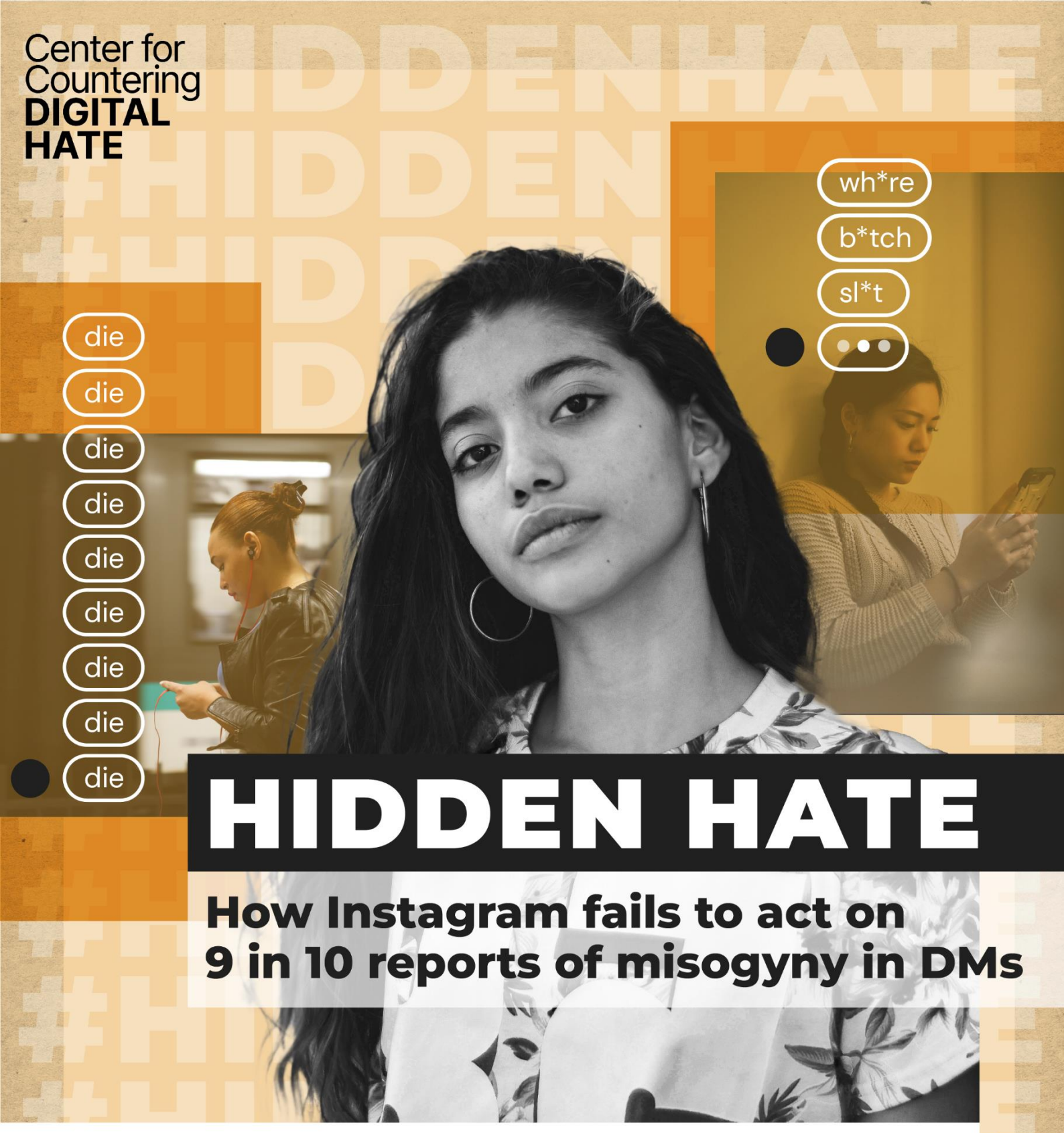
Center for Countering DIGITAL HATE



HIDDEN HATE

How Instagram fails to act on 9 in 10 reports of misogyny in DMs

With participation by
Sharan Dhaliwal, Bryony Gordon,
Amber Heard, Jamie Klingler + Rachel Riley

Center for Countering DIGITAL HATE

Contents

Published 6 April 2022

Center for Countering DIGITAL HATE

The Center for Countering Digital Hate is a US-headquartered international non-profit NGO that disrupts the architecture of online hate and misinformation.

Digital technology has changed forever the way we communicate, build relationships, share knowledge, set social standards, and negotiate and assert our societies' values.

Digital spaces have been colonized and their unique dynamics exploited by malignant actors that instrumentalize hate and misinformation. These movements are opportunistic, agile, and confident in exerting influence and persuading people.

Over time these malignant actors, advocating diverse causes - from hatred of women to racial and religious intolerance to science-denial - have formed a digital Counter-Enlightenment. The disinformation they spread to bolster their causes has socialized the offline world for the worse.

The Center's work combines both analysis and active disruption of these networks. CCDH's solutions seek to increase the economic, political, and social costs of all parts of the infrastructure - the actors, systems, and culture - that support and profit from hate and misinformation.

If you appreciate this report, you can donate to CCDH at www.counterhate.com/donate

In the United States, Center for Countering Digital Hate Inc is a 501(c)(3) charity.

In the United Kingdom, Center for Countering Digital Hate Ltd is a non-profit company limited by guarantee.

1 Introduction

For two decades, Silicon Valley has promulgated a self-serving fantasy that social media companies were founded to function as an extension of the public square. In this ideal, the online world allows communication and information to cross geographic and social divides and enable a diversity of voices to interact in ways that were impossible before the technology existed. But the fantasy is far from the reality.

Unlike the public square, the rules of the game on social media are rigged. The result is that rather than enhancing democracy, the reverse is true. Misogyny, racism, identity-based hate, and a whole host of inequalities persist. It polarizes society and strips communities of the ability to freely exercise their voices and take their seats at the table. This is not how a healthy democracy that respects human rights, and the dignity of all people should function. Reform is needed.

In this report, CCDH conducted several case studies in partnership with women with large Instagram followings to reveal how Meta, in its continued negligence and disregard for the people using its platforms whilst churning record profits, has created an environment where abuse and harmful content is allowed to thrive. This denies those being abused the ability to freely express themselves online.

After reporting on public gender-based violence and misogynistic abuse through posts directed at high-profile women, CCDH researchers have turned to an under-studied and even more unregulated facet of online abuse: the direct message (DM). This report uncovers the side of Instagram that is often unseen, but more often experienced first-hand by women who use social media: how harassment, violent threats, image-based sexual abuse can be sent by strangers, at any time and in large volumes, directly into your DMs without consent and platforms do nothing to stop it.

Instagram claim that they act on hate speech, including misogyny, homophobia, and racism; nudity or sexual activity; graphic violence; and threats of violence. But our research finds that Instagram systematically fails to enforce appropriate sanctions and remove those who break its rules.

CCDH has conducted a series of Failure to Act reports over several years – on platforms' failure to act on Covid-19 misinformation, identity-based hate, climate denial, and more. This report has one of the worst-ever failure rates of our reports.

Instagram failed to act on 90% of abuse sent via DM to the women in this study.

Further, Instagram failed to act on 9 in 10 violent threats over DM reported using its tools and failed to act on any image-based sexual abuse within 48 hours.

Women and civil society groups have long campaigned for platforms to take misogyny seriously and include gender-based violence as a violation of community standards. CCDH joined Ultraviolet, the Women's Disinformation Defense Project, and a coalition of dozens of groups in calling on platforms to do better and to stop failing women, BIPOC, and LGBTQ+ people by spreading hate and misinformation.[1] We know that women of color and the LGBTQ+ community disproportionately experience the worst of abuse online, simply by existing in digital spaces.[2] For many, the cost of being online is a torrent of abuse, slurs, threats of violence to their physical safety. The intended effect of the

abuse and the trauma of its constant barrage is simple: to drive women off platforms, out of public life, and to further marginalize their voices.

Online misogyny, made easy by platforms' functionality, has offline impacts in normalizing gender-based violence and harassment. In the absence of effective tools to stop the stream of harmful content, women have been forced to find their own solutions, often tailoring their content to avoid provoking abusers or avoiding posting altogether to reduce their visibility. Platforms' purported safety measures are both ineffective and shift the burden of preventing misogynistic and online gender-based violence to those who suffer the abuse. Instagram, and other mainstream platforms, have permitted the creation of a culture of intimidation, narrowing the parameters of people's freedom of speech and expression, thereby creating spaces that are safer for abusers than users.

Five women provided CCDH researchers with access to their direct messages on Instagram through a thorough data-download and direct access to their accounts. We sincerely thank those who gave their time, their trust, and their accounts to us for this research.

Social media is systemically and categorically failing women, just as they fail marginalized groups across the board. Meta and Instagram have the power to make changes to their platform functions and processes that allow misogynists free rein. Instagram has the moral imperative and the resources to fix the problems this report identifies. There is also a looming legal obligation. With the United Kingdom set to ban cyberflashing - the sending of explicit images without a recipient's consent - Instagram is at risk of millions of dollars' worth of fines and potential criminal prosecution if it continues to fail to act on image-based sexual abuse. Platforms must enforce their hate speech standards, strengthen reporting tools without putting the onus on those who are targeted, and fulfill the basic duty to put the safety of women and marginalized communities before profit. If they don't, legislators must act swiftly and decisively to put people before the profits of a few greedy, lazy tech billionaires.

Imran Ahmed
CEO and Founder
Center for Countering Digital Hate

2 Executive Summary

- To investigate abuse sent over direct message (DM) on Instagram, CCDH worked with five women with a total of 4.8 million followers on the platform:
    - Amber Heard, actress and UN Human Rights Champion
    - Rachel Riley, broadcaster and CCDH Ambassador
    - Jamie Klingler, co-founder of Reclaim These Streets
    - Bryony Gordon, award-winning journalist, and mental health campaigner
    - Sharan Dhaliwal, founder of South Asian culture magazine Burnt Roti

- Instagram's policies prohibit hate speech, including misogyny, homophobia and racism, nudity or sexual activity, graphic violence, threats of violence.

- New analysis of 8,717 DMs sent to participants shows that 1 in 15 DMs break Instagram's rules on abuse and harassment.
    - Researchers recorded 125 examples of image-based sexual abuse (IBSA), which the UK government recently announced would become illegal.
    - 1 in 7 voice notes sent to women were abusive, and Instagram allows strangers to place voice calls to women they don't know.
    - This analysis is based on Instagram 'data downloads' sent by Rachel Riley, Jamie Klingler, and Sharan Dhaliwal. Amber Heard and Bryony Gordon were not able to obtain full data downloads.

- Instagram is failing to act 9 in 10 abusive DMs reported using the platform's tools.
    - Instagram failed to act on 9 in 10 accounts sending violent threats over DM
    - Instagram failed to act on any image-based sexual abuse within 48 hours
    - Instagram failed to act on all accounts sending 'one-word' hatred
    - For this analysis, researchers worked with participants to record and report DMs from 253 abusive accounts. Amber Heard, Rachel Riley and Sharan Dhaliwal took part in this analysis, as they had not removed abusive DMs or blocked their senders.

- Researchers identified several systematic problems that Instagram must fix:
    - Users cannot report abusive voice notes that accounts have sent via DM
    - Users must acknowledge "vanish mode" messages to report them
    - Instagram does not automatically consider previous abusive messages
    - Instagram's "hidden words" feature is ineffective at hiding abuse
    - Users can face difficulties downloading evidence of abusive messages

- CCDH recommends that:
    - Instagram fixes its broken systems for reporting abuse
    - Instagram closes routes that strangers use to abuse women
    - Instagram invests in moderation and prioritizes direct abuse
    - Legislators must ensure platforms meet minimum standards on transparency, systems for reporting illegal behavior, and standards for action on reports of harmful messages

3 Participants

For this report, we worked with women with a public profile, an active Instagram account and experience of being abused on the platform. We approached over 50 women to participate in this report. Though some were unable to participate due to time constraints, others were reluctant to speak about misogynist abuse online out of concern that publicity around this report could make them a beacon for more abuse and negatively affect their mental health. Several women who use Instagram as a significant means to promote their personal brand or conduct commercial work expressed fears that the platform might punish them for criticism by deprioritizing their posts.

Amber Heard

1,152 Posts | 4.1M Followers | 55 Following

Amber Heard is an actress and activist. She joined Instagram in 2017 because "Everybody else had something to say about who I was and what I was up to and I was the only person not weighing in on that."

In 2018, Heard wrote an op-ed titled "I spoke up against sexual violence - and faced our culture's wrath. That has to change."[3] Following this, her ex-husband, the actor Johnny Depp, launched two legal actions, one of which is due to be heard in a Virginia court in April 2022. Discussions of the cases on social media often contain abuse towards Heard, as well as the promotion of conspiracy theories around the rate of women's false allegations of domestic abuse.

Rachel Riley - 535,000 followers

509 Posts | 538K Followers | 201 Following

Rachel is the co-presenter of Countdown, a television quiz series on the UK's Channel 4 station and its spin off comedy panel show, 8 Out of 10 Cats Does Countdown. She joined Instagram in 2017 "Partly for work reasons, to post about what I was doing and programs I was on, or what I was up to in my life. It's another way to connect with people, all of my friends were on it and it's a nice way to keep in touch."

Riley is a devoted Manchester United fan, supports many charities including those which promote STEM to young people, especially girls, and is mother to two young daughters.

She is a vocal campaigner against antisemitism, having first spoken out in 2018 against the antisemitic hate that she saw and experienced from members of the UK Labour Party under Jeremy Corbyn's leadership. Riley is an ambassador for the Center for Countering Digital Hate.

Jamie Klingler - 3,000 followers

7,920	3,541	1,938
Posts	Followers	Following

Primarily working in publishing and events, Jamie Klingler is a writer, activist, and pundit on women's safety. In response to the kidnap, rape and murder of Sarah Everard in March 2021 by a serving Metropolitan police officer Klingler co-founded campaign group Reclaim These Streets to organize a vigil in memory of Everard. The Met ordered the vigil to not go ahead under Coronavirus rules, warning each co-organizer with a fine of £10,000 if they attended. The group cancelled the vigil and raised £550,000 to establish the Stand With Us fund for grassroots anti-violence against women and girls organizations. In March 2022, the High Court found in favor of Reclaim These Streets' action against the Met on a claim that the force violated their human right to assemble. Klingler joined Instagram in 2018, she says, to post images of "me, my dog, the books I read. And now it's more activism stuff and some weight loss stuff, or me running."

Bryony Gordon - 201,000 followers

2,465	202K	3,363
Posts	Followers	Following

Bryony Gordon is an award-winning journalist, author and mental health campaigner. As well as writing a column for The Telegraph, she is the writer of five Sunday Times Bestselling books, including No Such Thing As Normal, Mad Girl and Glorious Rock Bottom. Her podcast, Mad World, has featured guests including Prince Harry, Mel B and Nadiya Hussain. In 2016, she founded Mental Health Mates, a peer support group that is now established across the country. She lives in south London, with her husband and daughter. She joined Instagram in 2012, she says, as "just a social thing". More recently, her intention for Instagram has been to "connect with readers users of Mental Health Mates... I see it as my portal to the zeitgeist!"

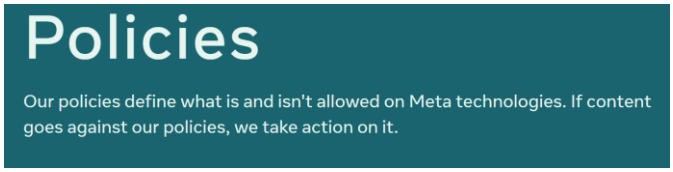
Sharan Dhaliwal - 8,390 followers

470	9,714	679
Posts	Followers	Following

Sharan Dhaliwal founded the UK's leading South Asian culture magazine Burnt Roti, a platform for young creatives to showcase their talent, and destigmatize topics around mental health and sexuality in a safe space. She is the Director of Middlesex Pride and creator of Oh Queer Cupid, a queer speed dating and comedy night. She has written for i-D, HuffPost, The Guardian and was on the list of global influential women for the BBC 100 Women 2019. Her debut non-fiction book, Burning My Roti, focuses on "sexual and cultural identity, body hair, colorism and mental health, and a particular focus on the suffocating beauty standards South Asian women are expected to adhere to." She joined Instagram in 2012 in order to, she says: "To post the best photo from a party on Instagram. But now it's more of a diary, and also a PR system for marketing me as a brand."

4 Instagram's policies on abusive DMs

Meta, Instagram's parent company, sets the same Community Standards across all its platforms, promising "if content goes against our policies, we take action on it."[4]



These policies clearly list the following categories as "content that is not allowed" on Instagram and Meta's other platforms:

- Hate speech, including misogyny, homophobia, racism and more[5]
- Bullying and harassment, including calls for death, suicide, or violence[6]
- Any displays of nudity or sexual activity[7]
- Violent and graphic content[8]

Meta states that its Community Standards "apply to everyone, all around the world, and to all types of content."[9] It is therefore clear that Meta's standards are intended to apply to direct messages (DMs) sent on Instagram, as well as other public content on the platform.

Where this report refers in general to "abusive" content or messages, it is referring to content that would violate one or more of the above policies.

Instagram promised to suspend or disable accounts that send abuse over DM

Following public pressure over accounts using the platform to send abuse to users, Instagram has acknowledged that "seeing abusive DMs in the first place takes a toll". In February 2021, it announced new measures aimed at "removing the accounts of people who send abusive messages", including:

- Blocking abusive DMs from sending further messages for a set period
- Disabling accounts that repeatedly send abusive DMs
- Disabling new accounts created to get around the above measures[10]

While Instagram states that "because DMs are private conversations, we don't proactively look for hate speech or bullying the same way we do elsewhere on Instagram."[11] However, Instagram does allow users to report DMs that breach its Community Standards.[12]

In giving evidence to the United States Senate Subcommittee on Consumer Protection, Product Safety and Data Security in December 2021, Adam Mosseri, Head of Instagram, said "I believe we try and respond to all reports, and if we fail to do so, that is a mistake that we should correct."[13]

5 1 in 15 DMs sent by strangers to high profile women break Instagram's rules

New research shows that 1 in 15 Instagram Direct Messages (DMs) sent by strangers to high profile women contain content that violates Instagram's own Community Standards.

This section of this report sets out to measure the scale and type of abuse sent by strangers to our participants through Instagram's DM feature which, once sent by an abuser, are received in a user's "Requests" inbox.

Instagram DMs allow strangers to send abuse to users

On Instagram, DMs sent by "strangers" - i.e., accounts a user does not follow - arrive in a subfolder called "Requests". Instagram sometimes refers to these messages as "message requests" or "DM requests".[14] Senders of "Requests" are not able to see if the recipient has read their message until the user "accepts" this message.

However, accounts do not need the recipient's permission to send "Requests". Our research identifies "Requests" as a source of abuse, and even threats of violence, sent by abusive accounts seemingly without any limits.

Instagram users can obtain a file record of messages from strangers received by their "Requests" inbox as part of the platform's data download feature. All this report's participants requested a data download from Instagram but not all were sent full data. Participants who had previously blocked abusive users could not access "Requests" sent to them by these users. **As a result, this section of the report analyses "Requests" sent to Rachel Riley, Jamie Klingler, and Sharan Dhaliwal.**

To assess the overall scale of abusive content that strangers have sent to these women via Instagram DMs, we combined the findings found in all audio, image and video messages sent to participants, all text messages sent to Klingler and Dhaliwal and a representative sample of messages sent to Riley.

This analysis of 8,720 text, audio, image, and video messages found 567 (6.5%) contained misogyny, image-based sexual abuse, other hatred and graphic violence.

Content Type	All content	Violating Content
Audio	142	20 (14.1%)
Image	6,059	374 (6.2%)
Text	1,889	65 (3.4%)
Video	630	108 (17.1%)
Total	**8,720**	**567 (6.5%)**

5.1 Instagram DMs are regularly used to send image-based sexual abuse

Analysis of data provided by Instagram to participants shows that Instagram DMs are regularly being used to send image-based sexual abuse (IBSA), a category of content that is illegal in some jurisdictions.

What is image-based sexual abuse (IBSA)?

The UK's End Violence Against Women Coalition (EVAW) defines image-based sexual abuse as "all forms of taking, making and sharing nude or sexual images without consent, including threats to share and altered images."[15] Image-based sexual abuse can take the form of "cyberflashing", where an abuser sends another person an image of their genitals without consent. This can include unsolicited "dickpics".[16]

In total, participants were sent 482 abusive image and video DMs from strangers, of which 125 fit the definition of image-based sexual abuse, including:

- 66 examples of users sending women pornography without consent
- 50 examples of men sharing images or videos of their genitals
- 9 examples of pornography edited to feature other faces, known as "fake porn"[17]

Instagram's Community Standards include a blanket ban on nudity and sexual content, with specific bans on "visible genitalia" or "imagery of sexual activity".[18] A reasonable reading of the rules would conclude that image-based sexual abuse is prohibited.

Moreover, image-based sexual abuse is already illegal in several jurisdictions, including Germany, France, Ireland, Canada and some US states.[19] The UK government has recently announced it will use its new Online Safety Bill to make cyberflashing a criminal offence.[20]

Riley says that knowing that Instagram accounts have sent her these images via Instagram DM "Turns my stomach," she tells CCDH. "It really makes me not want to go into my DMs at all because it's revolting. It's astounding to know that strangers are sending porn - it empowers them to know that it's gone to your inbox."

"On Instagram, anyone can privately send you something that should be illegal. If they did it on the street, they'd be arrested."

Dhaliwal, who says she has received "hundreds" of dickpics over the years from users she has blocked, explains that "It's a power play, it's less to do with having a sexual interaction, it's about them feeling they have power and can walk away from that saying, 'I did that'".

5.2 Serial cyberflashers accounted for a disproportionate amount of image-based abuse

Several users who sent image-based sexual abuse to participants were "serial cyberflashers" who repeatedly send women images of their genitals.

In total, serial cyberflashers were responsible for 39 of the 125 examples of image-based sexual abuse identified by researchers, equivalent to 31.2% of this type of abuse.

CCDH's researchers found within Dhaliwal's data proof that one man had sent her two images of his erect penis, adding the comments "Cute and cute 😍" and "Yea I like you".

Within Riley's data, CCDH's researchers found one man had sent her three images of his penis, and another man had sent 31 videos to Riley. In 26 of them, he is masturbating, including one video showing him masturbating onto a tablet computer displaying a photo of Riley. Another shows him naked in bed whispering "love you angel" as he masturbates.

Knowing that this man has sent her such a volume of extreme content again "turns my stomach", Riley says, before echoing experts on violence against women and girls who have long identified a link between flashing and sexual violence: "Someone can freely send you so many videos and carry on. But you know that if he's doing that, it's likely that it's not the only thing he's doing. We know in the real world that flashing is a big red flag."[21]

5.3 Instagram is allowing fake porn to be sent over DM

Riley received nine fake porn images. Three were of her head photoshopped onto the body of a nude model, two were of her head photoshopped into pornographic images depicting her engaging in a sex act. Three other images feature the faces or heads of other well-known women head photoshopped onto pornographic images. A final image is of two unknown men.

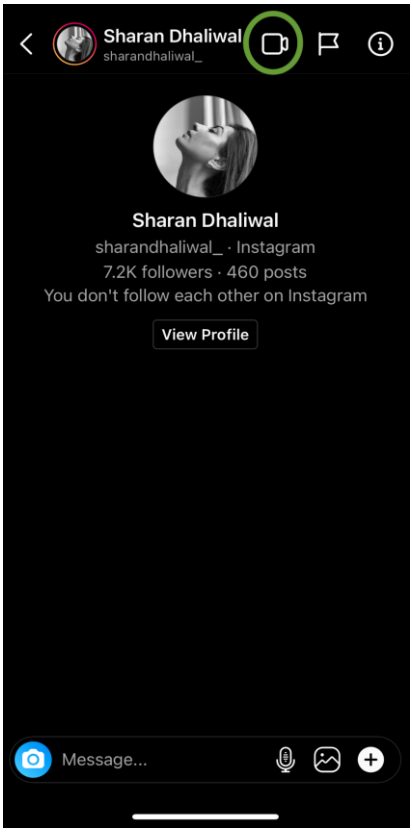
Riley was also sent by an Instagram account a link to a porn site with the URL extension "hate-wank-for-rachel-riley". The site reads: "The requested content is no longer available".

Instagram could and should stop abusers from sending image-based sexual abuse to women, Riley says: "It should be a default that you don't get porn sent into your DMs, and it should be an option to opt in if you want to see porn. [Instagram] have got the technology to identify it. What majority of Instagram users want porn in their DMs?"

5.4 One in seven voice notes sent to participants was abusive

Instagram's voice note feature enables accounts to send short audio recordings to others. This includes allowing strangers to send recordings to people who do not follow them.[22]

We reviewed a total of 142 audio files in the participants' data, and 20 contained abuse. Twelve contained misogynistic hate speech including one user calling Kamala Harris a "dumb satanic bitch". One was both ableist and misogynistic, with other identity-based hate voiced in five more audio files containing ableist, homophobic and racist - including antisemitic - hate speech. One man sent two unsolicited sexual comments to Riley.



Our participants were also subject to strangers attempting to video call them using Instagram's video call feature. Strangers are allowed to call people who do not follow them via a one-tap button in the top right corner of the chat window, circled in green here:

Riley's data showed that on five occasions, strangers had attempted to video call her. And Dhaliwal's data showed that strangers have attempted to video call her seven times. One account tried to call Dhaliwal after sending her two dickpics. Another attempted to video call Dhaliwal after messaging her 42 times in a 3-day period with comments such as "sex", "I like hairy" and "I wanna lick it".

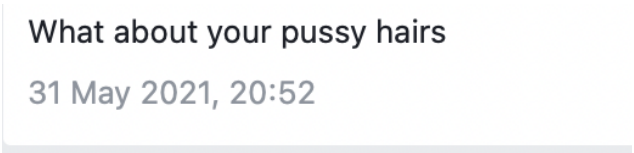
Dhaliwal tells CCDH that while "you can dissociate from most abuse", receiving a voice note feels particularly invasive because "when you hear their voice it becomes more real."

5.5 Most abusive texts were unsolicited sexual comments

While a lower proportion of text DMs were abusive compared to images, video, and audio DMs, most of these abusive texts were unsolicited sexual comments.

Researchers recorded a total of 65 abusive texts, of which 43 were unsolicited sexual comments, equivalent to just over two-thirds of all abusive text messages analyzed.

During an eight-day period, a stranger sent Dhaliwal 120 messages of "Can I LICK your feet" or "Can I kiss your feet", equating to 15 messages per day. She has also received several explicit messages about her body hair.

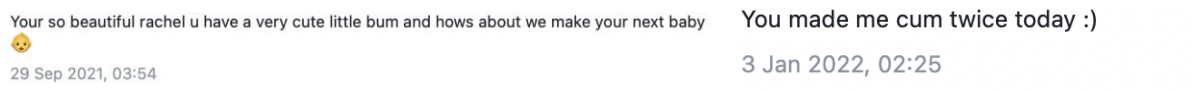


Example of unsolicited explicit comment sent to Dhaliwal[23]

As Dhaliwal explains it: "I've had a journey with my body hair. When I first started [posting images of herself showing her natural body hair] the messages I was getting were from other South Asian women being like 'Oh my god, yeah, I didn't want to remove

my arm hair, but my mum used to make me wax it.' And now it's men getting in touch saying they want to 'licky licky' my hair. I'm like "No, this isn't the conversation I wanted to have."

Abusive accounts sent Riley 26 unsolicited sexual comments, often late at night and detailing sexual fantasies about her.



Examples of unsolicited explicit comments sent to Riley[24]

Meanwhile Klingler received messages either asking directly for her to be a "sugar mommy" to younger men, or, as below, for sex.

Example of unsolicited explicit comment sent to Klingler[25]

She tells CCDH: "There's no sexual content on my Instagram whatsoever. There are no thirst traps, it's not sexy. I just think: 'Why are you emboldened to think this is a possibility?" There's a lack of boundaries. A lot of it makes me feel prudish. I feel like I'm too old for all of this."

6 Instagram is failing to act on 9 in 10 accounts that send abuse over DM

Instagram failed to act on 89.7% of the accounts who sent abuse via Instagram DM to our study's participants.

While the previous section of this report shows that a significant proportion of DMs sent by strangers are abusive, this section examines Instagram's performance in acting on abuse once users have reported it to them.

Working with access to participants' DMs, researchers logged abuse sent by 253 accounts and reported them using the Instagram app or website. An audit of abusive accounts revealed that 227 remained active at least a month after these reports were filed, representing Instagram's failure to act on 89.7% of reports sent to its moderators.

This finding is particularly concerning given that our research suggests half of abusive users go on to send further abusive messages when platforms fail to remove them.

In order to conduct this part of our report, CCDH's researchers selected participants who had neither deleted nor blocked all of the abuse were sent to them by strangers through Instagram DM. **These participants are Amber Heard, Rachel Riley and Sharan Dhaliwal.**

Abusive messages from strangers were identified by logging into participants' accounts and accessing their "Requests" inbox. Abusive messages were then logged and reported to Instagram. Logged messages were tagged according to the types of abuse they contained, and were tagged as "serial abuse" where it was clear from the conversation thread that the sender had repeatedly sent abusive messages. The table below gives an overview of how many unique accounts had sent particular categories of abuse.

Type of abuse	Accounts acted on	Not acted on	Total
Image-based sexual abuse	2 (50.0%)	2 (50.0%)	4
Die or kill yourself	1 (14.3%)	6 (85.7%)	7
Unsolicited sexual messages	4 (16.7%)	20 (83.3%)	24
"One-word" hatred	7 (12.7%)	48 (87.3%)	55
Serial abuse	10 (11.8%)	75 (88.2%)	85
Rape or sexual violence	0 (0.0%)	3 (100.0%)	3
Death threats	7 (11.3%)	61 (89.7%)	68
All violent threats	9 (11.2%)	79 (88.8%)	88
All abuse	**26 (10.2%)**	**227 (89.7%)**	**253**

An audit carried out 48 hours after reports were filed found an even worse rate of action, with Instagram failing to act on 99.6% of abuse within this timeframe.

Instagram has previously acknowledged that the "requests" inbox reserved for DMs from strangers "is where people usually receive abusive messages."[26] As well as promising to filter these messages, the platform has promised to act on reports of abuse.[27]

Adam Mosseri, Head of Instagram, told the United States Senate Subcommittee on Consumer Protection, Product Safety and Data Security last year that "I believe we try and respond to all reports, and if we fail to do so, that is a mistake that we should correct."[28]

Riley tells CCDH that she believes Instagram's moderation is so poor because Meta does not resource this department: "The galling thing is Instagram is pretending to care and it's clear there's no financial incentive for them to do it and nobody can see it, so they can get away with not putting any money in and not protecting people from this kind of material."

Instagram's failure to act on abuse forces women to manage abuse themselves

Riley and Heard both avoid Instagram direct messages, with Heard having "somebody else help me and act as a filter. Once in a while if it's really bad or really specific or if the death threats involve a plan I may or may not bring it to the police."

Both Heard and Riley are cognizant that their work is not dependent on them reading direct messages, so they can afford to avoid the platform.

However, other participants rely on Instagram for work. Klingler explains: "Press requests come in for me to talk about my activism." She blocks abusive messages rather than reporting them because, she says: "Instagram is not women-first about this, they're not safety-first about anything. Like with women being gang raped on the Metaverse, everything about it is a cesspool for us. They need pretty, young women - I don't put myself in that category - to put their content out to make the platform survive, but they're making pretty young women the targets."

Abusive messages sent to Gordon often reference her weight, she says: "I am a recovering bulimic who's been very open about eating disorders. People see a woman running a marathon in her underwear at 16 stone, someone who's said publicly 'I've had this body loathing' and their astonishing response is to loathe on that person's body." Her followers help report abusive comments on Instagram, but she calls this following is a "privilege" and knows this method won't work for direct messages. "It's not enough. There's got to be more accountability. Publicly, I talk about the nasty voices in my head, and I can confirm those nasty voices very easily on social media. It's very dangerous."

"When we look back on this period of our completely unboundaried use of social media we'll look back with the same horror as we do adverts with the Marlborough Man saying smoking is good for you."

Heard has attempted to take death threats to law enforcement in the US, however, "they told me I would need to bring them an invitation, a save-the-date in order for them to take that threat seriously. I've made many police reports but because it doesn't say 'I'm going to kill you at 2pm on Thursday', they won't investigate. I'm not kidding, it's ridiculous."

6.1 Instagram failed to act on accounts sending image-based sexual abuse

Researchers reported four examples of image-based sexual abuse to Instagram, which failed to act on all these messages within 48 hours.

Three of these messages were dickpics. In one case, an account sent three images of an erect and ejaculating penis, accompanied by the message "I don't care weather you put weigh on or not you just and will always make me feel like this babe...[sic]"**.**

The fourth account sent a pornographic video of a man masturbating on a woman's face.

More than a month after the messages had been reported, just two of the accounts that had sent dickpics had been removed, while the others remained active.

In response to Instagram's failure to act on image-based sexual abuse, Riley tells CCDH: "It's just embarrassing for Instagram isn't it, really? I just think all social media are not fit to regulate themselves, clearly. They've got the technology to identify that this content is revolting, but they still allow strangers to send it to you unsolicited".

Heard, who has previously spoken of being sexually assaulted by the time she was of college age, explains that despite having a large following and blue-tick verification on Instagram, "I have no recourse to address the kind of scale of attacks that I'm getting."[29]

The law on image-based sexual abuse

Image-based sexual abuse is already illegal in several jurisdictions, including Germany, France, Ireland, Canada, and some US states.[30]

In March 2022, the UK Secretary of State for Digital, Culture, Media, and Sport Nadine Dorries announced the introduction of a new offence of cyberflashing that will carry a two-year sentence: "The forthcoming Online Safety Bill will force tech companies to stop their platforms being used to commit vile acts of cyberflashing. We're bringing the full weight on individuals who perpetrate this awful behavior."[31]

The same Bill will introduce multi-million-dollar fines for tech companies that fail to prevent users from sending image-based sexual abuse.[32] At present, Instagram's failure to act promptly on such abuse could see it fall foul of this new legislation.

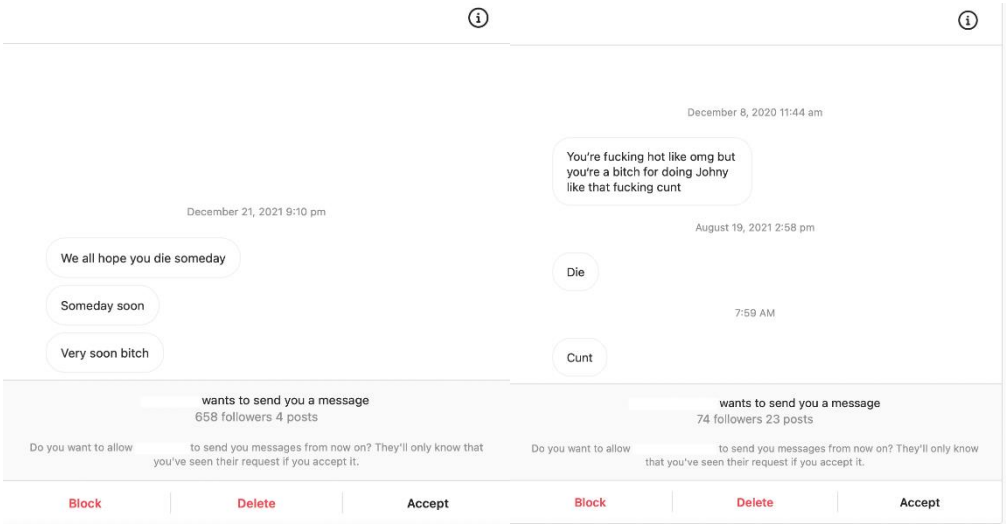
6.2 Instagram fails to act on 9 in 10 accounts sending violent threats over DM

Instagram allowed 9 in 10 abusers who sent violent threats to our participants to remain online, even after they were reported to Instagram using the platform's own tools.

Researchers reported a total of 88 violent accounts using Instagram's own tools, including 68 that had threatened to kill participants and 3 that had threatened sexual

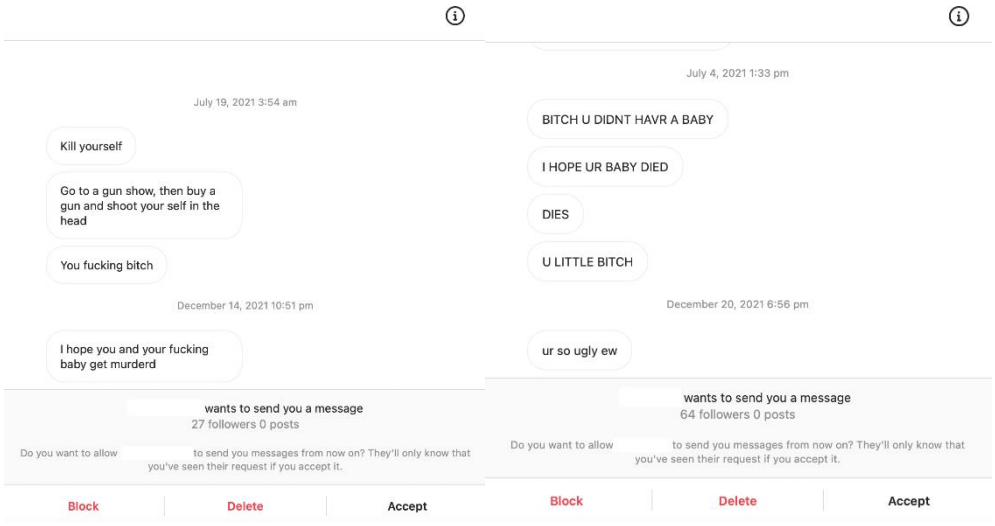
violence or rape. After 48 hours, the platform removed one abusive account, who had sent the message “die you stupid wh0re”.

Our second audit taken in March, more than a month after we had reported the abusive DMs, showed that 79 out of 88 accounts were still active, equivalent to 88.8% of accounts who had sent death threats. Accounts that have not been removed sent a range of abuse from “kys” (kill yourself) and “DIE”, to detailed descriptions of murder and suicide.



Examples of death threats sent by Instagram accounts[33]

A particularly disturbing finding of our study is that Instagram failed to remove some accounts that made death threats aimed at Heard’s family members, including her infant daughter.



Examples of death threats sent by Instagram accounts to Heard that referenced her family[34]

Meta’s Community Standards prohibit any content “calling for self-injury or suicide of a specific person or group of people.”[35] Instagram’s failure to apply these standards to DMs allows users to continue sending death threats and other violent threats to others.

Heard has become used to this level of abuse, she tells CCDH: “I’m not a sensitive person to these kinds of harassment. I’ve been dealing with it for a very long time.” Her method of dealing with it is to no longer use her own Instagram account.

Otherwise, her mental health is negatively impacted: “I know it’s baseless, it’s not rooted in reality, but it has real life effects. It increases my paranoia, my indignation, my frustration with lack of recourse or the ability to do anything about it. Social media is how we connect with one another today and that medium is pretty much off limits to me. That’s the sacrifice I made, the compromise, the deal I made for my mental health.”

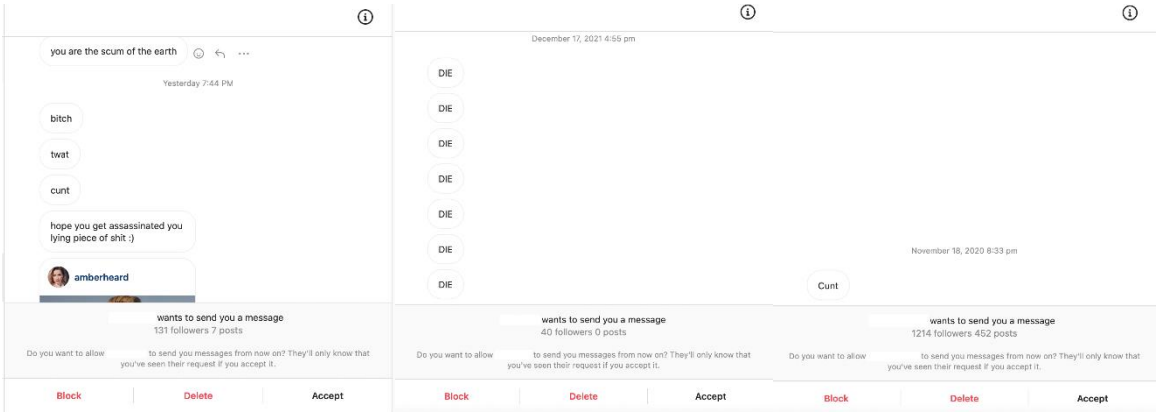
“Any time I see hate or misogynistic abuse online it makes me mad, even if it’s not directed at me. So, imagine if it’s directed at me. More importantly, directed at some of the worst things I have survived. I’ve survived an incredible amount and to come out of it and face the exact same patterns of abuse is ironic and hilarious on my best days and depressing on my worst.”

Instagram’s reporting function is, Heard says: “not user friendly, not intuitive, not common sense based, and the amount of work you have to go through to report someone - even when you exclude how not effective it is - is difficult. Take into consideration the efficacy then it becomes pointless. From the outside it’s looking like Instagram is doing nothing.”

6.3 Instagram is failing to act on clear hate speech

Instagram’s systems failed to remove obvious examples of hate speech sent to participants by strangers within 48 hours, even after they were reported to the platform.

Researchers reported 55 accounts that sent “one-word” misogynist hate such as “bitch”, “c*nt” or “whore” or one-word calls for participants to “die” or “KYS” meaning “kill yourself”.[36] All of these accounts were still active 48 hours after they were reported.



Examples of one-word misogyny sent by Instagram accounts[37]

In July 2021, Head of Instagram Adam Mosseri conceded that detecting hate speech on the platform “can be challenging because one word or emoji might be benign in one context and completely abhorrent in another.”[38]

Instagram might face obstacles deciphering the context surrounding gendered slurs used on its platform. However, the platform has also shown a failure to act on the simplest examples of misogynist hate speech sent by accounts using its direct messages feature.

7 Instagram makes it hard to measure and report abuse

In working with recipients of abuse over direct message for this report, researchers identified several problems with Instagram's features that make it hard for users to report abuse, block it out or retain evidence of what could be criminal messages.

This section outlines five ways in which Instagram's design enables abuse over DM:

1. Users cannot report abusive voice notes that accounts have sent via DM
2. Users must acknowledge "vanish mode" messages to report them
3. Serial abuse is not considered by Instagram's moderators in all territories
4. Instagram's "hidden words" feature is ineffective at hiding abuse from users
5. Users can face difficulties downloading evidence of abusive messages

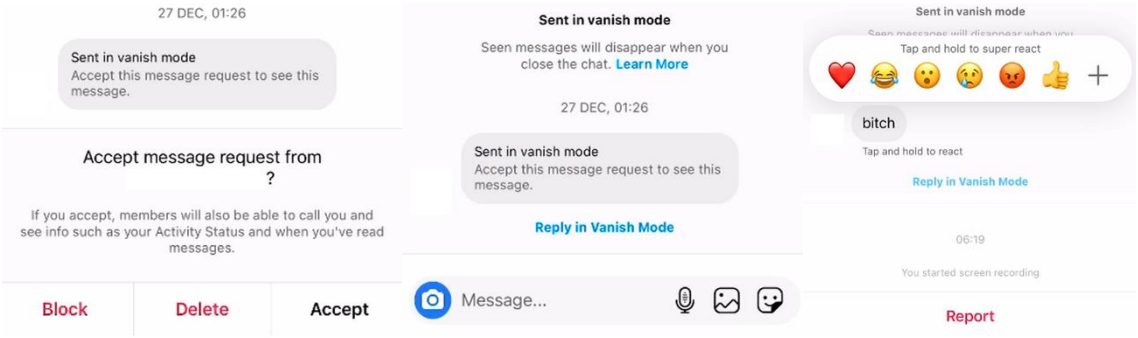
7.1 Voice notes cannot be reported

CCDH's research shows that 1 in 7 voice notes within our participants' data is abusive. According to Instagram, "You can report abusive photos, videos and messages that are sent to you".[39] This does not include voice notes as they cannot be reported.

During section 9 of this report, we could not report a voice note death threat that an account had sent to Heard saying: "You, I don't like you, you are bad people. Die! Die! Die! Die! DIE!".[40] We could, however, react to the post with an emoji. CCDH's researchers instead reported the account. It remained on Instagram by March 2022.

7.2 Vanish mode messages require users to face abusers to report them

One of few protections afforded by the "Requests" inbox is that the sender cannot see whether the user has read the message or not.[41] However, to open a "vanish mode" message to see and report abusive content, a user must "accept" the message.



Example of what a user sees when receiving and reporting a "vanish mode" message[42]

We reported this message and blocked the account so that it could not see that Heard's account had seen this message. By March 2022 the account remained online.

7.3 Instagram does not automatically consider previous abusive messages

To report a serial abuser who has sent more than one abusive message, many Instagram users are obliged to report every abusive message. According to its Help Center: "Instagram reviews up to 30 of the most recent messages sent in reported conversations that involve accounts based in the European Union."[43] It is unclear how many messages will be assessed by moderators after a report is made outside of the EU.

A third of the messages reported by CCDH using Instagram's own tools were sent by serial abusers. Instagram will not always identify this harmful abuse in many countries at present.

7.4 Instagram's "Hidden words" is ineffective, puts onus on victims to stop abuse

In April 2021 Instagram announced the opt-in "hidden words" function that, the platform states, filters "offensive or unwanted" into another folder.[44] However, it is ineffective.

Turning on the "hidden messages" function for Heard's account led to CCDH finding that phrases and words such as "f**k you", "bitch" and "f**king bitch" were not being hidden.

Klingler wonders why Instagram can identify abusive words but not act on the accounts sending them: "They're siloing the abuse and not doing anything about it. They want to keep account numbers up so they can keep growth up. But where does my protection and privacy start and abusive accounts' start?" Meanwhile Dhaliwal expects very little help from the "hidden words" feature as it won't filter images: "They're not going to put an alt text on their penis!"

7.5 Users' data is hard to access

Instagram promises users the ability to "access, rectify, port and erase your data".[45] However, our research found Instagram's inability to present this data clearly to users.

Instagram did not provide Heard her own data, emailing her to say: "a technical error prevented us from creating a file that you requested…we're sorry that this happened."[46]

No one was provided messages previously sent to them by blocked accounts. Instagram's "block" function is, the platform says, there to keep users safe from bullying and harassment.[47] However, blocking could make it harder to hold abusers to account.

For those who were provided their data by Instagram there was no clear way of identifying abusive senders. Usernames were scrambled and did not hyperlink to account pages. In some cases, the sender's name was listed as nothing other than "Instagram User" and in others the content of the message was entirely blank or the audio entirely inaudible.

8 Abuse over DM is having a chilling effect on women's speech

Participants of this report are aware that their high profile means they're not representative of the average woman Instagram user. However, each worry that the abuse they are sent online by misogynists face online could have a chilling effect on women's free speech. Heard tells CCDH: "The amount of people who might be in a similar situation to what I was back in 2015 who look at what has happened to me and decide to not act in the interests of their safety, or their voice is scary to think about."

She has also had to have frank conversations with women seeking to name their abusers: "Other women have wanted to come out and say something because the person who had hurt them was a person in the public eye and they thought it would be important that they maybe try to be the last victim."

"But I've had to have this conversation with people where I say, 'Look, I can't tell you it's gonna be ok.' I want to say, 'It'll be ok' and I do say that, but I also say 'You're not wrong in looking at me and thinking you don't want this for your life'.".

Heard is concerned about other people with less resources than her who are impacted by online abuse: "If I can't utilize this tool, if I can't open Instagram, if I can't engage at all, then what does it say about a person who doesn't have the emotional resources that I have, that come with age and experience?"

Meanwhile Riley worries about younger women being sent the same image-based sexual abuse that she receives in such an isolated format: "Teenage girls could receive this stuff while no-one else knows because it's behind closed doors. It's invasive and disgusting."

And Klingler is effusive about ensuring that whatever Meta does in response to this report, it benefits all women: "I don't want Meta to handle me with cotton gloves so that I'm safe while they don't bother about the safety of other women."

9 Recommendations

This report is a window into the harassment, abuse, and violence that is experienced globally by women - especially Black women and women of color, LBGTQ+ people and other systemically marginalized groups online - with a specific focus on those who have a high profile. The nature, volume, and repetition of the harmful content in our study shows the scale of the problem. The women who participated in our research, bravely sharing their stories, articulate the impact of this abuse. They are not alone. We know that this type of abuse of women online is widespread. There needs to be both investment, and system and regulatory changes to turn the tide of abuse and violence against women. We should not be propping up a system that allows this violence and abuse to flourish, and that acts to undermine women's freedom of expression, wellbeing, and safety. This is too important to ignore.

Despite the press releases and the spin, Big Tech companies like Instagram (Meta) choose not to act when they think no-one is watching and when there is no law or accountability requiring them to act. The promises that they've made in press statements and to politicians when asked to testify are simply empty platitudes.

This report shows once again a failure of Big Tech to respond and act on requests either for personal data or the content itself. The fact is they are failing to enforce their own terms and conditions 90% of the time.[48] This failure to act is consistent with our previous research that shows on receiving user reports, platforms are failing to act on:

- 87.5% of Covid and vaccine misinformation[49]
- 84% of content featuring anti-Jewish hate[50]
- 94% of users sending racist abuse to sportspeople[51]
- Users who repeatedly send hateful abuse.[52]

Everyone deserves better than this.

Through this study with these women, we're helping to name the impact and dangers of misogyny and violence against women and other marginalized communities online and joining the ongoing calls for change.

Fundamentally, Big Tech cannot be trusted to self-regulate or follow a voluntary code without teeth. The commitments that they've already made to users, the public and lawmakers have been broken. The harm they are perpetuating is real and widespread.

These recommendations are specific to our report, but they also include core features and principles that are generally applicable, and technical recommendations that should be built into Instagram's systems and processes to ensure the safety of other groups in society, including children, Black, Indigenous, and People of Color, LGBTQ+ folks, religious minorities, disabled people, and more.

Recommendations for Lawmakers

The report shows that core features of a regulatory framework and enforcement are both missing and needed for platforms like Instagram, including:

- **Transparency:** We know from our research, and others, that there are limits on the types of information that are currently publicly available, even with

knowledge and tools. This information is held by the companies, like Instagram, who can track and trace what is happening online and the impact in changes to algorithms, processes, or safety features. This information asymmetry has devastating impacts for individuals, communities, and society, particularly where the same vested interests who hold most of the information are making all the decisions and profiting from them.

In terms of this report, the findings support a need to ensure that key data, patterns, and trends about abuse online [and engagement with that content] can be easily accessed, identified, analyzed, and addressed. Not just the data that Big Tech decides to share publicly but the data that exposes the problems (and possible solutions).

A regulator should have the power to require this information to be given and enable it to be shared with experts, including independent researchers, academics, and civil society organizations. This could be through anonymized data. Only then will we really be able to understand the full extent of the harm being experienced and continue to identify effective intervention points and tools. Relying on studies released by the company or organizations funded by Big Tech is like asking Big Oil to produce a feasibility study on climate change.

- **Responsibility**: a clear, proactive duty of care that is placed on platforms to ensure that their services and products are safe for all users, including children, before those users are exposed to harm. This includes any substantial changes made to those products, services, and processes. The weight should not fall on individuals to address online harm, particularly where they don't have the privilege of access to the underlying causes of that harm or the resources to design interventions that would change the operating environment. There should be safety by design. Social media, like Instagram, should be safe before it is used, in the same way as we demand from those producing food, cars or pharmaceuticals. Clearly, unregulated Big Tech is not motivated to do this by itself - as evidenced in this report. Removing any general and unjustified exceptions to negligence law and rebalancing the investment that is spent on engagement with safety features will lead to a safer environment for all users, including children. People shouldn't unknowingly and unwillingly become a testing ground for new products - there is corporate responsibility that needs to be reflected in law.
- **Behavior that is illegal offline should be illegal online**: A lot of the content that we access and analyze in our reports, including this one, involves behavior that is illegal offline. This report shows that death threats, sexual violence and threats of violence are allowed to flourish unfettered on Instagram (Meta). If this behavior is criminalized offline, why should there be a double standard in the virtual world. Regulation should support this premise being realized.
- **Controls on legal but harmful content**: There is a large amount of content shared on platforms, including some of the DM content assessed in this report, which may fall short of a criminal standard but nonetheless is harmful to users because of its nature, intensity, or repetition or because some users may be more sensitive to that content, such as children or because it is targeted at people with particular characteristics or vulnerabilities. Much of this content has already been recognized as harmful content by the platforms themselves, which is why they

mention it in their terms and conditions and press releases. Problematic content that may fall in this category includes, for example, self-harm, posts glorifying eating disorders, or abuse messages. Legal but unwanted. Legal but harmful.

- **Complaints systems:** Clear, easy to access and responsive complaint systems - with oversight accountability of the platform by an independent government regulator. Our report shows significant failings in Instagram's current reporting pathways for DMs, and specific recommendations are outlined below.
- **Accountability**: An effective and resourced regulator, and process for appeal or prosecution through tribunals or a court, will help to ensure that core responsibilities are being met.
- **Consequences and offences**: Bad actors should not have free rein to abuse people online, but this is the status quo. Our work, as well as other research, has proved that deplatforming has been effective at reducing the online harm from bad actors. This could sit alongside a warning and suspension system. There are different ways that this could be actioned, including regulator direction powers or oversight powers with a positive obligation on platforms to ensure that they are acting. Consequences and offences also need to apply to companies and senior officials within the company - corporate liability for failure to discharge the responsibilities will help to ensure that the investment is made to improve the functions and processes and to put safety first. A similar model of incentives operates in health and safety law in many jurisdictions, and it is one of the underlying premises behind corporate manslaughter. In the US context, litigation (and the risk of litigation) is a strong motivator for shifting corporate behavior.

Recommendation: Ensure the regulatory framework is fit for purpose: transparency requirements, responsibility (duty of care), behavior that is illegal offline is illegal online, controls on legal but harmful content, effective complaints systems, accountability through an independent regulator and the courts, and consequences / offences for bad actors, platforms, and senior officials within the company.

Recommendations for Platforms and Regulators - Improving the Safety of Instagram DMs and Combating Misogynistic Abuse Online

CCDH is a member of the Women Disinformation Defense Project (WDDP), led by Ultraviolet. Ultraviolet, and the WDDP's work on misogyny, racism and tech platforms has paved the way for the conversation we are elevating in this report about the realities of abuse online for women and specifically for Black and other women of color.

You can find their work on this on their website at this link.

CCDH has signed up to these recommendations of UltraViolet for platform reform:

1. Broaden the definition of hate speech to include misogyny, bias and attacks against Black, Indigenous, and people of color, religious minorities, and transgender people, and gendered and racialized disinformation.
2. Create a clear, enforceable, and escalating process for reporting and removing hate speech, disinformation, and promotion of white supremacy and misogyny; frequent and severely abusive violators must be banned from the platform.
3. Support and protection for victims of harassment, hate, disinformation, and abuse must center the experiences of marginalized people and groups.
4. Create internal policies, training, and culture that address and acknowledge misogyny and the ways in which it intersects with other marginalized identities; provide staff and contractors access to mental health resources.

This further set of recommendations is about improving the safety for users on Instagram (Meta). The recommendations point specifically to the technical aspects of the platform and general process and safety features that should be available within/on the platform, with specific reference to DMs.

Given the current unresponsiveness to the problems raised and failure to self-regulate, public / political pressure and the use of statutory powers by regulators will be needed to ensure that these are implemented.

Principles for an effective complaints system: A complaints process for users on Instagram (and other platforms) should be accessible, easy to use, and responsive. This is not the case here, given that Instagram (Meta) is failing to act on 90% of misogynist abuse (including image-based abuse), image-based sexual abuse (within 48 hours), and violent threats[53].

Recommendation: Instagram (Meta) needs to invest more in their complaints systems and team to ensure that they are responsive to these complaints. This is a core element of the duty of care owed to users. There should be no lag time for responding to serious sexual abuse and it is inexcusable to fail to act on the other types of abuse, including violent threats.

Complaints pathway for abusive voice notes: Our research found that there is no current pathway to report abusive voice notes, despite 1 in 7 voice notes containing abuse.

Recommendation: Instagram (Meta) needs to fix its complaints system so that there is an easy to access pathway for users to report abusive voice notes, like any other type of content on Instagram.

Vanish mode DMs: Our research shows that to report "vanish mode" DMs from strangers, users must open them and in doing so inform the sender that it has been viewed. This is a problem for many reasons - including the fact that users must be exposed to harmful content to report it to Instagram (which results in harm, as discussed earlier in the report).

Vanish mode DMs also create a process for giving gratification, power, and control to the abuser by Instagram giving that person formal acknowledgement that their abusive content has been viewed by the recipient. In situations where a user is being harassed by an abuser, there is already an expectation based on experience that the unsolicited content being sent to them via DM will be harmful. Abusers are also likely to be using vanish mode to help prevent accountability. A recipient should not need to view that content to have it addressed by Instagram. Instagram is playing a systemic role in perpetuating the cycle of violence through giving power and control to the abuser.

Instagram's system for allowing DM vanish mode messages to be sent by strangers, and the process for reporting abusive messages, is not meeting basic duty of care principles.

Recommendations:

Vanish mode should not be a feature that is available when sending a DM to a stranger. The recipient should both have to accept a connection with that person and agree to opt into vanish messages before this feature can be used. Users should be able to opt out of accepting vanish mode messages from any person at any time.

A user should be able to report abusive messages (of any sort but including vanishing DMs) without having to view the abusive content each time. This is particularly important where a person has already made a complaint to Instagram about the abuser and Instagram should be on notice that the relationship and messages are both unwanted and harmful. Anything less results in perpetuating a cycle of abuse.

Instagram should disable read receipts for vanish mode.

Identifying serial abusers: The problems with Instagram being used by serial abusers, and perpetuating the cycle of abuse, have been canvassed above. Our research shows that serial abuse is a significant problem and that there are significant limitations with how the review process works for women subject to serial abuse.

At present, Instagram only promises to review up to 30 recent messages for reports involving accounts based in the EU. There is a question of whether this number of messages is sufficient (to avoid abusers gaming the system and given the tools that can be developed and used by Big Tech like Instagram to scan content feeds) but also why it is limited to the EU when the capability exists to apply it more generally than this, given that it is an acknowledged and widespread problem. The status quo situation appears to

be another example of Big Tech favoring profits over safety on their platforms, despite what they say to politicians or in their terms and conditions.

Recommendation: Given the issues outlined above, there needs to be a principled justification for any limitation on the number of posts that are reviewed when complaints are made, and best practices should be shared across jurisdictions. Online abusers and those subject to the abuse, don't just live in the EU and there should be a harmonization of safety standards and process, where possible.

Investment is needed by Instagram, to ensure that this is available and implemented well. This is likely to require a direction from a regulator, given their inability to self-regulate and fix known problems in this area, even where they have established systems in one jurisdiction.

Blocking abuse and the use of "hidden words": Instagram currently allows users to specify "hidden words" to hide abusive DMs, but any other abuse from the same sender will get through. This is problematic given the ability of abusers to game the system and is another example of unregulated harmful content slipping through a sieve with giant holes in it. At a principled level, the status quo puts all the responsibility onto the individual user rather than Instagram having a safe default position.

Recommendation: Overall, there needs to be a better system for filtering abusive DMs and abusive users. One option for targeting the limitations of hidden words is for Instagram to hide all DMs in a conversation if any one message contains "hidden words" specified by the user.

Give users access to all their data, including DMs from block accounts. Our research revealed that Instagram's data download feature simply doesn't work for some users. Even where participants obtained data downloads, they were missing DMs from accounts that they had blocked. This personal information will still be retained by the company on their systems and should be subject to privacy protections like any other piece of private information.

Recommendation: Instagram must fix these problems so that users are able to access and share evidence of any abuse they have received. This will be important evidential content, for example, in criminal proceedings, civil disputes (such as child custody or harassment cases), and preventative measures (such as trespass notices, domestic violence protection orders, or a no-contact / restraining orders). We anticipate that the status quo situation and evidence in our report will also be of interest to privacy regulators.

Appendix: Methodology

A Research for this report was carried out with the participation of a small number of high-profile women willing to share access to their Instagram DMs. Research focused on English-language content only, and was carried out using three methods:

1. Analysis of downloaded "DM requests" data
2. Audits of Instagram's failure to act on DM abuse
3. Qualitative interviews with participants

1 Analysis of downloaded "DM requests" data

All participants were asked to request a downloaded copy of the data that Instagram holds on them, following the platform's own instructions on accessing this data.[54]

The following participants were not provided with their full data, making it impossible to include them in our analysis:

- Amber Heard, who received "technical error" messages when requesting data[55]
- Bryony Gordon, whose decision to block abusive users meant their message data was not included in her data download, precluding analysis

The following participants were able to use this method to obtain data and share their "message requests" folder with researchers for analysis of abusive DMs from strangers:

- Rachel Riley
- Jamie Klingler
- Sharan Dhaliwal

This data was shared with researchers confidentially and will be deleted by the end of 2022.

The resulting "message requests" folder for each participant contained subfolders for each account that a stranger used to message them, which in turn contains any media sent by the account and HTML files recording text messages they sent.

For text messages, researchers manually examined all messages for Klingler and Dhaliwal, and sampled every 9th folder of messages for Riley whose "message requests" data contained 9,842 such folders.

For image, audio and video files, researchers searched the "message requests" folder for relevant file suffixes (for example, .jpeg or .mp4) and analyzed the resulting files.

Researchers logged the total number of messages analyzed and the number of messages that breached Instagram's policies on:

- Hate speech, including misogyny, homophobia, racism and more[56]
- Bullying and harassment, including calls for death, suicide, or violence[57]
- Any displays of nudity or sexual activity[58]
- Violent and graphic content[59]

This allowed for a final analysis showing the number and proportion of analyzed messages in each media category - text, image, audio, and video - that breached Instagram's policies.

2 Audits of Instagram's failure to act on DM abuse

CCDH researchers assessed Instagram's performance in acting on reports of abuse over DM by working with participants to log abusive messages, report them to the platform and finally audit what enforcement action Instagram took against abusive accounts.

The following participants could take part as they routinely delete abusive messages:

- Jamie Klingler
- Bryony Gordon

The following participants were able to share access to abusive DMs in the "Requests" folder with our researchers:

- Rachel Riley
- Amber Heard
- Sharan Dhaliwal

For these three participants, CCDH scrolled through messages in their "Requests" folders beginning with the messages Instagram placed at the top of the inbox and working backwards. Messages were accessed between 28 December 2021 and 28 February 2022.

For each post that contravened Instagram's community standards, researchers collected the following information in our database:

- Screenshot or recording of the message
- Date the message was sent
- Type of media used (text, image, audio, or video)
- URL link to the sender's account
- Types of abuse featured in the message

After logging messages in this way, researchers reported them using Instagram's reporting tools in its mobile app or website. Once a user selects to "report" an abusive message, a dialog box appears requiring the user to report abuse under a category. CCDH's researchers reported abuse as follows:

- Misogyny and other identity-based hate → Hate speech or symbols
- Image based sexual abuse → Nudity or sexual activity → nudity or pornography
- Violence and gore → Violence or dangerous organizations
- Self-harm imagery → Suicide or self-injury

As there was no option to report audio files, CCDH instead reported the account that sent an abusive voice note.

After 48 hours of reporting, CCDH checked if the 253 abusive accounts that had sent DMs remained online. We performed a second audit on 12 March 2022 ahead of publication.

3 Qualitative interviews with participants

CCDH spoke with participants off the record to find how each uses Instagram, what sort of abuse they are sent from strangers and how they respond. Although some participants' behaviors, such as blocking or deleting abusers meant we couldn't access this abuse, we still wanted to find out how it affects them. There is no rulebook for how to respond to this abuse, and we were keen to not invalidate any of our participants' own responses.

We then met with participants in spaces and at times where they could feel comfortable speaking freely on the record about their experiences of online abuse, sending them the results of our research ahead of time so they could take in our findings before responding.

We asked participants about:

- Their historical use of Instagram
- Their experience of abuse on Instagram
- How they handle abuse that they receive from strangers on Instagram
- How they handle their own safety online
- Their response to our findings

We transcribed the interviews manually, inserting select quotes into this report.

Published 6 April 2022

[1] Ultraviolet, https://weareultraviolet.org/wp-content/uploads/2021/11/Letter-to-platforms-from-75-nonprofits-about-misogyny.pdf

[2] "1 in 4 black Americans have faced online harassment because of their race or ethnicity", Pew Research Center, 25 July 2017, https://www.pewresearch.org/fact-tank/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/
"The State of Online Harassment", Pew Research Center, 13 January 2021, https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
End Online Violence Against Women, https://www.endviolenceagainstwomen.org.uk/online-abuse-during-covid-almost-half-of-women-have-experienced-online-abuse-during-pandemic/

[3] "I spoke up against sexual violence — and faced our culture's wrath. That has to change.", Amber Heard, Washington Post, 18 December 2018, https://www.washingtonpost.com/opinions/ive-seen-how-institutions-protect-men-accused-of-abuse-heres-what-we-can-do/2018/12/18/71fd876a-02ed-11e9-b5df-5d3874f1ac36_story.html

[4] "Policies", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies

[5] "Hate speech", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

[6] "Bullying and harassment", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/bullying-harassment/
"Violence and incitement", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/

[7] "Adult nudity and sexual activity", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/

[8] "Violent and graphic content", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/violent-graphic-content/
[9] "Facebook Community Standards", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/
[10] "An update on our work to tackle abuse on Instagram", Instagram, 11 February 2021, https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram
[11] "Introducing new tools to protect our community from abuse", Instagram, 21 April 2021, https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse
[12] "How do I report a message that was sent to me or stop someone from sending me messages on Instagram?", Instagram Help Center, retrieved 24 March 2022, https://help.instagram.com/568100683269916
[13] Donie O'Sullivan (CNN), Twitter, 0:54, 8 December 2021, https://twitter.com/donie/status/1468684487156457473
[14] "Continuing to Make Instagram Safer for the Youngest Members of Our Community", Instagram, 17 March 2021, https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community
"Introducing new tools to protect our community from abuse", Instagram, 21 April 2021, https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse
[15] "Joint VAWG Sector Principles for the Online Safety Bill", End Violence Against Women (EVAW) Coalition, 8 September 2021, https://committees.parliament.uk/writtenevidence/39075/pdf/
[16] "Modernising Communications Offences", Law Commission, 20 July 2021, p233, https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/07/Modernising-Communications-Offences-2021-Law-Com-No-399.pdf
[17] "Deepfake pornography could become an 'epidemic', expert warns", 27 May 2021, https://www.bbc.co.uk/news/uk-scotland-57254636
[18] "Adult nudity and sexual activity", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/
[19] "Ireland's image based sexual abuse laws compared to other countries", Her, 11 March 2021, https://www.her.ie/news/irelands-image-based-sexual-violence-laws-revenge-porn-513656
[20] "'Cyberflashing' to become a criminal offence", UK Government, 13 March 2022, https://www.gov.uk/government/news/cyberflashing-to-become-a-criminal-offence
[21] "Pressure on police to crack down as 'clear link' emerges between flashing and serious sexual offending", The Telegraph, 9 July 2021, https://www.telegraph.co.uk/news/2021/07/09/nearly-800-men-prosecuted-flashing-last-year-despite-growing/
"Don't look now", The Guardian, 19 April 2001, https://www.theguardian.com/world/2001/apr/19/gender.uk
[22] "Instagram is bringing voice messaging to your DMs", The Verge, 10 December 2018, https://www.theverge.com/2018/12/10/18134675/instagram-voice-messaging-direct-message-dm
[23] Messages sent by Instagram account to Sharan Dhaliwal and retrieved through her Instagram data download, 31 May 2021
[24] Messages sent by Instagram accounts to Rachel Riley and retrieved through her Instagram data download, 28 September 2021 and 3 January 2022
[25] Messages within Jamie Klingler's Instagram data download, 20 September 2021
[26] Instagram, 21 April 2021, https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse
[27] ibid.
See Section 4 for details of Instagram's promises to act on abuse

[28] Donie O'Sullivan (CNN), Twitter, 8 December 2021, 0:52, https://twitter.com/donie/status/1468684487156457473

[29] "I spoke up against sexual violence - and faced our culture's wrath. That has to change", Amber Heard, The Washington Post, 18 December 2018 https://www.washingtonpost.com/opinions/ive-seen-how-institutions-protect-men-accused-of-abuse-heres-what-we-can-do/2018/12/18/71fd876a-02ed-11e9-b5df-5d3874f1ac36_story.html

[30] "Ireland's image based sexual abuse laws compared to other countries", Her, 11 March 2021, https://www.her.ie/news/irelands-image-based-sexual-violence-laws-revenge-porn-513656

[31]Nadine Dorries, Twitter, 13 March 2022, https://twitter.com/NadineDorries/status/1502917422596210688

[32] "Online flashers face two years in prison under new law", The Telegraph, 13 March 2022, https://www.telegraph.co.uk/news/2022/03/13/online-flashers-face-two-years-prison-new-law/

[33] Messages sent by Instagram account to Amber Heard, 8 December 2020, 19 August 2021, 21 December 2021 and 28 December 2021

[34] Messages sent by Instagram account to Amber Heard, 4 July 2021, 19 July 2021, 14 December 2021, 20 December 2021

[35] "Bullying and Harassment", Meta, retrieved 24 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/bullying-harassment/

[36] Words identified as "one-word" abuse are: "bitch", "c*nt", "whore", "die", "KYS", "kill yourself", "hoe", "slut", "skank", "twat" and "slag".

[37] Messages sent by Instagram accounts to Amber Heard, 17 and 28 December 2021, Message sent by Instagram account to Rachel Riley 26 December 2021

[38] Adam Mosseri, Twitter, 22 July 2021, https://mobile.twitter.com/mosseri/status/1418246950319702016

[39] "How To Report Things", Instagram Help Center, retrieved 24 March 2022, https://help.instagram.com/2922067214679225

[40] Amber Heard's Instagram inbox, retrieved 7 February 2022

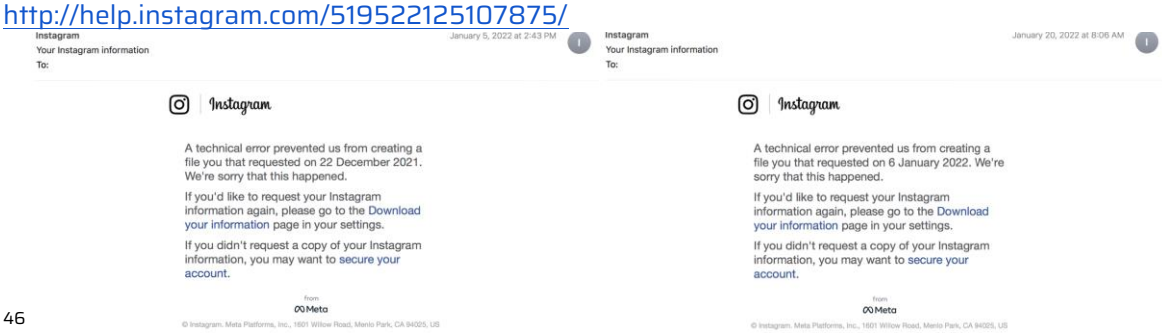
[41] Instagram, retrieved 24 March 2022

[42] Message sent by Instagram user to Amber Heard, 27 December 2021

[43]"How To Report Things", Instagram Help Center, retrieved 24 March 2022, https://help.instagram.com/2922067214679225

[44] "Introducing new tools to protect our community from abuse", 21 April 2021, https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse

[45] "Instagram Data Policy", Instagram, retrieved 31 January 2022, http://help.instagram.com/519522125107875/

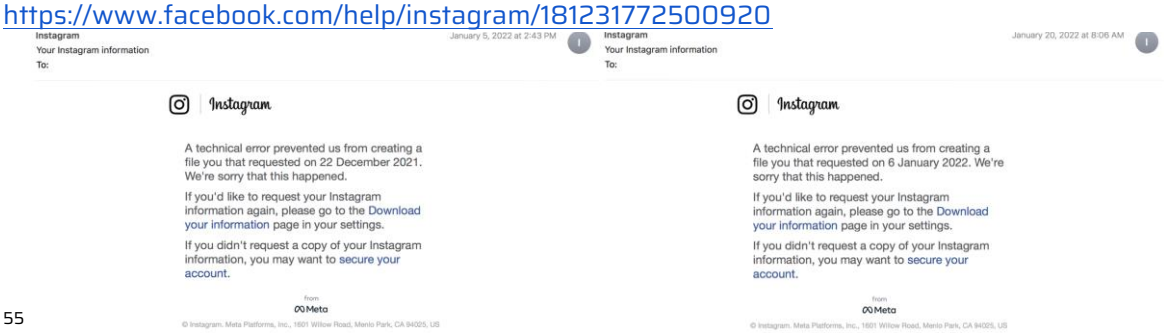


[46] Emails sent by Instagram in response to a request for a data download for Amber Heard's account, 5 January 2022 and 20 January 2022

[47] "Staying Safe", Instagram, retrieved 24 March 2022, https://help.instagram.com/1502695926736394/

[48] We found that 1 in 15 DMs that were studied broke Instagram's rules on abuse and harassment.;

[49] "Marketplace flagged over 800 social media posts with COVID-19 misinformation. Only a fraction were removed", CBC, 30 March 2021, https://www.cbc.ca/news/marketplace/marketplace-social-media-posts-1.5968539
[50] "Failure to Protect", Center for Countering Digital Hate, 30 July 2021, https://www.counterhate.com/failuretoprotect
[51] "Instagram fails to take down more than 94% of racist abuse accounts targeting England players after Euros", iNews, 15 July 2021, https://inews.co.uk/news/technology/instagram-racist-abuse-posts-england-players-after-euros-1102896
[52] "Twitter fails to remove 100 abusive misogynists", The Times, 13 January 2022, https://www.thetimes.co.uk/article/twitter-fails-to-remove-100-abusive-misogynists-z7nwg6d9t
[53] As above, this included 9 in 10 of accounts ending violent threats over DM, failing to act on any image-based sexual abuse within 48 hours and failing to act on all accounts sending 'one-word' misogynist abuse, failing to act on 90% of abuse sent over DM.
[54] "How do I access or review my data on Instagram?", Instagram, retrieved 23 March 2022, https://www.facebook.com/help/instagram/181231772500920



[55] Emails sent by Instagram in response to a request for a data download for Amber Heard's account, 5 January 2022 and 20 January 2022
[56] "Hate speech", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/
[57] "Bullying and harassment", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/bullying-harassment/
"Violence and incitement", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/
[58] "Adult nudity and sexual activity", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/
[59] "Violent and graphic content", Meta, retrieved 23 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/violent-graphic-content/