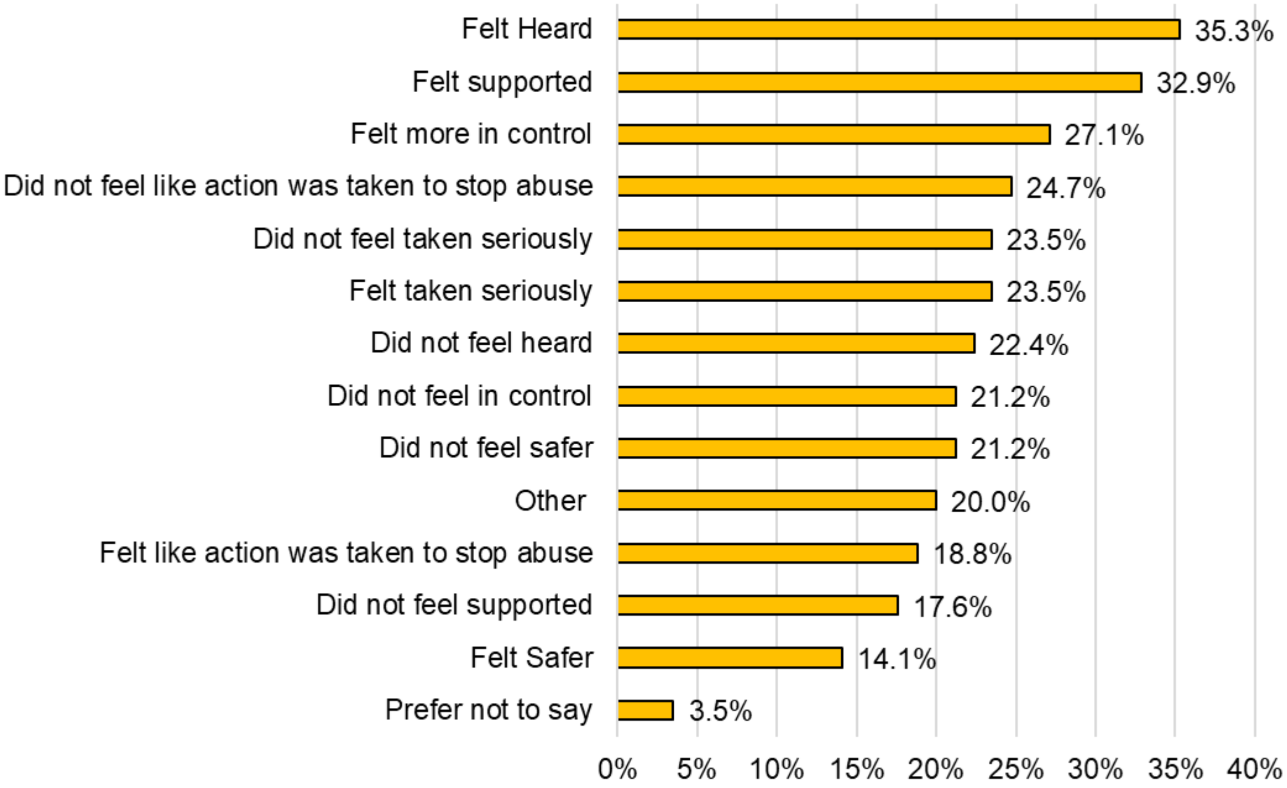


Figure 10: How participant felt as a result of reporting

As these responses indicate, the impacts of reporting were complex and varied, with both negative and positive impacts frequently being cited.

These responses were largely reflected in the findings from professional survey respondents, although the indicated outcomes from reporting were more mixed/positive among victim-survivor respondents than professionals, potentially pointing to differences between ‘community’ and ‘clinical’ populations[4]. Professional respondents indicated that service users were most likely to report to the police (73.3%), followed by social media admins or mods (56.7%). In terms of how respondents indicated how service users felt after reporting to social media admins or mods, the outlook was more negative than positive, with users feeling no action was taken to stop the abuse (36%), and not feeling supported (32%), in contrast to 8% feeling action was taken, and 16% feeling supported. The least frequent response was ‘felt heard’ (4%). Similarly, when reporting to police, service users were likelier to feel not heard (40.7%) and feel as though action was not taken (40.7%), in contrast to 14.8% feeling heard and 25.9% feeling as though action was taken. The least frequent response was ‘felt more in control’ (7.4%).

Another notable finding from the professional survey is the fact that, while a plurality of professional respondents had been in their roles from between 2 to 10+ years (66% overall), only 48% reported receiving training on online abuse.

[4] I.e., it is plausible that victim-survivors who seek support from professionals regarding their experiences are more likely to be experiencing ongoing, complex or challenging-to-resolve forms or patterns of abuse than those who disclose only to informal support sources and admins/moderators.

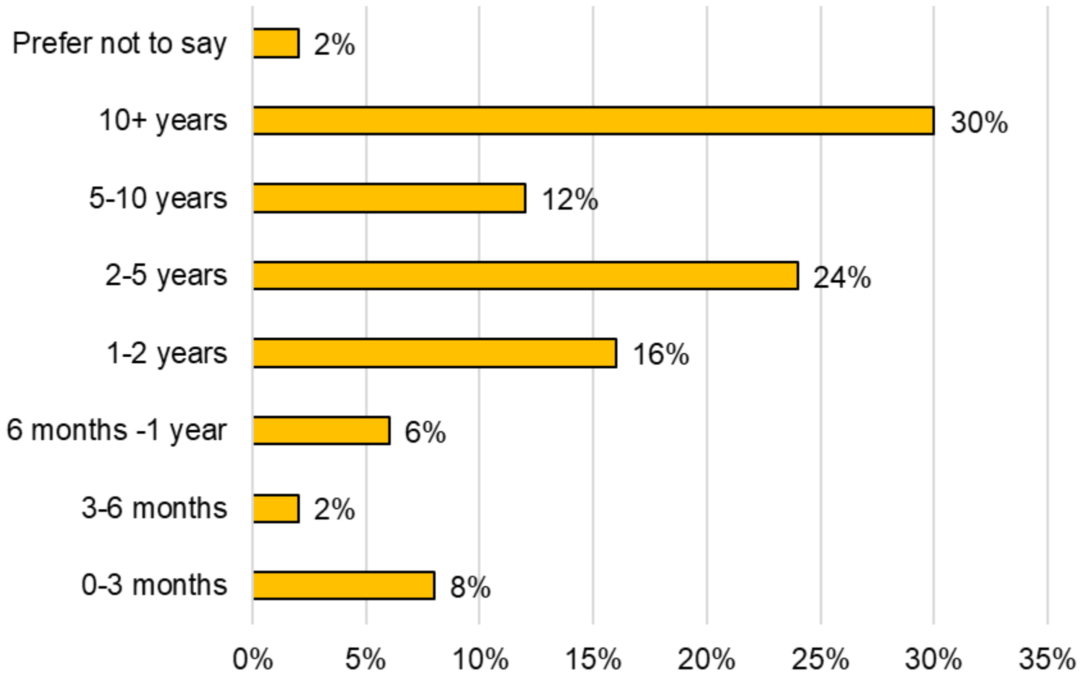


Figure 11: Professional survey respondents time in role

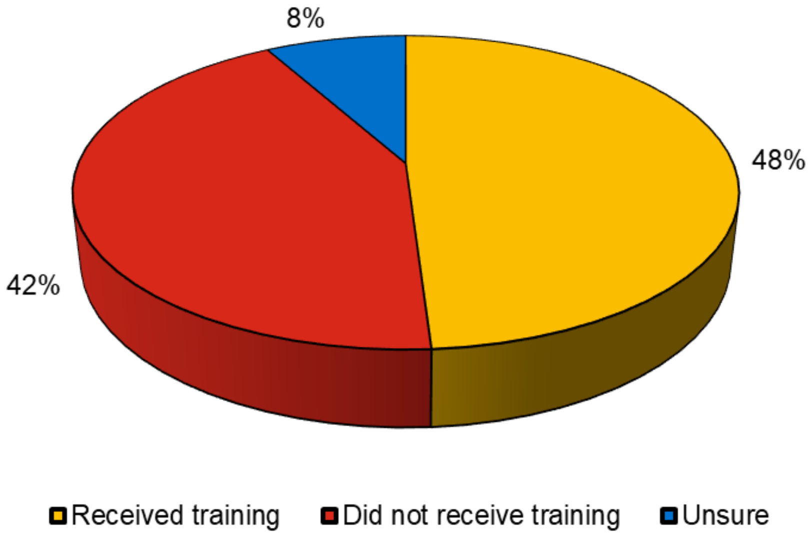
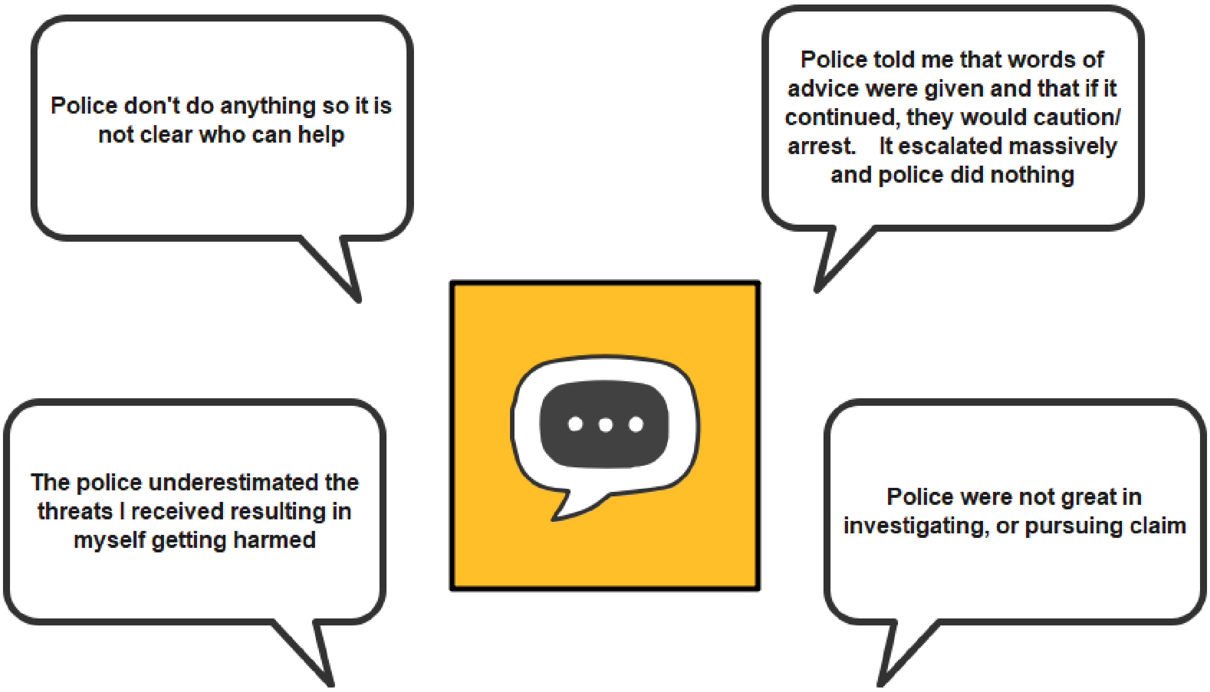


Figure 12: Professional survey respondents training

Professionals indicated a variety of barriers to giving support or to safeguarding. For example, 63% pointed to issues with legislation on online abuse, and, in accord with the findings regarding levels of training received, 63% indicated lack of relevant knowledge or training. 55% of respondents noted issues with social media company policies and processing.

Qualitative survey responses

The open-ended survey responses from victim-survivors provide additional context for these findings. The most frequently occurring descriptive code (*n* = 20) identified when analysing qualitative survey data related to negative or ineffective police responses:



Police don't do anything so it is not clear who can help

Police told me that words of advice were given and that if it continued, they would caution/ arrest. It escalated massively and police did nothing

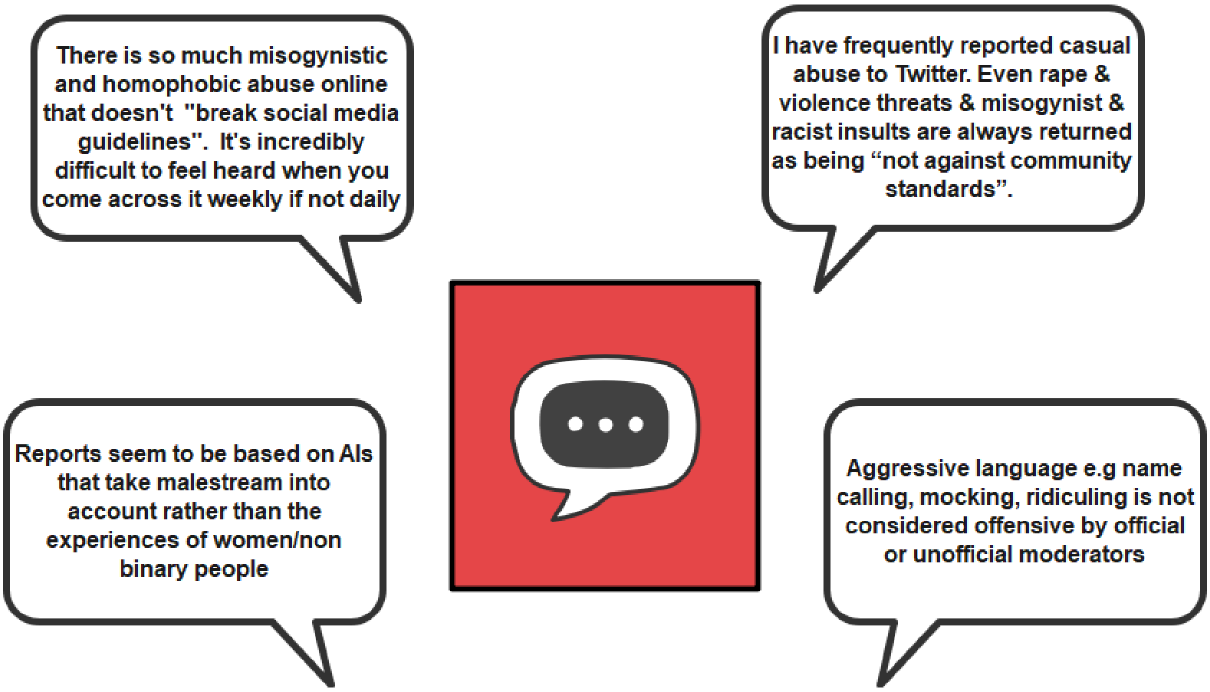
The police underestimated the threats I received resulting in myself getting harmed

Police were not great in investigating, or pursuing claim

Figure 13: Selected excerpts from victim-survivor survey respondents 27, 75, 86 and 144

Four survey respondents specifically cited what they perceived as unduly permissive or gender-biased community standards on platforms such as Twitter, which mean that hostile, harmful and abusive behaviours and content can go unchecked. One respondent also raised the prospect of algorithmic bias, with AIs informed by "*malestream*" epistemic and experiential norms being insufficiently attuned to the needs and experiences of women and non-binary individuals.

These comments regarding community standards and bias find parallels in the wider research literature on the development of digital cultural norms and "gendered technological hegemony" (Salter, 2017: 251). Salter (2017) argues that the "design of many online and social media platforms reflects foundational 'geek' conceptualizations of the internet as a 'new frontier' to be invaded and colonized through force and bravado [and that] these governing ideals have encoded combative modes of communication and laissez faire approaches to platform governance" (Salter, 2017: 251).



There is so much misogynistic and homophobic abuse online that doesn't "break social media guidelines". It's incredibly difficult to feel heard when you come across it weekly if not daily

I have frequently reported casual abuse to Twitter. Even rape & violence threats & misogynist & racist insults are always returned as being “not against community standards”.

Reports seem to be based on AIs that take malestream into account rather than the experiences of women/non binary people

Aggressive language e.g name calling, mocking, ridiculing is not considered offensive by official or unofficial moderators

Figure 14: Selected excerpts from victim-survivor survey respondents 7, 17, 23 and 85

12 participants felt that online abuse and TFVAWG are rarely taken seriously by gatekeepers such as police, with several identifying this perception as a deterrent or barrier to reporting, or even recognising that abuse is not “*normal*”:

People tend to make jokes about "dick pics" but unsolicited sexual pictures/texts can be really triggering for people who have suffered abuse/trauma and this should be taken more seriously

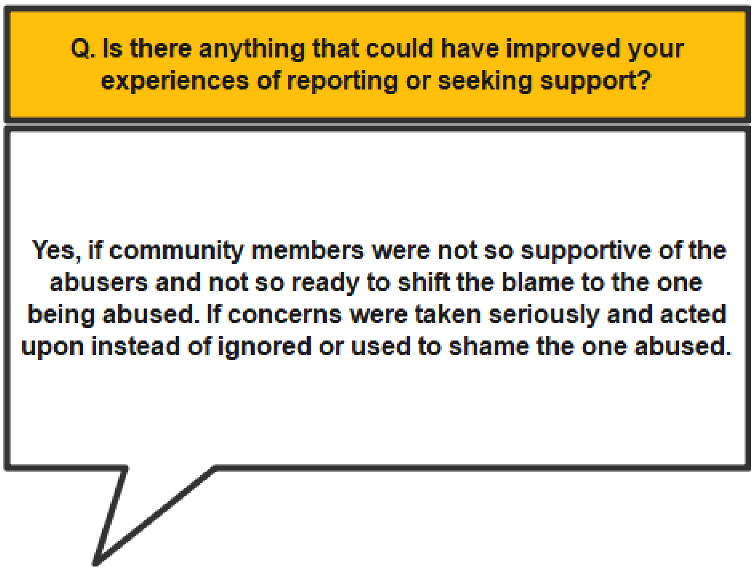
It is still not taken seriously enough by those who we report it to, from the police, to doctors and employers. There are massive gaps and this is why in some cases it leads to long term mental health issues and/or suicide

Women aren't taken seriously when receiving messages/pictures and unwanted harassment therefore it doesn't get reported as it's viewed as normal

My former partner gaslit me and made threats against himself online whenever we got into an argument - it doesn't feel like these sorts of things are taken seriously enough to report to anyone but friends and family

Figure 15: Selected excerpts from victim-survivor survey respondents 45, 57, 87, 115

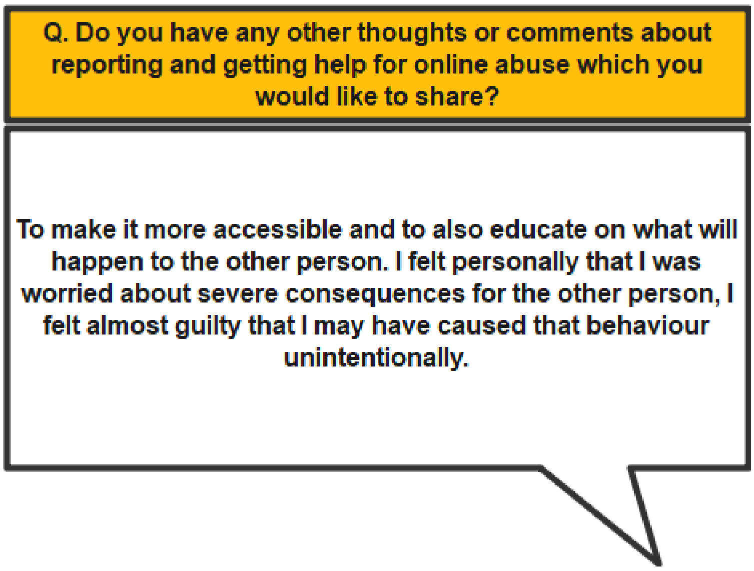
Four participants described self-blame, or a fear of being judged or blamed by others, as a potential barrier to reporting:



Q. Is there anything that could have improved your experiences of reporting or seeking support?

Yes, if community members were not so supportive of the abusers and not so ready to shift the blame to the one being abused. If concerns were taken seriously and acted upon instead of ignored or used to shame the one abused.

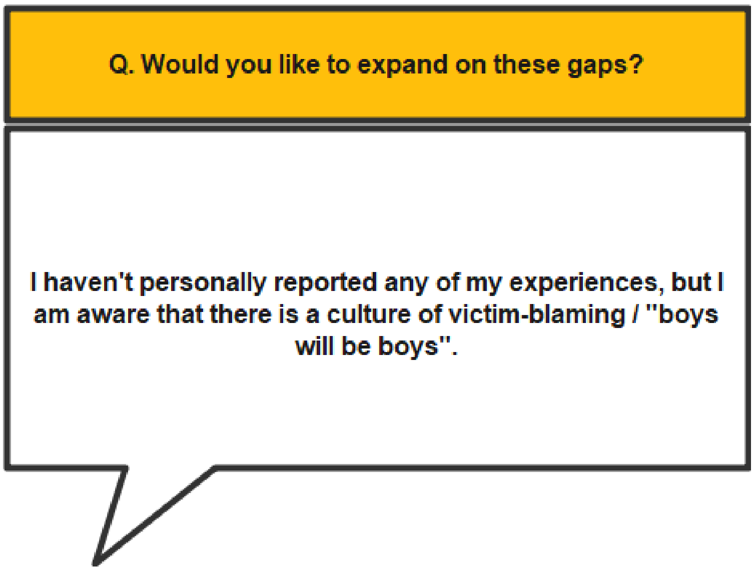
Figure 16: Victim-survivor survey respondent 19



Q. Do you have any other thoughts or comments about reporting and getting help for online abuse which you would like to share?

To make it more accessible and to also educate on what will happen to the other person. I felt personally that I was worried about severe consequences for the other person, I felt almost guilty that I may have caused that behaviour unintentionally.

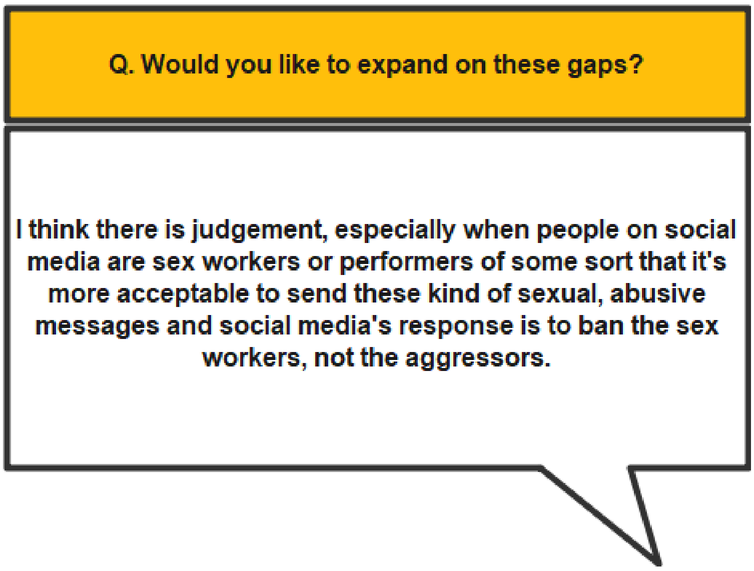
Figure 17: Victim-survivor survey respondent 37



Q. Would you like to expand on these gaps?

I haven't personally reported any of my experiences, but I am aware that there is a culture of victim-blaming / "boys will be boys".

Figure 18: Victim-survivor survey respondent 87



Q. Would you like to expand on these gaps?

I think there is judgement, especially when people on social media are sex workers or performers of some sort that it's more acceptable to send these kind of sexual, abusive messages and social media's response is to ban the sex workers, not the aggressors.

Figure 19: Victim-survivor survey respondent 94

Interviews

Interviews with victim-survivors and professionals suggest a number of significant barriers and challenges in relation to reporting and accessing support, including evidence collection and police responses. The most widely cited barriers across both sets of interviewees were police responses and the legal, jurisdictional and technological complexity of online abuse.

As with survey participants, ineffective or negative responses from police were identified as a major obstacle to pursuing a criminal justice response, with the perception that online abuse and TFVAWG were regarded as low (or lower) risk than other offences.

One interviewee who had experienced hacking and IBSA described feeling *“dismissed”* by police after being assessed as low risk:

> *So within 24 hours I went went to my local police first to make a report […]*
> *Police weren't super helpful in stopping these websites popping up with my photos. They kind of just assessed me as low risk and dismissed me. So I actually needed to find help through other organisations in the taking down of these websites. And that's when my friend recommended Revenge Porn Helpline, and they were incredible and a lot more efficient than police for sure. […]*
>
> *With the police, I filed a report online. They sent me a follow up email asking for more information, which I provided. And after that they wanted to talk to me on the phone for my welfare risk assessment, to see how I was. And it was a very short five minute conversation on the phone, basically asking me routine questions. At the end. She said, ‘You're low risk. There's nothing more we can do. It's out of our hands’ [laughs]. It was very- didn't feel as heartfelt as I wanted it to be or as supportive.*
>
> **‘Rosa’, victim-survivor**

Another interviewee recalled her experiences of reporting online abuse and harassment by a former partner, and the inconsistent responses she received from police across the country.

> *I'm like, ‘Look, I've got all these screenshots. All this, he's just harassing me’ and I go to the police. And the police here were quite good. It was actually a male police officer. he was like ‘It's clear-cut harassment, he's had multiple police warnings. You're still getting 60 calls and threats and him saying you're abusing him and all this crap’. But then it went to the police department where he was working, different part of the country. And there was a guy there was just like ‘Oh it's just child contact, though, isn't it? It's a family court issue. You'll have to go back to the family court". I really had to push for them to do anything about it. So it's just kind of hit and miss really.*
>
> **‘Elle’, victim-survivor**

The five professional interviewees shared the view that police do not regard online abuse is as a priority due to the perception that there is a lesser risk of physical violence.

This sense of being dismissed or deprioritised, perhaps owing to an underlying framework of assumptions about which forms or manifestations of abuse pose a significant threat, speaks to wider issues around how police theorise risk and triage cases when responding to GBV. While coercive control has been recognised as a criminal offence in England and Wales since 2015 (Crown Prosecution Service 2017), and stalking was introduced as a distinct criminal offence in 2012 (Crown Prosecution Service, 2018), there are still challenges in investigating and prosecuting crimes which are defined in terms of an ongoing course of conduct, and whose legal status hinges on linking together a pattern of behaviours which may, taken individually, appear minor or innocuous, resulting in high rates of attrition (Bird *et al*, 2021; Suzy Lamplugh Trust, 2021). Conceptions of risk and harm which centre physically violent acts may therefore miss the forest for the trees, eliding the significant cumulative harms to victim-survivors of offences such as IBSA or online stalking and harassment, and the potential for escalation/progression.

As discussed in section 2.1, given the impacts on mental health reported by survey participants, with almost one in four reporting feeling traumatised, and around one in five experiencing panic attacks, the survey findings suggest that those subjected to online abuse and TFVAWG can incur serious harms. While one cannot draw a straight line or equivalency between the various forms of abuse experienced by survey respondents and DA, the identified link between DA victims experiencing coercive control and subsequently completing suicide underlines that there are multiple forms of risk that need to be taken into account when responding to VAWG (Bates *et al*, 2021).

A third interviewee spoke about her experiences of reporting transnational online stalking:

I went to National Stalking Helpline first, as I said, just to get that information. They suggested to go to the police. So I did reach out to the police, but there was a big problem with it being overseas. They couldn't really do anything. They were just like, "Well, we can contact the [other country local] police and it's up to the [other country local] police if they want to take it further". So again, it just felt like, what was the point?

[…] I didn't make a formal report or anything in the end because I thought this is going to take so much time and I'm already overwhelmed by it.

'Maeve', victim-survivor

As Maeve's experience suggests, policing online abuse is associated with jurisdictional complexities because, while the internet extends perpetrators' reach and facilitates access to victims, there are currently a lack of legal mechanisms to tackle transnational victimisation (Salter, 2015).

Interviews with professionals pointed to similar issues regarding a lack of leverage with sites that are hosted outside the UK, and for whom acquiring a reputation for sharing non-consensual content is a selling point rather than a liability:

I think it's all about their reputation isn't it, they have to be seen as doing something […] The big pornography websites, they want to be known as caring about this stuff and doing something right. And they want to keep those really positive relationships going. I think the smaller websites are just- they're hosted overseas, so they just don't really care. Like we, you know, as part of our reporting, we say we're funded by the Home Office and this breaks the law, they can just ignore that, there's no repercussions for them by ignoring that. Also, a lot of these kind of horrible sites, a lot of them are actually built with the purpose of sharing intimate content without consent, these collector sites.

'Penny', professional

Other prominent challenges and obstacles identified by interviewees included the burden of collating evidence needed to document online abuse/TFVAWG (*n* = 5), which often fell on victim-survivors themselves, and the long and unpredictable afterlife of non-consensual content (*n* = 4). Hostile design – whereby sites hosting IBSA images employ 'hidden', inordinately time-consuming or degrading reporting routes for individuals seeking to get content removed – was also described by two professional interviewees specialising in online abuse and TFVAWG.

Following synthesis of the scoping review, survey and interview findings, researchers developed a simplified – and linearised[5] - visualisation of the barriers and challenges that may being encountered during victim-survivor journeys from the point of first experiencing (or discovering) the abuse to the aftermath of reporting, pictured on the next page.

[5] In actuality, given the issues associated with course of conduct offences such as cyberstalking and technology facilitated coercive control, and/or the re-sharing of IBSA content and harassment by perpetrators involved in 'collector culture', many victim-survivors' trajectories will be far less linear, and may incorporate multiple 'visits' to different stages and barriers. Additionally, disclosures to informal support sources and the wider community may take place on a 'need to know' basis (e.g., if perpetrators send IBSA to friends, work colleagues or family members as part of the pattern of offending)

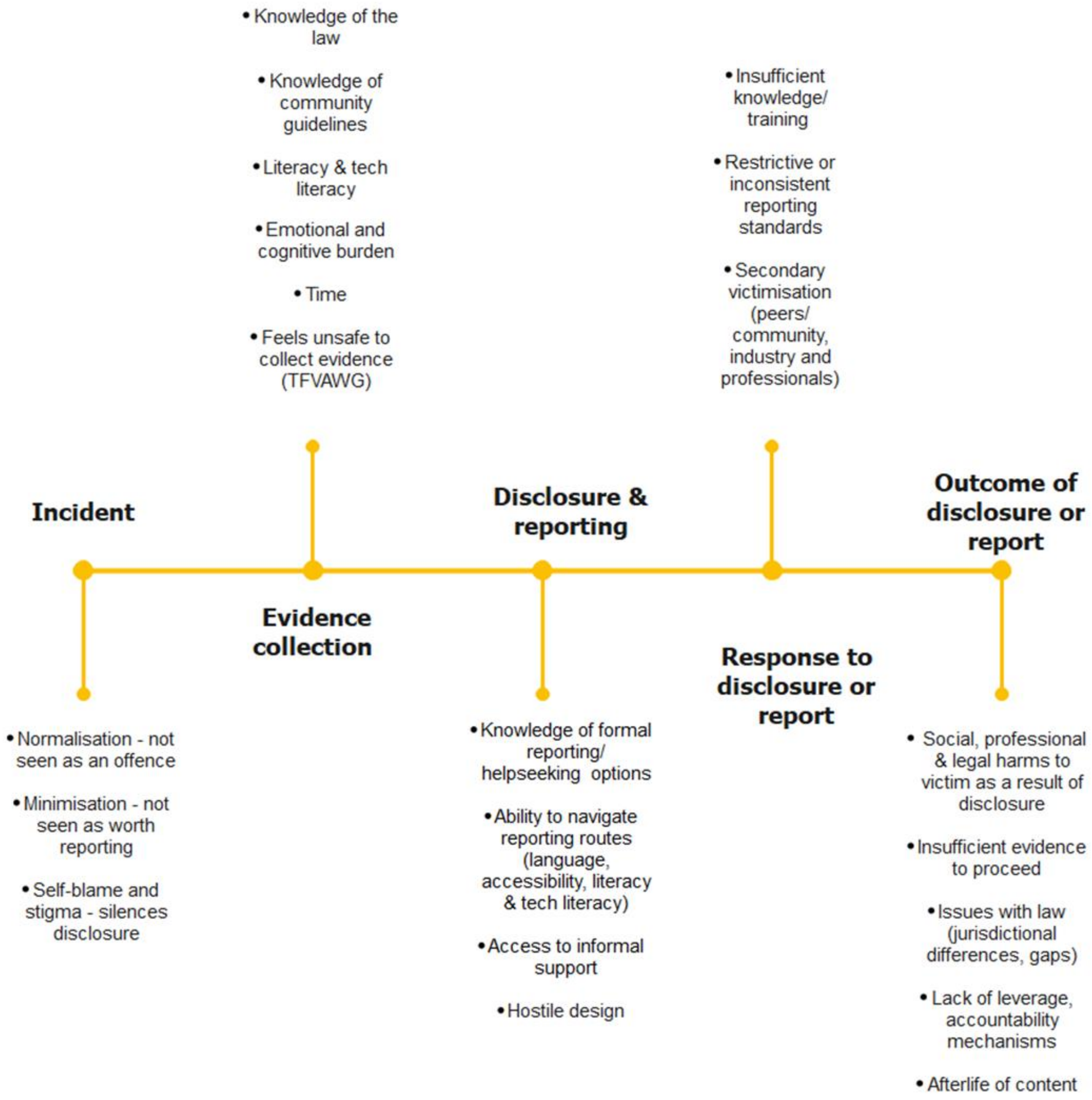

Incident
• Normalisation - not seen as an offence
• Minimisation - not seen as worth reporting
• Self-blame and stigma - silences disclosure
Evidence collection
• Knowledge of the law
• Knowledge of community guidelines
• Literacy & tech literacy
• Emotional and cognitive burden
• Time
• Feels unsafe to collect evidence (TFVAWG)
Disclosure & reporting
• Knowledge of formal reporting/ helpseeking options
• Ability to navigate reporting routes (language, accessibility, literacy & tech literacy)
• Access to informal support
• Hostile design
Response to disclosure or report
• Insufficient knowledge/ training
• Restrictive or inconsistent reporting standards
• Secondary victimisation (peers/ community, industry and professionals)
Outcome of disclosure or report
• Social, professional & legal harms to victim as a result of disclosure
• Insufficient evidence to proceed
• Issues with law (jurisdictional differences, gaps)
• Lack of leverage, accountability mechanisms
• Afterlife of content

Figure 20: Visualisation of victim-survivor journey

2.3 *“I think it might just help the user feel heard, that their experience does matter, that something's been done”*[6]: Co-Designing Minerva

Co-design interviews conducted with five victim-survivors flagged several key recommendations for developing and evaluating the Minerva tool.

There were broad areas of consensus among interviewees regarding what they would like to see from the Minerva tool, including:

- Signposting to emotional support and longer-term “*aftercare*” (‘Rosa’), including via a safe peer support forum or victim-survivor “*success stories*” (‘Cara’)
- Optional status updates or check-in calls at the user’s own pace (‘Rosa’, ‘Maeve’, ‘Cara’)
- Proactive digital security/pattern detection features, including flagging concerning patterns of behaviour (‘Elle’), gathering/linking evidence on serial perpetrators (‘Jemma’), or running a health check to alert the user to potential security issues (‘Rosa’)
- Customisable support with evidence collection, such as the ability to create, populate and annotate timelines or export bundles for court (‘Elle’)

[6] ‘Maeve’, victim-survivor and co-design interviewee

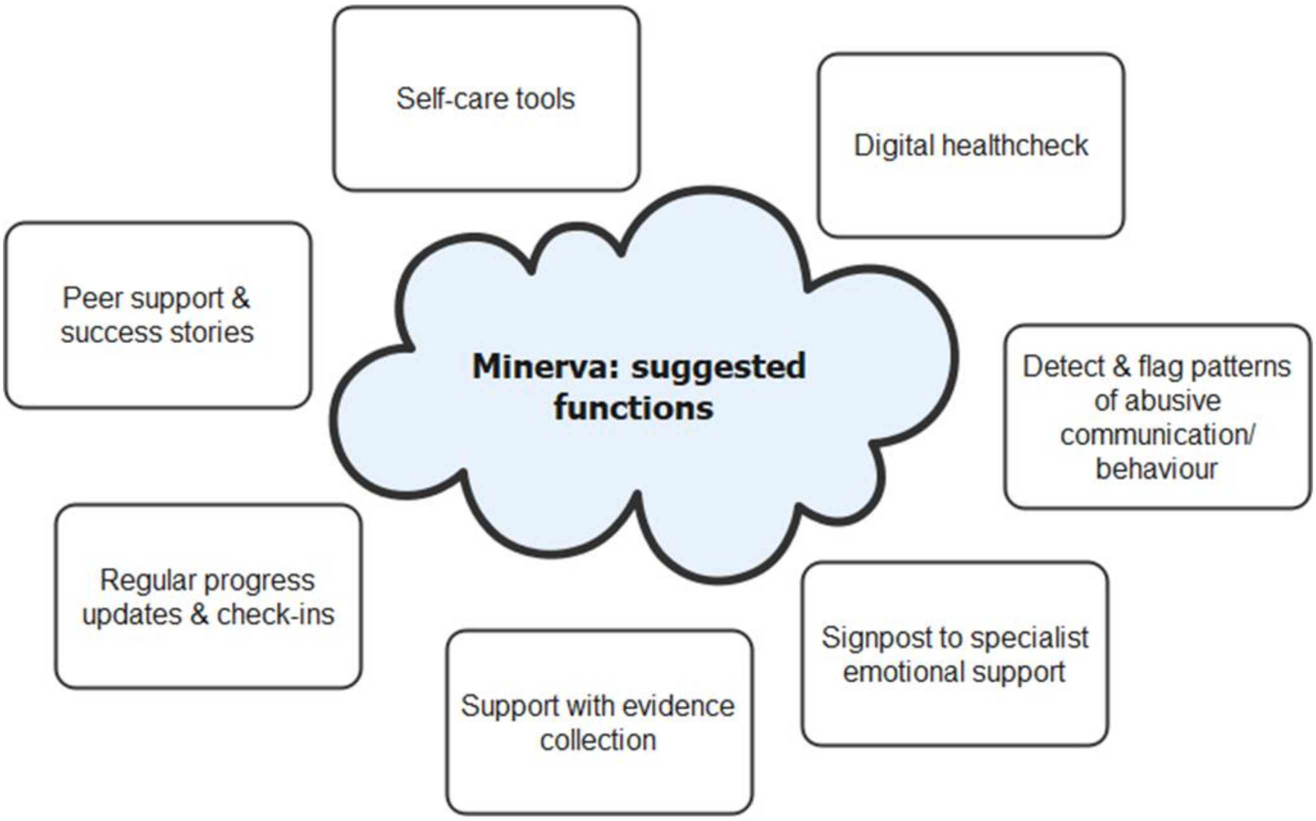

Self-care tools
Digital healthcheck
Peer support & success stories
Minerva: suggested functions
Detect & flag patterns of abusive communication/ behaviour
Regular progress updates & check-ins
Signpost to specialist emotional support
Support with evidence collection

Figure 21: Co-design interview findings: suggested functions for Minerva

Thematic analysis of the co-design interviews, in synthesis with the research findings as a whole, identified that these priorities for the tool were grounded in six core principles that promote victim-survivor agency and wellbeing:

Transparency and informed consent

> *Loss of control, you know, is one of the biggest problems. And so providing transparency and choice throughout is gonna be I would hope hugely beneficial to people who really feel like they don't have any agency right now*
> **'Jemma', co-design interviewee**

As discussed in section 2.1, loss of control is a defining characteristic of victimisation and, in some cases, of reporting and pursuing justice. Empowering users to understand and accept or reject the terms of engagement when using the Minerva tool through functions such as accessible informed consent procedures and a clear explanation about the security and limitations of the tool was identified as an important countermeasure to this loss, restoring a sense of autonomy.

Continuity

> *I think if you kind of put the ball in the user's court at the time of using that chat system, how often they want to hear from you, and then you can run a report weekly and be like, 'Oh I need to speak to this person this week and this person next week or whatever'. I think it might just help the user feel again, heard, that their experience does matter, that something's been done.*
> **'Maeve', co-design interviewee**

The survey and interview findings suggest that many victim-survivors who take the step of reporting online abuse or TFVAWG face ineffective and uninformed responses, contributing to a sense of feeling dismissed, devalued and unheard. Interviewees' suggestion of user-led status updates and follow-up contacts to ensure that victim-survivors are kept informed – or as informed as they want to be – about the progress of their report, suggests that Minerva can fulfil a 'holding' function, providing a sense of continuity and enabling users to feel heard.

Customisability and flexibility

> *I think that off the bat I'd want to be given the option of 'Do you want a person or do you want to deal with this, you know, through this sort of AI system?' Kind of having like that branch off immediately at the beginning would, I think, give people, the benefit of having that sort of privacy and anonymity if they want it […] Because I think it is a very individual thing when you're going through this and some people want to like sort of separate themselves from it as much as possible, and other people really want a human response. So yeah, I think if I was given that choice at the at the outset that I would really enjoy having that*
> '**Jemma', co-design interviewee**

Suggested points for adaptation included: different contact methods and frequency for receiving status updates, whether or not to share incoming information (e.g., about content being found), a welfare questionnaire to assess how the user would like to be supported on their journey.

Customisability is essential for meeting the needs of a range of users who may have very different lived experiences, patterns of victimisation, preferences and accessibility needs. Equally importantly, it affords users a sense of control, enabling them to engage on their own terms.

Security and respect for privacy

The Internet isn't always 100% safe, so there can be, just as how people can collect your private personal images they can just hack straight into this database. So I think there's always problems with hacking and safety, safe storage of information, because you wouldn't want any of that to come out and you want full confidentiality and anonymity of this information and just make it completely bullet-proof I guess

'Rosa', co-design interviewee

It's got to be super secure, if the abuser gets hold of your tool and sees what you're doing, there's a ton of compromised information that's then centralised in one place that potentially like exposes you

'Elle', co-design interviewee

Victim-survivors of online abuse and TFVAWG have first-hand knowledge of the internet's sharp edges. Unsurprisingly, concerns around protecting users' security and privacy and ensuring that users understood and trusted the security measures in place, were a recurrent discussion point, occurring across all five interviews.

Given victim-survivors' apprehension about data security and the potential for discovery and malicious use by perpetrators, having a clearly worded explanation of how data is secured and stored, as well as any potential risks associated with using the tool and how to mitigate these (i.e., if sharing a home or devices with a perpetrator) is a crucial recommendation.

Accessibility and Trauma-Informed Design

If someone is emotionally distraught in those moments, they're not gonna be able to understand jargon and, you know, they don't want to read loads of paragraphs

'Jemma', co-design interviewee

I think obviously these women that are looking for this help are going to be quite distressed in one way or another. So I feel like the pages need to be quite easy to navigate, you know, especially if the lady's in a state of panic and they're really needing some help

'Cara', co-design interviewee

Accessibility and trauma-informed design were identified as meta- or supra-values organising the other principles.

Trauma-informed practice (TIP) refers to a way of working that is designed to empower victim-survivors, and minimise the possibility of retraumatisation, by promoting transparency, agency, collaboration, victim-survivor voice and peer support, and eschewing unclear, coercive, directive or deceptive modes of engagement (Harris & Fallot, 2001). TIP originally emerged in healthcare contexts, in response to a growing evidence base on the prevalence and legacies of psychological trauma (Harris & Fallot, 2001), and works to promotes equal access to services for those with lived experiences of victimisation, marginalisation and adversity. TIP is intersectional by design, and attentive to cross-cutting forms of oppression and disadvantage, disavowing "one size fits all" model in favour of a needs-led approach (Kulkarni, 2019: 4).

Trauma-informed computing is an emerging concept which applies these core values to user experience of digital technologies, recognising that "experiences with trauma can both stem from technology and impact how one experiences technology" (Chen *et al*, 2022: 6).

Trauma-informed computing represents a formalised commitment to:

Improving the design, development, deployment, and support of digital technologies by explicitly acknowledging trauma and its impact, recognizing that digital technologies can both cause and exacerbate trauma, and actively seeking out ways to avoid technology-related trauma and retraumatisation

(Chen *at al.*, 2022:7)

In concrete terms, this involves considerations such as "ensuring technology artifacts, processes and organizations operate transparently, predictably, and reliably while providing users with the ability to make mistakes and corrections" (Chen *et al,* 8)

Interviewee observations that feed into this idea of trauma-informed and intersectional design included their suggestion to avoid the use of "*jargon*" or walls of text as part of the informed consent process and feedback that users may experience difficulties in reading or retaining information due to being in crisis mode, or due to language, literacy or tech literacy support needs.

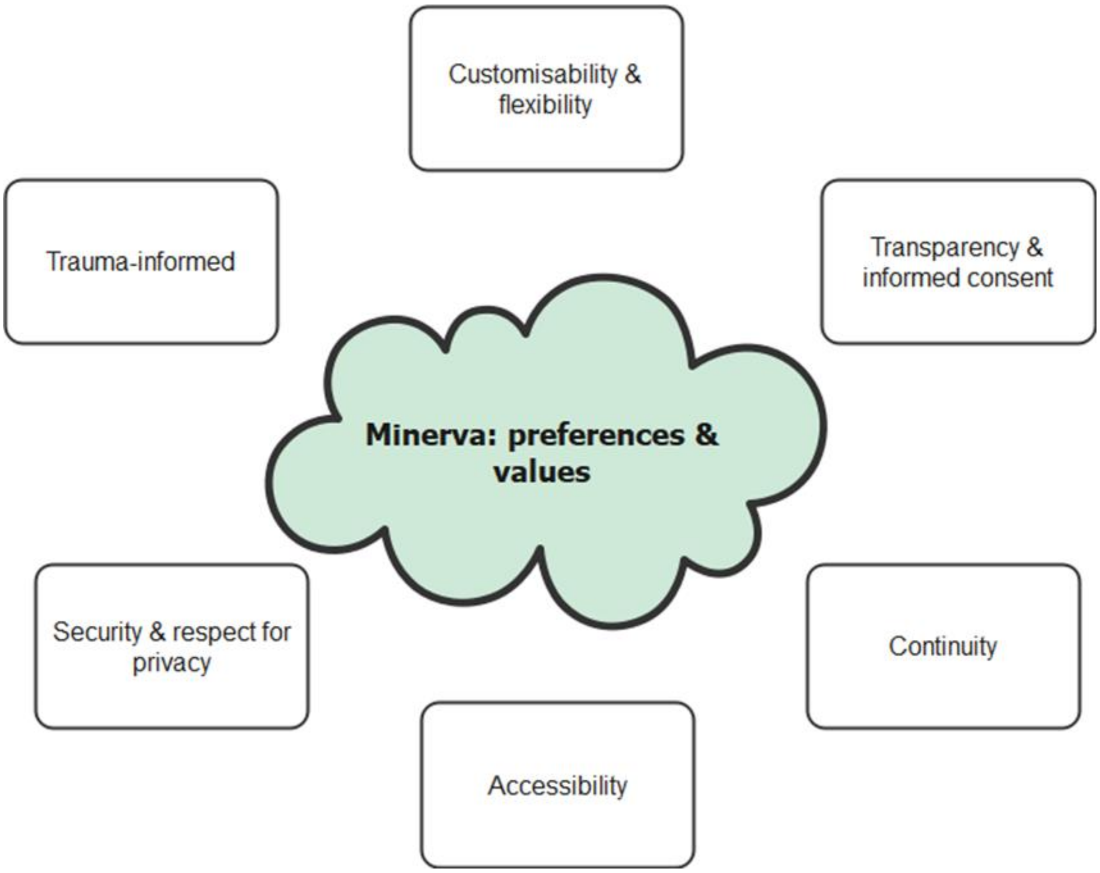

Customisability & flexibility
Trauma-informed
Transparency & informed consent
Minerva: preferences & values
Security & respect for privacy
Continuity
Accessibility

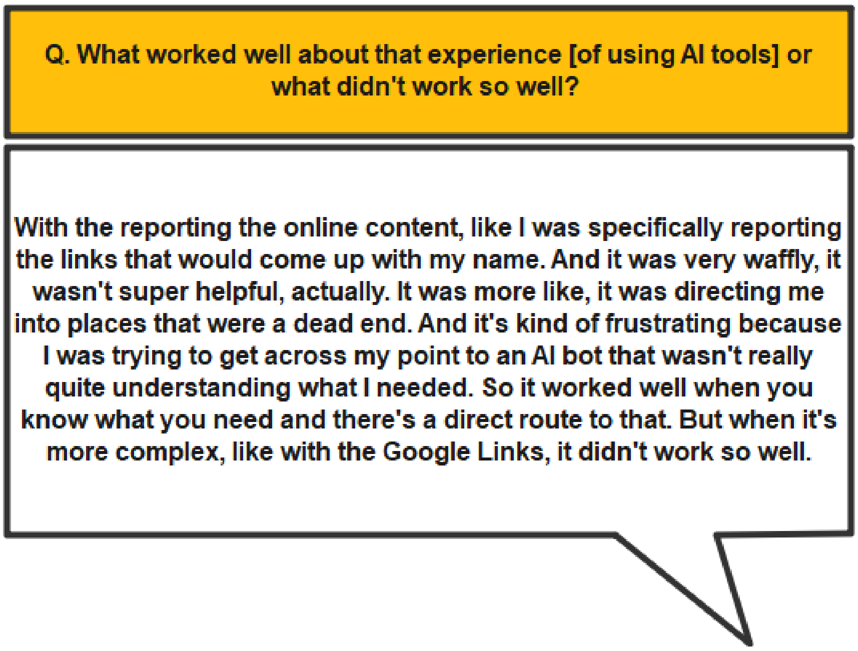
Figure 22: Co-design interview findings: Preferences and values

Co-design interviews also suggested perceived risks and barriers involved with the use of the Minerva tool.

2. 4 Other identified risks and challenges

Communicating complexity

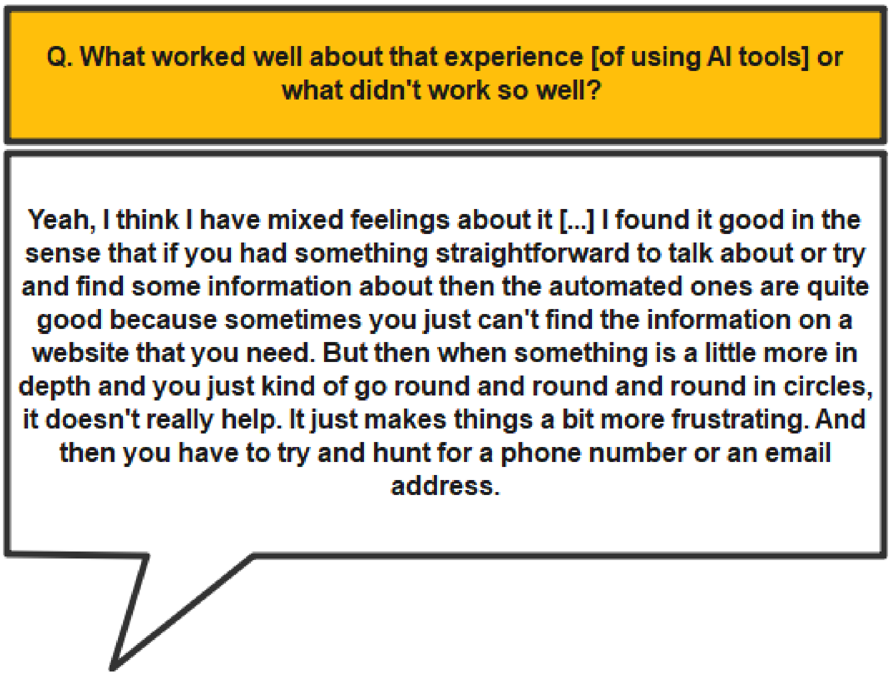
The primary findings suggest a degree of scepticism regarding AI-mediated communication about complex and sensitive topics, with four of the five interviewees expressing concerns about the risk of users in crisis feeling “*frustrated*” by conversational loops or dead ends, or of chatbots missing conversational red flags which indicate users may be at risk of offline harm. However, interviewees also felt that there were clear benefits to having access to a tailored AI tool, including its potential to offer streamlined, anonymous guidance with simpler enquiries and ‘FAQs’, and its ability to provide an immediate response out of hours.



Q. What worked well about that experience [of using AI tools] or what didn't work so well?

With the reporting the online content, like I was specifically reporting the links that would come up with my name. And it was very waffly, it wasn't super helpful, actually. It was more like, it was directing me into places that were a dead end. And it's kind of frustrating because I was trying to get across my point to an AI bot that wasn't really quite understanding what I needed. So it worked well when you know what you need and there's a direct route to that. But when it's more complex, like with the Google Links, it didn't work so well.

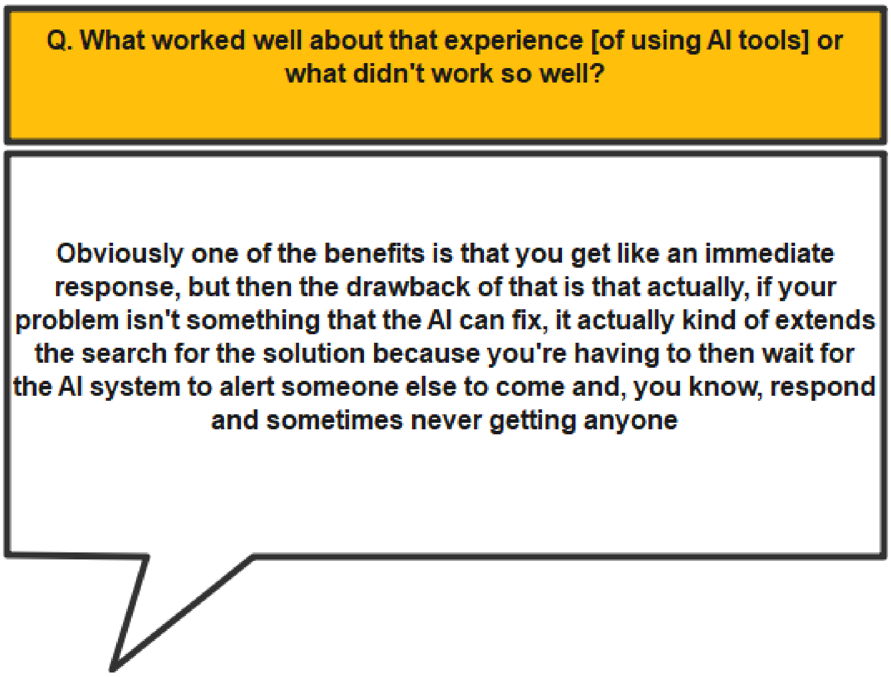
Figure 23: ‘Rosa’, co-design interviewee



Q. What worked well about that experience [of using AI tools] or what didn't work so well?

Yeah, I think I have mixed feelings about it [...] I found it good in the sense that if you had something straightforward to talk about or try and find some information about then the automated ones are quite good because sometimes you just can't find the information on a website that you need. But then when something is a little more in depth and you just kind of go round and round and round in circles, it doesn't really help. It just makes things a bit more frustrating. And then you have to try and hunt for a phone number or an email address.

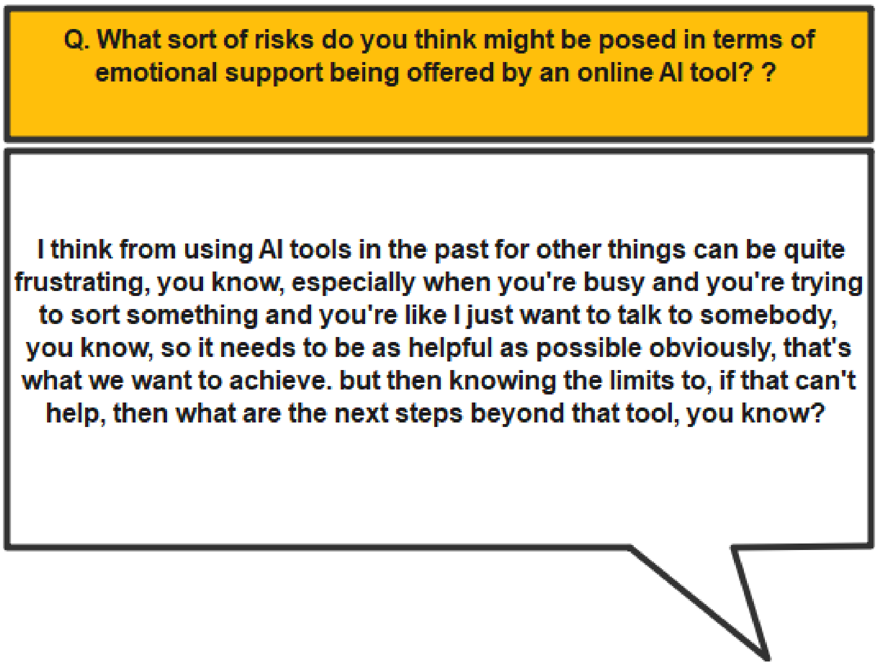
Figure 24: 'Maeve', co-design interviewee



Q. What worked well about that experience [of using AI tools] or what didn't work so well?

Obviously one of the benefits is that you get like an immediate response, but then the drawback of that is that actually, if your problem isn't something that the AI can fix, it actually kind of extends the search for the solution because you're having to then wait for the AI system to alert someone else to come and, you know, respond and sometimes never getting anyone

Figure 25: 'Jemma', co-design interviewee

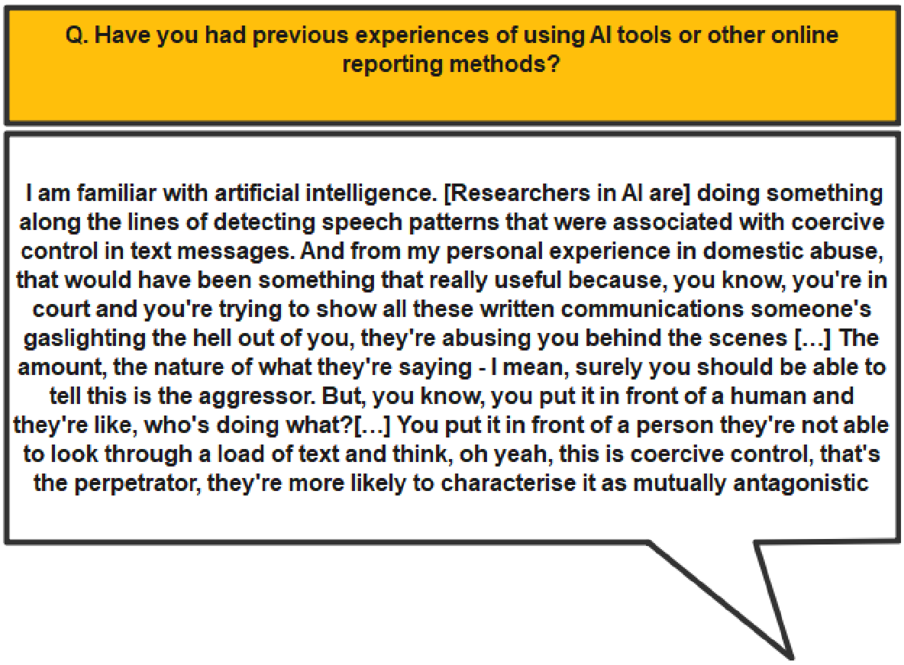


Q. What sort of risks do you think might be posed in terms of emotional support being offered by an online AI tool? ?

I think from using AI tools in the past for other things can be quite frustrating, you know, especially when you're busy and you're trying to sort something and you're like I just want to talk to somebody, you know, so it needs to be as helpful as possible obviously, that's what we want to achieve. but then knowing the limits to, if that can't help, then what are the next steps beyond that tool, you know?

Figure 26: 'Cara', co-design interviewee

Based on her professional knowledge and lived experience, one co-design interviewee felt that an AI tool might be less susceptible to the social and cognitive biases that can prevent human interlocutors from determining the directionality of abuse:



Q. Have you had previous experiences of using AI tools or other online reporting methods?

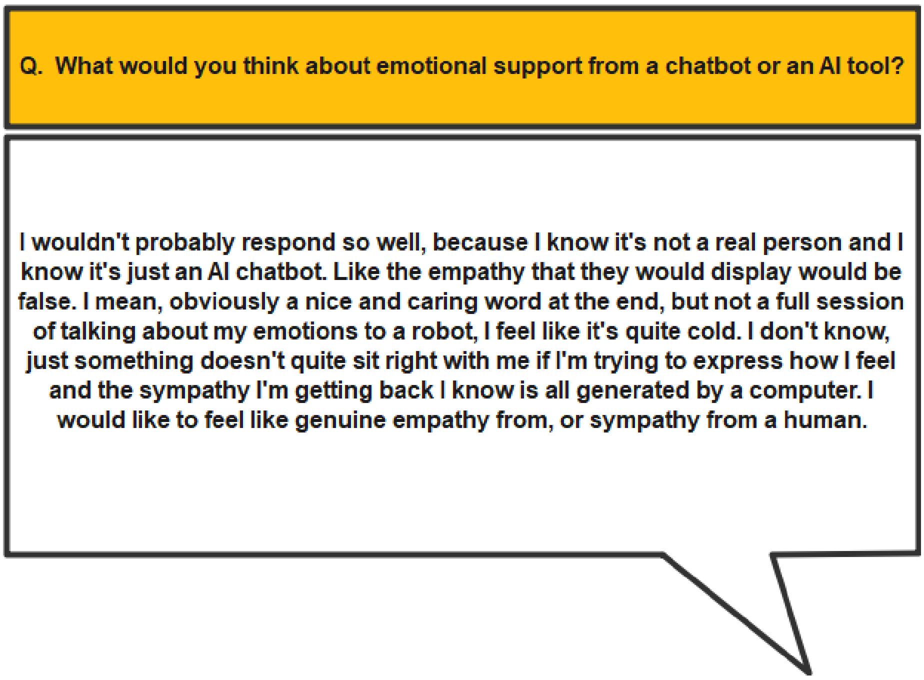
I am familiar with artificial intelligence. [Researchers in AI are] doing something along the lines of detecting speech patterns that were associated with coercive control in text messages. And from my personal experience in domestic abuse, that would have been something that really useful because, you know, you're in court and you're trying to show all these written communications someone's gaslighting the hell out of you, they're abusing you behind the scenes [...] The amount, the nature of what they're saying - I mean, surely you should be able to tell this is the aggressor. But, you know, you put it in front of a human and they're like, who's doing what?[...] You put it in front of a person they're not able to look through a load of text and think, oh yeah, this is coercive control, that's the perpetrator, they're more likely to characterise it as mutually antagonistic

Figure 27: 'Elle', co-design interviewee

These dual perspectives on AI tools' capacity to navigate the complex and emotionally charged terrain of online abuse/TFVAWG extended to the discussion of Minerva's potential emotional support functions.

‘Faking’ empathy

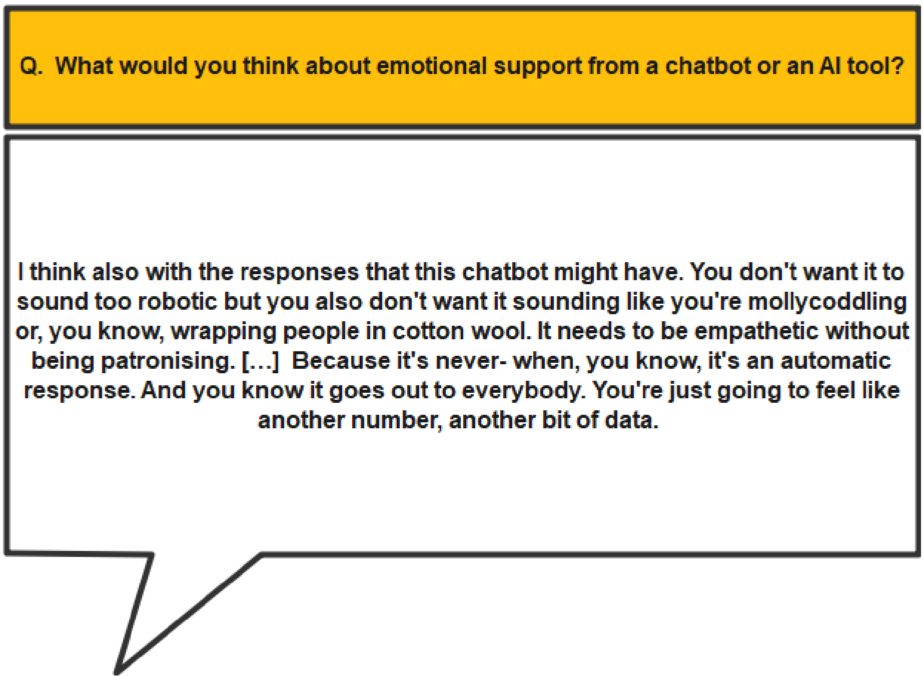
Three interviewees expressed doubts regarding the efficacy or acceptability of an AI tool providing emotional support as part of its functions, suggesting that it may come off as false, “*cold*” (Rosa) or “*robotic”* (Maeve) to users.



Q. What would you think about emotional support from a chatbot or an AI tool?

I wouldn't probably respond so well, because I know it's not a real person and I know it's just an AI chatbot. Like the empathy that they would display would be false. I mean, obviously a nice and caring word at the end, but not a full session of talking about my emotions to a robot, I feel like it's quite cold. I don't know, just something doesn't quite sit right with me if I'm trying to express how I feel and the sympathy I'm getting back I know is all generated by a computer. I would like to feel like genuine empathy from, or sympathy from a human.

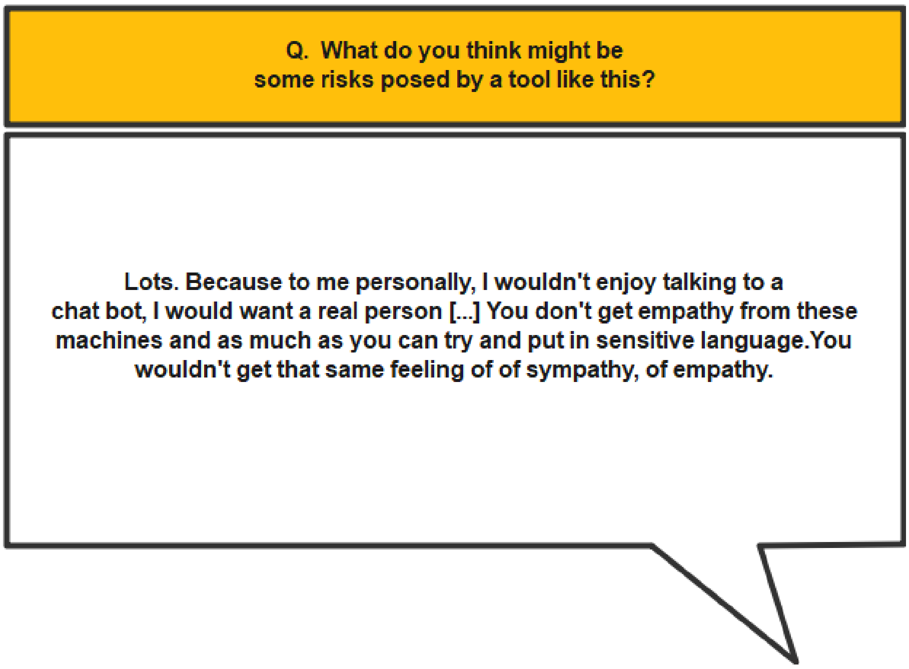
Figure 28: ‘Rosa’, co-design interviewee



Q. What would you think about emotional support from a chatbot or an AI tool?

I think also with the responses that this chatbot might have. You don't want it to sound too robotic but you also don't want it sounding like you're mollycoddling or, you know, wrapping people in cotton wool. It needs to be empathetic without being patronising. […] Because it's never- when, you know, it's an automatic response. And you know it goes out to everybody. You're just going to feel like another number, another bit of data.

Figure 29: 'Maeve', co-design interviewee



Q. What do you think might be some risks posed by a tool like this?

Lots. Because to me personally, I wouldn't enjoy talking to a chat bot, I would want a real person [...] You don't get empathy from these machines and as much as you can try and put in sensitive language.You wouldn't get that same feeling of of sympathy, of empathy.

Figure 30: 'Jemma', co-design interviewee

Given that this *a priori* feedback is based on the views of a small number of participants who may differ from other victim-survivors in significant respects, these responses cannot necessarily be taken as indicative of wider opinions, nor even as representative of the views of participants themselves should they have the opportunity to interact with the Minerva tool. However, they do reflect historic concerns, and ongoing debates, regarding the application of conversational AI (CA) and Human-Computer Interaction (HCI). For example, digital interventions employing CA have been heralded as a cost-effective way of promoting wider access to non-judgemental mental health support but have also met with ethical critique as some theorists argue that “despite being perceived as less-stigmatising, CAs might actually pose harm to users due to their limited capacity to re-create human interaction and to provide tailored treatment” (Ruane et al, 2019: 5).

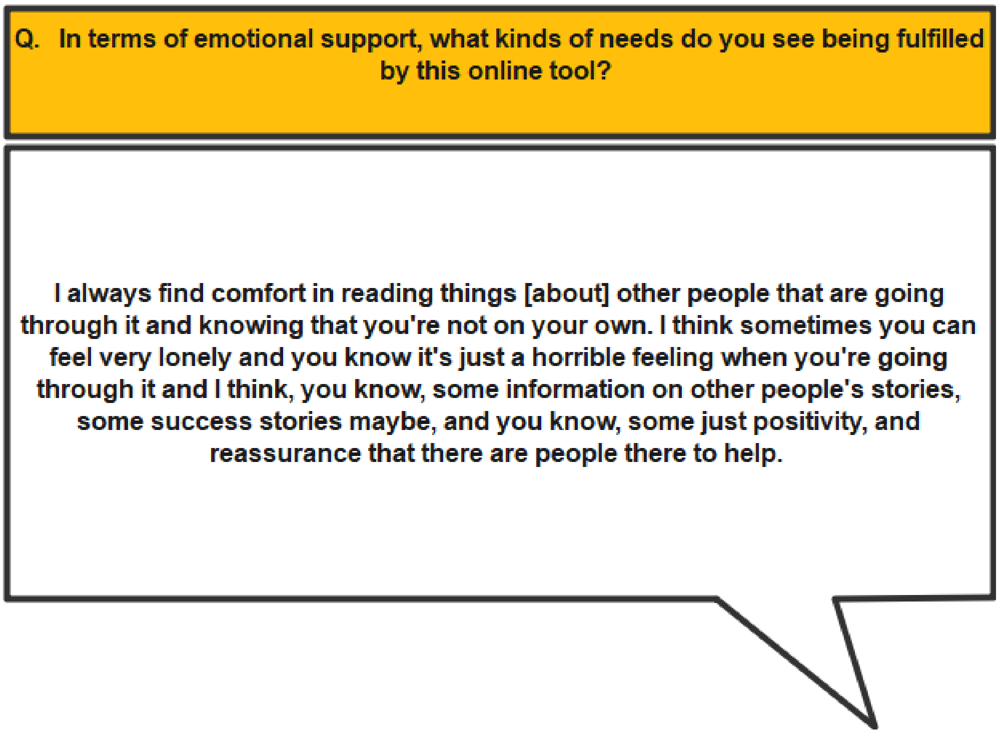
The negative perception among some participants regarding AI tools which are seen to be performing a sophisticated – yet “*cold*” or emotionless – simulation of empathy also finds parallels with research on unintended negative user responses to chatbots due to the “uncanny valley” effect (Mori *et al,* 2012). When conversing with chatbots, consumer research suggests that many users report experiencing a sense of “creepiness” or unease, theorised to be linked to the ambiguity or not-quite human-ness of the interaction (Rajaobelina *et al,* 2021: 2339). Transparency and careful expectation management regarding “an agent’s status as automatic (non-human) and the limits of its capabilities” have been identified as important steps in enabling users to make an informed choice about engaging, and to feel comfortable doing so (Ruane et al, 2019: 6). This recommendation also chimes with interviewees’ own emphasis on transparency and respect for user agency.

Peer support and self-care

While several participants expressed reservations about directly engaging on an emotional level with a chatbot, interviewees were receptive to the broader notion of accessing emotional or peer support facilitated by the Minerva tool.

Two interviewees suggested that Minerva could include the creation of a safe peer support forum for those with lived experience of online abuse/TFVAWG, offering a “*closed community*” to get *“support from other people who are going through similar things*” (Elle) and “*see how they may have dealt with the situation*” (‘Maeve’).

A third interviewee also cited the power of shared experience, and its role in helping those subject to abuse not to feel adrift and alone:



Q. In terms of emotional support, what kinds of needs do you see being fulfilled by this online tool?

I always find comfort in reading things [about] other people that are going through it and knowing that you're not on your own. I think sometimes you can feel very lonely and you know it's just a horrible feeling when you're going through it and I think, you know, some information on other people's stories, some success stories maybe, and you know, some just positivity, and reassurance that there are people there to help.

Figure 31: 'Cara', co-design interview

4. Conclusions and recommendations

Key findings

Based on scoping review findings, analysis of the survey and interview data and the wider literature, the key findings regarding our research questions were as follows:

RQ1. What forms of online abuse and TFVAWG are UK women experiencing, and what are the harms associated with these experiences?

Our findings and the wider evidence base indicate that online abuse and TFVAWG are prevalent among women in the UK and globally, and form part of a wider continuum of GBV, exerting a chilling effect on women's ability to access public spaces and exercise their freedom of expression and association, as well as adversely impacting their wellbeing.

- Survey respondents reported being subject to a range of forms of abuse, most commonly experiencing unwanted sexual messages (61%), cyberstalking or harassment (44%) or receiving unwanted violent or pornographic content (36.4%). Interviewees reported experiencing IBSA and technology facilitated stalking and coercive control.
- Survey respondents and interviewees experienced a range of harms associated with the abuse, including emotional, psychological, social, professional and behavioural impacts. Almost one in four survey participants

(23.5%) felt traumatised as a result of the abuse while around one in five (19.3%) reported experiencing panic attacks.

- Interviewees' experiences also point to the loss of control as a defining aspect of some patterns of victimisation; for example, ongoing IBSA or technology facilitated coercive control.

RQ2. What do women experiencing online abuse expect/want/need from an AI tool, and what outcomes are they looking for?

Analysis of co-design interview data suggests several main points of agreement between participants' stated desires, expectations and outcomes for an AI tool:

- Aftercare and emotional support, including access to safe/moderated peer communities and curated survivor narratives
- A sense of continuity and being 'held' throughout the reporting/help seeking journey, with user-defined opportunities for regular updates and check-ins from specialist professionals
- Proactive support with identifying threats, linking patterns of harmful behaviour/serial perpetration and staying safe online
- Bespoke evidence collection tools, including functions to create and annotate timelines and export evidence for criminal justice proceedings

Researchers identified a set of core values underlying this 'ask' for the tool, including trustworthiness and transparency, agency, continuity of care, choice, digital and psychological safety, and accessibility.

RQ3. What remedies are currently available to UK women experiencing online abuse, and what are the gaps and vulnerabilities within existing systems?

Scoping review, survey and interview findings indicate a dearth of specialist and longer-term emotional support options for UK women experiencing online abuse or TFVAWG. The findings also point to systemic issues with industry, criminal justice and legislative responses to online harms, as well as knowledge and training deficits within police, statutory and voluntary sectors (in part due to technology's ability to outpace legal remedies and organisational change). Widescale cultural issues such as victim-blaming and stigmatisation of women judged to have engaged in 'risky' behaviours such as consensually producing or sharing intimate images, and a minimisation of non-physical and online forms of abuse, were also commonly cited in the qualitative responses, including in relation to reactions from family, friends or community members.

Interviews with professionals pointed to promising practices such as reciprocal training between VAWG and online safety specialist services, suggesting opportunities for future collaborations, particularly in light of the inconsistent levels of specialist training across the voluntary and statutory sectors. For example, regular 'refresher' training to support VAWG services in navigating the evolving landscape of online harms, joint training programmes on the intersections between online abuse and other forms of VAWG, to deliver to police and statutory services.

- Survey findings are consistent with the wider evidence base on VAWG which indicates that a significant proportion of victim-survivors will choose to disclose informally or not at all. Survey respondents who disclosed their experiences of online abuse or TFVAWG most commonly disclosed to informal support sources such as family and friends (44.9%) or reported to social media moderators or admins (28.8%). Around one in four reported to police (25.4%). Additionally, almost a quarter of respondents (24.6%) indicated that they had never reported what had happened to anybody. This suggests that, as with other forms of GBV, there is a significant 'dark figure' of undetected perpetration and victim-survivors who have never received support regarding their experiences.
- Survey findings suggest that available platform/website reporting avenues were felt to offer unsatisfactory responses to reports, with references to unduly permissive or biased community guidelines which meant that harmful and abusive behaviours were tolerated, and hostile or indifferent design (e.g., 'hidden' reporting routes when trying to remove non-consensual intimate images).
- Negative or inconsistent police responses formed the single most common open-ended response category among both victim-survivor and professional survey participants, with respondents citing lack of knowledge, inaction and/or minimisation.
- Interviewee responses suggested a postcode lottery when it came to reporting, with good responses from some police forces and inadequate ones from others, underlining the need for regular evidence-informed training for frontline professionals whose work brings them into contact with women reporting online abuse/TFVAWG.

RQ4. What are the potential risks associated with developing and implementing such a tool and how can these be pinpointed, tracked and managed?

During the course of the research stage, four primary areas of risk were identified:

- **Security:** The safety and 'weaponisability' of the Minerva tool developed as a prominent theme, with concerns about the tool acting as a centralised repository of evidence, which could be intercepted, compromised or misused. Co-design interviewees expressed that the tool would need to be "*bulletproof*" ('Rosa') in order to instil a sense of trust in users.
- **Communicating complexity and detecting risk:** Co-design interviewees voiced concerns about the potential for users to get caught in unhelpful 'loops' if they initially select the wrong option, or for AI to fail to pick up on conversational red flags that indicate the user may be in danger.
- **Uncanny empathy:** Three co-design interviewees expressed reservations regarding the role of an AI in providing emotional support, framing it as cold

and inauthentic, and potentially off-putting to some users in need of assistance.

- **Accessibility:** Accessibility, including in relation to trauma-informed design, also emerged as a significant area of concern.

Recommendations

Based on synthesis of findings from the research stage, we identified the following recommendations for developing and refining the Minerva tool:

Collaborative and needs-led

Choice and customisability emerged as important values throughout the scoping review, the wider evidence base and experiential evidence from victim-survivors. An embedded recognition of, and respect for, the agency and individual preferences of the user is a key aspect of trauma-informed design (Chen *et al*, 2022).

This means fostering a collaborative rather than an "autocratic" user experience, with meaningful opportunities for user input, feedback and customisation throughout the process (Chen *et al,* 2022).

In practical terms, this could involve presenting clear choices about settings and how these can be modified (e.g., regarding notifications or updates), offering ongoing options to provide open-ended or survey feedback for tool improvement, and providing a range of options for collecting and exporting evidence.

Transparency

To promote informed consent and trust, honesty about the tool and its capabilities is crucial.

In interviews with online safety professionals, transparency and expectation management emerged as important aspects of their work, and of maintaining trust with service users.

To translate these principles to the Minerva tool, this could involve clearly explaining the limitations of the tool, detailing in lay-friendly language how user information will be collected, used, stored and shared, and stating that use of Minerva will not be able to guarantee a particular outcome. This is particularly important given the evidence base on wider systemic failings, which suggests that even 'perfect' evidence collection does not mean a case will progress.

Explaining the limits of user confidentiality regarding safeguarding, and the steps that would be taken if these limits were breached, is also ethically important.

Peer and emotional support

Based on interviewee feedback and the wider evidence base, researchers would recommend exploring safe opportunities for peer support via a range of options e.g., co-developed written, video or self-care resources to which users can be signposted, access to an asynchronous message board or strictly moderated online forum.

Based on preliminary feedback from interviewees, researchers would also recommend the Minerva tool offers a level of customisability regarding ‘emotional talk’ i.e., users can choose whether they want the chatbot to engage them on an emotional level and can revise this decision easily at any point by accessing settings.

Accessibility

Co-design interviewee feedback suggested concerns about accessibility were paramount, including the need for the tool to be readily useable by those who are in a state of emotional distress/trauma, and the importance of inclusivity regarding disability, sensory differences and language or literacy support needs.

Recommendations include clear customisability options regarding font, size and background colour, easy read options and language settings and ‘revisability’, where users can easily revisit settings and rectify errors in entering data.

5. Preliminary evaluation framework

Aim	Outcome	Indicator	Data source	When/how/who	How to use
To support users to report abuse	Users understand how to collect legally valid evidence	Reported level of understanding	Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g. target Likert rating 3.5, % of reports result in investigation or % content removal)
		Reported level of satisfaction with guidance	Survey & open-ended feedback		
		Outcome once submitted	Case timeline	Ongoing documenting/logging outcomes by users	
		Professional feedback	Survey & open-ended feedback	Ongoing feedback by professionals	
			Aggregated Minerva data	Quarterly exports by SWGfL	Quarterly monitoring to assess trends
			Aggregated comparator/baseline data (RPH, RHC)		
	Users know how and where to report	Reported level of knowledge	Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks

Aim	Outcome	Indicator	Data source	When/how/who	How to use
	different forms of abuse	Outcome of report Site/professional feedback	Case timeline Survey & open-ended feedback Aggregated Minerva data Aggregated comparator/baseline data (RPH, RHC)	Ongoing documenting/logging outcomes by users Ongoing feedback by professionals Quarterly exports by SWGfL	(e.g. target Likert rating 3.5, % of reports result in investigation or % content removal)
To promote user wellbeing	User feels emotionally supported	Reported sense of feeling emotionally supported Reported level of satisfaction with emotional support offered	Survey & open-ended feedback Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5)
	User feels heard/connected	Reported sense of feeling heard/connected Reported level of satisfaction with peer support options	Survey & open-ended feedback Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5)

Aim	Outcome	Indicator	Data source	When/how/who	How to use
		(or resources) offered			
	User feels more in control	Reported sense of control Reported level of satisfaction with user experience and options when using Minerva Reported level of satisfaction with process of reports/referrals from using Minerva	Survey & open-ended feedback Survey & open-ended feedback Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5)
	User feels safe	Reported sense of safety Reported level of satisfaction with Minerva tool in increasing safety	Survey & open-ended feedback Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5)

Aim	Outcome	Indicator	Data source	When/how/who	How to use
	User feels enabled/included	Reported accessibility of tool Reported level of satisfaction with accessibility & inclusion of the Minerva tool Satisfaction by demographic Outcome by demographic	Survey & open-ended feedback Survey & open-ended feedback Aggregated Minerva data Aggregated comparator/baseline data (RPH, RHC)	Ongoing feedback by users Quarterly exports by SWGfL	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5, demographic representativeness of user base) Quarterly monitoring to assess trends
To safeguard users from on- and offline harms	User is alerted to pattern of high-risk behaviour	User feedback on accuracy Report made? Safeguarding referral needed? Report outcome Professional feedback on accuracy	Survey & open-ended feedback Minerva logs/data Case timeline	Ongoing feedback by users Routine Minerva data capture Ongoing documenting/logging outcomes by users	Export monthly for review/comparison against benchmarks (e.g., target Likert ratings 3.5, % of reports or safeguarding referrals projected based on SWGfL baseline)

Aim	Outcome	Indicator	Data source	When/how/who	How to use
		User feedback on effectiveness	Survey & open-ended feedback Survey & open-ended feedback Aggregated Minerva data		Quarterly monitoring to assess trends
	User is supported with digital health check	User feedback on accuracy User feedback on usefulness Reported level of satisfaction with guidance	Survey & open-ended feedback Survey & open-ended feedback	Ongoing feedback by users	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5)
	User is flagged as potentially being in crisis	User feedback on accuracy Report made? Safeguarding referral needed?	Survey & open-ended feedback Minerva logs/data Minerva logs/data	Ongoing feedback by users Routine Minerva data capture	Export monthly for review/comparison against benchmarks (e.g., target Likert rating 3.5, % of reports or safeguarding referrals projected

Aim	Outcome	Indicator	Data source	When/how/who	How to use
		Report outcome			based on SWGfL baseline)
		User feedback on helpfulness	Survey & open-ended feedback	Ongoing feedback by users	
		Professional feedback on accuracy	Survey & open-ended feedback		
			Aggregated Minerva data		
					Quarterly monitoring to assess trends

6. References

Alexy, E.M., Burgess A.W., Baker, T. & Smoyak S.A. (2005) Perceptions of cyberstalking among college students. Brief Treatment and Crisis Intervention 5(3): 279–289.

Bates, L., Hoeger, K., Stoneman, M. & Whitaker, A. (2021) Vulnerability Knowledge and Practice Programme (VKPP): Domestic Homicides and Suspected Victim Suicides During the Covid-19 Pandemic 2020-2021. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1013128/Domestic_homicides_and_suspected_victim_suicides_during_the_Covid-19_Pandemic_2020-2021.pdf (Accessed: 15 July 2022)

Branch, K., Hilinski-Rosick, C.M., Johnson, E. & Solano, G.bra (2017) Revenge porn victimization of college students in the United States: An exploratory analysis. International Journal for Cyber Criminology 11(1): 128–142

Brem, M. J., Florimbio, A. R., Grigorian, H., Wolford-Clevenger, C., Elmquist, J. A., Shorey, R. C., Rothman, E. F., Temple, J. R., & Stuart, G. L. (2019). Cyber abuse among men arrested for domestic violence: Cyber monitoring moderates the relationship between alcohol problems and intimate partner violence. Psychology of Violence, 9(4), 410–418. https://doi.org/10.1037/ vio0000130

Campbell, J. K., Poage, S. M. C., Godley, S., & Rothman, E. F. (2020). Social anxiety as a consequence of non-consensually disseminated sexually explicit media victimization. Journal of Interpersonal Violence, Online first. https://doi.org/10.1177/0886260520967150

Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., Levy, K., Dell, N., McCoy, D., & Ristenpart, T. (2018). The spyware used in intimate partner violence. IEEE Symposium on Security and Privacy. https://www.ipvtechresearch.org/pubs/spyware.pdf

Chen, J.X., McDonald, A., Zou, Y., Tseng, E., Roundy, K.A., Tamersoy, A., Schaub, F., Ristenpart, T. and Dell, N., (2022), April. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In CHI Conference on Human Factors in Computing Systems (pp. 1-20).

Crown Prosecution Service (2018) Stalking and Harassment - Legal Guidance, Domestic abuse, Cyber / online crime. Available at: https://www.cps.gov.uk/legal-guidance/stalking-and-harassment (Accessed: 15 July 2022)

Crown Prosecution Service (2017) Controlling or Coercive Behaviour in an Intimate or Family Relationship - Legal Guidance, Domestic abuse. Available at: https://www.cps.gov.uk/legal-guidance/controlling-or-coercive-behaviour-intimate-or-family-relationship (Accessed: 15 July 2022)

Douglas, H., Harris, B.A. and Dragiewicz, M. (2019). Technology-facilitated domestic and family violence: Women's experiences. The British Journal of Criminology, 59(3), pp.551-570.

Dreißing, H., Bailer, J., Anders, A., Wagner, H. & Gallas ,C. (2014) Cyberstalking in a large sample of social network users: Prevalence, characteristics, and impact upon victims. Cyberpsychology, Behavior, and Social

Drouin, M,. Ross, J. & Tobin, E. (2015) Sexting: A new digital vehicle for intimate partner aggression? Computers in Human Behavior 50: 197–204.

Gamez-Guadix, M., Almendros, C., Borrajo, E., & Calvete, E. (2015). Prevalence and association of sexting and online sexual victimization among Spanish adults. Sexuality Research and Social Policy: A Journal of the NSRC, 12(2), 145–154. https:// doi.org/10.1007/s13178-015-0186-9.

Glitch & End Violence Against Women (2020) The Ripple Effect: Covid-19 and the Epidemic of Online Abuse. Available at: https://www.endviolenceagainstwomen.org.uk/wp-content/uploads/Glitch-and-EVAW-The-Ripple-Effect-Online-abuse-during-COVID-19-Sept-2020.pdf (Accessed: 05 May 2022

Harris, M. & Fallot, R.D. (2001). Envisioning a trauma-informed service system: a vital paradigm shift*. New directions for mental health services,* 89, pp.3-22.

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People’s Advisory Group (2019). Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support. Biomedical informatics insights, 11, 1178222619829083. https://doi.org/10.1177/1178222619829083

Kulkarni, S. (2019) Intersectional Trauma-Informed Intimate Partner Violence (IPV) Services: Narrowing the Gap between IPV Service Delivery and Survivor Needs. J Fam Viol 34, pp. 55–64. https://doi.org/10.1007/s10896-018-0001-5

Leitao, R. (2019). Anticipating smart home security and privacy threats with survivors of intimate partner abuse. In Designing Interactive Systems (DIS) Conference (pp. 527–539). https://doi.org/10.1145/3322276.3322366

Lenhart, A., Ybarra, M. & Price-Feeney M (2016) Online Harassment, Digital Abuse and Cyberstalking in America. URL (accessed 3 July 2017): https://www.datasociety.net/pubs/ oh/Online_Harassment_2016.pdf.

Maple, C., Short, E. and Brown, A. (2011). Cyberstalking in the United Kingdom: An analysis of the ECHO pilot survey. University of Bedfordshire.

McGlynn, C., Rackley, E., Johnson, K., Henry, N., Flynn, A., Powell, A., Gavey, N. & Scott, A. (2019) 'Shattering lives and myths : a report on image-based sexual abuse.', Project Report. Durham University; University of Kent.

Mori, M., MacDorman, K.F. and Kageki, N., (2012) The uncanny valley [from the field]. IEEE Robotics & automation magazine, 19(2), pp.98-100. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6213238 (Accessed: 15 July 2022)

Rajaobelina, L., Prom Tep, S., Arcand, M. and Ricard, L., 2021. Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. Psychology & Marketing, 38(12), pp.2339-2356.

Ruane, E., Birhane, A. & Ventresque, A. (2019) Conversational AI: Social and Ethical Considerations. In AICS (pp. 104-115). Available at: http://ceur-ws.org/Vol-2563/aics_12.pdf (Accessed: 15 July 2022)

Salter, M., 2018. From geek masculinity to Gamergate: the technological rationality of online abuse. Crime, Media, Culture, 14(2), pp.247-264.

Sylaska, K.M. and Edwards, K.M., 2014. Disclosure of intimate partner violence to informal social support network members: A review of the literature. Trauma, violence, & abuse, 15(1), pp.3-21.

Suzy Lamplugh Trust (2021) Police, Crime, Sentencing & Courts Bill: Written Evidence Submitted By Suzy Lamplugh Trust. Available at:

https://bills.parliament.uk/publications/41827/documents/377 (Accessed: 15 July 2022)

Tanczer, L.M., Lopez-Neira, I., Parkin, S., Patel, T., Danezis, G., (2018) Gender and IoT (G-IoT) Research Report: The rise of the Internet of Things and implications for technology-facilitated abuse. UCL. https://www.ucl.ac.uk/steapp/sites/steapp/files/giot-report.pdf

Tanczer, L.M., Patel, T., Parkin, S., Lopez-Neira, I., Slupska, J., (2019) Written Submission to the Online Harms White Paper Consultation. UCL. https://www.ucl.ac.uk/steapp/sites/steapp/files/online_harms_white_paper_consultation_resp onse_giot_june_2019_final.pdf

Vera-Gray, F. & Kelly, L. (2020) “Contested gendered space: public

sexual harassment and women’s safety work,” International Journal of Comparative and Applied Criminal Justice, 2020 https://www.tandfonline.

com/doi/full/10.1080/01924036.2020.1732435